Information Processing and Management 60 (2023) 103267

Contents lists available at ScienceDirect

Information Processing and Management

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

A qualitatively analyzable two-stage ensemble model based on machine learning for credit risk early warning: Evidence from Chinese manufacturing companies

Lu Wang [a], Wenyao Zhang [b,*]

[a] School of Information Management and Artificial Intelligence, Zhejiang University of Finance and Economics, China
[b] School of Economics and Management, Qilu University of Technology (Shandong Academy of Sciences), China

ARTICLE INFO

Keywords:
Credit risk early warning
Machine learning
Qualitative analysis
Ensemble model
Deep learning

ABSTRACT

Constructing ensemble models has become a common method for corporate credit risk early warning, while as to deep learning model with better predictive ability, there have been no fixed theoretical models formed in corporate credit risk early warning, as such models often fail to conduct further qualitative analysis of the results. Thus, this article builds a new two-stage ensemble model using a variety of machine learning methods represented by deep learning for corporate credit risk early warning, which can not only effectively improve the prediction performance of the model, but also qualitatively analyze the source of corporate credit risk from multiple angles according to the results. At first stage, the improved entropy method is used to reassign the instance weight in correlation degree based on grey correlation analysis. At second stage, this study adopts Bagging method to integrate multiple one-dimensional convolutional neural networks, and borrows idea of N-fold cross validation to expand the difference of the base classifier. Empirically, this article selects listed companies in Chinese manufacturing industry between 2012 and 2021 as datasets, including 467 samples with 51 financial indicators. The new ensemble model has the highest F1-score (87.29%) and G-mean (89.47%) among comparative models, and qualitatively analyzes corporate risk sources. Further, it also analyzes how to increase early warning effect from the angles of indicator number and time span.

1. Introduction

Financial risk is the possibility that economic subjects suffer losses due to the uncertain changes of various factors in the process of financial activities or investment. It has always been the most important research issue in risk management (Wang et al., 2020). As the main bodies of national economy, large companies are constantly threatened by financial risks, whose credit risk becomes one of the most important problems of financial risk, especially for listed companies. It will cause immeasurable impact on the entire national economy if they suffer from credit risk due to poor management, or even bankruptcy. Thus, how to better warn their credit risk and prevent them from the threat of bankruptcy shows a popular trend in academic circles.

In the era of big data, artificial intelligence algorithms and data mining technology have been widely used in corporate credit risk early warning with new models and methods introduced, such as artificial neural network (Fu et al., 2020; Jardin, 2016), support

* Corresponding author.
E-mail address: 20171482@qq.com (W. Zhang).

https://doi.org/10.1016/j.ipm.2023.103267
Received 22 September 2022; Received in revised form 10 December 2022; Accepted 3 January 2023
Available online 9 January 2023

vector machine (Dastile et al., 2020; Sun et al., 2021), and decision tree (Xia et al., 2017; Sun et al., 2018). However, each model or method has its disadvantages in algorithms, so that an increasing number of scholars encourage to ensemble two or more model algorithms and thus complement those respective shortcomings. In this sense, there have been some deficiencies in corporate credit risk early warning issues that need to resolve. First, in the selection of early warning models, deep learning has better predictive ability and has been successfully applied to speech recognition, natural language processing and automatic driving, but there has been less studies of its application in corporate credit risk, nor even the formation of mature theoretical models or frameworks. Especially specific to the processing of financial indicator data characterized by time series, it has not gotten scholars' attention, so it is worth applying deep learning to this area. Second, although machine learning has strong fitting or prediction ability, most of which are black box in models and methods, unable to make qualitative analysis further. At present, many scholars have begun to notice it and tried to build interpretable models, however, it is still in the preliminary stage in corporate credit risk early warning as there has been less literature on the impact of financial indicators on prediction results. Consequently, if managers intend to prevent risks, they need to point out risk sources, so that it is particularly important to analyze qualitatively financial indicators affecting early warning results.

According to the above discussion, the research objective of this article is to build one model capable of both increasing prediction accuracy and qualitatively analyzing risk source and early warning effect. Therefore, a new two-stage ensemble model is proposed on the basis of multiple machine learning for corporate credit risk early warning with financial indicator data in the form of time series. At the first stage, it ranks financial indicators data from every company according to grey correlation analysis to sort out the best indicators that can reflect risks personalized for each company, so as to establish a dynamic indicator system. It can not only analyze the source of credit risk qualitatively for each company, but also traces the causes of credit risk in the industry by summarizing common indicators of all companies. More importantly, as there exists the shortcoming that each instance is given the same weight when calculating grey correlation degree, the entropy method is improved to reassign the weight and make its distribution more reasonable and effectively increase the accuracy of grey correlation analysis to judge correlation degree of each indicator. At the second stage, an ensemble model is created by means of Bagging method on the basis of convolutional neural network (CNN) specific to personalized financial indicators, in a way to further improve prediction accuracy of corporate credit risk. Additionally, since financial indicators in time series are regarded as panel data, this study applies one-dimensional CNN to process it. Further, given Bagging method requires the difference of base classifiers to be as large as possible, this study draws on the idea of N-fold cross validation to classify training sets as much as possible to maximize the difference of base classifiers trained.

Based on the above, this study will solve the following research questions by adopting four separate experimental methods to compare and analyze financial panel data divided into two separate groups of manufacturing-listed companies with and without credit risk.

Q.1: Regarding the attribute dimension of panel data, is it verified that the more financial indicators, the more prediction accuracy improved, and that can the source of risk be qualitatively analyzed?

Q.2: Regarding the time dimension of panel data, is it verified that the larger the time span of the selected data, the more prediction accuracy improved?

Q.3: Compared with the original algorithm without improvement, do the improved grey relation analysis and convolutional neural network really improve the prediction performance?

Q.4: Does the new two-stage ensemble model have the most superior predictive performance compared to traditional methods?

The rest of the article is organized as follows. Related work is analyzed in Section 2. Methodology is presented and an improved two-stage ensemble model is proposed in Section 3. Section 4 describes the experiments. Section 5 analyzes comparative results. Section 6 summarizes our results and illustrate implications and limitations, followed by future research works.

2. Related work

Sample index system and early warning model are two major research directions in credit risk prediction. Furthermore, as most of credit risk early warning models cannot interpret the results well, it leads to the trend that qualitative analysis conducted on the basis of prediction results becomes a research focus.

For sample index system, feature selection and feature extraction are commonly used approaches. Specifically, feature selection is to choose a subset of features out of the whole feature set whereas a new set of features is created through feature engineering from its extant features (Dastile et al., 2020). For instance, Liang et al. (2015) confirmed the effect of feature selection and summarized that the performance of prediction is not always improved by performing feature selection. Brown and Mues (2012) indicated that the performance of credit scoring model can be enhanced by removing redundant features. Chi and Hsu (2012) and Oreski and Oreski (2014) adopted meta-heuristic approaches for feature selections. For feature extraction, techniques mostly refer to rough set technique (Wang et al., 2012; Chen & Li, 2010), stepwise (Brown & Mues, 2012), genetic algorithm (Chi & Hsu, 2012) and PCA (Han et al., 2013). For example, Chi and Hsu (2012) constructed a dual credit scoring model for management of mortgage accounts by combining the bank's behavior scoring model and credit bureau scoring model via genetic algorithm to select essential variables. Chen and Cheng (2013) built hybrid models based on rough set classifiers to set up rules for credit rating decision in the global banking industry.

Similar to the idea of dimensionality reduction in feature selection, grey relational analysis is a quantitative means to describe and compare the changing development of a system and judge the degree of correlation between various factors (Yang & Liu, 2016; Wu et al., 2022). Thus, grey relational analysis is often used to find the most suitable sample indexes. However, it has some deficiencies, in this sense, many scholars have carried out researches to improve them. For instance, Chen et al. (2021) established an emergency decision model with grey wolf optimization algorithms and grey relational analysis embedded according to case-based reasoning. Wu et al. (2016) incorporated the concept of grey relational degree into the technique's ideal solution for preference order by similarity to ideal solution. Hu (2020) considered the prediction of bankruptcy as a grey system problem to create multivariate models of grey prediction on account of GM (1, N) for the prediction of bankruptcy.

