IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 16, 2021, accepted May 17, 2021, date of publication May 25, 2021, date of current version June 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3083490

A Novel Method for Credit Scoring Based on Cost-Sensitive Neural Network Ensemble

WIROT YOTSAWAT[1], **PAKAKET WATTUYA**[1,2], **AND ANONGNART SRIVIHOK**[1]
[1]Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok 10903, Thailand
[2]Center of Artificial Intelligence Innovation for Healthtech, Faculty of Science, Kasetsart University, Bangkok 10903, Thailand

Corresponding author: Pakaket Wattuya (fscipkw@nontri.ku.ac.th)

This work was supported by the Department of Computer Science, Faculty of Science, Kasetsart University, Thailand.

ABSTRACT Most existing studies on credit scoring adapted a concept of classifier ensemble for solving an imbalanced dataset. They apply resampling methods to generate multiple training subsets for constructing multiple base classifiers. However, this approach leads to several problems that degrade the classification performance, such as problems of information loss, model overfitting, and computational cost. Thus, we propose a novel ensemble approach for developing a credit scoring model based on a cost-sensitive neural network, called Cost-sensitive Neural Network Ensemble (CS-NNE). In the proposed approach, multiple class weights are adapted to original training data, enabling the multiple base neural networks to consider imbalanced classes. Following this approach, a high diversity of multiple base classifiers without consequent problems can be achieved. The approach's effectiveness is evaluated on five real-world credit datasets. Among them is a loan-requesting dataset provided by a financial institution in Thailand. The remaining datasets are publicly available and widely used by several existing studies. The experimental results showed that the proposed CS-NNE approach improves the predictive performance over a single neural network based on imbalanced credit datasets, e.g., Thai credit dataset, by achieving 1.36%, 15.67%, and 6.11% Area under the ROC Curve (AUC), Default Detection Rate (DDR), and G-Mean (GM), respectively, and achieving the best Misclassification Cost (MC). The proposed CS-NNE approach can effectively solve a class of imbalance problems and outperform many existing models. The prediction model can well compromise between classes of default (bad credit applicants) and non-default (good credit applicants), whereas existing approaches preferred a class of non-default over default loans (having high specificity and low DDR), resulting in NPL.

INDEX TERMS Credit scoring, neural network, cost-sensitive learning, ensemble, imbalanced dataset.

I. INTRODUCTION

A credit scoring model is a statistical analysis tool that determines the creditworthiness of a loan applicant by estimating the probability of default based on historical data [1]. It is used by lenders to decide whether to accept or reject borrowers. By leveraging of credit scoring model, the number of loan defaulters has rapidly increased, particularly during the financial crisis. In Thailand, the Bank of Thailand reported that in Q4, 2020, the total gross NPLs in commercial banks is more than 523 million baht, while the number of new-entry NPLs is more than 53.9 thousand accounts (source: https://www.bot.or.th/).

The credit scoring model is leveraged under credit policies, which is chanced depending on challenge environments. For example, during the COVID-19 pandemic that directly affects the applicants' creditworthiness [2], lenders will maintain the dept to prevent more increasing current NPLs. Besides, many situations affect the applicants' creditworthiness, such as economic conditions, financial crisis, and political situations. The credit scoring model should be flexible depending on the credit policies, and adjustable/compromise between acceptance and rejection rate. Most credit scoring models lack flexibility because they try to enhance the accuracy of the classification model dominated by good credit. To fill the gap, this study proposes a flexible credit scoring model that can adjust between acceptance and rejection by setting the appropriate attention between the classes of borrowers and can reach the credit policies.

To develop a credit scoring model, financial institutions collect borrowers' information and utilize it through statistical or machine learning techniques. The collected data

The associate editor coordinating the review of this manuscript and approving it for publication was Tomasz Trzcinski.

often encounters some problems, e.g., it may be indiscernible from noisy and outlier data, the appearance of an imbalanced problem, and an issue of asymmetric cost matrix due to the misclassification cost on defaulters, which is greater than the misclassification cost on non-defaulter loans. These problems directly decrease the credit scoring model's performance. Besides, the choices of the techniques for constructing the models are considered an important issue, affecting the performance of the credit scoring system.

To tackle an imbalanced problem in credit scoring data, many studies employed resampling techniques, such as under-sampling and over-sampling [3]–[8]. The major disadvantage of resampling techniques is led to the overhead cost and the other consequent problems, e.g., 1) information may be lost using under-sampling techniques, 2) the final model may be overfitted using over-sampling techniques, 3) the original data distribution may be changed, and 4) the model is more complex and it has high computational cost. Assigning different costs to the training instances is the most efficient approach to deal with the class of imbalance problems [9]–[13] and avoid the major problems after applying the resampling techniques.

Most credit scoring models focused only on maximizing predictive accuracy based on the profit generated by interest. However, little attention has been paid to default detection, although mistaken prediction on default generating cost was much higher than that on non-default loans. Thus, these models will result in huge NPLs due to bad borrowers' misclassification. In this study, the proposed credit scoring model consider default and non-default loans, which are not considered by other credit scoring models. The proposed model can prevent some situations, such as suffering a great loss due to a failing decision on loan granting, especially when classifying 'bad' borrowers as 'good' borrowers.

Inspired by the above studies, we propose a CS-NNE approach for developing a credit scoring model. The proposed approach can address the problems in the credit scoring task and improve the performance of credit scoring model. Auto-encoder Outlier Detection (AEOD), cost-sensitive learning, and neural network ensemble techniques are combined for several reasons. First, AEOD is used to detect and eliminate noisy and outlier data. It is also used to solve the model-overfitting problem. Second, cost-sensitive learning is employed to address an imbalanced problem and asymmetric cost issue. Third, multiple class weights are assigned to the neural network algorithm to achieve accurate and diverse base learners for ensemble classification. Finally, the results generated through neural networks are combined using a voting mechanism to generate the final output. The contributions of this study are as follow:

1) Introducing a novel cooperative of cost-sensitive learning and neural network ensemble for credit scoring, called CS-NNE.
2) Introducing a novel diversifying technique to form a neural network ensemble by assigning multiple class weights to based classifiers.
3) The proposed CS-NNE addresses imbalanced datasets without the need for any complex rebalancing tasks.

The remainder of the paper is structured as follows: Section II details related previous literatures on credit scoring. Section III describes the proposed framework. Section IV presents the details of experiment. Based on the observations and experiments, the results and discussion are presented in Section V. The last section draws conclusions and future research directions.

II. LITERATURE REVIEW

A. TRADITIONAL CREDIT SCORING MODELS

Traditional credit scoring models have been developed using statistical or machine learning methods, such as Support Vector Machine (SVM) [14], [15], Logistic Regression (LR) [16]–[18], Decision Tree (DT) [4], [19], and Neural Network (NN) [6], [20]–[24] based on a single or ensemble approach. Ma *et al*. [25] obtained that boosting method outperforms competitive state-of-the-art classification algorithms based on extracted features from phone usage data and individual's app usage behaviors. Abid *et al*. [16] found that the LR model outperforms Linear Discriminant Analysis (LDA) based on a Tunisian commercial bank dataset. Alaraj *et al*. [17] obtained that the LR model outperforms NN and SVM using various performance indicators based on the loan dataset of a public commercial bank in Jordan. Nalić and Martinovic (2020) [18] investigated the LR model for credit scoring, and obtained that the LR model shows better prediction confidence and accuracy than DT, naïve Bayes (NB), and SVM. However, most studies assumed that misclassification on different classes have a consistent cost. In the real economic world, the cost associated with granting some loans for a customer who defaults on the loan is far greater than the cost (opportunity loss) associated with rejecting some loans from a customer who may have successfully repay the loan [11], [26]–[29]. Thus, cost-sensitive learning models for credit scoring need further investigation.

Among the existing modern machine learning methods, NN models are widely used due to the models provide competitive classification ability against other methods. In 2017, Eletter and Yaseen [27] proposed a comparison of predictive results on credit scoring using NN, LDA, and CART based on Jordanian commercial banks dataset. They obtained the NN model provided the highest accuracy and the lowest estimated misclassification cost. However, the model is based on a single classifier, which can improve the model by employing ensemble methods. In 2018, Dželihodžić *et al*. [21] proposed the performance improvement of credit scoring model using bagging neural network and utilizing the Bosnian commercial bank dataset and two publicly available datasets. Their model achieved high performance compared to the benchmark models. However, the diversity of the ensemble should be considered, and an asymmetric misclassification problem was unsolved. In 2010, Eletter *et al.* [30] applied a backpropagation neural network approach to loan application evaluation in Jordanian commercial bank. Tsai [31] proposed

a novel hybrid technique by binding between cluster analysis and classifier ensembles for credit scoring classification. The study showed that integrating between self-organizing maps and multilayer perceptron (MLP) classifier ensembles achieved the best performance. Zhao *et al.* [20] proposed an improved MLP with 1 hidden layer and $6-39$ hidden nodes based on backpropagation to improve the credit scoring performance. Their proposed method achieved higher accuracy than their literature search indicated. However, even though NN exhibits significant accuracy advantages, there have not been investigated based on the cooperation of cost-sensitive learning, neural network, and ensemble for developing credit scoring models.

In the case of outlier and noisy data being indiscernible in datasets, several studies illustrated that by eliminating noisy and outlier data at the preprocessing step, the predictive models' performance is improved [32]–[34]. Xia [35] proposed integration between outlier removal and gradient-boosting algorithm for credit scoring on peer-to-peer lending datasets. The researcher found that outlier removal significantly outperformed the benchmark models. Besides, the computational cost showed great potential for handling large-sized datasets. Wei *et al*. [8] combined the outlier removal method and classification algorithm to develop a credit scoring model called backflow learning. It was relearned the misclassified data points and combined the prediction of based learners by a two-layer ensemble. The results showed that backflow learning with an outlier removal outperforms other models and can improve the credit scoring model performance. Thus, it is essential to eliminate noisy and outlier data, which can be performed before developing models. In this study, we use AEOD [36] to eliminate noisy and outlier data.

B. IMBALANCED DATASETS HANDLING

Cost-sensitive learning is a type of learning which considers misclassification costs [37]. It can be used to solve two data mining problems: an imbalanced dataset and an issue of the asymmetric cost of misclassification. Cost-sensitive learning minimizes the total cost of misclassification [37]. The category of the cost-sensitive learning can be separated into indirect and direct cost-sensitive methods [11], [37]. The indirect cost-sensitive method builds a cost-sensitive classifier by preprocessing the training data through resampling techniques or postprocessing the results through threshold-moving, while the direct method constructs a cost-sensitive learning algorithm by assigning different misclassification costs into the learning process [11], [38].

