Applied Soft Computing Journal 98 (2021) 106852

Contents lists available at ScienceDirect

Applied Soft Computing Journal

journal homepage: www.elsevier.com/locate/asoc

A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique

Feng Shen [a,b,*], Xingchao Zhao [a], Gang Kou [c,**], Fawaz E. Alsaadi [d]

[a] School of Finance, Southwestern University of Finance and Economics, Chengdu 611130, PR China
[b] Fintech Innovation Center, Southwestern University of Finance and Economics, Chengdu 611130, PR China
[c] School of Business Administration, Southwestern University of Finance and Economics, Chengdu 611130, PR China
[d] Department of information Technology, Faculty of Computing and IT, King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 19 November 2019
Received in revised form 14 October 2020
Accepted 22 October 2020
Available online 28 October 2020

Keywords:
Risk management
Credit risk evaluation
Improved SMOTE
LSTM network
Ensemble model

ABSTRACT

In recent years, research has found that in many credit risk evaluation domains, deep learning is superior to traditional machine learning methods and classifier ensembles perform significantly better than single classifiers. However, credit evaluation model based on deep learning ensemble algorithm has rarely been studied. Moreover, credit data imbalance still challenges the performance of credit scoring models. Therefore, to go some way to filling this research gap, this study developed a new deep learning ensemble credit risk evaluation model to deal with imbalanced credit data. First, an improved synthetic minority oversampling technique (SMOTE) method was developed to overcome known SMOTE shortcomings, after which a new deep learning ensemble classification method combined with the long-short-term-memory (LSTM) network and the adaptive boosting (AdaBoost) algorithm was developed to train and learn the processed credit data. Then, area under the curve (AUC), the Kolmogorov–Smirnov (KS) and the non-parametric Wilcoxon test were employed to compare the performance of the proposed model and other widely used credit scoring models on two imbalanced credit datasets. The experimental test results indicated that the proposed deep learning ensemble model was generally more competitive when addressing imbalanced credit risk evaluation problems than other models.

1. Introduction

Since the sub-prime mortgage crisis in 2008 and the release of the Basel III Accord, credit risk management has become a key concern at financial institutions, with many quantitative credit risk evaluation models having been developed to accompany Fintech developments to avoid financial losses. In recent years, machine learning and data mining techniques have been introduced to improve financial decision-making prediction accuracy, which has lowered credit analysis costs, enabled faster loan decisions, guaranteed payment collections, and provided adequate risk mitigation; therefore, efficient, accurate credit risk evaluation models are vital for banks and financial institutions as any small improvements can return greater profits [1].

Credit risk evaluations can essentially be treated as binary classification problems. To construct the classification model, the observed socio-economic variables or attributes are first computed, after which the new credit applicants are processed using the already trained model and then categorized into predefined credit classes. In the past, the most dominant statistical methods have been discriminant analysis and logistic regression; however, the comprehensive comparisons between discriminant analysis and logistic regression models in [2] indicated that these statistical approaches may have estimation bias due to violations in the underlying assumptions, such as the independence of the explanatory variables. Because machine learning has been found to have good data classification performances, machine learning techniques such as artificial neural networks (ANN) [3], support vector machines (SVM) [4], and classification and regression trees (CART) [5] have become more common, especially in increasingly complex credit risk systems; some widely used comparisons are given in [6]. While machine learning techniques significantly outperform traditional statistical models, several previous studies have argued that no single classifier produces the best results in all cases. The latest developments in [7] and [8] used hybrid and ensemble models to address credit risk evaluation problems, and demonstrated that these advanced models were more accurate and more able to reduce costs and increase efficiency compared to traditional individual classification methods.

[*] Corresponding author at: School of Finance, Southwestern University of Finance and Economics, Chengdu 611130, PR China.
[**] Corresponding author.
E-mail addresses: shenfeng1213@gmail.com (F. Shen), kougang@swufe.edu.cn (G. Kou).

https://doi.org/10.1016/j.asoc.2020.106852
1568-4946/© 2020 Published by Elsevier B.V.

However, when faced with big data for credit scoring applications, traditional statistical and machine learning methods have been found to be difficult to reveal the complex relationship between credit data variables. Therefore, deep learning methods have been applied to predict financial market prediction problems (see the review of Omer et al. [9]) such as long short-term memory (LSTM) network. The LSTM network, which has three special multiple cell gates, has been proven to be able to effectively mine the interrelationships between credit data variables; for example, a good education often indicates a good company job, which further guarantees a higher customer income. More recently, Zhang et al. [10] applied a single LSTM network to peer-to-peer lending credit scoring evaluations, with the results indicating a higher overdue credit classification accuracy than statistical or machine learning models. Wang et al. [11] used borrower online operating behavior data and proposed a consumer credit scoring method based on an attention mechanism LSTM, and found that the proposed solution effectively increases the predictive accuracy. Therefore, because of its strong learning abilities, in this paper, an LSTM network was employed as the base classifier.

Compared with traditional statistical and classical machine learning models, significant improvements have been observed in the financial forecasting applications of deep learning technology [12], however, deep learning technology has not been extensively applied to credit scoring. It has been argued that its performance may be affected by data scale and data imbalance [13]. Therefore, how to further improve the performance of credit scoring model based on deep learning technology is an important research problem.

Previous research has identified credit scoring as a class imbalance problem with uneven sample distributions [14], in which the number of non-risky applicants is usually much larger than that of credit defaulters. Even though machine learning and data mining approaches have been applied to commercial decision-making, class imbalance problems still pose significant challenges to classification model validity [15]. There have been three main solutions offered for imbalanced data classification problems: data-level solutions, cost-sensitive learning solutions, and ensemble solutions [16].

Random oversampling (ROS), random under-sampling (RUS), and the SMOTE algorithm [17] have been widely used to rebalance imbalanced datasets for data-level solutions. However, as the ROS and RUS methods rebalance the original data by either oversampling the minority samples or removing the majority samples, overfitting or underfitting can occur. Therefore, compared to the ROS and RUS methods, the SMOTE technique has proven superior for imbalanced classification problems. However, there are also obvious limitations to the SMOTE technique as it creates the minority samples by randomly selecting the nearest neighbors, which can generate a significant amount of noise. Further, the Euclidean distance in the original SMOTE algorithm does not consider the relationship between various variables, and it could be affected by the correlation between the variables. To overcome the SMOTE shortcomings, Abdi and Hashemi [18] proposed a novel Mahalanobis distance-based oversampling (MDO) technique that generates synthetic samples for the minority classes without any class decomposition, with the samples obtained through the MDO maintaining the same Mahalanobis distance from the corresponding class mean. To some extent, the MDO method resolved the nearest neighbor random selection strategy problem in the SMOTE technique; however, because the MDO focuses on the selection of the nearest neighbors and considers each minority sample to be appropriate for generating new minority samples, it could be argued that extra boundary minority samples could be created using the MDO method. As the boundary between defaulters and non-risky applicants in credit scoring problems commonly overlaps and the outliers in the imbalanced credit data further challenge the effectiveness of traditional data-level based solutions, the overlapping and abnormal samples generated by the SMOTE or MDO methods could increase the difficulty of classification.

Besides data-level solutions, cost-sensitive learning solutions have been applied to deal with data imbalance problems. Cost-sensitive learning removes the fitting bias on the majority classes by modifying the learners' loss function, with the minority misclassification costs typically being greater than the majority misclassification costs. Xia et al. [19] proposed a cost-sensitive boosted tree loan evaluation model that incorporated cost-sensitive learning and extreme gradient boosting (XGBoost) to enhance the discrimination capabilities for potential default borrowers, which proved superior in addressing the imbalance problem. Marcin et al. [20] introduced a Cost-sensitive Global Model Tree (CGMT) that applied a fitness function that minimized the average misprediction costs, finding that the application of a specialized evolutionary algorithm to model the tree induction resulted in significantly more accurate predictions than the other methods. However, cost-sensitive learning solutions for imbalanced classification could encounter difficulties in practical application. First, in many cases, as the misclassification costs in the credit data are difficult to estimate, they are commonly predefined depending on the users' experience, which means that the specific calculations for the misclassification costs are often subjective. Further, the underlying assumptions for cost sensitive learning methods commonly treat the misclassification costs as being stationary; however, this assumption is frequently violated in reality because of the complexity of credit scoring issues; for example, the loan defaulter costs are highly dependent on variables that evolve over time such as interest rates and inflation rates.

More recently, ensemble solutions have become popular. Galar et al. [21] explored several common class-imbalance data classification approaches, and found that ensemble solutions were superior to the other methods, which was also found by Yu et al. [14]. Kim, et al. [22] found that the AdaBoost algorithm was effective in addressing data imbalance problems as it was able to consider sample distributions. Recent studies have found that hybrid methods combined with ensemble algorithms and data-level solutions proved superior for imbalanced data classification. Song et al. [23] evaluated several imbalanced classifiers for credit risk prediction using a multi-criteria decision making (MCDM)-based method, and proved that the SMOTEBoost-based model was more effective for imbalanced data classification than other methods. Chen et al. [24] introduced a novel oversampling procedure in which the sampling probability distribution was reweighted using a boosting algorithm, and found that the ensemble method had a superior performance compared to other benchmark methods. Significant improvements have been achieved with ensemble methods being applied in credit scoring, however, the classification performances of ensemble models could be affected by the accuracy and diversity of the base learners.

Overall, because of the superior credit scoring prediction capabilities of the LSTM network and the superior performances of the combination of data-level solutions and ensemble models for imbalanced classification, this paper proposes a novel deep ensemble learning model based on an improved SMOTE technique for imbalanced credit scoring. Therefore, this paper makes the following research contributions.

First, to improve the performance of the credit scoring model in dealing with unbalanced data, an improved SMOTE method is proposed for imbalanced credit data processing. The original SMOTE technique commonly considers each minority sample as appropriate for the generation of new minority samples, but noisy samples can be synthesized if inappropriate original minority classes are selected. The core idea of the improved SMOTE method is to select appropriate original minority classes rather than employing each minority sample as a candidate. Further, the high linear correlations between the credit sample variables can reduce the effectiveness of the original SMOTE technique. Because of its covariance distribution structure, the Mahalanobis distance has been found to be an effective method for computing the distance between a sample and a set of observations [25]. Therefore, it is employed in this study to choose the appropriate minority class samples from Mahalanobis distance. Then, the selected minority class candidates are used to synthesize the new samples by oversampling and k-Nearest Neighbor methods. Compared to the SMOTE algorithm, the proposed improved SMOTE method excludes the noise/outliers and the overlapping instances, which improves its imbalanced credit data processing performance.