In early warning model of credit risk, statistical learning model, machine learning model and deep learning model are commonly used approaches, wherein linear discriminant analysis (Goerigk et al., 2022), logistic regression (So et al., 2016) and naïve bayes classifier (Sinnl, 2022) are major methods to build statistical learning model. For instance, Goerigk et al. (2022) regarded the problems of two-stage robust optimization as games between an adversary and a decision maker. They proved NP- hardness for an extensive scope of problems but notified that the special case in which adversarial costs of first and second stages are equal are solvable in polynomial time. So et al. (2016) raised a method related to fuzzy logistic regression to credit scoring to predict the probability of a company's default. Sinnl (2022) proposed a novel integer linear programming formulation for problems based on Benders decomposition, which had a nice combinatorial structure where linear programs did not need to be solved to divide Benders cuts.Support vector machine (Zhang et al., 2022), K-nearest neighbor (Henley & Hand, 1996), decision tree (Dumitrescu et al., 2021), and artificial neural network (Bhattacharya et al., 2017) are common methods used to construct machine learning model. For example, Zhang et al. (2022) employed Gaussian RBF kernel due to its good prediction effect to analyze the complex mechanism of financial performance in terms of profitability, business solvency, and development ability on company's green behavior. Henley and Hand (1996) adjusted k-nearest-neighbour (k-NN) method to address credit scoring issue by means of Euclidean distance metric. Dumitrescu et al. (2021) proposed an interpretable and high-performance method of credit scoring based on penalized logistic tree regression using information from decision trees to improve overall performance. Also, Bagging and Adaboost are ensemble algorithms often used to build credit risk early warning model. For instance, Youn and Gu (2010) established a two-stage hybrid model by using logistic regression model and artificial neural network to predict credit risk of Chinese small- and medium-sized enterprises. Zhu et al. (2019) combined multi-boosting technique with random subspace to raise up an improved hybrid ensemble model and RS-Multi Boosting model.

The application of intelligent algorithm and early warning effect is greatly enhanced by constructing deep learning model. Huang et al. (2020) analyzed the studies on how deep learning model is applied to affect key banking and finance areas and provided a systematic assessment of preprocessing and input data. Further, CNNs have been broadly employed in the analysis of data in time series and image processing. Yin et al. (2022) utilized convolution neural network to build risk early warning model of supply chain finance. Kim and Cho (2019) adopted deep CNN to present an architecture for extracting automatically features and improve the repayment performance in P2P lending. Further, CNN is integrated with other models together to increase overall prediction performance. Yoichi and Naoki (2020) put forward a novel method to realize transparent and concise datasets with heterogeneous properties for credit scoring by creating a fully-connected one-dimensional layer of CNN integrated with recursive-rule extraction algorithm with decision tree. Dai (2022) applied the improved CNN to financial credit to verify its characteristics and prediction accuracy with data sourced from stock market in China. Xia et al. (2022) reported a prediction model of hybrid spatiotemporal money laundering in accordance with long short-term memory and graphed CNN to investigate the dependency among diverse money laundering transactions.

Extensive and in-depth researches have been conducted in academia and industry to improve the explanatory ability of machine learning models and also with a series of approaches proposed. Datta et al. (2016) developed a family of Quantitative Input Influence (QII) measures which gain the influence level of inputs on system outputs. Agarwal et al. (2021) introduced an interpretable approach by new relevant input variable measures utilized for abandoning iteratively redundant input feature vectors, thereby causing overfitting of noisy data decreased and accuracy of tests improved. Moraffah et al. (2020) comprehensively presented a survey on interpretable models from the causal perspectives of machine learning approaches and problems. Bussmann et al. (2021) came up with an interpretable artificial intelligence model able to be applied in the management of credit risks and in the measurement of risks. Mi et al. (2020) reviewed the extant explainable methods specifically by separating them into interpretable methods with external co-explanation and interpretable methods with self-explanatory model. Kliegr et al. (2021) analyzed the extent to which human understanding of interpretable machine learning models are impacted by cognitive biases, especially following logical rules inferred from data. ElShawi et al. (2021) formulated four basic quantitative measures to evaluate interpretability technique quality, i.e., bias detection, similarity, trust, and execution time and comprehensively assess local model agnostic interpretability techniques.

Complex models are often given preference to analysis due to high complexity of data in financial area. Such models are not only difficult to be interpreted, but also difficult to analyze further prediction results. Especially in credit risk early warning, different researchers have varied ideas in building models, thus requiring diverse interpretability from models and interpretable methods put forward with respective emphasis. Bussmann et al. (2021) constructed a new model with correlation networks applied to Shapley values and thus, predictions of artificial intelligence can be grouped in light of the similarity among basic explanations. Park et al. (2021) employed LIME algorithm to measure and analyze the importance of every data point feature in prediction models of bankruptcy. Moscato et al. (2021) started with a benchmarking analysis on some credit risk scoring models commonly used to forecast whether loans are allowed to be repaid in a P2P platform, as well as employed three methods with the best interpretability through diverse interpretable artificial intelligence instruments. Shin (2021)) investigated how the interpretability of artificial intelligence influence users' attitude and trust towards AI by conceptualizing causability as one of its factors and as an essential cue of algorithms and by testing how trust influences user's perceived effect of AI-oriented products and services. Peng et al. (2021) proposed an explainable framework of artificial intelligence with local and global interpretation of auxiliary diagnosis of hepatitis, helpful to clarify the transparency of complex models. Gramespacher and Posth (2021) demonstrated how machine learning methods are adapted to concrete demands of credit evaluation and how it optimizes for an economic target function rather than for accuracy in the circumstance of strongly asymmetric costs of wrong predictions. Zhang et al. (2022) raised up an interpretable artificial intelligence method comprising an interpretable frame for FDP as well as a whole process ensemble method, whose ensemble algorithm shows high accuracy from feature selection to predictor construction.

In this sense, grey relational analysis is helpful to ordering indicators by relevance, retaining the most effective relevant indicators, and decreasing data dimensions to develop an index system. However, it is disadvantaged by unreasonable weight coefficients so that it needs to be improved further. Furthermore, CNN is taken as a typical deep learning algorithm useful to building early warning model. However, further studies are needed in analyzing how to further improve the superior performance of CNNs through ensemble models with data. In addition, the existing studies about explanation of models only stay in one method utilized to explain models or indicator data, rather than integrating explainable methods into the prediction process, nor even in consideration of enhancing prediction results through explainable conclusions, i.e., qualitative analysis on how to improve the prediction results of models.