By the indirect cost-sensitive methods, in 2018, He *et al*. [5] and Sun *et al*. [7] introduced the idea of generated training subsets using different resampling rates for ensemble classifiers to develop a credit scoring model. Their results were superior to other comparative algorithms. Although different resampling rates resolve imbalanced problems, they might lead to other problems, such as computational cost, information loss, original data distribution change, and model over-fitting. This study uses different costs to diversify based classifiers without applying any resampling method. Thus, the consequent problems generated by resampling methods do not occur, and the empirical results showed high performance. Khemakhem *et al*. [6] investigated the relevance and performance of sampling methods integrated with LR, NN, and SVM using an imbalanced dataset. In their experiment, Tunisian commercial bank data were utilized, and an imbalanced data issue was addressed through random oversampling (ROS) and SMOTE. They suggested that the combination of the classification and resampling techniques can improve the default detection rate. The indirect cost-sensitive methods are mostly constructed at the data preprocessing step using resampling techniques, which may be attributed to other consequent problems, which are the disadvantage of indirect cost-sensitive learning. The direct cost-sensitive methods may have more value and need further consideration. Thus, this study investigates direct cost-sensitive learning based on original data distribution.

In direct cost-sensitive methods, forming new training sets using a resampling method is not required. Thus, original data distribution will be kept. Some researchers stated that direct cost-sensitive learning required further consideration. Alejo *et al*. [39] improved MLP predictive performance using cost-sensitive learning based on a single classifier. However, recent studies showed that ensembles of classifiers achieved better results than single methods for such task [5], [15], [19], [28], [31], [40]–[44]. Bahnsen *et al*. [45] developed a classification model using cost-sensitive decision tree by incorporating different example-dependent costs to impurity measure and pruning criteria. The model showed superior performance on baseline models in terms of training time and cost savings. However, cost-sensitive decision trees are sensitive to training data patterns because a small difference in the training data can cause a different tree model and result in different classification [46]. Shen *et al*. [11] developed a credit scoring model using cost-sensitive logistic regression method. The model optimized the hyper-parameters of LR by using a particle swarm optimization algorithm, which effectively reduces error rates and total misclassification cost and improves the overall model's performance. Logistic regression is limited by some statistical assumptions. For example, collinearity among independent variables should not occur, and the independent variables are linearly related to the log odds. NN is not sensitive to data and does not require any statistical assumptions. It is suitable for continuous variables, which provide a competitive prediction ability. Besides, NN is a nonlinear model, where associations between credit attributes are complex and nonlinear. Thus, NN is an appropriate choice for credit scoring prediction problems. However, the standard NN is cost-insensitive learning, which is affected by imbalanced problems. This study investigates cost-sensitive learning and NN simultaneously. The empirical study showed that the proposed direct cost-sensitive method outperforms indirect cost-sensitive methods using resampling methods to form new training sets.

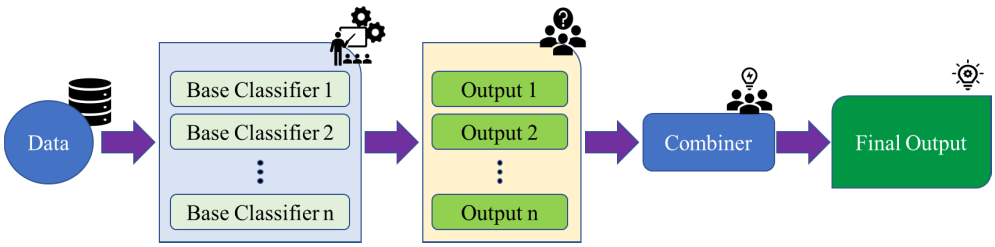

FIGURE 1. General concept of ensemble classifier.

Wang *et al*. [47] extended the traditional multi-instance learning method to a cost-sensitive version using boosting technique, where different costs are reconsidered in each iteration of boosting. Xia *et al*. [38] proposed a cost-sensitive boosted tree model to improve the performance of classification model. The model combines cost-sensitive learning with Extreme Gradient Boosting (XGBoost) to predict the probability of default in peer-to-peer lending. Wang *et al*. [13] proposed a model for multiple classes of credit grading based on peer-to-peer lending data. They concluded that cost-sensitive classifiers (DT, Random Forest (RF), LR, and SVM) can significantly reduce the total cost of misclassification. The previous works have shown that direct cost-sensitive methods are potential techniques for addressing credit scoring. Thus, this study is focused on the directed cost-sensitive learning model.

C. ENSEMBLE CLASSIFICATION METHODOLOGY

A single classifier is successful in some specific datasets. The ensemble classification is an effective methodology for enhancing the classification performance of an individual method by cooperating with the limited strength of each base classifier to be a powerful classifier. Ensemble classifiers are divided into two categories: homogeneous and heterogeneous. The homogeneous ensemble builds a classifier using the same algorithm based on the different data (either in the instance or feature spaces), such as bagging and boosting. The heterogeneous or stacking ensemble constructs a classifier using different based classification algorithms. The general concept of ensemble classifier is illustrated in Figure 1.

Some aspects of bagging, boosting, and stacking ensemble need carefully considered before using. First, bagging forms new sub-training sets, which depend on random sampling with replacement. Thus, some observations may be randomly selected by multiple times, while other observations may be excluded. It is highly probable that the randomly selected observations will be in all sub-training sets, which can decrease performance degradation since the diversity of base learners is insufficient. Second, boosting requires weak learners to boost into strong learners by combining various sequential learnings. The boosting mechanism is sensitive to noise and outlier data because noise and outliers will have much larger residual errors than others. In some cases, boosting leads to an overfitting problem by the large weight of misclassifying examples, which has been selected into sub-training sets. Third, the stacking ensemble performs on multiple classification algorithms. Thus, it is a more complex setting than the homogeneous ensemble, which employs only one classification algorithm. In this study, we introduce a novel high diversity classification ensemble approach using the concept of cost-sensitive learning. We obtain diverse base classifiers and address an imbalanced problem, simultaneously.

There are several studies proposed credit scoring models based on ensemble techniques. For example, Paleologo *et al*. [48] extended the advantage of the bagging approach, where the training subsets are formed by random sampling to address a class of imbalanced problems. Dželihodžić *et al*. [21] used the advantage of the bagging approach to improve credit scoring models. Luo [49] compared the performance of the bagging approach using DT, SVM, K-nearest Neighbor (KNN), and MLP based on an imbalanced and large dataset. They obtained that bagging-KNN was more suitable than other methods for large and imbalanced datasets in credit scoring. Finlay [50] investigated the performance of various multiple classifiers and concluded that bagging and AdaBoost outperform other ensemble methods. Tsai *et al*. [51] conducted a comprehensive study comparing classifiers ensemble methods for three public credit scoring datasets. They found that the boosting DT achieves the best performance. Existing studies on ensemble classification to construct credit scoring model mainly employ random sampling technique for generating base learners. They focused on how to construct accurate base learners, how to choose the best base learners, or how to bind the results of base learners for better ensemble performance. However, the diversity of the constructed based learners has been slightly attention and need more consideration. This restrictive weakness may decrease the overall performance of classification model because the effective performance of ensemble methods requires that base learners have diversity in their predictions [15].

III. PROPOSED COST-SENSITIVE NEURAL NETWORK ENSEMBLE

The main concept of the proposed CS-NNE is to utilize different class weights to diversify the models and generate base classifiers' results. This section describes the conceptual framework of the proposed CS-NNE, as shown in Figure 2. The proposed CS-NNE consists of two phases: Phase-I data preprocessing and Phase-II classification and evaluation.

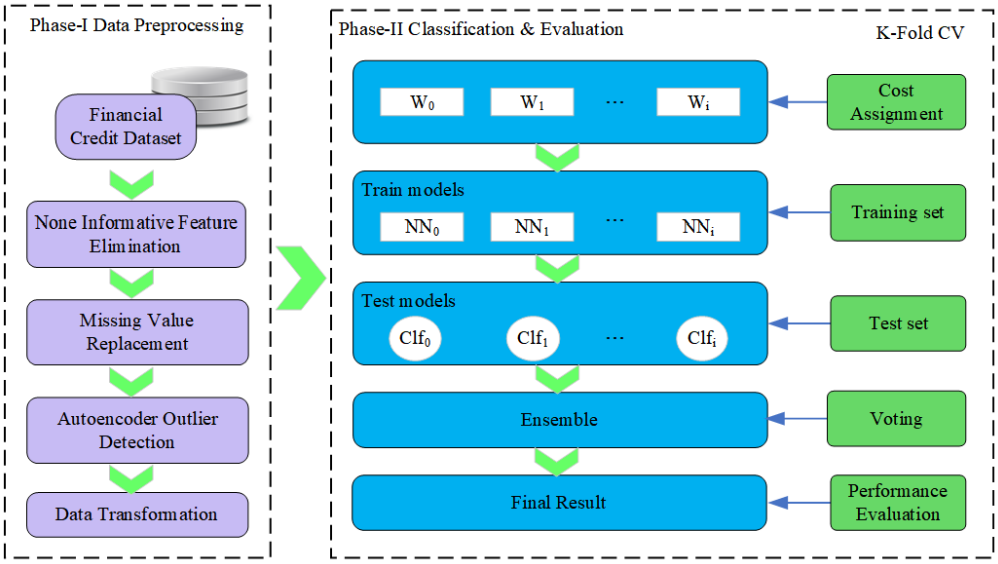

FIGURE 2. Conceptual framework for proposed CS-NNE.

A. PHASE-I DATA PREPROCESSING

Data preprocessing is a crucial step in machine learning and data mining tasks before constructing a model. It can enhance the performance of models based on accuracy and time complexity. To prepare the data for the classification algorithm, a representative and consistent dataset is required by data preprocessing tasks. Data preprocessing steps consist of data cleaning, outlier removal, and data transformation.

1) Data cleaning: The irrelevant features are eliminated and missing values are handled.
2) Outlier removal: Some machine learning algorithms are sensitive to outlier such as KNN, Boosting. To enhance the classification accuracy, outlier detection and removal method will be utilized by autoencoder method as called AEOD.
3) Data transformation: Some classification algorithms require numerical data such as SVM, NN. Thus, one hot encoder is applied to convert the categorical input features. Besides, different ranges of numerical attributes are normalized into a small fixed range.

B. PHASE-II CLASSIFICATION AND EVALUATION

For effective ensemble classification, accurate and divers base classifiers are essential since they can achieve better outcomes. This study employs neural network since it achieves accurate classifiers. Besides, divers' classifiers are reached by the different from class weights assigned to the neural networks.