Second, a deep learning ensemble model is constructed based on the LSTM network and the AdaBoost algorithm. Previous research has found that the performance of ensemble model depends on the accuracy and diversity of the base classifiers. Besides, ensemble credit scoring models based on deep learning algorithms have rarely been studied. To fill this gap, in this study, the LSTM network, which has proven to be superior in identifying the intercorrelations between the credit data variables, is employed as the base classifier for developing the ensemble credit scoring model. The AdaBoost ensemble algorithm is then used to avoid overfitting of the LSTM network and provide more robust and reliable prediction results.

Third, to verify the effectiveness of the proposed method, systematic experiments were designed in this study, and the prediction results of the proposed deep learning ensemble model were comprehensively compared with widely used benchmark methods. Furthermore, a nested cross-validation was performed for all the used classification models to diminish the influence of probability on the results. Lastly, to achieve more robust and convincing conclusions, a non-parametric Wilcoxon statistical significance test was conducted on the proposed model and the benchmark models, and the empirical study indicates that our proposed learning method was more superior to benchmark methods.

The remainder of this paper is organized as follows. Section 2 reviews the relevant SMOTE technique, LSTM neural network, and the AdaBoost algorithm. Section 3 constructs the new deep learning ensemble credit risk evaluation model. Section 4 details the experimental design, including the credit datasets, hyperparameter optimizations, performance measurement metrics, and significance test method. Section 5 analyzes and tests the effectiveness and superiority of the proposed method, and conclusions are given in Section 6.

2. Methodology

2.1. Synthetic minority oversampling technique

SMOTE is a popular and effective method for addressing class imbalance problems in many domains [17]. The core idea of SMOTE is the synthesis of extra minority samples based on the feature space similarities between the existing minority instances. Specifically, given imbalanced data T, for each minority class instance $x_i \in T$, the SMOTE algorithm first finds the K nearest neighbors for x_i using Euclidean distance, after which one of the K nearest neighbors is randomly chosen and the feature vector difference between x_i and its corresponding nearest neighbor calculated. Finally, the feature vector difference multiplies a stochastic number and adds the new vector to x_i. The mathematical formulation for synthesizing a new minority sample is shown in Eq. (1):

$$x_{new} = x_i + \left(x_i^k - x_i\right) \times \delta \tag{1}$$

where x_i^k is one of the nearest neighbors to x_i, and δ is a random value belongs to (0, 1).

Therefore, the synthesized minority instance x_{new} is a point along the line segment joining x_i and its nearest neighbor x_i^k.

2.2. Long–short-term memory network

The LSTM network, which is a special type of recurrent neural network architecture consisting of memory cells and gate units [26], has been found to be superior in mapping interrelationships between variables. LSTM network applications assigned to financial forecasting tasks have achieved excellent performances in recent years [27].

The LSTM network adopted in this study was proposed by Gers et al. [28] and is one of the more popular variants that are based on the typical schematic LSTM diagram using forget gates. An LSTM network maintains the previous state information over long sequences using a core memory cell design C_t, which efficiently allows the gradient to flow over a long time, thereby easing the "vanishing gradient" problem. The input information processed by the input gate i_t and the forget gate f_t flows into memory cell C_t, and the state information, which is regulated by the output gate o_t, then flows to the other LSTM blocks. The mathematical formulation for the memory cell and gate unit results are shown in Eqs. (2)–(6):

At every time step, the combination of the current time input vector x_t and the hidden state h_{t-1} from the previous step are transformed to an LSTM cell unit, and then calculated using a logistic sigmoid function as shown in Eq. (2):

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}C_{t-1} + b_i) \tag{2}$$

where σ denotes the logistic sigmoid function, W_{xi}, W_{hi}, and W_{ci} are the separate weight vectors for each input connecting two components, C_{t-1} is the cell state from the previous step, and b_i is the bias vector for the input gate unit.

The forget gate in LSTM structures defines which information is to be removed from the cell state. The output for the forget gate is calculated using Eq. (3):

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}C_{t-1} + b_f) \tag{3}$$

where b_f denotes the bias for the forget gate, W_{xf}, W_{hf}, and W_{cf} are the separate weight vectors for each input connecting two components. As can be seen in Eq. (3), the sigmoid function produces values between zero and one; if the output value for the forget gate is close to zero, then the previous memory is forgotten, while a value of one indicates that everything stored in the previous memory block is remembered.

The cell state C_t is then updated using Eq. (4):

$$C_t = f_t \odot C_{t-1} + i_t \odot tanh(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}C_{t-1} + b_c) \tag{4}$$

where b_c denotes the bias vector and the operator $\odot$ denotes the Hadamard (element-wise) product.

Finally, the LSTM block output is generated as shown in Eqs. (5) and (6) (see Gers et al. [26] for specific details):

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}C_{t-1} + b_o) \tag{5}$$

$$h_t = o_t \odot tanh(C_t) \tag{6}$$

where b_o denotes the bias for the output gate, W_{xo}, W_{ho}, and W_{co} are the separate weight vectors for each input connecting two components.

2.3. AdaBoost algorithm

As one of the most widely used techniques for imbalanced credit data scoring, the AdaBoost algorithm has been studied extensively [29]. The basic idea of AdaBoost is to repeatedly train a weak learner and at the same time adaptively update the training dataset weights. At the beginning, all training samples are assigned a uniform weight, and based on the performance of the base learner, at the end of a training epoch, the AdaBoost algorithm reweights the instances, where the weights of the correctly classified instances are lessened, and the weights of the incorrectly classified instances are increased. From this process, a sequence of base learners endowed with different voting weights is obtained for a predefined number of iterations. The final ensemble model is then formatted using a linear combination of the voting weights and the trained weak learners.

3. Proposed deep learning ensemble model with improved SMOTE approach

3.1. Improved SMOTE approach for imbalanced credit data processing

To improve the SMOTE's algorithmic performances, the Euclidean distance used in the classical SMOTE technique was combined with the Mahalanobis distance to synthesize the minority samples. The Mahalanobis distance has been found to be an effective method for computing the distance between one sample and a set of observations because of its covariance distribution structure [25]. Unlike the Euclidean distance, the Mahalanobis distance places importance on the linear correlations between the random variables (for example, the job position correlating with salary level in credit data), and the distance measured by the Mahalanobis distance does not suffer from any correlation disturbances between the variables.

The core idea in the proposed improved method is as follows. First the Mahalanobis distance between the minority samples and the majority class distribution is computed, and the minority instances with distances larger than the predefined Mahalanobis distance level are selected for consideration. To find the most similar, nearest neighbors for the selected minority samples, the Euclidean distance method is then applied to calculate the distances between the chosen minority samples. Finally, the new minority instances are generated using Eq. (1).

Given a credit dataset T = $\{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$, with the C_{min} and C_{maj} that belong to T respectively representing the minority class partition and the majority class partition. To determine each minority instance of $x_i \in C_{min}$, the similarity between x_i and C_{maj} is first calculated using Eq. (7):

$$D_m\left(x_i, C_{maj}\right) = \sqrt{\left(x_i - \mu\right)^T S^{-1}\left(x_i - \mu\right)}, \tag{7}$$

where S is the covariance matrix, and μ is the mean of the distribution for C_{maj}.

Then, as is shown in Eq. (8), the minority samples with Mahalanobis distance that has fallen into the predefined bias interval are chosen for the new minority sample generation:

$$\varphi_1 \le D_m\left(x_i, C_{maj}\right) \le \varphi_2, \tag{8}$$

where φ_1 and φ_2 are the predefined distance bias intervals for $[min(D_m), max(D_m))$. Note that if the bias φ_1 equals $min(D_m)$ and φ_2 reaches $max(D_m)$, then the improved SMOTE method proposed in this study is as the same as the standard SMOTE algorithm.

As the proper settings for the threshold φ_1 and φ_2 may be different for specific algorithms, in this study, a grid search method is employed to obtain the optimal threshold settings. Then, the Euclidean distance is applied to find the nearest neighbors for the selected samples. Consider two N-dimensional samples; $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,N})$ and $x_j = (x_{j,1}, x_{j,2}, \ldots, x_{j,N})$; the Euclidean distance between x_i and x_j is given in Eq. (9):

$$D_e\left(x_i, x_j\right) = \sqrt{\sum_{n=1}^{N}(x_{i,n} - x_{j,n})^2}, \tag{9}$$

where $\varphi_1 \le D_m\left(x_i, C_{maj}\right) \le \varphi_2$, and $\varphi_1 \le D_m\left(x_j, C_{maj}\right) \le \varphi_2$.

The rest of the procedure for synthesizing the new minority instances is the same as for the SMOTE algorithm described in Section 2.1. Fig. 1 shows the differences between the classic SMOTE method and the proposed improved SMOTE method.

The blue balls, red balls, and green triangles respectively represent the original majority classes, the original minority classes, and the synthetic minority classes. In Fig. 1(b), there is a large quantity of noise because of the extra boundary samples generated from the SMOTE technique's random synthesis strategy. However, as shown in Fig. 1(c), the improved SMOTE method selectively generates the minority classes that could possibly enhance the classification algorithmic performances.

3.2. Integrating the multiple LSTM classifiers into a strong ensemble output

The AdaBoost algorithm is employed to construct a high-quality ensemble system that sequentially combines the diverse LSTM based networks and endows the well-performing weaker classifiers with larger voting weights.

The main steps for developing the proposed deep learning ensemble classification model are as follows.

Given a credit dataset that consists of n training samples, S = $\{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_n, y_n)\}$, x_i is the input credit applicant instance feature space and y_i is the credit risk evaluation label (good/bad credit conditions). LSTM networks are then used as the base learners in the ensemble model, with the output prediction being the base learner ensemble, which is denoted F = $\{L_1, L_2 \ldots, L_M\}$. Each base classifier was trained on the training dataset as described in Section 2.2.

For simplification purposes, without a loss of generality, suppose that the weight distribution over these samples at the mth boosting iteration is denoted D_m, which is initially endowed with an identical value 1/n at the first iteration, with the total prediction error for the current weak classifier over the training data being calculated using Eq. (10):

$$\varepsilon_m = \sum_{1}^{n} D_m(i) \times \begin{cases} 1 \; if \; L_m(x_i) \neq y_i \\ 0 \; if \; L_m(x_i) = y_i, \end{cases} \tag{10}$$

where y_i is the observed label for input sample x_i, ε_m is the classification error for the current classifier, and L_m is the trained LSTM network at iteration m.