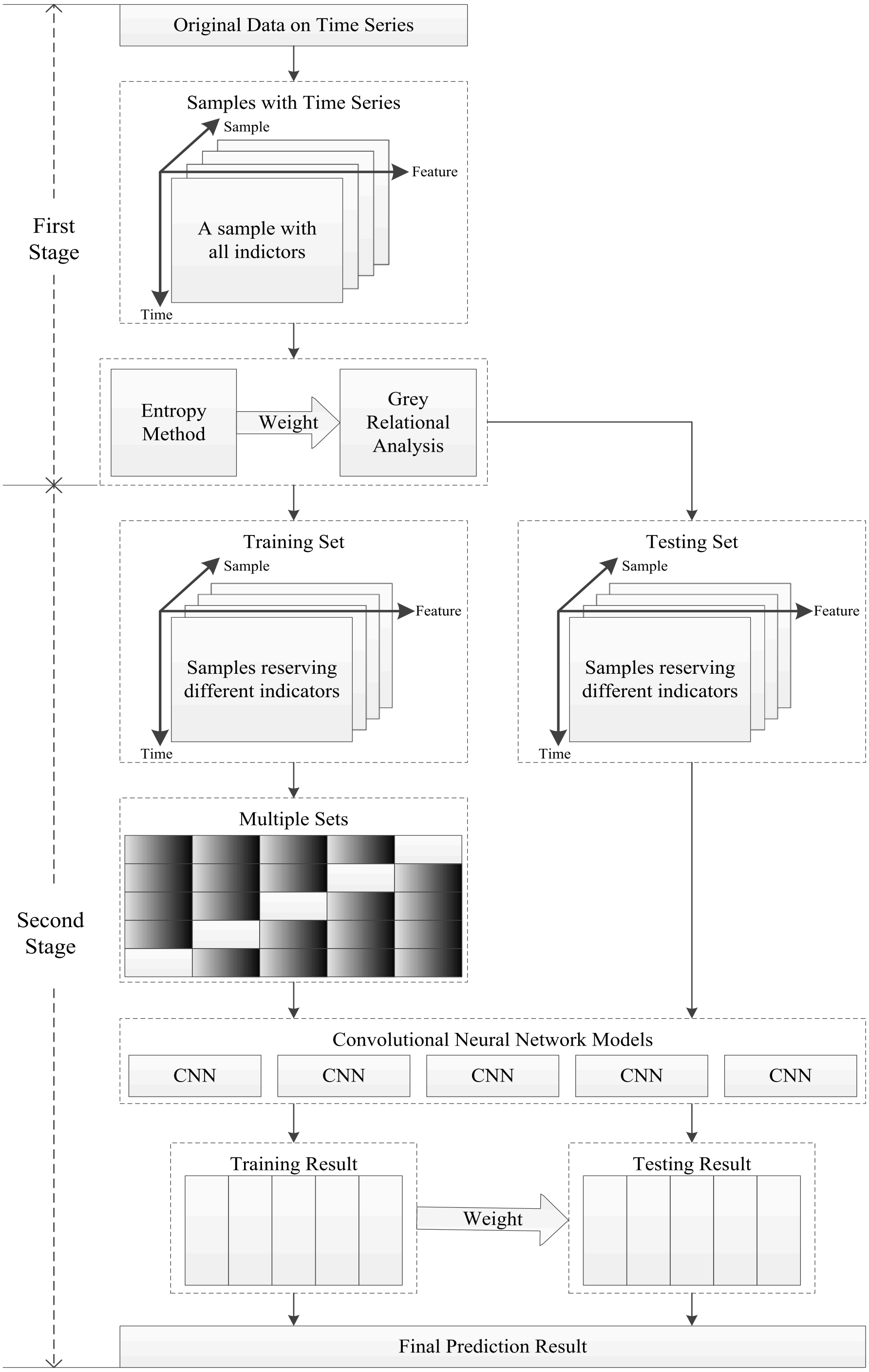


Fig. 1. Framework and calculation process of prediction model.

3. Methodology

3.1. A qualitatively analyzable two-stage ensemble model

This paper built a two-stage integration model based on multiply machine learning to predict corporate credit risk, which not only enhances prediction accuracy, but analyzes the risk source and prediction result qualitatively. This study regarded financial indicators of sample companies of each year as the data in time series. According to the criteria of judging whether an enterprise has credit risk, the grey relational analysis improved by entropy weight method is used to dynamically select the most suitable indicators to build a personalized indicator system. By summarizing these indicators, it was able to qualitatively analyze the source of risks for sample companies and even the entire industry and how qualitative analysis could effectively improve the prediction accuracy through selecting the numbers of indicators and time span. Then, the personalized indicator system was calculated by CNN algorithm integrated by Bagging for enhancing accuracy to achieve better prediction result, as shown in Fig. 1.

Specifically, in the first stage, weighted value was determined by the improved entropy method, and then was introduced into grey relational analysis, so as to dynamically determine the most suitable financial indicators as sample. In the second stage, all samples were further separated into training and test samples. In training samples, according to the thought of 5-fold cross-validation, they were divided into five equal parts, with a different part retained each time and other four parts used to train CNN, and in this sense, to build base classifiers with large difference. This study also utilized Bagging method to integrate 5 CNNs to gain final prediction results. The data in test set was brought in the CNN trained to obtain final weighted results through weight matrix formulated by training results. In next sections, the details of each part will be explained further.

3.2. Grey relational analysis with improved entropy weight

Financial indicators can be reflective of whether an enterprise has credit risk, but it is very vague for the distinction between financial indicators of enterprises with credit risk and with normal operation. Also, given financial indicators reveal different aspects of enterprise's operations, for the purpose of analyzing the enterprises' causes of credit risks, it is necessary to analyze financial indicators qualitatively. Hence, grey relational analysis is more approriate for handling this problem since this is subject to grey system problems. However, there may be unreasonable weight allocation in it, thus this paper adopts entropy weight method to further solve this problem.

3.2.1. Entropy method to determine weight

Entropy method is a method used to determine the weight of indicators by judgment matrix which consists of evaluation indicator value under objective conditions. It is able to eliminate the subjectivity of each factor's weight as maximum as possible, thereby making evaluation results more objective. Basic calculation steps of the entropy method are shown as follows:

(1) Construct an input matrix X with n evaluation indicators of m schemes:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} \tag{1}$$

(2) Normalize the input matrix **X**: x_{ij}represents the element of the ith row and the jth column

$$x_{ij}^{'} = \frac{x_{ij} - min(x_j)}{max(x_j) - min(x_j)}, \ (i = 1, \ 2, \dots, \ m; \ j = 1, \ 2, \dots, \ n) \tag{2}$$

(3) Calculate the entropy value of the jth evaluation indicator:

$$e_j = -\frac{1}{\ln(m)} \sum_{i=1}^{m} p_{ij} \ln(p_{ij}), \ (j = 1, \ 2, \dots, \ n) \tag{3}$$

Wherein, p_{ij} is the proportion of the ith scheme of the jth evaluation indicator. But to make $ln(p_{ij})$ meaningful, $p_{ij} \neq 0$. In order to meet the conditions, p_{ij} computational formula should be:

$$p_{ij} = \frac{1 + x_{ij}^{'}}{\sum_{i=1}^{m} (1 + x_{ij}^{'})}, \ (j = 1, \ 2, \dots, \ n) \tag{4}$$

(4) Determine the entropy weight of each evaluation indicator:

$$\omega_j = \frac{1 - e_j}{n - \sum_{j=1}^{n} e_j}, \quad \sum_{j=1}^{n} \omega_j = 1 \tag{5}$$

3.2.2. Grey relational analysis

Grey relational analysis refers to an effective statistical method for evaluating the influence of multiple factors (Kadier et al., 2015) and for judging if they are closely correlated by geometric similarity between standard data array (evaluation metric) and several correlation data array (evaluation object). This similarity is also called grey relational degree, whose effect is to manifest the degree of correlation between curves. Compared with other statistical methods that generally require a large number of sample data, grey relational analysis can better solve the problem of small sample size, whose procedures are shown as follows:

(1) Determine sequence type:

Determine the standard data array, i.e., evaluation metric column:$Y = y(k) = [y(1),\ y(2), \ldots, y(n)],\ k = 1,\ 2, \ldots, n$

Determine the correlation data array, i.e., evaluation object column:$X = x_i(k) = \begin{bmatrix} x_1(1) & \cdots & x_1(n) \\ \vdots & \ddots & \vdots \\ x_m(1) & \cdots & x_m(n) \end{bmatrix},\ i = 1,\ 2, \ldots, m;\ k = 1,\ 2, \ldots,$

n

(2) Dimensionless processing of data:

$$x_i^{'}(k) = \frac{x_i(k) - min\ x_i}{max\ x_i - min\ x_i}, \quad (i=1,\ 2, \ldots,\ m;\ k=1,\ 2, \ldots,\ n) \tag{6}$$

(3) Calculate correlation coefficient, which is to compare the correlation between each element of correlation data array and its corresponding element of standard data array:

$$\xi_i(k) = \frac{\underset{i}{min}\underset{k}{min}|y(k) - x^{'}_i(k)| + \rho \cdot \underset{i}{max}\underset{k}{max}|y(k) - x^{'}_i(k)|}{|y(k) - x^{'}_i(k)| + \rho \cdot \underset{i}{max}\underset{k}{max}|y(k) - x^{'}_i(k)|} \tag{7}$$

Wherein, $\rho \in (0,\ 1)$ is called grey resolution coefficient, generally valued 0.5.