1) COST-SENSITIVE NEURAL NETWORK

Artificial neural network with backward propagation is widely used in many domain problems. The standard neural network is a fully connected structure between each layer. It consists of multiple levels of nonlinear operations and hidden layers. The hidden layers can be single or multi-nonlinear layers, which are between the input and output layers. Neural network learns input data through a function $f(\cdot) : R^d \rightarrow R^o$, where d is the number of input neural nodes, and o is the number of output neural nodes. Given a set of leftmost input layers representing the features $X = x_1, x_2, \ldots, x_d$, neurons in the hidden layer transform the values from the input layer with weighted linear summation $w_1x_1 + w_2x_2 + \ldots + w_dx_d$. Then, a nonlinear activation function $g(\cdot) : R \rightarrow R$ maps the summation for either classification or regression.

The standard neural network refers to the cost-insensitive learning. Several studies have shown that the class of imbalanced problems generate unequal contributions to the Mean Square Error (MSE) during the training process [39]. Given a training set with binary classes ($C = 2$) of size $N = \sum_c^C n_c$, where n_c is the number of instances in class c, the MSE for class c can be calculated as follows:

$$MSE_c = \frac{1}{N} \sum_{i=1}^{n_c} \left(Y_i' - Y_i\right) \tag{1}$$

where Y_i' is the predicted output, and Y_i is the actual output of the network for the instance i. Thus, the overall MSE can be illustrated in the term of each class as follows:

$$MSE = \sum_{c=1}^{C} MSE_c = MSE_1 + MSE_2 \tag{2}$$

when the class is imbalanced, $n_1 \ll n_2$, then the $MSE_1 \ll MSE_2$ and $\|\nabla MSE_1\| \ll \|\nabla MSE_2\|$, where the operator ∇ represents the gradient of the error function. Consequently, $\|\nabla MSE\| \approx \|\nabla MSE_2\|$. Thus, ∇MSE is not always the best minimizer of MSE in both classes. The unequal contribution on MSE can be compensated by modifying a misclassification cost function (γ) as follows:

$$MSE = \sum_{c=1}^{C} \gamma_c MSE_c = \gamma_1 MSE_1 + \gamma_2 MSE_2 \tag{3}$$

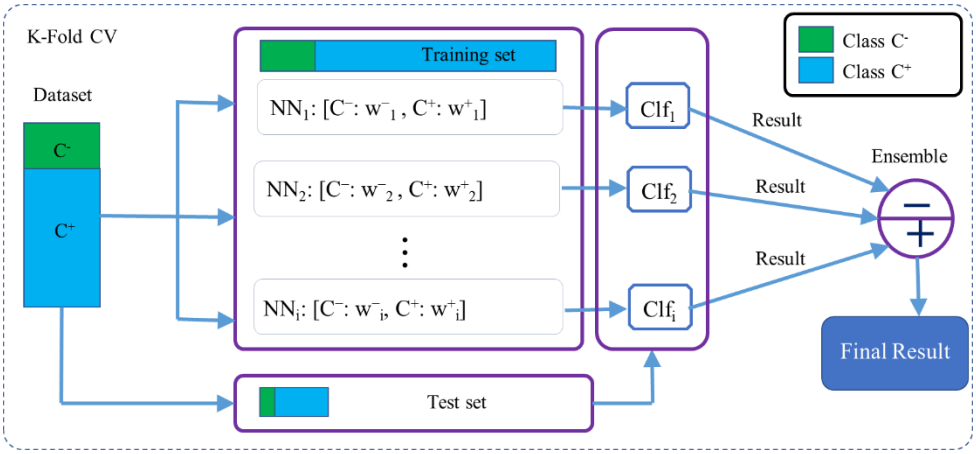

FIGURE 3. Class weights setting for base classifiers. Denoted that *Clf* is a base classifier to form an ensemble model. C^+ and C^- represent class default and non-default, respectively. w^+ and w^- are class weights setting to class C^+ and C^-, respectively. i is the numbers of based classifiers.

As mentioned above, cost-sensitive learning does not change the data distribution directly. It considers misclassification costs during the training process. It considers a weight on each instance according to the specified costs. It means that instances of the minority class which holds a higher misclassification cost, are assigned proportionally high weights.

In the proposed method, several classifiers are trained with different costs. The ReLU function is used as an activation function in each hidden layer. In the output layer, the Softmax activation function is used to determine the neuron output values as the default or non-default loans. Adaptive moment estimation (Adam) optimizer [52] is used to optimize the gradient descent. It performs well in practice and compares favorably to other stochastic optimization functions [52].

2) NEURAL NETWORK GENERATION METHOD BASE ON COST-SENSITIVE LEARNING

The proposed CS-NNE employs cost-sensitive neural networks by assigning different costs as class weights to the neural networks. The cost can be initialized by the proportion of classes as follows:

$$w_i = \frac{n}{C * n_i} \tag{4}$$

where n is the number of instances in the training set, C is the number of classes, and w_i is the cost, which will be a class weight of class i. w_i is a key success parameter to generate the diverse base learners of the proposed ensemble method. The bigger cost allows the base classifier to pay more attention to instances from its class than another one. Different costs for network ensemble are adjusted and input to neural networks as hyper-parameters that can be turned. The cost assignment for neural networks ensemble is shown in Figure 3.

3) ENSEMBLE CLASSIFICATION METHODOLOGY

We employ the success of NN and ensemble approach in credit scoring. Many studies have shown that ensemble classifiers can produce near-optimal results [5], [15], [19], [28], [31], [40]–[44]. Ensemble approaches are confused base homogeneous or heterogeneous classifier into a single result. In this study, the majority voting technique has been used. The majority voting technique combines based classifiers by adding their results. The final output is based on the highest score, as shown in the following equation.

$$c^* = arg_i \, max \sum_{j=1}^{NC} v_{i,j} \tag{5}$$

where v is the probability result of classifier j predicted as class i, and NC is the number of classifiers. For example, the instance X, if base classifier Clf_j (j $= 1, 2, \ldots, NC$) predicts its class label as $y = c^*$, a binary variable $v_{i,j}$ is set to be 1, else 0. Thus, the voting result represents the probability that X belongs to class c^*. According to the voting rule, the final output of instances X is the class with the highest votes. The algorithm of the proposed CS-NNE is shown in Figure 4.

To evaluate the models, the performance of the proposed CS-NNE is compared with the other state-of-the-art techniques used for credit scoring, such as KNN, LDA, LR, DT, SVM, and NN. Among the benchmark models, popular ensemble methods, such as RF [53], XGBoost [54], Bagging [55], and AdaBoost [56], are included.

IV. EXPERIMENTS

In this section, the experimental setting is presented. It includes the datasets and data preprocessing, experimental setup, and performance measurements.

A. DATASET AND DATA PREPROCESSING

In this study, the proposed approach is evaluated by utilize five real-world credit datasets. The first dataset is loan-requesting data provided by a financial institution in Thailand (Thai credit). For confidentiality reasons, the name of the financial institution has not been provided. The raw data of Thai credit consists of 147,620 records of consumers

Algorithm: CS-NNE

Input: training set, test set, W and the number of classifiers I.

Step 1: Initial $Ens = \{\}$ #for store base classifiers

Step 2: For i=0 to I

$W_i = [w_i^+, w_i^-]$

Input training set to train Cost-Sensitive Neural Network Clf_i base on the cost W_i

Ens U $\{Clf_i\}$

Step 3: For each base classifier in Ens

Test the base classifier by test set

Output: Credit evaluation result by CS-NNE model.

FIGURE 4. Algorithm of CS-NNE (for binary classification).

who had been given a loan. Among them, 15.14% of consumers were considered default loans because borrowers fail to pay back more than three months. Thus, the dataset was imbalanced. In the dataset, each instance consists of more than 100 attributes, such as customer ID, account status, account number, age, gender, marital status, education, religion, main occupation, income, customer type, restructure flag, branch owner, branch description, product type description, loan commitment, address, and telephone number. However, many attributes, such as customer ID, account number, account status, address number, code, and description of bank branches are not very informative for loan default prediction. Thus, these attributes were discarded to avoid the problem of model over-fitting.

For data cleaning on the Thai credit dataset, attributes with a single value of more than 99%, or missing values of more than 30%, were eliminated. After cleaned data, 21 remaining features were used, as shown in Table 1. Missing values were replaced by the mean or median value of the entries depending on the attribute's data type. All categorized features were transformed to binary numeric features using the one-hot-encoder. Thus, a feature with k different nominal values was transformed into k binary features. Then, all numerical features were normalized into an interval from 0 to 1. The autoencoder method has been used to remove outliers by 10%.

The remaining four datasets used to evaluate the proposed approach are German, Polish, Australian, and Lending club. German, Polish, and Australian datasets are collected from the UCI repository, and the Lending club dataset provided by Kaggle. The data preprocessing used in the Thai credit dataset is applied to all datasets, except AEOD is only applied to the Lending club dataset. The details of the datasets are presented in Table 2. For the Lending club dataset, we randomly selected from the Lending club data 2015–2018 consisting of more than 1M records. We focused on the customers whose loan status is either "*Fully Paid*" or "*Charged Off*". The selected customers are applied to AEOD for 10% removal. After removing irrelevant features, the 22 remaining features consists of loan_amnt, fico_score, sub_grade, int_rate, term, installment, annual_inc, emp_length, home_ownership, cr_hist_len, dti, delinq_2yrs, inq_last_6mths, open_acc, pub_rec, revol_bal, revol_util, total_acc, purpose, zip_code, verification_status, and initial_list_status.

B. EXPERIMENTAL SETUP

In this study, the credit scoring models were constructed by utilizing real-world training and test datasets. The experiment was conducted using Windows 10 operating system with an Intel Core i7 7500 CPU and 8 GB of RAM. Python version 3.6 was used in the computer and other associated libraries. Imbalanced-learn version 0.4.3 and PyOD version 0.7.4 were also used to address the imbalanced problem and outlier detection, respectively. Tensorflow version 1.4.0 and Keras version 2.1.2 libraries were used in the experimental study to develop the fully connected neural network. Scikit-learn library version 0.20.0 was used to apply several well-known algorithms, such as LDA, SVM, LR, KNN, DT, RF, and other ensemble methods. XGBoost version 0.90 was used for a specific XGBoost classifier.

On the Thai credit training set, Bayesian hyper-parameter optimization was used on XGBoost for hyper-parameters turning because of the curse of dimensionality problem. A grid search is used to determine the optimal hyper-parameters for other classification algorithms. The optimizers perform with the given parameters space as presented in Table 3. The best parameters were selected for credit scoring and set to base classifier for Bagging and AdaBoost. Excluding parameter setting, the other parameters were set to the default value with respect to the common model in the literature. On the remaining four publicly training sets, hyper-parameters were applied by the default value. Besides, the cost of CS-NNE was adjusted through the empirical study based on the training data.