Then, the training dataset weight distribution is updated based on the classification performance of the current hypothesis so that the correctly classified samples are assigned lower weights and the misclassified samples are assigned higher weights. The formula for the updating process is shown in Eq. (11):

$$D_{m+1}(i) = \frac{D_m(i)}{Z_m} \exp\left(-\partial_m \times y_i \times L_m(x_i)\right), \tag{11}$$

where Z_m is a normalization constant that ensures that the weights $D_{m+1}(i)$ have a proper distribution, and ∂_m is the voting weight for the trained classifier L_m. The mathematical expressions for Z_m and ∂_m are shown in Eqs. (12) and (13):

$$Z_m = \sum_{1}^{n} D_m(i) \exp(-\partial_m \times y_i \times L_m(x_i)) \tag{12}$$

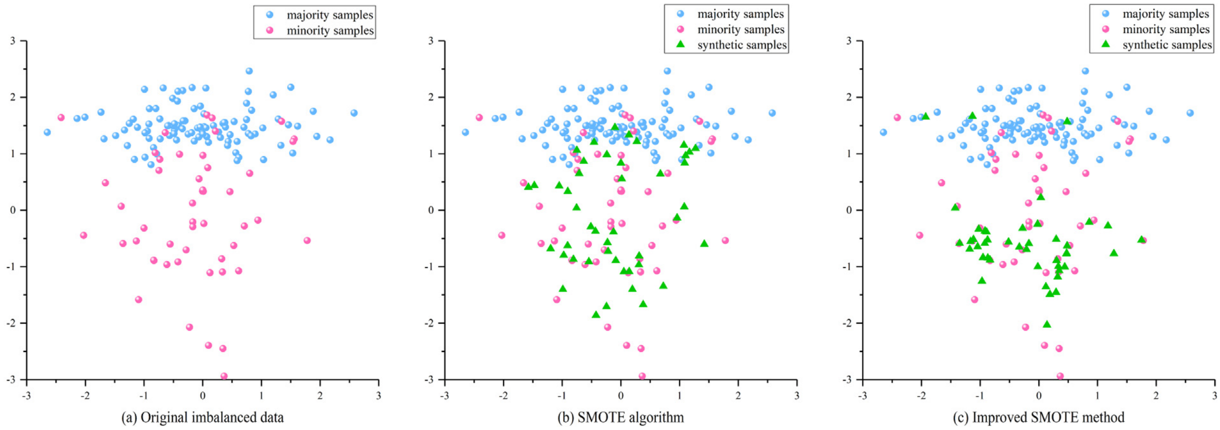


Fig. 1. Sketches for the classic SMOTE method and the improved SMOTE method.

$$\partial_m = \frac{1}{2}\ln\left(\frac{1-\varepsilon_m}{\varepsilon_m}\right). \tag{13}$$

When M iterations are processed, the ensemble is composed of M weak classifiers. As is formulated in Eq. (14), the final AdaBoost classification result is a combination of the classification results weighted by ∂_m:

$$F(x) = \text{sign}\left(\sum_{1}^{M} \partial_m \times L_m(x)\right), \tag{14}$$

where function *sign*(x) represents a sign function, the formulation for which is shown in Eq. (15):

$$sign(x) = \begin{cases} 1, & if\ x > 0 \\ 0, & if\ x = 0 \\ -1, & if\ x < 0. \end{cases} \tag{15}$$

Pseudo code of the deep learning ensemble model is shown in Fig. 2.

The schematic illustration for this ensemble procedure is shown in Fig. 3. The credit dataset is split into a training and a test subset in the first data processing stage and the improved SMOTE method is applied to synthesize the new minority class samples, after which the LSTM networks are trained over the rebalanced training subset and integrated with the AdaBoost algorithm. Finally, the predictions obtained from the base learners for the test subset are aggregated using a weighted voting method and the final classification results determined using the proposed deep learning ensemble model.

4. Empirical study

The main objective of this study was to examine the performance of the improved SMOTE-based deep learning ensemble credit risk evaluation model. Therefore, in this section, the used credit datasets are first detailed, each of which has a different imbalance ratio, Sections 4.2 and 4.3 describe the implementation process for all the classification models in this study, and Sections 4.4 and 4.5 respectively outline the performance evaluation metrics and the statistical significance test method. Finally, Section 4.6 designs the experiment and proposes the research questions.

4.1. Credit dataset

Two credit datasets with different imbalance ratios were employed in this study to test the performance of the proposed classification method on imbalanced credit data. A brief description of the credit datasets is given in this section.

Table 1
Summary of the two credit datasets.

Dataset	Number of instances	Good credit	Bad credit	Number of features	Imbalance ratio
German data	1000	700	300	24	2.33
Taiwan data	30,000	23,364	6636	24	3.52

This study used a German credit dataset and a Taiwanese personal loan dataset, each of which was obtained from the UCI machine learning repository. These credit datasets were used for three main reasons. First, as an ensemble classification model combined with an improved SMOTE method is developed in this study to evaluate credit risk for real-world loans, all credit risk evaluation focused datasets are easily accessible from the UCI website. Second, as the main aim of this study was to verify the effectiveness of the proposed method in handling imbalanced data, these datasets were naturally imbalanced. Finally, German and Taiwanese credit datasets are often used to test classification model performances, which makes it easier to compare the classification performance of the proposed ensemble model with other benchmark models. The detailed information for these two public datasets can be found in [30], and the summary information is given in Table 1.

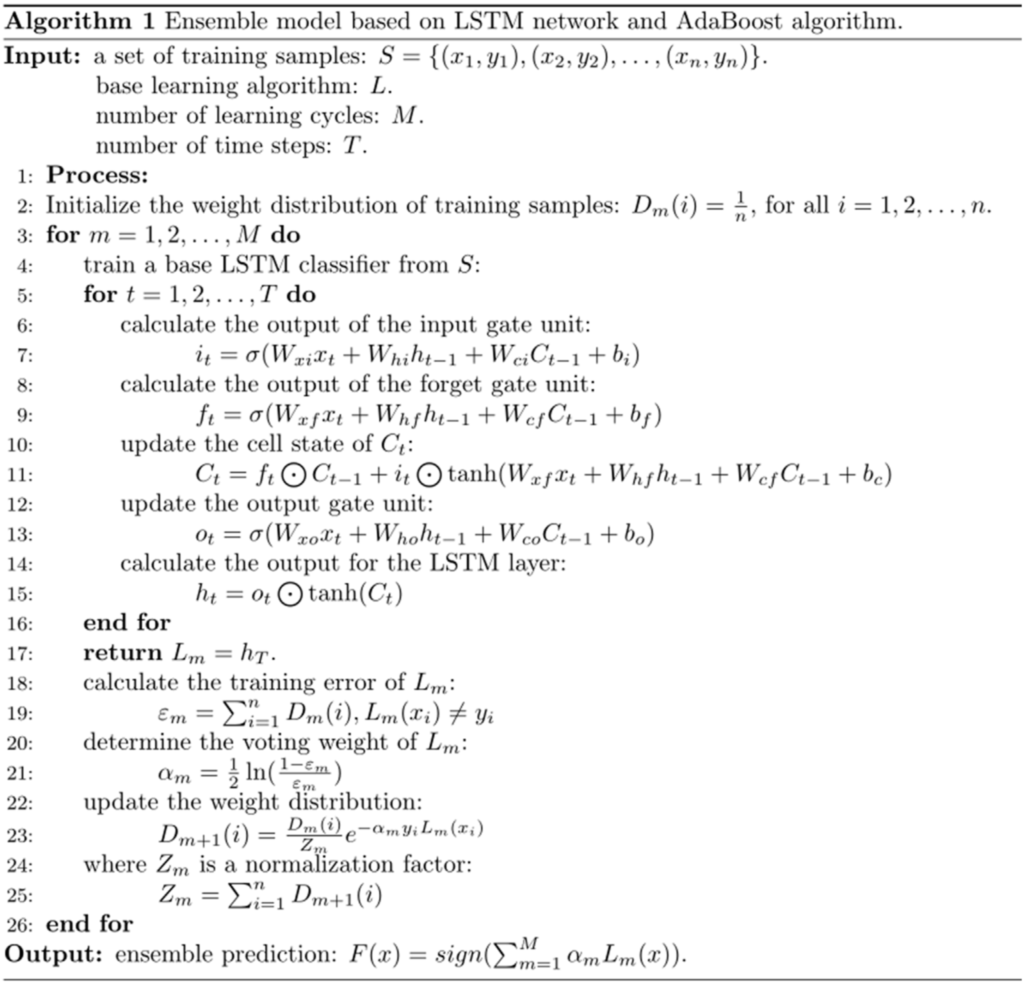
4.2. Benchmark learners

Some widely used classification methods were also employed in this study for comparison: logistic regression (LR), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), k-nearest neighbors (KNN), classification and regression tree (CART), naïve Bayes (NB), support vector machine (SVM), and artificial neural network (ANN). As each of these methods have shown significant success, there have been several comparative studies [6].

In this paper, all the above benchmark models were completed using the Scikit-learn toolbox [31], and "Keras", a deep learning package based on TensorFlow, was used as the deep learning framework for the LSTM network.