(4) Calculate grey relational degree, that is, the correlation degree between each correlation data array and standard data array:

$$r_i = \frac{1}{n}\sum_{k=1}^{n} \xi_i(k), \quad i = 1, 2, \ldots, m \tag{8}$$

(5) Sorting of grey relational degree, judged by the sorting order of relational degree between each correlation data array and standard data array. The closer r gets to 1, the better the correlation is.

3.2.3. Grey relational degree with improved entropy-based weight

When in the calculation of weights by the entropy method, if different indicators expressed the similar entropy values, it indicated that those indicators provided similar amount of useful information, i.e., they had similar entropy weights. However, if different indicators were calculated in terms of entropy weights following Eq. (5), they had slightly different entropy values, entropy weights might change exponentially.

For example, the entropy values of four indicators of a scheme were 0.999, 0.998, 0.997, 0.996, respectively, whose entropy weights calculated by Eq. (5) were 0.1, 0.2, 0.3, 0.4, respectively. That was a difference of 0.003 between the entropy value of the first indicator and that of the last indicator and a difference of four times between entropy weights of both indicators.

As a result, given the deficient original method, this study improves entropy weight Eq. (5) as follows:

$$\omega_j = \frac{\sum_{i=1}^{n} e_i + 1 - 2 \cdot e_j}{\sum_{j=1}^{n}\left(\sum_{i=1}^{n} e_i + 1 - 2 \cdot e_j\right)}, \quad \sum_{j=1}^{n} \omega_j = 1 \tag{9}$$

When grey relational degree was computed according to Eq. (8), all n instances showed the same weights by default, which was not in line with actual circumstances. Weights were correspondingly given to different instances so as to get more realistic results. Based on

Algorithm 1
Grey relational analysis with improved entropy-based.

```
Input: The samples in time series
Output: Indicators of every sample after being sorted
Select the data of one sample in time series
For $i = 1, 2, \ldots$, k
    def grey Relational Analysis(Sample, Standard Column)
      Calculated by procedure 1 to procedure 3 following grey relational analysis
      def Improved Entropy (Sample)
          Calculated by procedure 1 to procedure 3 following entropy method
          Entropy weight calculated by Formula (9)
        Calculate relational degree according to Formula (10)
        Sort indicators according to grey relational analysis procedure 5
```

this idea, entropy weight can be introduced into the calculating process of grey relational degree, with weight correspondingly assigned. Therefore, we improved the Eq. (8) as follows:

$$r_i = \sum_{k=1}^{n} \omega_k \cdot \xi_i(k), \quad i = 1, 2, \ldots, m \tag{10}$$

Among them, ω_k is the result computed according to Eq. (9). The process of calculation is shown in Algorithm 1 as follows, where k is the number of samples, and Sample is the data of a single sample in the form of time series:

3.3. Convolutional neural network

CNN was first coined by LeCun et al. (1989) and has been widely applied in recognition and time series fields (Seo & Shin, 2019; Sezer & Ozbayoglu, 2018). It is a multi-layer feed-forward network, whose neurons between layers can be connected, whose loss function needs to be defined, and which is optimized by the gradient descent method. Generally, it is composed of an input layer, a convolutional layer, a pooling layer, and a fully connected layer. It is the widespread practice that 2-dimensional CNN is used to deal with image. But it needs to be adjusted for the data of financial indicators. Consequently, this study regards financial indicators that occurred every year as the data in time series, and uses 1-dimensional CNN to predict corporate credit risk, whose structure is shown in Fig. 2 as follows:

It is assumed that the data constituted by m financial indicators and n years, namely $[n \times m]$ matrix, are put into the input layer. Because of 1-dimensional CNN, "filter" can only set one parameter, namely the width of filter, this paper sets it as k. So filter is $[k \times m]$ matrix. Thus, it can obtain convolution layer as $[j \times i]$ matrix. j is decided by the value of n and whether it is set as padding. Pooling can only slide down in each column. If the size of pooling is set as h, it can obtain pooling layer as $[r \times i]$ matrix. r is decided by is decided by the value of j and h. After passing through multiple convolution layers and pooling layers, it enters the fully connected layer and is calculated by the Softmax activation function to get the result from the output layer, namely the classification result.

3.3.1. An input layer for processing data

As shown in Fig. 2, the matrix represents the data of financial indicators of sample companies over the years in the input layer, where n rows signify the data generated in n years, the data of every year has m columns, standing for m financial indicators. There should be as many matrices $[n \times m]$ of input data as there are sample companies.

3.3.2. A convolutional layer for processing data

A convolutional layer can be obtained after multiple filters processing the input data. As the 1-dimensional CNN, these filters treat each row - the data of financial indicators that occurred in one year, as a whole, and scroll down along the row of the entire matrix - the time axis. Actually, a filter is a matrix inclusive of weight arrays that are in need of being optimized. The convolutional layer is operated in this way that, such the filter is applied to the input matrix and to acquire the sum of products of elements of filter weights and time series, and then is passed and transformed through an activation function. Usually, this activation function can be a ReLU (Rectified Linear Unit) or a Sigmoid function, so that the sum of products of corresponding elements of two matrices is activated and mapped to play the role of convolution.

In Fig. 2, the width of filter is set as k, then the black-filled $[k \times m]$ matrix is a filter, scrolling down along the dotted line, and a column vector of $[j \times 1]$ is formed. j is decided by the value of n and whether it is set as padding. If a total of i filters with different initial values are constructed, a matrix with the value of $[i \times j]$ is obtained in the convolutional layer. Since i filters are randomly initialized, they all extract diverse information from the input matrix.

3.3.3. A pooling layer for processing data

The pooling layer is usually connected to reduce the complexity of data output and prevent data overfitting. As the 1-dimensional CNN, it is operated in the same motional direction of the filter in the convolutional layer to complete feature mapping of downsampling along the time axis. This mapping can be max-pooling or average-pooling. Wherein max-pooling is to select the largest value at each position in the local window as the result of the mapping, while average pooling is to select the average value of each position in the

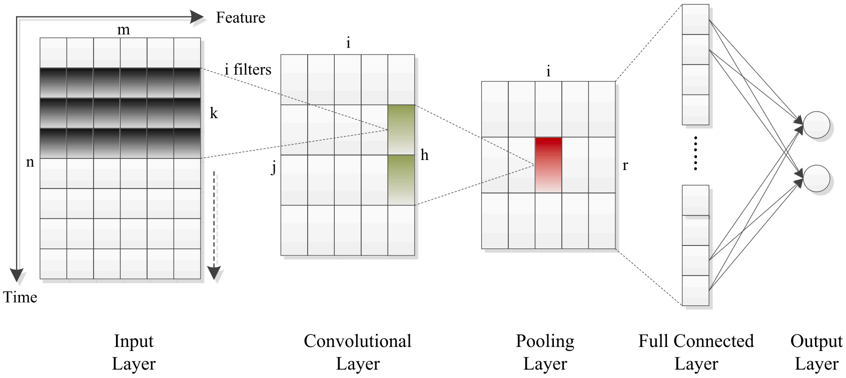


Fig. 2. Structure of one-dimensional CNN.

local window as the result of mapping. It is worth noting that no additional parameters are introduced into this layer.

As shown in Fig. 2, as the same movement direction of the filter, pooling moves down along a column of the convolutional layer. Let's set the size of pooling as *h*, and the maximum or average values of *h* data are calculated as the pooled eigenvalues. Each column in the convolutional layer is pooled in this way, and finally a pooling layer matrix of $[r \times i]$ is obtained, and *r* is decided by the value of *j* and *h*.

3.3.4. A fully connected layer for processing data

To predict if a company has credit risk, it is necessary for pooled characteristics to be combined into a binary classifier through the fully connected layer. As shown in Fig. 2, this process is similar to the multilayer perceptron. It was finally passed to the output layer through full connected layer. As the result has only two outputs - the probability values belonging to each category. To ensure the result in the interval [0, 1], the activation function that connects the output layer selects the Softmax function:

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_{c=1}^{2} e^{z_c}} \tag{11}$$

Where, z_i is the output value of the *i*th node, and *c* is the number of output nodes - number of categories.