C. PERFORMANCE MEASURES

The measurements are evaluated to compare several classifiers. In this experiment, the classification performance was measured using evaluation terms consisting of accuracy, AUC, DDR and non-default detection rate (specificity), and G-Mean based on the confusion matrix (Figure 5). These performance measures are selected because the measures cover

TABLE 1. Attribute description and statistics of thai credit dataset.

Attribute	Description	Count/Mean	%	Attribute	Description	Count/Mean	%
Occupation	1: Registered business owner	6,522	4.91	Count of phone number	[0, 3]	2.37	-
	2: State enterprise	9	0.01	Monthly income (THB)	1: <10,000	41,609	31.32
	3: Company employee	28,057	21.12		2: 10,000 - 14,999	26,113	19.65
	4: Freelance	12,238	9.21		3: 15,000 - 29,999	41,117	30.95
	5: Specific profession	2,358	1.77		4: 30,000 - 49,999	17,747	13.36
	6: Farmer	1,233	0.93		5: 50,000 - 99,999	4,245	3.20
	7: Non-registered business owner	18,789	14.14		6: 100,000 - 199,999	930	0.70
	8: Government officer	3,326	2.50		7: >200,000	1,097	0.83
	9: Other	60,326	45.41	Loan amount (THB)	[3,000, 120,000,000]	125,929.82	-
Religion	1: Islam	52,249	39.33	Interest rate on the loan	[0.25, 18.75]	11.21	-
	2: Buddhism	79,922	60.16	Purpose	1: Consume	36,605	27.55
	3: Other	687	0.52		2: Medical treatment /welfare	3,244	2.44
Age (years)	[22, 85]	40.17	-		3: Home	6,116	4.60
Gender	1: Male	46,358	34.89		4: SME business	30,341	22.84
	2: Female	86,500	65.11		5: Peddler/stall	27,839	20.95
Marital status	1: Single	48,461	36.48		6: Multi-purpose	23,835	17.94
	2: Married	53,172	40.02		7: Other	4,878	3.67
	3: Cohabitation	22,606	17.02	Term (months)	[2, 420]	62.93	-
	4: Divorce	5,075	3.82	Grace day (days)	1: 0	22,466	16.91
	5: Widow	3,544	2.67		2: 7	110,383	83.08
Education	1: Primary education	38,112	28.69		3: 30	9	0.01
	2: Secondary education	44,699	33.64	Payment (THB)	[100, 1,150,000]	2,248.46	-
	3: Diploma	14,542	10.95	Selling price (THB)	[5,230, 306,009,706]	279,987.94	-
	4: Bachelor degree	33,798	25.44	Account receivable (THB)	[0, 293,359,706]	242,209.69	-
	5: Master degree	1,608	1.21	Ledger balance (THB)	[0, 114,544,853]	103,454.23	-
	6: Doctorate	99	0.07	Restructured flag	1: Yes	2,734	2.06
Area	1: Southern	75,407	56.76		2: No	130,124	97.94
	2: Central	46,796	35.22	Collateral requires	1: Yes	20,516	15.44
	3: North Eastern	3,641	2.74		2: No	112,342	84.56
	4: Northern	2,147	1.62	Class	0: Non-default/Good	112,742	84.86
	5: Eastern	4,867	3.66		1: Default/Bad	20,116	15.14

TABLE 2. Details of datasets used for evaluating the performance of CS-NNE.

Dataset names	#Attributes	#Instances	#Good:#Bad	IR*	Source
Australian credit	14	690	383:307	1.25	UCI
German credit	24	1,000	700:300	2.33	UCI
Polish	64	7,027	6,756:271	24.93	UCI
Lending club	22	88,890	64,377:15,623	4.12	Kaggle
Thai credit	21	132,858	112,742:20,116	5.60	Commercial bank in Thailand

* IR is an imbalanced ratio.

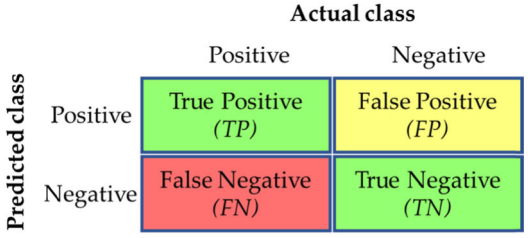

FIGURE 5. Confusion matrix.

almost all aspects of model performance and they are widely used to measure the performance of credit scoring models.

1) The accuracy (Acc) measures the overall true predicted value in all classes. However, the accuracy value alone cannot indicate model performance due to the imbalanced problem. Thus, it only indicates the overall classification accuracy of the dataset.

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{6}$$

2) The sensitivity or default detection rate (DDR) measures the proportion of actual default loans that are correctly detected by models as bad applicants. A number of NPLs will be small when the DDR is high. On the other hand, NLPs will be increasing when the DDR reduces. The DDR indicator is very important due to the misclassification on default loan leads to NPLs.

$$DDR = \frac{TP}{TP + FN} \tag{7}$$

TABLE 3. Hyper-parameters setting for each classification algorithm.

Method	Parameters	Parameter space
SVM	gamma	0.001, 0.01, 0.1
	cost	1, 10, 100, 500
*K*NN	number of nearest neighbors	3, 5, 7, 9, 11
DT	minimum number of samples required to split	range(10,500,20)*
	minimum number of samples at a leaf node	range(2,100,2)
	maximum depth of the tree	range(10,25,1)
NN	hidden layer	[n_feature], [n_feature/2], [n_feature,n_feature], [n_feature,n_feature/2], [n_feature/2,n_feature/2], [n_feature,n_feature,n_feature], [n_feature,n_feature,n_feature/2], [n_feature,n_feature/2,n_feature/2], [n_feature/2,n_feature/2,n_feature/2]
	batch size	64, 128
	epoch	300, 500
Bagging	maximum number of estimators	5, 7, 9, 11
	sample ratio	0.8, 0.9, 1.0
XGBoost	maximum tree depth	range(4,15,1)
	learning rate	range(0.01,1.0)
	number of boosting estimators	range(100,500,20)
	subsample ratio	range(0.8, 1.0)
	column subsample ratio	range(0.8 ,1.0)
	gamma	range(0, 0.1)
RF	number of features randomly sampled	4, 8, 12, 16, 20
	number of trees	100, 500, 1,000
CS-NNE	numbers of base classifiers	3, 5, 7, 9, 10, 11

*range(a,b,[c]) represents a range of numbers starting with a, stop before b and increment by c.

3) The specificity (Spec) measures the proportion of non-default loans which are correctly classified by models as good applicants. Large Spec indicates the profit gaining generated by the interest of granted loans, while small Spec means the opportunity loss from potential interest which the loan is likely to produce.

$$Spec = \frac{TN}{TN + FP} \tag{8}$$

4) The Area Under the ROC Curve (AUC) measures the classification ability of entire sample and the balance of classified samples simultaneously [57]. Thus, it can be considered as more appropriate measurement in an imbalanced credit scoring [41]. The AUC is derived from the area under the ROC curve which plot between True positive rate (or DDR) at the y-axis and False positive rate at the x-axis. The score of AUC ranges from 0 to 1 which value near 1 refers to the model has high accuracy.

5) The Geometric Mean (GM) is an evaluation indicator constructed by Spec and DDR. The higher GM indicates the balance of classification performance between the classes without any class dominated. It is reasonable and good performance indicator in the binary classification model. The GM is computed by the following equation:

$$GM = \sqrt{Spec \times DDR} \tag{9}$$

If the values of those indicators for a classifier are greater than the other ones, its classification performance would be better. Besides, for financial institutions, the risk brought by the credit scoring models is measured by Misclassification cost (MC). According to Feng *et al.* [28], the relative cost ratio is used to estimate the MC of the models and evaluate the performance of credit scoring models in financial terms. The MC is calculated using the following equation:

$$MC = \frac{FN}{TP + FN} \times P(0) \times 5 + \frac{FP}{FP + TN} \times P(1) \times 1 \tag{10}$$

where $P(0)$ and $P(1)$ are the prior probability of non-default and default loans on the test set, respectively. According to the literature [27]–[29], [58]–[60], the cost is set to 5 for the misclassification on a defaulter (as a good customer), more costly than the misclassification on a non-defaulter (as a bad customer).

V. EXPERIMENTAL RESULTS

In this study, real datasets were used. Four of the five datasets are imbalanced. In the experiments, benchmark classification algorithms were implemented. This section presents the experimental results and analysis. We divided this section into three parts: A) and B) present the results of private dataset (Thai credit) and public datasets (German, Polish, Australian, and Lending club), respectively, and C) discuss the results.

A. RESULTS ON THE THAI CREDIT DATASET

The proposed CS-NNE was compared with the benchmark models, such as individual and ensemble models. To reduce the bias-related to the random sampling of the training and test sets, the experiment was set to five-fold cross-validation so that all samples are selected into both training and testing sets. The average of indicators (Acc, AUC, Spec, DDR, and GM) is presented in this section with MC comparison based on the five-fold experiment.

1) PERFORMANCE COMPARISON BETWEEN CS-NNE AND SINGLE CLASSIFIERS

Table 4 presents the approximated parameters for single classifiers, which are set for ensemble classifiers. Table 5 presents the detailed performance of the proposed CS-NNE and other state-of-the-arts classification models. Based on the literature, RUS and SMOTE are used for preprocessing the training sets to address the imbalanced problem.

TABLE 4. The approximated parameters setting found on single methods.

Method	Parameters	Approximated Parameters
SVM	gamma	0.01
	cost	500
*K*NN	number of nearest neighbors	11
DT	minimum number of samples required to split	90
	minimum number of samples at a leaf node	30
	maximum depth of the tree	17
NN	hidden layer	[n_feature, n_feature/2]
	batch size	64
	epoch	300

TABLE 5. Performance comparison of the proposed CS-NNE with other single classifiers based on before and after resampling.