4.3. Hyper-parameter optimization

To ensure the effectiveness and comparability of the experiment, it is necessary to carefully set the hyper-parameters for each classifier. If decision-maker use traditional cross-validation methods to estimate the model hyper-parameters and then uses those hyper-parameters to fit a model to the whole dataset, this approach is likely to be biased, and there is the possibility



```
Algorithm 1 Ensemble model based on LSTM network and AdaBoost algorithm.
Input: a set of training samples: $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$.
       base learning algorithm: $L$.
       number of learning cycles: $M$.
       number of time steps: $T$.
Process:
Initialize the weight distribution of training samples: $D_m(i) = \frac{1}{n}$, for all $i = 1, 2, \ldots, n$.
for $m = 1, 2, \ldots, M$ do
    train a base LSTM classifier from $S$:
    for $t = 1, 2, \ldots, T$ do
        calculate the output of the input gate unit:
            $i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}C_{t-1} + b_i)$
        calculate the output of the forget gate unit:
            $f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}C_{t-1} + b_f)$
        update the cell state of $C_t$:
            $C_t = f_t \bigodot C_{t-1} + i_t \bigodot \tanh(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}C_{t-1} + b_c)$
        update the output gate unit:
            $o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}C_{t-1} + b_o)$
        calculate the output for the LSTM layer:
            $h_t = o_t \bigodot \tanh(C_t)$
    end for
    return $L_m = h_T$.
    calculate the training error of $L_m$:
        $\varepsilon_m = \sum_{i=1}^{n} D_m(i), L_m(x_i) \neq y_i$
    determine the voting weight of $L_m$:
        $\alpha_m = \frac{1}{2}\ln(\frac{1-\varepsilon_m}{\varepsilon_m})$
    update the weight distribution:
        $D_{m+1}(i) = \frac{D_m(i)}{Z_m} e^{-\alpha_m y_i L_m(x_i)}$
    where $Z_m$ is a normalization factor:
        $Z_m = \sum_{i=1}^{n} D_{m+1}(i)$
end for
Output: ensemble prediction: $F(x) = sign(\sum_{m=1}^{M} \alpha_m L_m(x))$.
```

Fig. 2. Pseudo code of the proposed deep learning ensemble model.

of over-fitting the model selection criterion. To avoid this bias, model selection must be treated as an integral part of the model fitting process and performed afresh each time the model is fitted to new sample data [32]. Therefore, as suggested in [33] and [34], a nested cross-validation is performed to choose the hyper-parameters and estimate the resulting model performance. As shown in Fig. 4, the $k \times n$-fold nested cross-validation method involves an outer k-fold cross-validation loop in which the training subset is used to search and optimize the hyper-parameters and the testing subset is used to estimate the performance of the classification models. At the same time, a grid search method that exhaustively explores all parameter combinations advances into the inner n-fold cross-validation loop taking one fold as the verification set and the other folds as the training set. Finally, the test fold samples in the outer loop are classified by the optimal classifier, which performs best in the inner loop cross-validation.

To fairly compare the proposed method with some widely used models and achieve convincing results, all algorithms employed in this study were independently optimized for each credit dataset, and all classification models were performed for six times to diminish the influence of probability on the results. Table 2 summarizes the search space for the algorithms involved in this study.

Table 2
Search space of the hyper-parameters for the used algorithms in experimental study.

Algorithm	Search space
LR	C ∈ [0.1,1.0]; penalty ∈ {'L1','L2'}
LDA	Shrinkage ∈ [0,1]
QDA	Regularization parameter ∈ [0,1]
KNN	Number of nearest neighbors ∈ [1,10]
NB	No hyper-parameters need tuning
CART	Maximum depth ∈ [5,10]; minimum sample leaf ∈ [1,10]
SVM	C ∈ $[2^{-5}, 2^5]$; γ ∈$[2^{-5}, 2^5]$
ANN	Hidden layer sizes ∈ [5,100]; activation ∈ {'relu', 'logistic', 'tanh'}
LSTM	Hidden layer sizes ∈ [5,100]; batch size ∈ $[2^1, 2^{10}]$; epochs ∈ {10, 100}
SMOTE	Number of nearest neighbors ∈ [1,10]
Improved SMOTE	$\varphi_1, \varphi_2 \in [\varphi_{min}, \varphi_{max}]$; number of nearest neighbors ∈ [1,10]

4.4. Evaluation measures

Accuracy is widely employed as a performance evaluation indicator for classification models in credit scoring problems; however, when the datasets are unbalanced, accuracy suffers because of the bias towards the majority class [35]. Therefore, it is critical to choose the appropriate measurement metrics to comprehensively evaluate the model effectiveness and guide the classifier learning. In this study, two widely used metrics for credit scoring problem, the area under the curve (AUC) and the Kolmogorov–Smirnov statistic (KS), were employed to evaluate the credit scoring model performances.

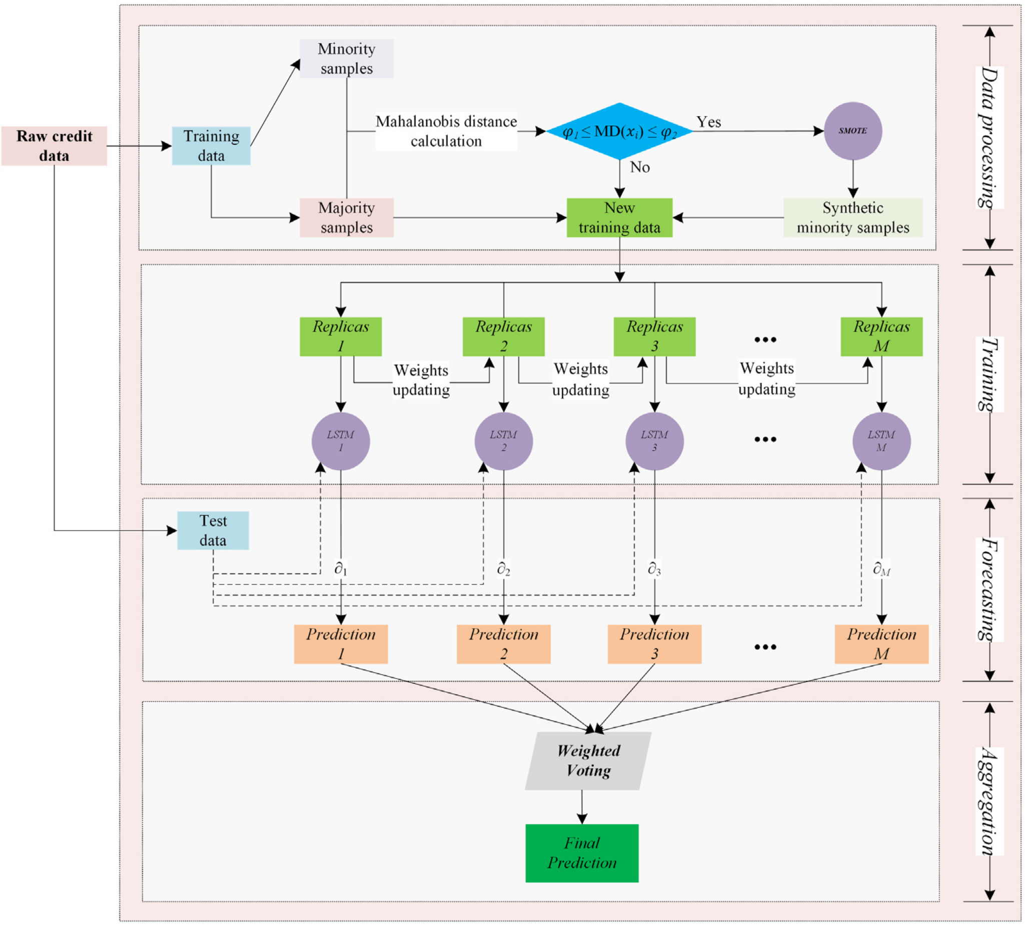


Fig. 3. Flowchart for the proposed deep learning ensemble model.

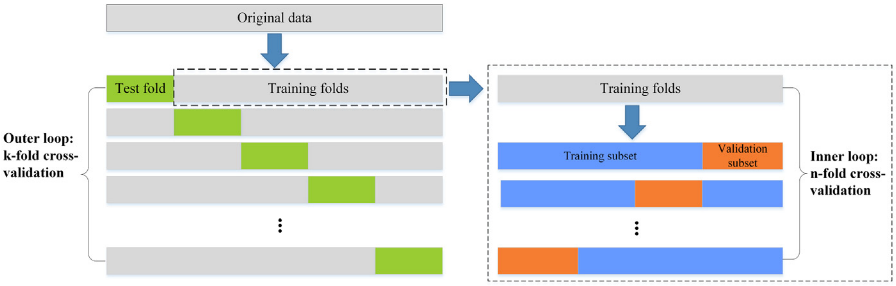


Fig. 4. Illustration of $k \times n$-fold nested cross-validation.

AUC is an extensively used discrimination capability measurement based on the receiver operating characteristic curve (ROC) [36], with its value being equal to the area under the ROC curve. The AUC ranges from "0" (indiscernible) to "1" (perfectly discernible), where 0.5 indicates the prediction performance of a random classifier. The x-axis of the ROC curve represents a false-positive rate (computed as 1-specificity), and the y-axis represents a true-positive rate sensitivity.

KS is a commonly used discrimination evaluation indicator for credit scoring [37]. First, the data samples are sorted from low to high based on the predicted default probability, after which the cumulative true positive rate and the cumulative false positive rate under each default rate are calculated. Finally, the maximum value of the difference between the two values is obtained as the KS value; the larger the KS value, the stronger the model's ability to distinguish between default borrowers and on-time repayment borrowers.

4.5. Statistical significance tests

To obtain a convincing conclusion, experimental statistical evaluations were used to test the statistical significances of the performance differences [38]. Non-parametric rather than parametric tests were employed as the assumptions in parametric tests tend to be violated when comparing classification models. Therefore, following the recommendation in Sun et al. [5], a non-parametric Wilcoxon test was employed to compare the performances of each pair of models for each evaluation measure. If the statistical p-value of the Wilcoxon test was lower than a certain significance level, such as 5% for example, then the null hypothesis that determined that there were no performance differences between the two models was rejected at a 5% level, with the performance differences between the two models being considered to be significantly different at the 5% significance level.

Table 3
Performance comparisons between the individual classification models (%).

Classifier	German credit data		Taiwan credit data	
	AUC	KS	AUC	KS
LR	78.52 ± 2.74	47.56 ± 5.51	71.73 ± 0.92	37.68 ± 1.39
LDA	78.76 ± 2.62	48.16 ± 4.43	71.75 ± 0.69	37.69 ± 1.25
QDA	77.50 ± 3.14	45.71 ± 6.59	70.23 ± 0.75	34.88 ± 1.26
KNN	71.19 ± 3.38	34.64 ± 6.29	70.55 ± 0.95	31.94 ± 1.56
NB	75.27 ± 2.98	43.49 ± 5.55	73.62 ± 0.84	**39.88** ± **1.36**
CART	71.64 ± 3.33	36.29 ± 5.07	73.29 ± 0.95	37.49 ± 1.70
SVM	76.77 ± 2.31	44.55 ± 4.20	70.46 ± 1.28	38.37 ± 1.13
ANN	76.05 ± 3.32	44.17 ± 5.71	71.14 ± 0.70	37.18 ± 1.18
LSTM	**79.58** ± **3.40**	**48.63** ± **6.31**	**74.51** ± **1.30**	39.64 ± 2.30

Note that the average score and standard deviation were reported in this study; the optimal value for each column was marked by bold and underline.

4.6. Experimental design

The main purpose of this study was to verify the effectiveness and superiority of our proposed approach for imbalanced dataset credit scoring; therefore, this section outlines the three main research questions and explains the experiments that were designed to solve them.