3.4. Bagging model ensemble algorithm

Bagging is a typical parallel ensemble learning method in machine learning. Bagging constructs multiple datasets for training base classifiers by parallel sampling, and then integrates the prediction results of each base classifier through a combination strategy to obtain the final output. To guarantee strong generalization ability, it is necessary for the bagging model to ensure: (1) the degree of differentiation of the base classifier as much as possible; (2) sufficient training of the base classifier. However, these two conditions are mutually exclusive: if the difference of the base classifier is satisfied, the data of training sets cannot be overlapped, thus limiting the number of samples in every training set and affects the training of the base classifier. In this sense, the sampling process should have certain constraints and satisfy both conditions in order to ensure generalization ability of the bagging model.

By borrowing the idea of N-fold cross-validation, this study develops different datasets for training multiple base classifiers. It separates original training sets according to categories, and divides the data of each category into five equal parts, with one different part set aside for each category every time and the remaining four parts integrated as the training sets. The final prediction result of each base classifier is achieved via the voting method. This secures the difference in training base classifiers and guarantees each base classifier is trained to the greatest extent under the premise of limited samples. Therefore, this model's calculation process can be seen in Algorithm 2.

Algorithm 2
CNN ensemble model using Bagging.

```
Input: Sample data set with personalized indicator time series
Output: credit risk prediction results
Divide sample data set into training and testing sets
According to the idea of 5-fold cross-validation, divide training sets into categories first and partition the data into five parts for each category, and finally
conduct the fusion of subsets under different categories.
def CNN model()
        Construct CNN according to CNN construction process of each layer
For $i = 1, 2, \ldots, 5$
        Use training$_i$ to train CNN$_i$ and get precision rate of training$_i$
      Bring testing sets into CNN$_i$ to obtain evaluation indicators
According to training$_{1-5}$ precision rate, get evaluation indicators of testing sets weighted to obtain final results
```

4. Experiments

4.1. Sample data and indicators selected

To solve research questions, we made reference to the approach of special treatment (ST) implemented on manufacturing-listed companies in stock trading in stock exchange due to their prominent issues in credit risk, and thus classified samples into a ST sample and a normal sample. The former referred to companies with credit risk if they had negative net profit in two consecutive years, while the latter was companies with business normally operated. Therefore, this paper divides the listed companies into two categories according to whether they have credit risks, which is a binary problem. The ST listed companies are chosen as samples with credit risks, and the normal listed companies are regarded as samples without credit risks. Additionally, given the length of listing time and the release time of annual financial statements, the samples with credit risk were those that have been listed for over five consecutive years and were specially treated for the first time between 2012 and 2021, and normal samples referred to those that have been listed for over five consecutive years and have never been specially treated. A total of 121 listed companies specially treated were selected as samples with credit risk, and 346 listed companies with normal operation were selected as normal samples. All of these data are from Shanghai Stock Exchange and Shenzhen Stock Exchange in China. To this end, this is a binary classification problem for this paper.

51 financial indicators reflecting credit risk were initially selected from their annual financial statements, covering solvency, operating capacity, profitability, growth capacity, cash flow capacity and stockholder's profitability. These financial indicators were subjected to T-test and Mann-Whitney U test, respectively, 10 indicators with no significant difference were eliminated, and 41 financial indicators shown in Table 1 were finally selected as sample indicators of the early warning model. More specifically, the sample data includes the occurrence data of 41 sample indicators in annual report of each company from the date of listing to the date of excerpt. Since each company has been listed for at least 15 years, this constitutes no less than 15 items per sample of data in time series. The data all comes from commercial dataset GuoTaiAn and CCER Economic and Financial Research Database.

4.2. Evaluation indicators

The numbers of companies with or without credit risk were not the same in the samples, resulting in the fact that the prediction result obtained from the overall sample was incapable of mirroring the model's judgment on every category and even of comprehensively measuring the model's superiority. To avoid such issues, the expression of some evaluation indicators selected in this study is described below.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{12}$$

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

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$F1 - score = \frac{\left(1 + \beta^2\right) \cdot \frac{TP}{TP+FN} \cdot \frac{TP}{TP+FP}}{\beta^2 \cdot \frac{TP}{TP+FN} + \frac{TP}{TP+FP}} \tag{15}$$

$$G - mean = \sqrt{\frac{TP}{TP + FN} \cdot \frac{TN}{TN + FP}} \tag{16}$$

As shown in Table 2, TP stands for the correctly judged number of samples with credit risk, FN for the wrongly judged number of samples with credit risk, FP for the wrongly judged number of samples with normal operation, TN for the correctly judged number of

Table 1
Financial indicators for credit evaluation.

Type	Financial indicators
Solvency	Current ratio(X1), Quick ratio(X2), Cash ration(X3), Cash coverage ratio(X4), Asset-liability ratio(X5), Fixed assets ratio(X6), Intangible Asset Ratio(X7), Tangible Assets To Total Assets(X8), equity to fixed assets (X9)
Operating capacity	Inventory turnover(X10), Fixed asset turnover(X11), Total assets turnover(X12)
Profitability	Return on equity(X13), return on total assets(X14), Net profit margin on total assets(X15), Net profit margin on current assets(X16), Operating Profit Ratio(X17), rate of return on sale(X18), cost-profit ratio(X19), Total operating cost ratio(X20), sales and expense ratio (X21), Administration Expense ratio(X22), Financial Expense Ratio(X23), Total Profit Cost Ratio(X24)
Growth capacity	Total Assets Grow Rate(X25), owner's equity Grow Rate(X26), Total operating income Grow Rate(X27), Total Operating Cost Grow Rate (X28), net profit Grow Rate(X29), net asset per share Grow Rate(X30)
Cash flow capacity	Rate of capital accumulation(X31), Cash Rate of Sales(X32), Operating Income Cash Coverage(X33), Operating Profit Cash Coverage (X34), Net Profit Cash Coverage(X35), total Asset Cash Coverage(X36)
Stockholder's profitability	Net Asset Per Share(X37), Undivided Profit Per Share(X38), Total Operating Revenue Per Share(X39), Earnings per share(X40), Operation Profit Per Share(X41)

samples with normal operation.

It was known from the formulas of those evaluation indicators that F1-score and G-mean comprised four basic indicators. Only when TP and TN values were larger and FN and FP values smaller, two evaluation indicators can be larger. Hence, it was concluded that F1-score and G-mean were more comprehensive evaluation indicators.

4.3. Experimental design

This article is to build a two-stage ensemble model to process the data in time series. In the first stage, it adopted the improved grey relational analysis. The listed enterprises are used as samples by ST to judge whether the enterprise has credit risk, whose judgment criteria are directly associated with the enterprise's net profit. However, net profit was treated as a reference index, while numerous financial indicators were arranged abiding by correlation degree with the reference index. In the second stage, Bagging was used to integrate multiple CNN models to get prediction results ultimately.

This study employed four methods for comparative analysis in order to solve four research questions proposed. First, the two-stage ensemble model was used to rank the financial indicators in the order of correlation from strong to weak, and financial indicators with the strongest correlation were utilized to qualitatively analyze the risk source for corporates and industry. In addition, prediction results were compared when different numbers of financial indicators were added in the model to qualitatively analyze whether more indicators could obtain better prediction result.

Second, specific to sample data in time series, we chose financial indicators of different time spans to construct sample data sets and still employed the novel two-stage ensemble model to carry out classification prediction. In this way, prediction results under diverse time spans were compared to qualitatively analyze if the more the historical data, the better the prediction result.

Third, to prove the effectiveness of the improved methods in the novel model, three comparative analysis models were constructed against improved grey correlation analysis and CNN integrated with bagging. They were: (1) the two-stage ensemble model built by original grey relational analysis bagging-improved CNN, named Model 2. The five-fold cross-validation method was still adopted to construct classifiers; (2) the two-stage ensemble model created by grey relational analysis with improved entropy weight and original CNN, called Model 3; the two-stage ensemble model set up by original grey relational analysis and original CNN, referred to Model 4.