Method	Preprocessing	Acc	AUC	Spec	DDR	GM
SVM	-	93.21	93.71	**98.92**	61.21	77.81
	RUS	81.74	91.33	81.19	84.85	83.00
	SMOTE	90.80	94.23	91.99	84.12	87.97
LR	-	88.48	86.44	98.46	32.53	56.56
	RUS	74.87	83.12	74.32	77.95	76.11
	SMOTE	75.74	84.41	75.19	78.83	76.99
*K*NN	-	83.57	73.96	94.12	24.46	47.97
	RUS	69.63	77.69	75.44	68.59	71.93
	SMOTE	88.22	89.73	89.85	79.08	84.29
LDA	-	84.39	80.99	94.70	26.64	50.22
	RUS	73.59	81.28	73.21	75.71	74.45
	SMOTE	73.51	81.35	73.06	76.00	74.52
DT*	-	93.96	94.59	98.12	70.62	83.24
NN	-	**95.20**	96.66	98.44	77.04	87.08
	RUS	90.04	95.27	90.45	87.73	89.08
	SMOTE	93.18	96.48	94.28	87.01	90.57
CS-NNE	-	93.52	**98.02**	93.67	**92.71**	**93.19**

* RUS and SMOTE are not applied to the training set for DT since they do not theoretically affect the DT mechanism.

Based on original training and resampled training sets by RUS and SMOTE, the experimental results revealed that the proposed CS-NNE model outperforms traditional credit scoring models, such as LR, SVM, *K*NN, LDA, DT, and single NN in the terms of AUC, DDR, and GM. Single classifiers maximize the overall accuracy without considering the minority class of imbalanced datasets. Based on imbalanced dataset, the DDR of LR, *K*NN, and LDA were less than 50% (Table 5). It means poorly accurate on the detection of default loans. This problem can be addressed using resampling methods. The results showed that resampling methods improve the performance of single classifiers on DDR and GM. However, the average accuracy and Spec were reduced.

From Table 5, all traditional models (except SVM-RUS, SVM-SMOTE, NN-RUS, and NN-SMOTE) show that they will generate very high NPLs, as indicated by DDR. CS-NNE can compromise between profit gain and NPLs generated by non-defaulters and defaulters, respectively. Although the CS-NNE model decreases a small number of profits generated by loan interest, it will reduce a vast number of NPLs, which critically decreases the financial profit. The compromising between profit and NPLs provided by CS-NNE is better than SVM-RUS, SVM-SMOTE, NN-RUS, and NN-SMOTE indicated by GM, and CS-NNE achieves the highest AUC.

As the predictive performance of single classifiers, NN outperforms other compared methods on most measurements. It is reasonable to use NNs as the base learners for homogeneous ensemble because a better base learner is preferred to corporate the ensemble approach. Thus, the proposed method outperforms other single models since it is based on the accuracy of the base learners, as confirmed by the results in Table 5.

2) PERFORMANCE COMPARISON BETWEEN CS-NNE AND ENSEMBLE CLASSIFIERS

The efficiency of ensemble approaches depends on the diversity among base learners, the performance of the base learners, and the combining techniques. Popular ensemble methods, such as RF, XGBoost, Bagging, and AdaBoost have been implemented to evaluate the influence of diversity on ensemble performance. The diversity of these ensemble methods employs random sampling (either in the samples or feature spaces). Besides, SMOTE and RUS are used to address imbalanced problems for these ensemble methods. The proposed CS-NNE model employs cost-sensitive learning to address the imbalance and diversity of ensemble. Table 6 presents the best searching parameters for ensemble classifiers. Besides, the CS-NNE model based on nine optimal classifiers is set to the cost of class non-default and default by [1:1.5], [1:3.5], . . . , [1:17.5], justified by the empirical study. The empirical results are illustrated in Table 7.

This study enhances the classification performance of the credit scoring model on DDR with the competitive average predictive results. Table 7 shows that the proposed CS-NNE model achieves the best AUC, DDR, and GM compared to other ensemble models. RF achieves the best accuracy, while several classifiers, such as RF, NN-Bagging, DT-AdaBoost, and DT-Bagging are dominated by non-default loans. They also unsuccessfully detect the default samples that did not achieve the aims of the predictive models and the financial institution's needs. RF, DT-AdaBoost, and NN-Bagging results show closed performance.

TABLE 6. Approximated parameters setting found on ensemble methods.

Method	Parameters	Approximated Parameters
DT-Bagging	maximum number of estimators	11
	sample ratio	0.9
NN-Bagging	maximum number of estimators	9
	sample ratio	0.9
XGBoost	maximum tree depth	5
	learning rate	0.2569855283603088
	number of boosting estimators	400
	subsample ratio	0.9394390590412429
	column subsample ratio	0.8506194042405409
	gamma	0.03988488636589157
RF	number of features randomly sampled	20
	number of trees	500
CS-NNE	numbers of base classifiers	9

TABLE 7. Performance comparison of the proposed method with other ensemble approaches.

Methods	Acc	AUC	Spec	DDR	GM
XGBoost	93.84	95.33	96.94	76.47	85.93
RF	**95.74**	97.86	**99.09**	77.00	87.35
DT-AdaBoost	94.95	96.32	98.23	76.58	86.73
DT-Bagging	94.11	96.14	98.71	68.36	82.14
NN-Bagging	95.64	97.37	98.87	77.48	87.52
NN-Bagging-SMOTE	91.08	96.22	91.46	88.90	90.17
NN-Bagging-RUS	94.34	97.49	95.41	88.36	91.82
NN-Bagging (1:5)*	93.84	97.35	94.59	89.67	92.09
CS-NNE	93.52	**98.02**	93.67	**92.71**	**93.19**

* NN-Bagging (1:5) is set the misclassification cost of base learners on class non-default and default by 1 and 5, respectively.

From Table 7, we observe that the DT-based ensemble method (XGBoost, RF, DT-Bagging, and DT-AdaBoost) and NN-Bagging generate NPLs more than the NN-based ensemble with addressing an imbalanced problem using direct or indirect cost-sensitive learning. The training cost-sensitive learning on NN-Bagging with RUS, SMOTE, and fixed cost (1:5) showed the approximate predictive performance, which is improved by the traditional NN-Bagging on DDR and GM. It means the NPLs will be reduced using cost-sensitive learning to assess applicants' creditworthiness. The proposed CS-NNE achieves better performance than NN-Bagging with RUS, SMOTE, and fixed cost (1:5) on some reasons. For example, 1) CS-NNE reduces the NPLs by 4.35%, 3.81%, and 3.04% generated from NN-Bagging with RUS, SMOTE, and fixed cost (1:5), respectively. 2) CS-NNE does not require any complex resampling methods to address an imbalanced problem. However, the imbalanced problem can be addressed without any consequent problems generated by resampling methods. 3) CS-NNE performs on the original training distribution. However, bagging forms sub-training sets by random sampling with replacement. Thus, its computational cost will be more.

3) MC COMPARISON

Misclassification on default and non-default classes is a significantly different cost. Errors on predictive defaults generate higher costs than those associated with predictive non-default because financial institutions lose potential interest and the principal that was granted to borrowers. The trustworthy estimates of the MC are a complicated task. According to [27]–[29], [58]–[60], the relative cost ratio of MC is 1:5 for non-default and default loans, respectively. According to (10), MC depends on the misclassification of both classes with different costs. The lowest MC means that the model can decrease misclassification costs from potential interest and the principal on the defaulters.

The CS-NNE model does not achieve the highest Spec because it does not dominate by the majority class and the overall classification ability is efficient. Thus, the proposed CS-NNE model can reduce misclassification costs from potential credit default for lenders. In Figure 6, the CS-NNE model showed the best performance on MC by 0.109, indicating that the costs generated by opportunity loss and NPLs will be minimized by CS-NNE.

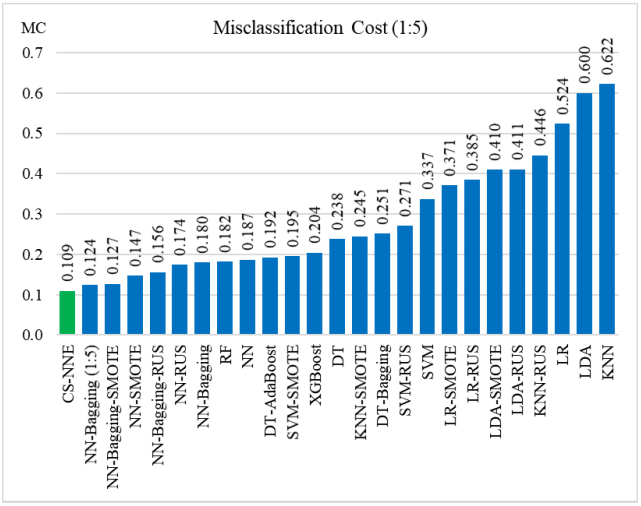

FIGURE 6. Comparison of MC on different models.

B. RESULTS ON PUBLIC DATASETS

This section compares the results of CS-NNE with other classifiers and previous studies on publicly available datasets (German, Polish, Australian, and Lending club).

1) PERFORMANCE COMPARISON BETWEEN CS-NNE AND OTHER CLASSIFIERS

The performance of different credit scoring models is evaluated by utilizing K-fold cross-validation. Besides, five-fold cross-validation is used on the Lending club dataset, while the rest datasets are used for ten-fold cross-validation. By default-parameters, the performance evaluation of various classifiers and CS-NNE is presented in Table 8.

Table 8 shows that the CS-NNE model can compromise the predictive accuracy between classes for the four datasets. CS-NNE achieves the best performance on DDR for the German, Polish, and Lending club. It means that the CS-NNE model is suitable for addressing imbalanced dataset problems occurring in credit scoring problems. Using the CS-NNE model, financial institutions can decrease the loss generated by NPLs since it can detect defaulters better than other models. However, RF achieves the best performance on the Australian dataset, which is considered a class of balanced dataset. The RF slightly performs better than the CS-NNE model.

2) PERFORMANCE COMPARISON WITH PREVIOUS STUDIES

As described above, CS-NNE is suitable for addressing imbalanced datasets. Thus, we present comparative results obtained from previous studies on credit scoring in terms of AUC, DDR, and GM. Table 9 shows the comparison of the CS-NNE performance with previous studies based on imbalanced datasets consisting of German, Polish, and Lending club datasets.

Considering the German dataset in Table 9, the AUC score provided by CS-NNE is inconsistent with the highest AUC provided by MHS-RF [61]. However, MHS-RF unsuccessfully detects default customers as well as the models provided by NNBag [21] and CS-Bagging-CS- CART [62]. Compared with BP-ANN-PSO [57], the CS-NNE model can compromise the accuracy between classes better than BP-ANN-PSO, as indicated by GM. These models will generate a large number of NPLs. There are a few studies, which do not report the Spec and DDR in their studies. Thus, we cannot compare the insight's results.

Based on the Polish dataset, which has a high imbalance ratio, CS-NNE achieves the best performance on DDR by improving more than 20%, while Spec is still higher than 92%. Although other compared models will make high profits generated by loan interest, they will make very high NPLs, which critically damage the net profit.