RQ01: Does the LSTM network perform better on the credit datasets than the other widely used methods?

RQ02: Is the proposed improved SMOTE method superior to the classical SMOTE technique for credit datasets that are differently imbalanced?

RQ03: Is the proposed novel deep learning ensemble method combined with a data-level solution and ensemble algorithm the most appropriate method for imbalanced credit data scoring?

To answer RQ01, the raw credit datasets were first employed to train and test the nine individual classification models, after which classification performance comparisons and a statistical significance test were conducted for the two indicators.

To answer RQ02, the two credit datasets were first processed using the classical SMOTE method and the improved SMOTE method, after which the nine base learners were applied to the rebalanced datasets to explore the classification abilities. Finally, the statistical significance of the performance differences for each classifier was independently tested.

To answer RQ03, the proposed deep learning ensemble model was independently trained and tested with each of the two credit datasets, and the other eight base classifier ensembles were also tested for comparison. Then, the results of all classification methods were compared using the AUC and KS indicators and statistical significance tests performed to verify the superiority of the ensemble method.

5. Experimental results and analysis

The classification scores achieved using the different learning methods from the experimental design are presented and compared in Tables 3–11. Note that each model was performed using 6 times 5 × 2 nested cross-validation methods. Therefore, 30 prediction results of each model were obtained and analyzed.

5.1. Results for research Question 1

In this experiment, the performances of all the nine individual classifiers were evaluated and a significance test conducted to verify whether the LSTM outperformed the other widely used methods.

The results in Table 3 give the average AUC and KS scores for the nine algorithms and the standard deviations. As can be seen in Fig. 5, in general, the LSTM network had higher evaluation scores than the other individual classifiers for all two credit datasets. The AUC analysis showed that the LSTM achieved the best scores for all datasets (79.58%, 74.51%), with the differences between the LSTM and the suboptimal classifiers respectively being 0.82% for German credit data, and 0.89% for Taiwan credit data, which indicated that the LSTM network had good credit risk modeling abilities. The KS metric analysis found that the LSTM network performed quite well with respective ranks of 2nd and 1st for the Taiwan and German credit datasets. The differences in the KS scores between the LSTM network and the suboptimal German and Taiwan classifiers became smaller when the sample size increased. Fig. 5 highlights several interesting findings: (1) the average AUC indicator scores for all models for all credit datasets were much larger than the corresponding KS scores; (2) the deviations for the Taiwanese data in the AUC and KS indicators for all classification methods were much smaller than for the German dataset and the AUC score deviations were much smaller than the KS score deviations.

To comprehensively understand the effectiveness of the LSTM network compared to the other widely used methods, statistical significance tests based on the non-parametric Wilcoxon test were also conducted. As shown in Tables 4–5, in terms of the AUC indicator, the LSTM network performed significantly better than any of the other individual classifiers except the LR and LDA models under a 90% confidence level for all the two credit datasets, and there shows a 99% confidence level between the LSTM mode and the other models in the Taiwan dataset; for the KS indicator, the LSTM significantly outperformed the KNN, CART, SVM and ANN, under 99% confidence level over the two datasets. Note that the tests show no statistical significance when comparing the LSTM network with the LR, LDA and QDA in German data and the NB model in Taiwan dataset; however, any small credit scoring improvements could result in significantly greater profits for financial institutions. This finding again proved that the LSTM network was a promising credit scoring research direction, and the improvements obtained by advanced technology could especially promote the development of small-sized financial enterprises [39,40].

5.2. Results for research Question 2

To answer RQ02, the original credit datasets in this experiment were first processed using both the traditional SMOTE method and the improved SMOTE method, after which the nine classifiers were trained and tested using the processed datasets to determine the classification results.

The classification result comparisons shown in Tables 6 and 7 indicate that the single classifier scores for the processed credit data using the SMOTE method had modestly enhanced AUC scores; however, as shown in Fig. 6(a), the AUC scores for the NB, CART, SVM, and the LSTM for the German dataset were lower when the training data were processed using the SMOTE method. Further, when the SMOTE method was employed, the QDA, NB, SVM and LSTM had worse performances for the Taiwan credit data. However, all classifiers except the NB and LSTM classifiers performed better over the German dataset using the SMOTE method for the KS index.

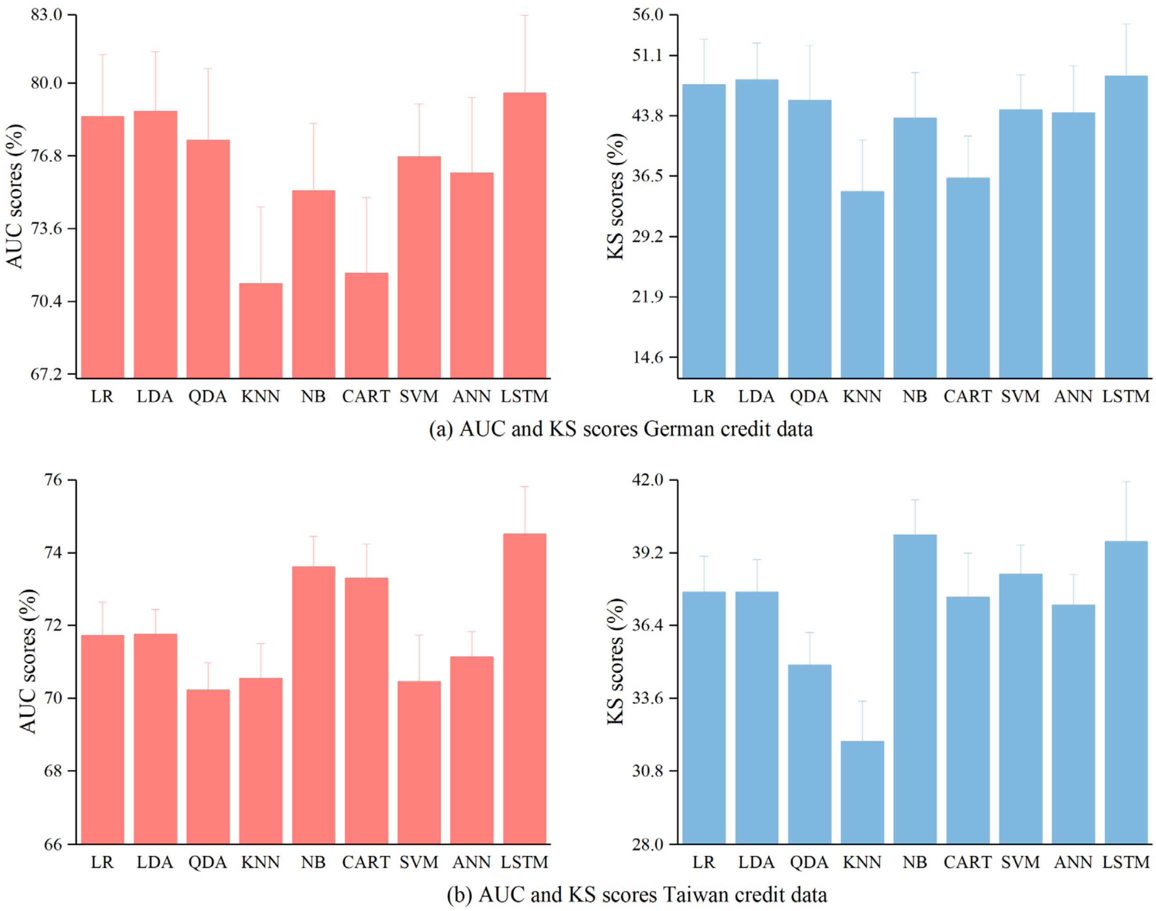
Therefore, based on the results for RQ02, an improved SMOTE method combined with Mahalanobis distance and Euclidean distance is proposed to deal with imbalanced credit data. To comprehensively understand the effectiveness of the proposed method, the results of the classifiers trained with the improved SMOTE approach were compared to the original data.

From Table 7, Fig. 6(b), and (c), there were three main findings. First, as the proposed method was superior to the SMOTE algorithm in terms of the AUC and the KS scores, it is a much

Table 4
Statistical significance test results for all classifiers for the German credit data.

	LR	LDA	QDA	KNN	NB	CART	SVM	ANN	LSTM
Panel A. AUC									
LR		−0.57	−1.65*	−4.78***	−3.51***	−4.78***	−2.52**	−2.48**	−1.40
LDA			−1.60	−4.78***	−3.86***	−4.78***	−3.16***	−3.32***	−1.37
QDA				−4.68***	−2.93***	−4.06***	−1.08	−1.57	−1.96**
KNN					−4.33***	−0.46	−4.51***	−4.39***	−4.60***
NB						−3.24***	−2.15***	−0.81	−3.90***
CART							−4.37**	−3.86***	−4.69***
SVM								−0.71	−3.24***
ANN									−3.30***
LSTM									
Panel B. KS									
LR		−0.38	−1.48	−4.57***	−2.49**	−4.72***	−2.60***	−1.80*	−0.71
LDA			−1.57	−4.78***	−3.42***	−4.78***	−3.03***	−2.58***	−0.43
QDA				−4.28***	−1.47	−4.00***	−0.98	−0.75	−1.40
KNN					−4.46***	−1.22	−4.19***	−4.54***	−4.64***
NB						−3.75***	−1.04	−0.48	−3.09***
CART							−4.35***	−3.92***	−4.56***
SVM								−0.19	−2.73***
ANN									−2.89***
LSTM									

The *z*-statistics along with significance levels were reported in this study.
*Represent significance at the 10% level.
**Represent significance at the 5% level.
***Represent significance at the 1% level.



(a) AUC and KS scores German credit data

(b) AUC and KS scores Taiwan credit data

Fig. 5. Classification performances for all the single models over the original credit datasets.

suitable method for dealing with credit risk evaluation imbalance problems. Second, the improved SMOTE method was found to be more effective for the classifiers tested on the German dataset than for the Taiwan credit dataset, and third, the results indicated that the enhancements were much better for the KS indicator than for the AUC indicator.

To ensure the conclusions were convincing, statistical significance tests were performed on each classifier. For each classifier, the AUC scores and the KS scores achieved with the SMOTE

Table 5
Statistical significance test results for all classifiers for the Taiwan credit data.