Finally, as new methods are still emerging in prediction of corporate credit risk, no consensus has been reached on the best model or method. Thus, this study chooses commonly used models as comparison models, such as SVM (Support Vector Machine), DT(Decision Tree), RF(Random Forest), MLP(Multilayer Perceptron) and LSTM(Long Short-Term Memory). They are also used as comparison models after integrating by Bagging. Among which, the parameters of SVM were $C = 1$, $g = 1$, with Polynomial kernel function adopted; MLP had three layers with 50 units on each layer; LSTM had three layers with 50 units on each layer. The novel two-stage ensemble model was further verified in its superiority through the analysis on comparative models.

To eliminate chances in the analysis results, our study repeated to operate each model 50 times and compared and analyzed the mean values and standard deviations of 50-time experiments, and the evaluation indicators listed were average results obtained after 50 experiments.

5. Comparative analysis of results

5.1. Qualitative analysis from the indicator dimension

Financial indicators were successively brought into the model for verification according to Q1. Moreover, this study adopted grey relational analysis with improved entropy weight to sort out all financial indicators based on their correlation with net profit from strong to weak, and added them step by step into the model in sorted order to compare and summarize the prediction accuracy in different numbers of financial indicators. Thus, the prediction results obtained with different numbers on testing sets were shown in Table 3.

Longitudinal data was compared in Table 3 to show the results obtained by 41 financial indicators being selected in order following the size of correlation to construct numbers of attributes of different samples. For five evaluation indicators, it demonstrated an overall trend of increasing first, then decreasing, and recovered slightly afterwards, no matter of which evaluation indicator. Fig. 3 reported the curves of five evaluation indicators, which were obtained by the datasets built by different numbers of indicators.

The curves went roughly the same way that increased first and then decreased. In particular, F1-score and G-mean were able to generally assess the performance of the model when it deals with unbalanced class data as their curves were of similar shape – increase first and then decrease, and hence they showed the same increase and the same decrease trend. All this also proved that the model was added with different numbers of financial indicators in order from strong to weak, its performance would be enhanced with a certain

Table 2
Confusion matrix for binary classification.

		Prediction Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

Table 3
Prediction results of different numbers of financial indicators.

No.	Acc	Rec	Pre	F1-score	G-mean	No.	Acc	Rec	Pre	F	G
1	92.02	77.55	90.48	83.52	86.79	22	87.23	61.22	85.71	71.43	76.83
2	91.49	75.51	90.24	82.22	85.64	23	85.64	57.14	82.35	67.47	73.94
3	92.02	75.51	92.50	83.15	85.95	24	84.57	53.06	81.25	64.20	71.25
4	92.55	75.51	94.87	84.09	86.27	25	85.64	59.18	80.56	68.24	74.97
5	93.09	77.55	95.00	85.39	87.43	26	82.98	57.14	71.79	63.64	72.54
6	92.55	77.55	92.68	84.44	87.11	27	81.38	59.18	65.91	62.37	72.66
7	94.15	79.59	97.50	87.64	88.89	28	82.45	57.14	70.00	62.92	72.26
8	94.15	81.63	95.24	87.91	89.70	29	82.45	51.02	73.53	60.24	69.08
9	93.09	83.67	89.13	86.32	89.81	30	84.04	59.18	74.36	65.91	74.11
10	91.49	79.59	86.67	82.98	87.27	31	85.11	57.14	80.00	66.67	73.66
11	90.96	79.59	84.78	82.11	86.94	32	86.17	57.14	84.85	68.29	74.22
12	90.96	77.55	86.36	81.72	86.14	33	86.70	61.22	83.33	70.59	76.54
13	88.83	75.51	80.43	77.89	84.04	34	86.70	65.31	80.00	71.91	78.45
14	87.23	71.43	77.78	74.47	81.42	35	85.11	63.27	75.61	68.89	76.62
15	89.89	71.43	87.50	78.65	82.98	36	86.70	65.31	80.00	71.91	78.45
16	87.77	69.39	80.95	74.73	80.87	37	85.64	63.27	77.50	69.66	76.92
17	84.57	59.18	76.32	66.67	74.40	38	84.57	57.14	77.78	65.88	73.39
18	87.23	61.22	85.71	71.43	76.83	39	86.70	65.31	80.00	71.91	78.45
19	86.17	63.27	79.49	70.45	77.22	40	85.11	65.31	74.42	69.57	77.55
20	87.23	61.22	85.71	71.43	76.83	41	85.64	63.27	77.50	69.66	76.92
21	87.23	59.18	87.88	70.73	75.82						

number of indicatorswere added until it reached up to a peak value (about ten indicators). However, when the number continued to increase after that, its performance would not increase but would be gradually reduced. When the number of indicators is further increased to 31 or more, its performance rises slightly, but limited in range. Further, comparative results between F1-score and G-mean of 41 indicators on training set and testing set were shown in Fig. 4.

As compared in Fig. 4, as the numbers of selected financial indicators increased, the results of evaluation indicators were close to 100% in training set, while the results in the testing set were in the state of rising first, then falling down, and slightly uplifting afterwards. This is because after indicators were added one by one according to their respective correlation from strong to weak, the performance of the model was improved due to the gradual addition of important information. However, with the number of indicators increased, the problem of overfitting appeared, resulting in degradated performance of the model. Later on, because of indicators sufficiently increased, more information was introduced to slightly improve the model, but it was very limited to improve the overall prediction effect of the model with its ability. Therefore, it was not the more accurate prediction result came up with the more financial indicators chosen. Instead, selecting some of the most suitable financial indicators contributed to facilitating prediction effect of the base classifier to reach the best state.

Specifically, for evaluation indicators of F1-score model, the entire curve could achieve ideal prediction result when choosing financial indicators between five and ten. The peak values of prediction results of the base classifier was reached especially when seven to nine financial indicators were selected, namely, 87.64%, 87.91% and 86.32%, and then prediction performance was gradually decreased as the numbers of financial indicators were increased. Moreover, prediction results of the base classifier were generally higher than 82% in datasets with less than five financial indicators, while prediction result became lower than 80% in datasets with over ten indicators. When the number of indicators exceeded 31, the predictive performance was improved but in very limited range.

For evaluation indicators of G-mean model, the ideal prediction result was achieved when five to ten financial indicators were selected. The peak values predicted by the base classifier were obtained especially when seven to nine financial indicators were selected, namely 88.89, 89.70 and 89.81%. The overall performance would be decreased when exceeding the optimal number of indicators selected. Also, the prediction results that were realized by datasets with less than five financial indicators were much better than that obtained with over ten financial indicators, when exceeding 31 indicators, the performance was recovered but in very limited range. These results of analysis were similar with F1-score model. It further confirmed that it was not the fact that the more the number

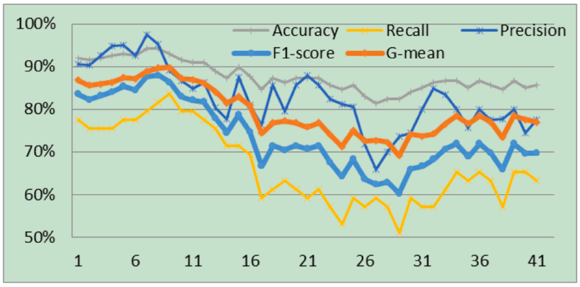


Fig. 3. Evaluation values of different number of financial indicators.

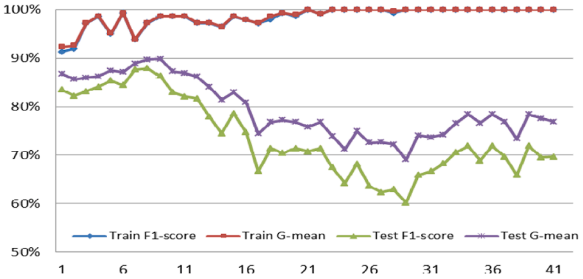


Fig. 4. F1-score and G-mean on training set and testing set.

of samples selected, the better the prediction performance of classifier, instead, it should select a certain number of the most suitable financial indicators to achieve the desired effect.