Based on the Lending club dataset, the first and second models with the highest NPLs are RF [60] and DM–ACME [63], respectively. RF-RUS [64] shows the best on compromising accuracy between classes. However, comparing to RF-RUS [64], CS-NNE is less than 1% on the entire compared measurements, which is insufficiently different from the overall performance. EBCA+PSO [5] achieves the best AUC. However, the GM score is very low, indicating that EBCA+PSO gives very low on either Spec or DDR.

As described above, most previous studies lead to NPLs when datasets are imbalanced. The proposed CS-NNE model can solve the problem and improve predictive performance on default loans. Thus, using the CS-NNE model, the financial institutions can reduce the loss on the principal, which the loan is probably to be an NPL. Besides, the net profit partially depends on the loss given default.

C. DISCUSSION

In credit scoring systems, any small improvement in the performance of credit scoring models can significantly improve future savings and has important commercial implications. It can be achieved using ensemble methods [3], [4], [6]. This experiment integrated cost-sensitive learning with the NN ensemble approach. In this study, state-of-the-art models were benchmarked with classical techniques. Based on the experimental results in Table 5, the overall accuracy rates are generally high on the original imbalanced dataset and for the different classifiers. This is because the dataset is highly dominated by non-default loans around 85% and underrepresented by the number of remaining default loans. The classifiers show a huge weakness on correctly detecting the minority observations. These results confirmed the studies by Loyola-González *et al.* (2016) and Sihem Khemakhem *et al.* [6], Loyola-González *et al.* [76]. Their studies suggested that the training process is dominated by the majority class, while the minority class is misclassified even though the classification models generate a high overall accuracy.

He *et al.* [5] and Sun *et al.* [7] suggested the idea of generating diverse training subsets by different resampling rates for ensemble classifiers to construct credit scoring models. Although different resampling rates resolve an imbalanced

TABLE 8. Performance comparison of CS-NNE with different methods on public datasets.

Dataset	Measure	NN	DT	SVM	LR	RF	AdaB-DT*	Bag-DT*	Bag-NN*	XG-Boost	CS-NNE
German	Acc	74.80	68.80	77.20	77.10	76.50	71.00	74.40	77.80	**79.00**	74.40
	AUC	76.46	63.05	79.62	79.01	**80.20**	65.57	75.22	79.85	79.60	80.11
	Spec	86.00	77.43	90.86	89.29	**92.57**	79.14	85.14	90.43	90.29	75.43
	DDR	48.67	48.67	45.33	48.67	39.00	52.00	49.33	48.33	52.67	**72.00**
	GM	65.73	62.04	59.73	65.39	61.14	63.62	61.34	65.42	64.89	**73.63**
Polish	Acc	97.15	95.62	96.14	95.99	97.64	95.80	97.71	96.94	**98.11**	91.30
	AUC	87.85	76.12	75.91	67.59	89.66	75.67	85.29	87.64	**95.26**	88.62
	Spec	99.50	97.25	**100.00**	99.79	99.84	97.48	99.70	99.56	99.72	92.01
	DDR	38.78	54.99	0.00	1.10	42.83	53.86	47.98	31.77	57.94	**73.80**
	GM	61.10	72.29	0.00	5.74	64.08	71.67	67.79	54.46	75.53	**82.14**
Australian	Acc	84.78	79.42	84.78	85.51	**86.81**	79.47	85.07	85.65	85.94	84.93
	AUC	90.43	79.37	91.79	92.09	**92.49**	68.96	90.98	91.36	92.48	91.31
	Spec	86.41	79.89	83.51	84.85	88.52	**92.38**	85.88	86.93	87.76	86.40
	DDR	82.75	78.85	86.30	**86.31**	84.71	34.04	84.06	84.04	83.73	83.06
	GM	84.40	79.18	84.77	85.46	**86.48**	56.05	84.89	85.32	85.64	84.63
Lending club	Acc	79.65	70.29	80.48	80.57	80.47	70.45	77.34	**80.60**	80.29	63.61
	AUC	66.90	54.38	67.78	70.79	69.72	54.57	64.27	**70.91**	69.69	70.82
	Spec	96.22	80.49	**99.99**	98.33	98.48	80.63	91.22	98.21	97.40	67.41
	DDR	11.36	28.26	0.08	7.39	6.27	28.52	20.12	8.03	9.79	**62.69**
	GM	32.96	47.69	1.96	26.92	24.83	47.94	42.84	28.06	30.87	**65.00**

* AdaB-DT, Bag-DT and Bag-NN is the abbreviation of AdaBoost-DT, Bagging-DT, Bagging-NN, respectively.

problem, they might lead to extra cost and other problems. This study used different costs to diversify based classifiers without applying any resampling method. Thus, the consequent problems generated by resampling methods did not occur. As shown in Table 7, different costs could diversify and improve the performance of based classifiers.

In statistical techniques, LR is still considered the industry-standard method for constructing credit scoring models. Some researchers showed the superiority of LR over ML techniques, such as SVM and DT [16]–[18], [77]. However, recent studies have demonstrated that ML techniques achieve better than statistical techniques in tackling financial problems, including credit scoring. As shown in Table 5, NN predictive performance outperformed the statistical methods and other single methods on most criteria. It conforms to Mundra *et al.* [78] research. To form an ensemble, based methods should be competent classifiers, which provide high accuracy and variance. Besides, NN provided high accuracy and other measurements. Different costs are employed to achieve a high variance of predictive models. Thus, it is reasonable to integrate the powerful NN for ensemble classifiers, as implemented in this study.

Recently, ensemble classifiers have employed increasing on credit scoring models. Malekipirbazari and Aksakalli [60] recommended RF as an effective algorithm to build a credit scoring model. RF has been used as a benchmark of ensemble algorithms in many studies [8], [60], [79]. Compared with the RF model, the proposed CS-NNE model performs better based on DDR, GM, and MC. This is because the proposed CS-NNE is constructed using accurate based-classifiers, which achieve better than decision tree-based methods. Besides, an ensemble is employed to increase the predictive performance using different costs for diversification. The results presented in Table 7 showed the proposed CS-NNE achieves better than RF methods based on DDR and GM with 15.71% and 5.84%, respectively.

In the Bagging ensemble scheme, many high variance based-methods are aggregated to form a final decision [55]. The Bagging-NN method was reported as an alternative and effective tool for credit scoring models by Dželihodžić *et al.* [21], called NNBag. The difference between the proposed CS-NNE and NNBag is as follows. First, NNBag employs attribute selection techniques to enhance the performance of the credit scoring model based on small and imbalanced datasets. In the proposed CS-NNE method, AEOD is utilized to enhance the credit scoring model without any informative feature elimination based on a large and imbalanced dataset, occurring in the real world of credit scoring. Second, NNBag creates ten-based neural network classifiers and combines results using the majority voting based on Weka environment. In the proposed method, the optimal nine NN classifiers have been created based on Python. Finally, the variance of based classifiers in the NNBag model depends on a 0.9 randomly resampling ratio with replacement. Thus, some data points might be represented multiple times, while some informative instances might be excluded by the Bagging mechanism. However, the variance of the proposed CS-NNE relies on different

TABLE 9. Performance comparison of CS-NNE with prior works over public datasets.

Dataset	Techniques	Year	Acc	AUC	Spec	DDR	GM
German	CSVM-RBF [14]	2015	77.10	69.23	-	-	-
	Decorate+ LR [41]	2017	77.40	79.37	-	-	-
	XGBoost-TPE [44]	2017	77.34	-	-	-	-
	EBCA+PSO [5]	2018	-	80.02	-	-	62.03
	CS-Bagging-CS-CART [62]	2018	-	-	89.13	41.63	-
	CART+BPSO [65]	2018	78.00	73.92	91.71	46.00	-
	Bstacking [66]	2018	78.66	79.48	-	-	-
	NNBag [21]	2018	76.70	77.00	86.00	51.00	-
	MOPSO-CS [67]	2019	75.45	-	83.03	57.76	-
	MHS-RF [61]	2020	75.60	80.53	92.29	36.67	-
	Overfitting-Cautious [68]	2020	77.72	80.34	-	-	-
	BP-ANN-PSO [57]	2020	76.60	80.04	79.47	66.86	63.57
	mg-GBDT [69]	2021	77.15	79.29	91.86	-	-
	CS-NNE	2021	74.40	80.11	75.43	72.00	73.63
Polish	IFNA+ backflow XGB [8]	2019	97.46	91.98	-	-	-
	Bag-C4.5 [70]	2019	-	92.30	99.00	5.02	-
	ABoost(C4.5) [70]	2019	-	87.00	99.70	53.10	-
	V-GANs [71]	2019	-	73.79	-	52.05	-
	BLOF-RF [72]	2021	98.10	94.88	-	-	-
	CS-NNE	2021	91.30	88.62	92.01	73.80	82.14
Lending club	RF (#68,000*) [60]	2015	78.00	71.00	88.00	31.00	-
	EBCA+PSO (#95,633*) [5]	2018	-	73.07	-	-	1.64
	RF-RUS (#66,376*) [73]	2018	69.20	69.00	71.70	58.20	65.00
	CatBoost (#11,467*) [74]	2019	79.59	63.33	-	-	-
	CatBoost (#26,288*) [74]	2019	77.16	62.66	-	-	-
	CatBoost (#26,384*) [74]	2019	75.23	61.62	-	-	-
	DM–ACME (#70,860*) [63]	2020	72.31	66.97	76.78	46.07	60.09
	XGBoost (#1,347,681*) [75]	2020	63.6	67.4	64.5	60.2	-
	RF-RUS (#462,378*) [64]	2021	64.00	71.70	68.00	63.00	65.60
	CS-NNE (#88,890*)	2021	63.61	70.82	67.41	62.69	65.00

* The number of instances.

costs based on the dataset without noisy and outlier data. Bagging neural networks were also implemented as benchmarked models. As shown in Table 7, CS-NNE outperforms the Bagging neural networks indicated by AUC, DDR, and GM before and after applying resampling techniques to the training set.

Terko *et al*. (2019) [80] suggested that the XGBoost model fails to outperform recommended approaches for solving credit scoring problems. Although the overall accuracy score of the XGBoost model is satisfactory, a high number of misclassifications based on defaults is higher than the detected default. From Table 7, the empirical results are consistent with the study by Terko *et al*. XGBoost showed a high value of accuracy with high Spec. It means that XGBoost can accurately discover on non-default loans, which can increase the profit gained for financial institutions by the potential interest. However, the low value of DDR generated by XGBoost critically damages the financial profit. Compare with the proposed CS-NNE, XGBoost provides DDR less than CS-NNE by 16.24%, while CS-NNE provides a non-default prediction rate less than XGBoost by 3.27%.