	LR	LDA	QDA	KNN	NB	CART	SVM	ANN	LSTM
Panel A. AUC									
LR		−0.32	−4.04***	−3.38***	−4.56***	−4.39***	−3.63**	−2.46**	−4.74***
LDA			−4.68***	−4.45***	−4.76***	−4.37***	−3.84**	−3.01***	−4.70***
QDA				−1.14	−4.78***	−4.78***	−0.96	−3.63***	−4.78***
KNN					−4.78***	−4.72***	−0.52	−2.62***	−4.76***
NB						−1.04	−4.72***	−4.76***	−2.77***
CART							−4.76***	−4.76*	−3.61***
SVM								−2.25**	−4.78***
ANN									−4.76***
LSTM									
Panel B. KS									
LR		−0.34	−4.51***	−4.78***	−3.98***	−0.05	−2.33**	−1.49	−3.55***
LDA			−4.70***	−4.78***	−4.33***	−0.55	−2.15**	−1.45	−3.67***
QDA				−4.64***	−4.78***	−4.19***	−4.78***	−4.29***	−4.72***
KNN					−4.78***	−4.74***	−4.78***	−4.78***	−4.78***
NB						−4.12***	−3.79***	−4.60***	−0.44
CART							−2.23**	−1.20	−3.90***
SVM								−3.09***	−2.79***
ANN									−4.23***
LSTM									

The z-statistics along with significance levels were reported in this study.
*Represent significance at the 10% level.
**Represent significance at the 5% level.
***Represent significance at the 1% level.

Table 6
Performance of the classification models using the SMOTE algorithm (%).

Classifier	German credit data		Taiwan credit data	
	AUC	KS	AUC	KS
LR	78.59 ± 2.50	**48.38** ± **4.89**	71.77 ± 0.75	37.15 ± 1.07
LDA	**78.77** ± **2.33**	48.16 ± 4.43	71.98 ± 0.74	37.13 ± 1.22
QDA	77.52 ± 3.89	46.37 ± 7.29	70.03 ± 0.86	34.61 ± 1.31
KNN	71.38 ± 3.60	34.70 ± 6.21	71.50 ± 0.69	33.56 ± 1.19
NB	74.27 ± 4.29	42.76 ± 7.03	72.25 ± 0.99	37.90 ± 1.65
CART	71.25 ± 3.11	36.49 ± 4.59	**74.22** ± **0.94**	37.28 ± 1.40
SVM	75.90 ± 3.68	44.94 ± 6.02	62.82 ± 8.57	25.33 ± 9.40
ANN	77.16 ± 3.08	46.12 ± 5.98	71.71 ± 0.81	36.89 ± 1.34
LSTM	78.28 ± 3.86	46.72 ± 6.34	73.47 ± 3.13	**40.51** ± **5.15**

Note that the average score and standard deviation were reported in this study; the optimal value for each column was marked by bold and underline.

Table 7
Performance of the classification models using the improved SMOTE algorithm (%).

Classifier	German credit data		Taiwan credit data	
	AUC	KS	AUC	KS
LR	78.75 ± 3.26	48.45 ± 6.23	72.28 ± 0.88	37.35 ± 1.38
LDA	78.87 ± 2.73	47.40 ± 4.85	72.12 ± 0.87	37.25 ± 1.31
QDA	77.78 ± 3.29	46.25 ± 6.14	70.20 ± 0.91	35.66 ± 1.35
KNN	71.59 ± 3.34	35.87 ± 6.00	71.51 ± 0.75	33.70 ± 1.65
NB	75.36 ± 2.78	43.74 ± 4.37	72.43 ± 0.66	38.18 ± 1.43
CART	72.57 ± 3.49	38.25 ± 5.34	74.43 ± 0.84	37.86 ± 1.75
SVM	77.10 ± 2.96	44.94 ± 5.26	70.66 ± 1.00	38.28 ± 1.18
ANN	77.85 ± 3.36	47.50 ± 5.46	72.42 ± 2.33	38.22 ± 3.12
LSTM	**79.90** ± **2.41**	**50.36** ± **4.81**	**74.75** ± **1.05**	**39.80** ± **1.29**

Note that the average score and standard deviation were reported in this study; the optimal value for each column was marked by bold and underline.

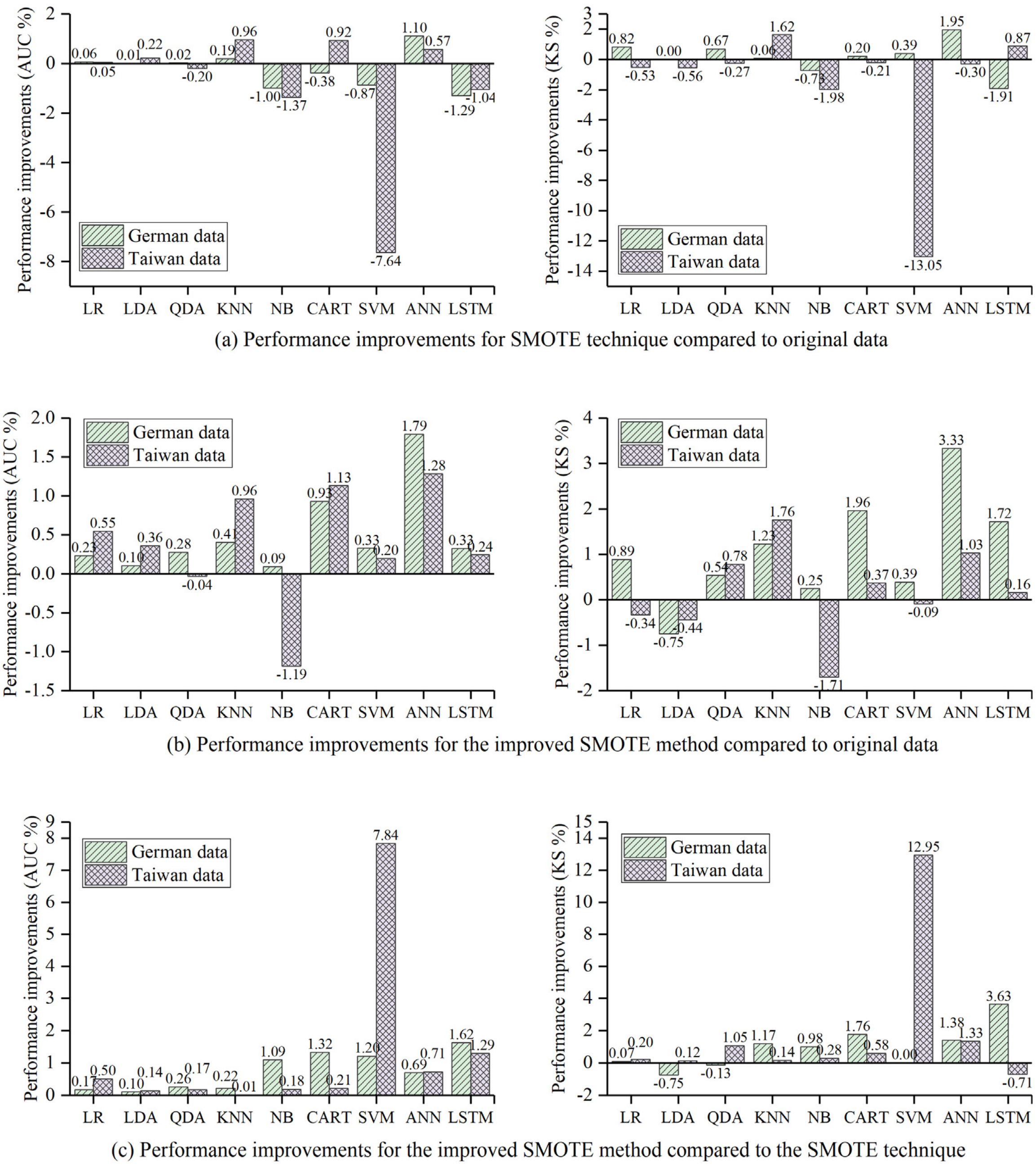
method and the improved SMOTE method were tested. As can be seen in Table 8, the classifiers trained using the improved SMOTE method had better AUC and KS scores than the classifiers trained using the traditional SMOTE method. For the AUC indicator, the performance differences between the improved SMOTE and the traditional SMOTE methods were significant for the NB, CART, and SVM models for the German credit data, and the LR, SVM, ANN, and LSTM classifiers were significantly enhanced for the Taiwan credit data; furthermore, the KS scores for the improved SMOTE method and the SMOTE algorithm were significant for the CART and the LSTM for the German credit dataset, and significant improvements were observed for the QDA, SVM, and ANN for the Taiwan credit data. Generally, the improved SMOTE method introduced in this study proved to be efficient and practical in addressing the imbalanced credit data and was proven to be superior to the traditional SMOTE algorithm.

5.3. Results for research Question 3

In this study, a new deep learning ensemble model with an improved SMOTE approach was proposed for credit risk evaluation. To ensure fair comparison, all the other individual classification methods were also ensembled with an AdaBoost frame and the improved SMOTE algorithm. Table 9 and Fig. 7 show the classification performances for the ensemble models, from which some conclusions were made in respect of RQ03.

As shown in Table 9, compared to the other ensemble approaches, the LSTM ensemble model had significantly better AUC and KS indicator scores for all two credit datasets, which indicates a superiority of the application for the proposed method in credit scoring. However, it is easy to find that the KNN model performed badly over the German and the Taiwan datasets regarding the AUC and the KS scores. It is worth pointing out that the standard deviations of AUC indicator for the ensemble models were relatively smaller than that of KS indicator.

To further explore the capabilities of the proposed novel ensemble method, the classification performance comparisons between the ensemble models and the single models are given in this section. Fig. 7 shows the performances for the ensemble models compared to the single learners using the different data processing methods. For convenience, the AUC scores and KS scores were plotted for these classification methods in the same subfigure in Fig. 7. For the AUC and KS scores, the proposed ensemble model based on the LSTM network and the improved SMOTE method outperformed the other classification methods except the CART on all two credit datasets. It was interesting to find that compared to the Taiwanese credit dataset, all model performances obtained from the German credit data were much better, which was possibly because the credit risk complexity was greater in the Taiwanese datasets. In general,



(a) Performance improvements for SMOTE technique compared to original data

(b) Performance improvements for the improved SMOTE method compared to original data

(c) Performance improvements for the improved SMOTE method compared to the SMOTE technique

Fig. 6. Performance improvement using the improved SMOTE algorithm.

Table 8
Statistical significance test results for RQ02 over two credit datasets.