In order to qualitatively analyze risk sources, this study carried on the statistics on the frequency of sorted financial indicators of sample companies in accordance with different sample types. Since the most ideal prediction result was achieved when five to ten indicators were selected, this study only made statistics on the to five to ten indicators with the highest frequency of occurrence, respectively, as shown in Table 4. It seemed no matter how many indicators were selected from sample companies with credit risk, four financial indicators of earnings per share, net profit margin of total assets, net profit margin of current assets, and operation profit per share had always the highest correlation with net profit.

No matter how many indicators were selected from sample companies with normal operation, three financial indicators of earnings per share, operation profit per share, and total profit cost ratio had always the highest correlation with net profit. Additionally, no matter what type of sample company is concerned, two financial indicators of earnings per share and operation profit per share were always highly correlated with net profit, signifying that these two indicators played a key role in judging the type of sample companies.

Some commonalities are found from comparative analysis that earnings per share, net profit margin on total assets, net profit margin on current assets, operation profit per share, total profit cost ratio originating from profitability and stockholder's profitability, all maintain a high degree of correlation to net profit compared to other indicators, as shown in Fig. 5. Meanwhile, this correlation shows a downward trend in companies with credit risk: earnings per share > net profit margin on total assets > net profit margin on current assets > operation profit per share; in companies with normal operation: earnings per share > operation profit per share > total profit cost ratio. It is worth notifying that, no matter what type of sample companies, earnings per share and operation profit per share from stockholder's profitability always maintain a high correlation with net profit, which proves that both of them exert an essential impact on judging the type of sample companies.

This can be explained, credit risk is directly related to corporate profitability. It refers to the risk concerned with the ability of the financial institution to get money back and of the company to pay debts (Hotchkiss & Altman, 2006). Profitability is understood as a company's ability to obtain profits as well as the ability of funds or capital appreciation. It is usually expressed as the amount and level of a company's income balance in a certain period of time. The higher the profit margin, the stronger the profitability. Cost is the price that a company pays for making profits. The smaller the cost, the better the cost control and the stronger the profitability. Moreover, profitability is not only a basic condition for the survival and development of business, but it is also crucial for shareholders because they receive income in the form of dividends. Profit is the central issue that all parties inside and outside the company care about. It is the source of funds for investors to obtain investment income and for creditors to collect principal and interest. It is the concentrated expression of operators' performance and management efficiency. The increase in profits pushes up stock price, which in turn makes shareholders gain capital. Meanwhile, the stronger the profitability is, the more cash flow it brings, the more strengthened the solvency. As a consequence, profitability comprehensively mirrors if the company has the ability to take risks, while the strength of the relationship between profitability and net profit can reflect if the company has credit risk. Based on the above, we can qualitatively analyze risk sources for every company in light of personalized indicator system. The same analysis about other cases will not be replicated due to the large number of samples.

5.2. Qualitative analysis from time dimension

Chronologically, this study collected annual financial indicators generated by sample companies year by year to create sample data in time series. Sample data was organized by year from near to far. For instance, a sample company was specially treated (ST) in year T, financial indicators of the sample company from listing year to year T-2 were collected in datasets, then financial indicators of one year (year T-2) were chosen to develop sample data, while financial indicators of two years (year T-2 and year T-3) were selected to build sample data. As this study has limitations in sample companies that financial indicators of six years at most were used to build panel datasets. The most ideal result was obtained by sample indicators at the number from seven to nine, thus such three numbers of indicators should be selected for each case, and mean values and standard deviations of prediction results were taken, as shown in Table 5.

By longitudinally comparing mean values of evaluation indicators of different years, it can be found that most of evaluation indicators could get higher values when selecting financial indicators of three years to construct datasets. In particular, the best effect was gotten in datasets built by financial indicators of three years for F1-score and G-mean which better reflected the model's

Table 4
Statistical results for different numbers of sample indicators.

Sample type	No. of financial indicators	The first in correlation	The second in correlation	The third in correlation	The fourth in correlation	The fifth in correlation	The sixth in correlation	The seventh in correlation	The eighth in correlation	The ninth in correlation	The tenth in correlation
Companies with credit risk	5	X40	X15	X16	X41	X24					
	6	X40	X15	X16	X41	X24	X14				
	7	X40	X15	X16	X41	X18	X24	X14			
	8	X40	X15	X16	X41	X18	X17	X24	X14		
	9	X40	X15	X41	X16	X18	X17	X14	X24	X19	
	10	X40	X15	X41	X16	X17	X18	X14	X24	X19	X29
Companies without credit risk	5	X40	X41	X38	X24	X15					
	6	X40	X41	X24	X38	X15	X19				
	7	X40	X41	X24	X15	X38	X19	X16			
	8	X40	X41	X24	X15	X38	X19	X16	X18		
	9	X40	X41	X24	X15	X19	X38	X18	X16	X14	
	10	X40	X41	X24	X19	X15	X18	X16	X38	X17	X14

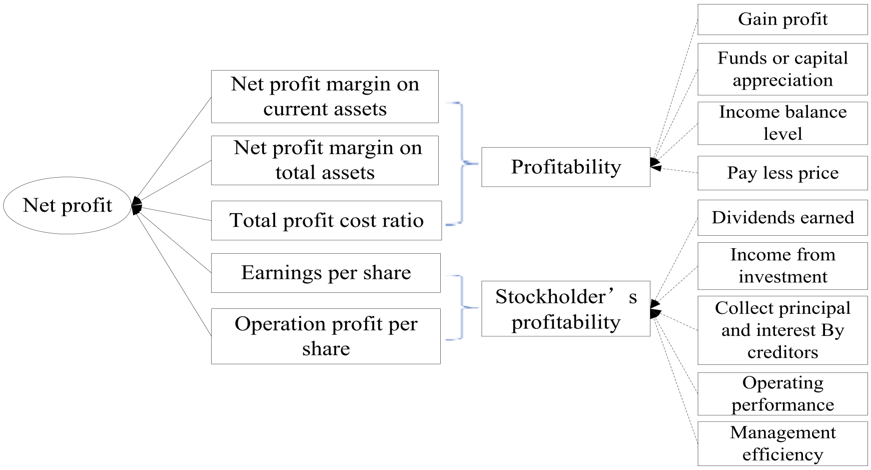

Net profit
Net profit margin on current assets
Net profit margin on total assets
Total profit cost ratio
Earnings per share
Operation profit per share
Profitability
Stockholder' s profitability
Gain profit
Funds or capital appreciation
Income balance level
Pay less price
Dividends earned
Income from investment
Collect principal and interest By creditors
Operating performance
Management efficiency

Fig. 5. Credit risk predicted by the correlation between financial indicators and net profit.

performance. To find rules more intuitively, the curves constructed by these five evaluation indicators with different performance were shown in Fig. 6.

Most of five evaluation indicators got the best results in sample data built in year 2 and in year 3. Especially, the highest values of F1-score and G-mean of the third-level model were 87.29% and 89.47%, respectively, from datasets created by financial indicators of three years. In the remaining years, F1-score values were all around 88%, while G-mean values between 85% and 86%, slightly lower than the corresponding values in three years. This study based on evaluation indicators of five models created by datasets built by financial indicators of three years, the remaining years were subtracted from their corresponding evaluation indicators, the results were plotted in Fig. 7. The basic results of five evaluation indicators obtained from datasets of three years were that Acc was 93.80%, Rec 81.63%, Pre 93.96%, F1-score 87.29% and G-mean 89.47%. However, only 2-year Rec value and 6-year Pre value were positive, higher than that of evaluation indicators of three years. The rest were all negative values, lower than that of evaluation indicators of three years. It can thus be seen that the most ideal prediction results cannot obtained by using all historical data that occurred for model's calculation, but historical data selected in a near period of time as input data of the model will effectively improve the its prediction. This also verifies that data is time-effective and data that is too old will affect the judgment of the model.

In addition, Fig. 8 shows longitudinally comparative standard deviations of evaluation indicators of different years. Most changes were less than 0.02 no matter evaluation indicators of which year or which model. Especially for the comprehensive evaluation indicators of F1-score and G-mean, most changes were less than 0.01. The relatively large standard deviations all appeared in evaluation indicators of Precision, but were all within the acceptable range, with maximum value only 0.0592. This indicated that no matter financial indicator data of how many years selected, they had no effect on the model's performance unless seven to nine financial indicators were selected to build final panel data.