In summary, the researchers assumed that the proposed CS-NNE can enhance the performance of DDR with the competition of the overall classification results. We investigated the performance of single and multiple classifiers for credit scoring utilizing an imbalanced and large dataset. The results showed that most ensemble algorithms have better performance than the single classifiers. This finding is consistent with previous comparative studies [21]. Besides, the proposed CS-NNE outperforms other techniques due to two reasons: 1) using NN based classifiers since NN provides better performance among other techniques, and 2) using different costs to diversify based classifiers. In the learning aspect, the CS-NNE does not require any resampling method to form diverse training subsets. Besides, the proposed CS-NNE minimizes MC, which reaches the financial institution needs. Thus, the proposed approach has great potential in credit scoring applications.

VI. CONCLUSION

Credit scoring model has become a powerful tool for banks and other financial institutions to assess the creditworthiness of applicants. Thus, the classification performance of the credit scoring model is essential to maximize the profitability of financial institutions. This research proposed a novel credit scoring model using a cost-sensitive neural network ensemble as called CS-NNE. We used autoencoder outlier detection to remove noisy and outlier data during data preprocessing. The proposed model can help financial institutions to detect the probability of default before granting loans. To verify the efficiency of the proposed CS-NNE model, individual and ensemble methods were implemented as benchmark models e.g., DT, LR, LDA, SVM, *K*NN, RF, Bagging, AdaBoost, and XGBoost. We utilize five real-world credit scoring datasets. One of them is a loan-requesting dataset provided by a financial institution in Thailand, while the remaining four datasets are publicly available and widely used by several existing studies.

It is essential to address the imbalanced problem in the training set to avoid bias to a majority class. Benchmark models are implemented by utilizing balanced and imbalanced datasets. The comparative results showed the superiority of the proposed CS-NNE over the benchmark individual and ensemble methods based on AUC, DDR, GM, and MC. Compare with a single NN based on the Thai credit dataset, the proposed CS-NNE model can improve AUC, DDR, and GM by 1.36%, 15.67%, and 6.11%, respectively, and achieving the best MC by 0.109. Besides, we obtained that using appropriate costs to form the flexible CS-NNE is better than complex resampling techniques. The superiority of the proposed CS-NNE is confirmed on imbalanced datasets, such as German, Polish, and Lending club datasets. Thus, the proposed model is suitable for assessing the applicants' creditworthiness and can be used as an alternative successful technique for credit scoring system, especially in a challenging environment, such as a financial crisis period.

Although we investigated the cost-sensitive neural network ensemble on imbalanced datasets, some aspects need further investigation, such as the cost-sensitive ensemble based on other algorithms and the effect of imbalanced ratios on original datasets. In future studies, further study will be conducted using other machine learning algorithms, novel approaches for data preprocessing, and optimizing cost-sensitive learning parameters.

REFERENCES

[1] D. Tripathi, D. R. Edla, A. Bablani, A. K. Shukla, and B. R. Reddy, "Experimental analysis of machine learning methods for credit score classification," *Prog. Artif. Intell.*, pp. 1–27, Mar. 2021.

[2] A. Horvath, B. S. Kay, and C. Wix, "The COVID-19 shock and consumer credit: Evidence from credit card data," FEDS, New York, NY, USA, Work. Paper 2021-008, Jan. 8, 2021. [Online]. Available: https://ssrn.com/abstract=3613408

[3] V. García, A. I. Marqués, and J. S. Sánchez, "Improving risk predictions by preprocessing imbalanced credit data," in *Proc. Int. Conf. Neural Inf. Process.* Berlin, Germany: Springer, 2012, pp. 68–75.

[4] Y.-C. Chang, K.-H. Chang, H.-H. Chu, and L.-I. Tong, "Establishing decision tree-based short-term default credit risk assessment models," *Commun. Statist.-Theory Methods*, vol. 45, no. 23, pp. 6803–6815, 2016.

[5] H. He, W. Zhang, and S. Zhang, "A novel ensemble method for credit scoring: Adaption of different imbalance ratios," *Expert Syst. Appl.*, vol. 98, pp. 105–117, May 2018.

[6] S. Khemakhem, F. Ben Said, and Y. Boujelbene, "Credit risk assessment for unbalanced datasets based on data mining, artificial neural network and support vector machines," *J. Model. Manage.*, vol. 13, no. 4, pp. 932–951, Nov. 2018.

[7] J. Sun, J. Lang, H. Fujita, and H. Li, "Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates," *Inf. Sci.*, vol. 425, pp. 76–91, Jan. 2018.

[8] S. Wei, D. Yang, W. Zhang, and S. Zhang, "A novel noise-adapted two-layer ensemble model for credit scoring based on backflow learning," *IEEE Access*, vol. 7, pp. 99217–99230, 2019.

[9] P. Cao, B. Li, D. Zhao, and O. Zaiane, "A novel cost sensitive neural network ensemble for multiclass imbalance data learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Aug. 2013, pp. 1–8.

[10] O. Loyola-Gonzalez, J. F. C. O. Martinez-Trinidad, J. A. Carrasco-Ochoa, and M. Garcia-Borroto, "Cost-sensitive pattern-based classification for class imbalance problems," *IEEE Access*, vol. 7, pp. 60411–60427, 2019.

[11] F. Shen, R. Wang, and Y. Shen, "A cost-sensitive logistic regression credit scoring model based on multi-objective optimization approach," *Technol. Econ. Develop. Economy*, vol. 26, no. 2, pp. 405–429, Nov. 2019.

[12] L. Zhang, H. Ray, J. Priestley, and S. Tan, "A descriptive study of variable discretization and cost-sensitive logistic regression on imbalanced credit data," *J. Appl. Statist.*, vol. 47, no. 3, pp. 568–581, Feb. 2020.

[13] H. Wang, G. Kou, and Y. Peng, "Multi-class misclassification cost matrix for credit ratings in peer-to-peer lending," *J. Oper. Res. Soc.*, vol. 72, no. 3, pp. 923–934, 2021.

[14] T. Harris, "Credit scoring using the clustered support vector machine," *Expert Syst. Appl.*, vol. 42, no. 2, pp. 741–750, Feb. 2015.

[15] H. Xiao, Z. Xiao, and Y. Wang, "Ensemble classification based on supervised clustering for credit scoring," *Appl. Soft Comput.*, vol. 43, pp. 73–86, Jun. 2016.

[16] L. Abid, A. Masmoudi, and S. Zouari-Ghorbel, "The consumer Loan's payment default predictive model: An application of the logistic regression and the discriminant analysis in a tunisian commercial bank," *J. Knowl. Economy*, vol. 9, no. 3, pp. 948–962, Sep. 2018.

[17] M. Ala'raj, M. Abbod, and M. Radi, "The applicability of credit scoring models in emerging economies: An evidence from jordan," *Int. J. Islamic Middle Eastern Finance Manage.*, vol. 11, no. 4, pp. 608–630, Nov. 2018.

[18] J. Nalić and G. Martinovic, "Building a credit scoring model based on data mining approaches," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 30, no. 2, pp. 147–169, Feb. 2020.

[19] A. Chopra and P. Bhilare, "Application of ensemble models in credit scoring models," *Bus. Perspect. Res.*, vol. 6, no. 2, pp. 129–141, Jul. 2018.

[20] Z. Zhao, S. Xu, B. H. Kang, M. M. J. Kabir, Y. Liu, and R. Wasinger, "Investigation and improvement of multi-layer perceptron neural networks for credit scoring," *Expert Syst. Appl.*, vol. 42, no. 7, pp. 3508–3516, May 2015.

[21] A. Dželihodžić, D. Đonko, and J. Kevrić, "Improved credit scoring model based on bagging neural network," *Int. J. Inf. Technol. Decis. Making*, vol. 17, no. 6, pp. 1725–1741, Nov. 2018.

[22] K. Bastani, E. Asgari, and H. Namavari, "Wide and deep learning for peer-to-peer lending," *Expert Syst. Appl.*, vol. 134, pp. 209–224, Nov. 2019.

[23] G. Chi, M. S. Uddin, M. Z. Abedin, and K. Yuan, "Hybrid model for credit risk prediction: An application of neural network approaches," *Int. J. Artif. Intell. Tools*, vol. 28, no. 5, Aug. 2019, Art. no. 1950017.

[24] L. Munkhdalai, J. Y. Lee, and K. H. Ryu, *A Hybrid Credit Scoring Model Using Neural Networks and Logistic Regression*. Singapore: Springer, 2020, pp. 251–258.

[25] L. Ma, X. Zhao, Z. Zhou, and Y. Liu, "A new aspect on P2P online lending default prediction using meta-level phone usage data in China," *Decis. Support Syst.*, vol. 111, pp. 60–71, Jul. 2018.

[26] L.-J. Kao, C.-C. Chiu, and F.-Y. Chiu, "A Bayesian latent variable model with classification and regression tree approach for behavior and credit scoring," *Knowl.-Based Syst.*, vol. 36, pp. 245–252, Dec. 2012.

[27] S. F. Eletter and S. G. Yaseen, "Loan decision models for the jordanian commercial banks," *Global Bus. Econ. Rev.*, vol. 19, no. 3, p. 323, 2017.