	LR	LDA	QDA	KNN	NB	CART	SVM	ANN	LSTM
Panel A. German credit dataset									
AUC	−0.44	−0.35	−0.28	−0.01	−1.76*	−1.68*	−1.88*	−1.22	−1.61
KS	−0.25	−0.97	−0.06	−0.53	−1.04	−1.65*	−0.30	−0.99	−1.87*
Panel B. Taiwan credit dataset									
AUC	−2.15**	−0.61	−1.35	−0.26	−0.79	−1.06	−4.78***	−2.11**	−1.84*
KS	−0.63	−0.40	−2.58*	−0.42	−0.57	−1.43	−4.78***	−2.27**	−0.63

The *z*-statistics along with significance levels were reported in this study.
*Represent significance at the 10% level.
**Represent significance at the 5% level.
***Represent significance at the 1% level.

the proposed ensemble architecture with the AdaBoost algorithm and the improved SMOTE method was found to be superior and more effective in enhancing the performances of the single classifiers under imbalanced circumstances, and therefore could

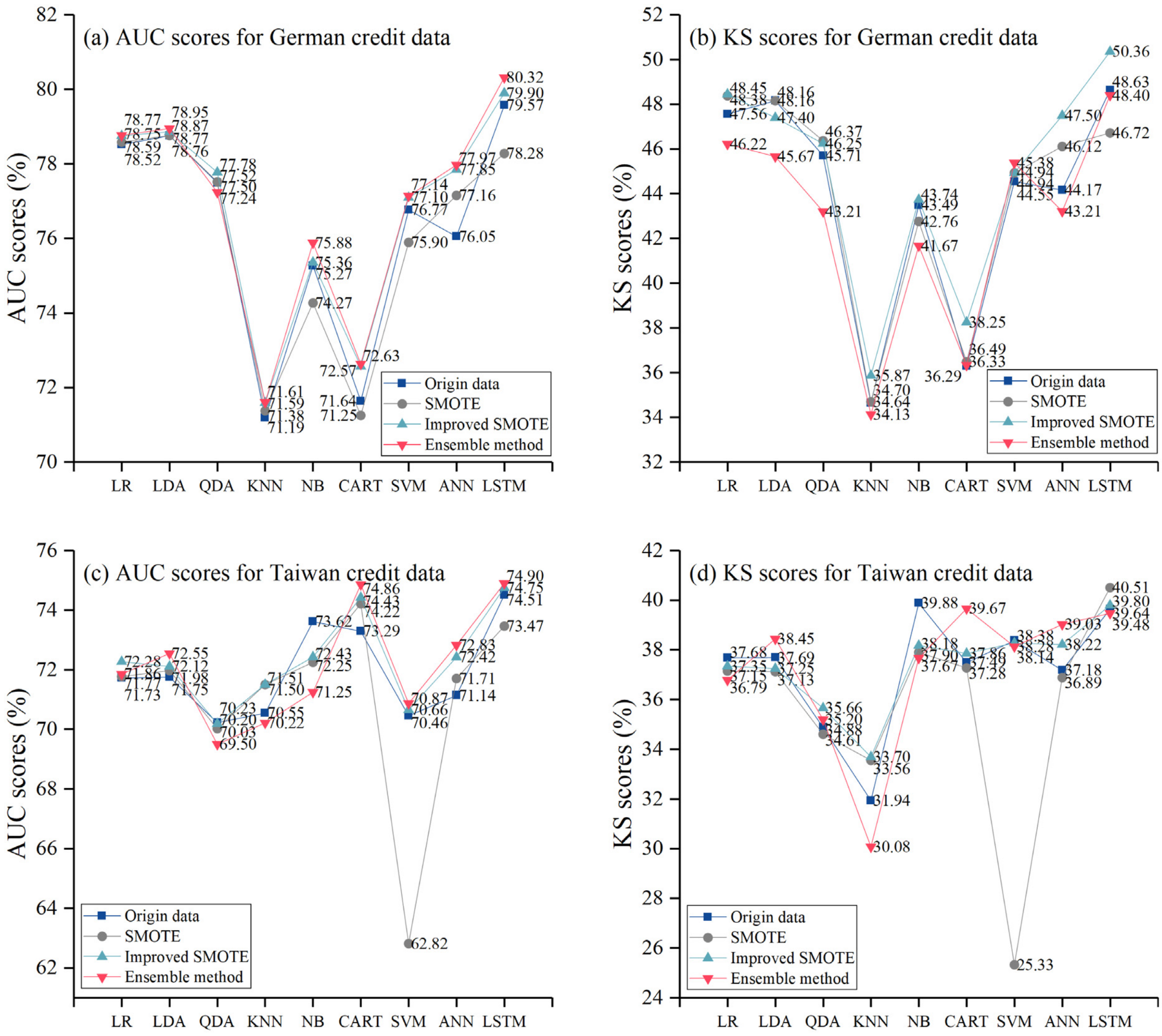


Fig. 7. Performance of all the classification models trained with different methods for the two credit datasets.

Table 9
Performance of the ensemble classification models.

Base	German credit data		Taiwan credit data	
Classifier	AUC	KS	AUC	KS
LR	78.77 ± 1.93	46.22 ± 5.04	71.86 ± 0.72	36.79 ± 1.29
LDA	78.95 ± 2.34	45.67 ± 4.29	72.55 ± 1.29	38.45 ± 3.28
QDA	77.24 ± 2.71	43.21 ± 6.11	69.50 ± 0.60	35.20 ± 1.31
KNN	71.61 ± 3.21	34.13 ± 5.16	70.22 ± 0.99	30.08 ± 2.66
NB	75.88 ± 2.71	41.67 ± 5.17	71.25 ± 0.70	37.67 ± 1.74
CART	72.63 ± 3.41	36.33 ± 5.43	74.86 ± 2.28	**39.67** ± **4.49**
SVM	77.14 ± 3.07	45.38 ± 5.58	70.87 ± 4.02	38.14 ± 6.51
ANN	77.97 ± 3.49	43.21 ± 5.78	72.83 ± 1.86	39.03 ± 3.62
LSTM	**80.32** ± **2.09**	**48.40** ± **5.29**	**74.90** ± **0.79**	39.48 ± 1.55

Note that the average score and standard deviation were reported in this study; the optimal value for each column was marked by bold and underline.

be a prospective credit risk management application for financial institutions.

To comprehensively exploit RQ3, the non-parametric Wilcoxon test was conducted between the proposed ensemble deep learning method and the other ensemble models, the z-statistics, and the significance levels for the AUC indicators and the KS indicators, are in Tables 10–11. The statistical significance tests also proved that the proposed ensemble method based on the LSTM network and the improved SMOTE technique outperformed all the other ensemble models over the AUC and the KS indicators. Because of the differences in the samples sizes and the imbalance ratios for these two datasets, the performance differences for the proposed ensemble model and the other models were less significant for the Taiwan dataset than for the German dataset. Any small improvements in the credit risk evaluations could significantly reduce potential loan losses. Faced with problems associated with big data and the high credit data imbalance ratio in real-world credit scoring, the proposed method provides a prospective application for credit risk management in financial institutions.

6. Conclusion

In this study, an improved SMOTE method to rebalance credit data was developed and a new combined LSTM neural network and AdaBoost algorithm ensemble classification model was proposed for credit risk evaluations. To avoid the original SMOTE algorithm's indiscriminate selection and improve minority class data generation efficiency, the Mahalanobis distance between the minority samples and the majority class distribution were first measured, after the minority samples were selected by the predefined Mahalanobis distance interval to synthesize the new

Table 10
Statistical significance tests for the comparison of ensemble methods performance for the German data.

Base learner	LR	LDA	QDA	KNN	NB	CART	SVM	ANN	LSTM
Panel A. AUC									
LR		−0.50	−2.07**	−4.76***	−3.63***	−4.49***	−2.15**	−1.02	−3.01***
LDA			−2.09**	−4.76***	−3.90***	−4.66***	−2.18**	−1.40	−1.97**
QDA				−4.70***	−1.37	−4.51***	−0.03	−0.73	−4.02***
KNN					−4.66***	−0.69	−4.51***	−4.76***	−4.74***
NB						−3.47***	−1.43	−2.64***	−4.25***
CART							−3.79***	−4.21***	−4.76***
SVM								−0.85	−3.69***
ANN									−2.79***
LSTM									
Panel B. KS									
LR		−0.59	−2.03**	−4.76***	−2.91***	−4.16***	−0.74	−1.81*	−1.74*
LDA			−1.70*	−4.76***	−3.44***	−4.45***	−0.31	−1.64	−2.15**
QDA				−4.45***	−0.42	−4.17***	−1.49	−0.18	−3.32***
KNN					−4.51***	−1.41	−4.68***	−4.64***	−4.66***
NB						−3.24***	−2.29**	−0.92	−3.66***
CART							−4.12***	−3.86***	−4.56***
SVM								−1.35	−2.21**
ANN									−3.18***
LSTM									

The *z*-statistics along with significance levels were reported in this study.
*Represent significance at the 10% level.
**Represent significance at the 5% level.
***Represent significance at the 1% level.

Table 11
Statistical significance tests for the comparison of ensemble methods performance for the Taiwan credit data.

Base learner	LR	LDA	QDA	KNN	NB	CART	SVM	ANN	LSTM
Panel A. AUC									
LR		−3.32***	−4.78***	−4.35***	−2.73***	−4.58***	−1.35	−2.40**	−4.78***
LDA			−4.78***	−4.45***	−4.00***	−4.04***	−1.96**	−0.36	−4.72***
QDA				−2.79***	−4.78***	−4.78***	−1.86*	−4.72***	−4.78***
KNN					−3.55***	−4.78***	−0.83	−4.76***	−4.78***
NB						−4.78***	−0.36	−3.59***	−4.78***
CART							−3.55***	−3.53***	−1.49
SVM								−2.33**	−3.92***
ANN									−3.73***
LSTM									
Panel B. KS									
LR		−3.03***	−4.41***	−4.76***	−2.03**	−2.83***	−1.00	−2.58***	−4.62***
LDA			−4.08***	−4.78***	−0.85	−0.98	−0.36	−0.75	−1.76*
QDA				−4.68***	−4.64***	−4.17***	−2.15**	−3.96***	−4.76***
KNN					−4.74***	−4.78***	−4.21***	−4.78***	−4.78***
NB						−2.05**	−0.26	−1.78*	−3.86***
CART							−0.73	−0.22	−0.40
SVM								−0.61	−0.94
ANN									−0.38
LSTM									

The *z*-statistics along with significance levels were reported in this study.
*Represent significance at the 10% level.
**Represent significance at the 5% level.
***Represent significance at the 1% level.

minority samples. Because of the superior success of LSTM applications for financial market predictions, the LSTM network was employed as basic learner for credit scoring, an AdaBoost algorithm utilized to develop a deep learning ensemble model, and systematic experimental comparisons designed and conducted to prove the effectiveness of the proposed method.