[28] X. Feng, Z. Xiao, B. Zhong, J. Qiu, and Y. Dong, "Dynamic ensemble classification for credit scoring using soft probability," *Appl. Soft Comput.*, vol. 65, pp. 139–151, Apr. 2018.
[29] X. Ye, L.-A. Dong, and D. Ma, "Loan evaluation in P2P lending based on random forest optimized by genetic algorithm with profit score," *Electron. Commerce Res. Appl.*, vol. 32, pp. 23–36, Nov. 2018.
[30] S. F. Eletter, S. G. Yaseen, and G. A. Elrefae, "Neuro-based artificial intelligence model for loan decisions," *Amer. J. Econ. Bus. Admin.*, vol. 2, no. 1, pp. 27–34, Jan. 2010.
[31] C.-F. Tsai, "Combining cluster analysis with classifier ensembles to predict financial distress," *Inf. Fusion*, vol. 16, pp. 46–58, Mar. 2014.
[32] J. Zou, J. Zhang, and P. Jiang, "Credit card fraud detection using autoencoder neural network," 2019, *arXiv:1908.11553*. [Online]. Available: http://arxiv.org/abs/1908.11553
[33] M. Rezapour, "Anomaly detection using unsupervised methods: Credit card fraud case study," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 11, pp. 1–8, 2019.
[34] Y. Lu, X. Leng, K. Xu, W. Luan, W. Yang, and J. Li, "Outlier detection for power data based on contractive auto-encoder," in *Proc. Int. Conf. Adv. Inf. Sci. Syst.*, Singapore. New York, NY, USA: ACM, Nov. 2019, Art. no. 23.
[35] Y. Xia, "A novel reject inference model using outlier detection and gradient boosting technique in peer-to-peer lending," *IEEE Access*, vol. 7, pp. 92893–92907, 2019.
[36] J. Chen, S. Sathe, C. Aggarwal, and D. Turaga, "Outlier detection with autoencoder ensembles," in *Proc. SIAM Int. Conf. Data Mining*, 2017, pp. 90–98.
[37] C. X. Ling and V. S. Sheng, "Cost-sensitive learning," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA, USA: Springer, 10, pp. 231–235.
[38] Y. Xia, C. Liu, and N. Liu, "Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending," *Electron. Commerce Res. Appl.*, vol. 24, pp. 30–49, Jul. 2017.
[39] R. Alejo, V. Garćia, A. I. Marqúes, J. S. Śanchez, and J. A. Antonio-Veĺazquez, "Making accurate credit risk predictions with cost-sensitive MLP neural networks," in *Advances in Intelligent Systems and Computing* (Management Intelligent Systems), vol. 220. Berlin, Germany: Springer, 2013, pp. 1–8, doi: 10.1007/978-3-319-00569-0_1.
[40] F. N. Koutanaei, H. Sajedi, and M. Khanbabaei, "A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring," *J. Retailing Consum. Services*, vol. 27, pp. 11–23, Nov. 2015.
[41] J. Abellán and J. G. Castellano, "A comparative study on base classifiers in ensemble methods for credit scoring," *Expert Syst. Appl.*, vol. 73, pp. 1–10, May 2017.
[42] A. De Caigny, K. Coussement, and K. W. De Bock, "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees," *Eur. J. Oper. Res.*, vol. 269, no. 2, pp. 760–772, Sep. 2018.
[43] M. Zięba, S. K. Tomczak, and J. M. Tomczak, "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction," *Expert Syst. Appl.*, vol. 58, pp. 93–101, Oct. 2016.
[44] Y. Xia, C. Liu, Y. Li, and N. Liu, "A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring," *Expert Syst. Appl.*, vol. 78, pp. 225–241, Jul. 2017.
[45] A. C. Bahnsen, D. Aouada, and B. Ottersten, "Example-dependent cost-sensitive decision trees," *Expert Syst. Appl.*, vol. 42, no. 19, pp. 6609–6619, Nov. 2015.
[46] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. New York, NY, USA: Springer, 2013.
[47] X. Wang, S. Matwin, N. Japkowicz, and X. Liu, "Cost-sensitive boosting algorithms for imbalanced multi-instance datasets," in *Advances in Artificial Intelligence* (Lecture Notes in Computer Science), vol. 7884. Berlin, Germany: Springer, 2013, pp. 174–186, doi: 10.1007/978-3-642-38457-8_15.
[48] G. Paleologo, A. Elisseeff, and G. Antonini, "Subagging for credit scoring models," *Eur. J. Oper. Res.*, vol. 201, no. 2, pp. 490–499, Mar. 2010.
[49] C. Luo, "A comparison analysis for credit scoring using bagging ensembles," *Expert Syst.*, Jun. 2018, Art. no. e12297. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/exsy.12297
[50] S. Finlay, "Multiple classifier architectures and their application to credit risk assessment," *Eur. J. Oper. Res.*, vol. 210, no. 2, pp. 368–378, Apr. 2011.
[51] C.-F. Tsai, Y.-F. Hsu, and D. C. Yen, "A comparative study of classifier ensembles for bankruptcy prediction," *Appl. Soft Comput.*, vol. 24, pp. 977–984, Nov. 2014.
[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[53] H. T. Kam, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, vol. 1995, vol. 1, pp. 278–282.
[54] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
[55] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.
[56] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.
[57] R. Zhang and Z. Qiu, "Optimizing hyper-parameters of neural networks with swarm intelligence: A novel framework for credit scoring," *PLoS ONE*, vol. 15, no. 6, Jun. 2020, Art. no. e0234254.
[58] D. West, "Neural network credit scoring models," *Comput. Oper. Res.*, vol. 27, nos. 11–12, pp. 1131–1152, Sep. 2000.
[59] K. B. Schebesch and R. Stecking, "Support vector machines for credit scoring: Extension to non standard cases," in *Innovations in Classification, Data Science, and Information Systems*. Berlin, Germany: Springer, 2005, pp. 498–505.
[60] M. Malekipirbazari and V. Aksakalli, "Risk assessment in social lending via random forests," *Expert Syst. Appl.*, vol. 42, no. 10, pp. 4621–4631, Jun. 2015.
[61] R. Y. Goh, L. S. Lee, H.-V. Seow, and K. Gopal, "Hybrid harmony search–artificial intelligence models in credit scoring," *Entropy*, vol. 22, no. 9, p. 989, Sep. 2020.
[62] M. Saidi, M. E. H. Daho, N. Settouti, and M. E. A. Bechar, "Comparaison of ensemble cost sensitive algorithms: Application to credit scoring prediction," in *Proc. ICAASE*, 2018, pp. 56–61.
[63] Y. Song, Y. Wang, X. Ye, D. Wang, Y. Yin, and Y. Wang, "Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in P2P lending," *Inf. Sci.*, vol. 525, pp. 182–204, Jul. 2020.
[64] V. Moscato, A. Picariello, and G. Sperlí, "A benchmark of machine learning approaches for credit score prediction," *Expert Syst. Appl.*, vol. 165, Mar. 2021, 113986.
[65] R. F. Malik and H. Hermawan, "Credit scoring using classification and regression tree (CART) algorithm and binary particle swarm optimization," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 6, pp. 5425–5431, 2018.
[66] Y. Xia, C. Liu, B. Da, and F. Xie, "A novel heterogeneous ensemble credit scoring model based on bstacking approach," *Expert Syst. Appl.*, vol. 93, pp. 182–199, Mar. 2018.
[67] Y. Guo, J. He, L. Xu, and W. Liu, "A novel multi-objective particle swarm optimization for comprehensible credit scoring," *Soft Comput.*, vol. 23, no. 18, pp. 9009–9023, Sep. 2019.
[68] Y. Xia, J. Zhao, L. He, Y. Li, and M. Niu, "A novel tree-based dynamic heterogeneous ensemble method for credit scoring," *Expert Syst. Appl.*, vol. 159, Nov. 2020, Art. no. 113615.
[69] W. Liu, H. Fan, and M. Xia, "Step-wise multi-grained augmented gradient boosting decision trees for credit scoring," *Eng. Appl. Artif. Intell.*, vol. 97, Jan. 2021, Art. no. 104036.
[70] V. García, A. I. Marqués, and J. S. Sánchez, "Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction," *Inf. Fusion*, vol. 47, pp. 88–101, May 2019.
[71] H. Mansourifar, L. Chen, and W. Shi, "Virtual big data for GAN based data augmentation," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Los Angeles, CA, USADec. 2019, pp. 1478–1487.
[72] W. Zhang, D. Yang, S. Zhang, J. H. Ablanedo-Rosas, X. Wu, and Y. Lou, "A novel multi-stage ensemble model with enhanced outlier adaptation for credit scoring," *Expert Syst. Appl.*, vol. 165, Mar. 2021, Art. no. 113872.
[73] A. Namvar, M. Siami, F. Rabhi, and M. Naderpour, "Credit risk prediction in an imbalanced social lending environment," *Int. J. Comput. Intell. Syst.*, vol. 11, no. 1, pp. 925–935, 2018.
[74] Y. Xia, L. He, Y. Li, N. Liu, and Y. Ding, "Predicting loan default in peer-to-peer lending using narrative data," *J. Forecasting*, vol. 39, no. 2, pp. 260–280, Mar. 2020.

[75] M. J. Ariza-Garzon, J. Arroyo, A. Caparrini, and M.-J. Segovia-Vargas, "Explainability of a machine learning granting scoring model in peer-to-peer lending," *IEEE Access*, vol. 8, pp. 64873–64890, 2020.
[76] O. Loyola-González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, and M. García-Borroto, "Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases," *Neurocomputing*, vol. 175, pp. 935–947, Jan. 2016.
[77] G. Nie, W. Rowe, L. Zhang, Y. Tian, and Y. Shi, "Credit card churn forecasting by logistic regression and decision tree," *Expert Syst. Appl.*, vol. 38, no. 12, pp. 15273–15285, Nov. 2011.
[78] S. Mundra, A. Mundra, A. Sharma, A. Mohini, and A. Yadav, "Analyzing credit defaulter behavior for precise credit scoring," *Int. J. Adv. Sci. Technol.*, vol. 28, no. 12, pp. 247–255, 2019.
[79] D. Tripathi, D. R. Edla, R. Cheruku, and V. Kuppili, "A novel hybrid credit scoring model based on ensemble feature selection and multilayer ensemble classification," *Comput. Intell.*, vol. 35, no. 2, pp. 371–394, May 2019.
[80] A. Terko, E. Zunic, D. Donko, and A. Dzelihodzic, "Credit scoring model implementation in a microfinance context," in *Proc. 27th Int. Conf. Inf., Commun. Autom. Technol. (ICAT)*, Oct. 2019, pp. 1–6.

WIROT YOTSAWAT was born in Songkhla, Thailand, in 1984. He received the B.Sc. degree in computer science from the School of Informatics, Walailak University, Thailand, in 2007, and the M.Sc. degree in computer science from the Faculty of Science, Kasetsart University, Thailand, in 2014, where he is currently pursuing the Ph.D. degree in computer science.

His research interests include data mining, machine learning, and image processing.

PAKAKET WATTUYA received the B.Sc. degree in computer science from the Faculty of Science, Kasetsart University, Thailand, in 2000, the M.Eng. degree in computer engineering from the Faculty of Engineering, Kasetsart University, in 2004, and the Dr. rer. nat. degree in computer science from Westfälische Wilhelms-Universität Muenster, Germany, in 2010.

She is currently an Assistant Professor with the Department of Computer Science, Kasetsart University. Her research interests include image processing, computer vision, machine learning, and deep learning.

ANONGNART SRIVIHOK received the B.Sc. degree in microbiology from the Faculty of Science, Chulalongkorn University, Thailand, in 1978, the M.S. degree in engineering sci-computer science from the University of Mississippi, USA, in 1984, and the Ph.D. degree in information systems from Central Queensland University, Australia, in 1998.

She has been an Associate Professor with the Department of Computer Science, Kasetsart University. Her research interests include data mining, machine learning, decision support systems, and knowledge management.