Compared to other widely used credit scoring techniques, the proposed learning method showed excellent credit scoring performances; therefore, it has a wide range of application scenarios in the field of credit scoring. This method could assist in the development of an advanced internal credit scoring system for banks and other financial institutions, which could reduce risks and increase profits.

Future work will focus on three main aspects: an exploration of the applicability and generality of the proposed the improved SMOTE method on high imbalanced credit datasets; the application of the deep learning ensemble method to multi-class rather than binary class assessment; and the employment of other deep learning algorithms with excellent capabilities such as the CNN, RNN, and the DBN as the ensemble model base learners, and using other ensemble approaches such as bagging and stacking to construct the credit scoring models.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (no. 72001178; no. U1811462; no. 71910107002; no. 71725001) , the Humanities and Social Sciences Foundation of the Ministry of Education of China (no. 17YJC630119) , and Chinese National Funding of Social Sciences (no. 19ZDA092). It was also supported by Applied Basic Research Program of Sichuan Province (no. 2020YJ0042), the Fundamental Research Funds for the Central Universities (no. JBK2003020), and the project of Research Center for System Sciences and Enterprise Development (no. Xq20B08).

References

[1] X. Dastile, T. Celik, M. Potsane, Statistical and machine learning models in credit scoring: A systematic literature survey, Appl. Soft Comput. J. (2020) http://dx.doi.org/10.1016/j.asoc.2020.106263.

[2] M. Mihalovic, Performance comparison of multiple discriminant analysis and logit models in bankruptcy prediction, Econ. Sociol. 9 (2016) 101–118, http://dx.doi.org/10.14254/2071-789x.2016/9-4/6, M. Mihalovic, 2016.

[3] Xiaobing Huang, Xiaolian Liu, Yuanqian Ren, Enterprise credit risk evaluation based on neural network algorithm, Cogn. Syst. Res. 52 (2018) 317–324, http://dx.doi.org/10.1016/j.cogsys.2018.07.023.

[4] Yue Wu, Yunjie Xu, Jiaoyang Li, Feature construction for fraudulent credit card cash-out detection, Decis. Support Syst. 127 (2019) 113155, http://dx.doi.org/10.1016/j.dss.2019.113155.

[5] Jie Sun, Jie Lang, Hamido Fujita, Hui Li, Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates, Inform. Sci. 425 (2018) 76–91, http://dx.doi.org/10.1016/j.ins.2017.10.017.

[6] S. Lessmann, B. Baesens, H.V. Seow, L.C. Thomas, Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, European J. Oper. Res. 247 (2015) 124–136, http://dx.doi.org/10.1016/j.ejor.2015.05.030.

[7] Paweł Pławiak, Moloud Abdar, U. Rajendra Achary, Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring, Appl. Soft Comput. 84 (2019) 105740, http://dx.doi.org/10.1016/j.asoc.2019.105740.

[8] Aleum Kim, Sung-Bae Cho, An ensemble semi-supervised learning method for predicting defaults in social lending, Eng. Appl. Artif. Intell. 81 (2019) 193–199, http://dx.doi.org/10.1016/j.engappai.2019.02.014.

[9] O.B. Sezer, M.U. Gudelek, A.M. Ozbayoglu, Financial time series forecasting with deep learning: A systematic literature review: 2005–2019, Appl. Soft Comput. J. (2020) http://dx.doi.org/10.1016/j.asoc.2020.106181.

[10] Zhang Yishen, Wang Dong, Chen Yuehui, Shang Huijie, Tian Qi, Credit risk assessment based on long short-term memory model, Intell. Comput. Theor. Appl. 70 (2017) 0–712, http://dx.doi.org/10.1007/978-3-319-63312-1_62.

[11] C. Wang, D. Han, Q. Liu, S. Luo, A deep learning approach for credit scoring of Peer-to-Peer lending using attention mechanism LSTM, IEEE Access 7 (2019) 2161–2168, http://dx.doi.org/10.1109/access.2018.2887138.

[12] B.M. Gupta, S.M. Dhawan, Deep learning research: Scientometric assessment of Global Publications Output during 2004-17, Emerg. Sci. J. 3 (1) (2019) 23–32, http://dx.doi.org/10.28991/esj-2019-01165.

[13] L. Munkhdalai, T. Munkhdalai, K.H. Ryu, GEV-NN: A deep neural network architecture for class imbalance problem in binary classification, Knowl.-Based Syst. (2020) http://dx.doi.org/10.1016/j.knosys.2020.105534.

[14] Lean Yu, Rongtian Zhou, Ling Tang, Rongda Chen, A DBN-based resampling SVM ensemble learning paradigm for credit classification with imbalanced data, Appl. Soft Comput. 69 (2018) 192–202, http://dx.doi.org/10.1016/j.asoc.2018.04.049.

[15] S. Crone, S. Finlay, Instance sampling in credit scoring: an empirical study of sample size and balancing, Int. J. Forecast. 28 (2012) 224–238, http://dx.doi.org/10.1016/j.ijforecast.2011.07.006.

[16] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, Prog. Artif. Intell. 5 (2016) 221–232, http://dx.doi.org/10.1007/s13748-016-0094-0.

[17] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artificial Intelligence Res. 16 (2002) 321–357, http://dx.doi.org/10.1613/jair.953.

[18] L. Abdi, S. Hashemi, To combat multi-class imbalanced problems by means of over-sampling techniques, IEEE Trans. Knowl. Data Eng. 28 (2015) 238–251, http://dx.doi.org/10.1109/tkde.2015.2458858.

[19] Yufei Xia, Chuanzhe Liu, Nana Liu, Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending, Electron. Commer. Res. Appl. 24 (2017) 30–49, http://dx.doi.org/10.1016/j.elerap.2017.06.004.

[20] Marcin Czajkowskia, Monika Czerwonkab, Marek Kretowski, Cost-sensitive Global Model Trees applied to loan charge-off forecasting, Decis. Support Syst. 74 (2015) 57–66, http://dx.doi.org/10.1016/j.dss.2015.03.009.

[21] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches, IEEE Trans. Syst. Man Cybern. C (Appl. Rev.) 42 (2012) 463–484, http://dx.doi.org/10.1109/tsmcc.2011.2161285.

[22] M.J. Kim, D.K. Kang, H.B. Kim, Geometric mean based boosting algorithm with oversampling to resolve data imbalance problem for bankruptcy prediction, Expert Syst. Appl. 42 (2015) 1074–1082, http://dx.doi.org/10.1016/j.eswa.2014.08.025.

[23] Yongming Song, Yi Peng, A MCDM-based evaluation approach for imbalanced classification methods in financial risk prediction, IEEE Access 7 (2019) 84897–84906, http://dx.doi.org/10.1109/access.2019.2924923.

[24] S. Chen, H. He, E. Garcia, Ramoboost: ranked minority oversampling in boosting, IEEE Trans. Neural Netw. 21 (2010) 1624–1642, http://dx.doi.org/10.1109/tnn.2010.2066988.

[25] Raúl Ruiz de la Hermosa González-Carratoa, Wind farm monitoring using mahalanobis distance and fuzzy clustering, Renew. Energy (2018) http://dx.doi.org/10.1016/j.renene.2018.02.097.

[26] S. Hochreiter, Schmidhuber. J., Long short-term memory, Neural Comput. 9 (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[27] Ha Young Kim, Chang Hyun Won, Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models, Expert Syst. Appl. 103 (2018) 25–37, http://dx.doi.org/10.1016/j.eswa.2018.03.002.

[28] F.A. Gers, N.N. Schraudolph, J. Schmidhuber, Learning precise timing with LSTM recurrent networks, J. Mach. Learn. Res. 3 (2002) 115–143, http://dx.doi.org/10.1162/153244303768966139.

[29] Stewart Jones, David Johnstone, Roy Wilson, An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes, J. Bank. Financ. 56 (2015) 72–85, http://dx.doi.org/10.1016/j.jbankfin.2015.02.006.

[30] A. Asuncion, D. Newman, UCI Machine Learning Repository, 2007 Irvine, CA: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830, https://dl.acm.org/doi/10.5555/1953048.2078195.

[32] Gavin C. Cawley, Nicola L.C. Talbot, On over-fitting in model selection and subsequent selection Bias in Performance Evaluation, J. Mach. Learn. Res. 11 (2010) 2079–2107, https://dl.acm.org/doi/10.5555/1756006.1859921.

[33] M.C. Robinson, R.C. Glen, A.A. Lee, Validating the validation: reanalyzing a large-scale comparison of deep learning and machine learning models for bioactivity prediction, J. Comput. Aided Mol. Des. (2020) http://dx.doi.org/10.1007/s10822-019-00274-0.

[34] Lin Shi, Johan A. Westerhuis, Johan Rosén, Rikard Landberg, Carl Brunius, Variable selection and validation in multivariate modelling, Bioinformatics 35 (2019) 972–980, http://dx.doi.org/10.1093/bioinformatics/bty710.

[35] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, Gong Bing, Learning from class-imbalanced data: Review of methods and applications, Expert Syst. Appl. 73 (2017) 220–239, http://dx.doi.org/10.1016/j.eswa.2016.12.035.

[36] Salvatore Carta, Anselmo Ferreira, Diego Reforgiato Recupero, Marco Saia, Roberto Saia, A combined entropy-based approach for a proactive credit scoring, Eng. Appl. Artif. Intell. 87 (2020) 103292, http://dx.doi.org/10.1016/j.engappai.2019.103292.

[37] Petr Teply, Michal Polena, Best classification algorithms in peer-to-peer lending, N. Am. J. Econ. Finance 51 (2020) 100904, http://dx.doi.org/10.1016/j.najef.2019.01.001.

[38] V. Garcia, A.I. Marques, J.S. Sanchez, An insight into the experimental design for credit risk and corporate bankruptcy prediction systems, J. Intell. Inf. Syst. 44 (2015) 159–189, http://dx.doi.org/10.1007/s10844-014-0333-4.

[39] R. Tsaih, Y.-J. Liu, W. Liu, Y.-L. Lien, Credit scoring system for small business loans, Decis. Support Syst. 38 (1) (2004) 91–99, http://dx.doi.org/10.1016/S0167-9236(03)00079-4.

[40] Yohannes Worku, Mammo Muchie, The uptake of E-Commerce services in johannesburg, Civ. Eng. J. 5 (2) (2019) 349–362, http://dx.doi.org/10.28991/cej-2019-03091250.