5.3. *Comparative analysis between improved and original ensemble models*

In the data preprocessing part, the new model proposed in this paper improved grey relational analysis in order to find the most suitable sample indicators more accurately; in the classifier construction part, it integrates bagging with CNN to improve prediction accuracy of the classifier. The model based on unimproved algorithm was used as a comparative model for comparative analysis to verify whether the improvement process could effectively improve the performance of the new model in processing data of financial indicators in time series. As a result, three comparative analysis models were constructed respectively, namely Model 2 (grey relational analysis + CNN_bagging), Model 3 (improved grey relational analysis + CNN) and Model 4 (grey relational analysis + CNN). In order to keep consistent with the optimal operation process of the new model, three comparison models in this section still selected three numbers - severn, or eight, or nine financial indicators - as the sample. Each comparison model has been respectively calculated 50

Table 5
Prediction results of different years.

Years	Accuracy		Recall		Precision		F1-score		G-mean	
	Mean%	S.D.	Mean%	S.D.	Mean%	S.D.	Mean%	S.D.	Mean%	S.D.
1	93.44	0.0111	80.95	0.0118	91.33	0.0592	85.69	0.0183	88.66	0.0026
2	92.55	0.0053	82.99	0.0118	87.77	0.0118	85.31	0.0106	89.22	0.0075
3	93.80	0.0053	81.63	0.0165	93.96	0.0310	87.29	0.0090	89.47	0.0062
4	92.73	0.0081	78.91	0.0118	92.26	0.0460	85.01	0.0126	87.75	0.0016
5	92.73	0.0111	79.59	0.0000	91.55	0.0438	85.12	0.0191	88.03	0.0068
6	93.26	0.0061	78.23	0.0118	95.08	0.0238	85.82	0.0121	87.81	0.0073

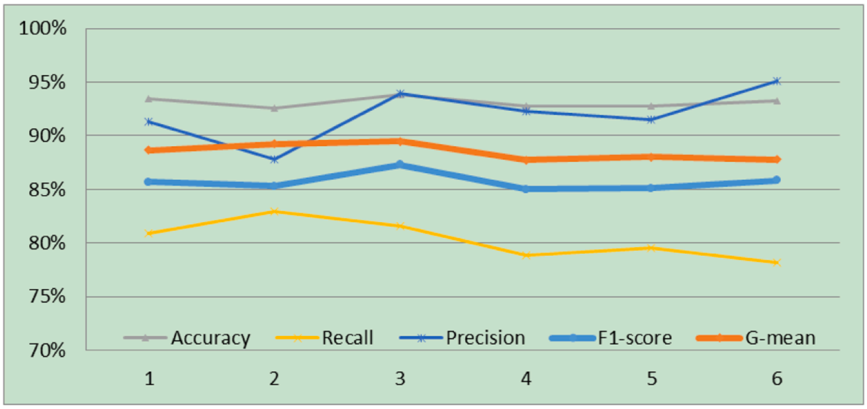


Fig. 6. F1-score and G-mean of indicators of different years.

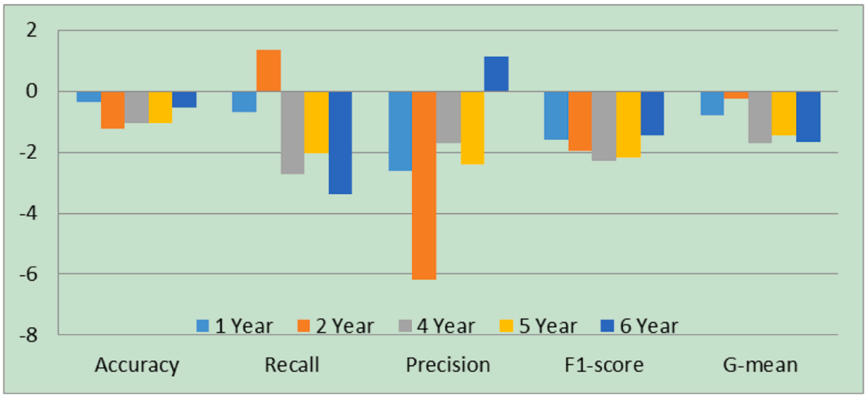


Fig. 7. The difference of means of evaluation measures of different years.

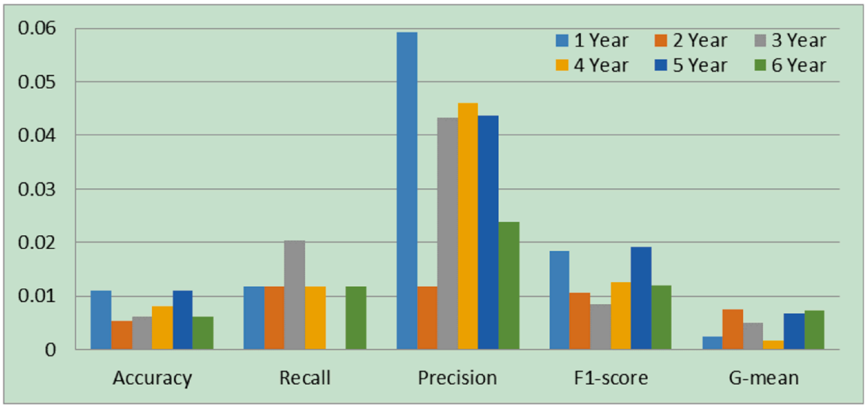


Fig. 8. The S.D. of evaluation measures of different years.

times in the condition of three numbers of indicators to achieve mean values and standard deviations of evaluation indicators. The results were depicted in a class boxplot, where median values were replaced by mean values in Table 6.

In evaluation indicators of five models in Fig. 9, our new model has the better maximum, minimum and mean values than other three models. Especially for F1-score and G-mean, the optimal values of the new model were 90.11 and 91.14%, the mean values were 87.29 and 89.47%, the minimum values were higher than 85.11 and 88.57%. All these values were significantly higher than corresponding values of other three models. Additionally, it can be found that the results of Model 2 are generally better than those of Model 3 and Model 4 compared to the results from Model 2 to Model 4, while the results of Model 3 are slightly superior than those of Model 4, especially in terms of F1-score and G-mean. Hence, the new model proposed has even better and superior performance with both improved methods fused.

Moreover, T-test and non-parametric Wilcoxontest were adopted for paired samples in 50 times of experiments to investigate if the results of four models were significantly different, as shown in Table 7. That was a way to infer whether the results of comparative analysis were generated by chance. The values marked with * illustrated that there was no significant difference at the 1% level. It can be seen that, except for evaluation indicators Recall and Precision of the new model and Model 2, all the other indicators were considered to be significantly different from each other at the 1% level. Consequently, it was clear that the new model's performance was better than the other three models indeed.

Standard deviations of evaluation indicators of different models were compared in Fig. 10. No matter which model or which evaluation indicator, most changes were less than 0.02, especially for F1-score and G-mean, all the changes were less than 0.02. The relatively large standard deviations all appeared in evaluation indicators of Precision, but were all within the acceptable range, with maximum value only 0.0350. This demonstrated that the results calculated were relatively stable in any model with floating range limited.

5.4. Comparative analysis among different models

This study took some commonly used intelligent algorithms as comparative models for comparative analysis in order to prove that our new model was able to more accurately predict corporate credit risk when facing data of financial indicators in time series. These comparison models included SVM, DT, RF, MLP and LSTM, and these models became the ensemble model after Bagging. Since most of them had no capacities to deal with data in time series, we extracted data of financial indicators of the latest year in every sample company's dataset as sample data for comparative analysis.

It was common to reduce dimensions when there were too many data characteristics. The principal component analysis (PCA) was used to reduce data dimensions, improve the accuracy of each comparative model, and make the comparison results more credible. Therefore, this study employed PCA to process the sample data of the comparison models with SPSS software. The results of KMO and Bartlett tests are shown in Table 8. The KMO value is greater than 0.7, and the results are valid and better. 12 factors are finalized, that is, the dimensions of comparison models input is 12. This means the sample data is reduced from original 41 dimensions to 12 dimensions.

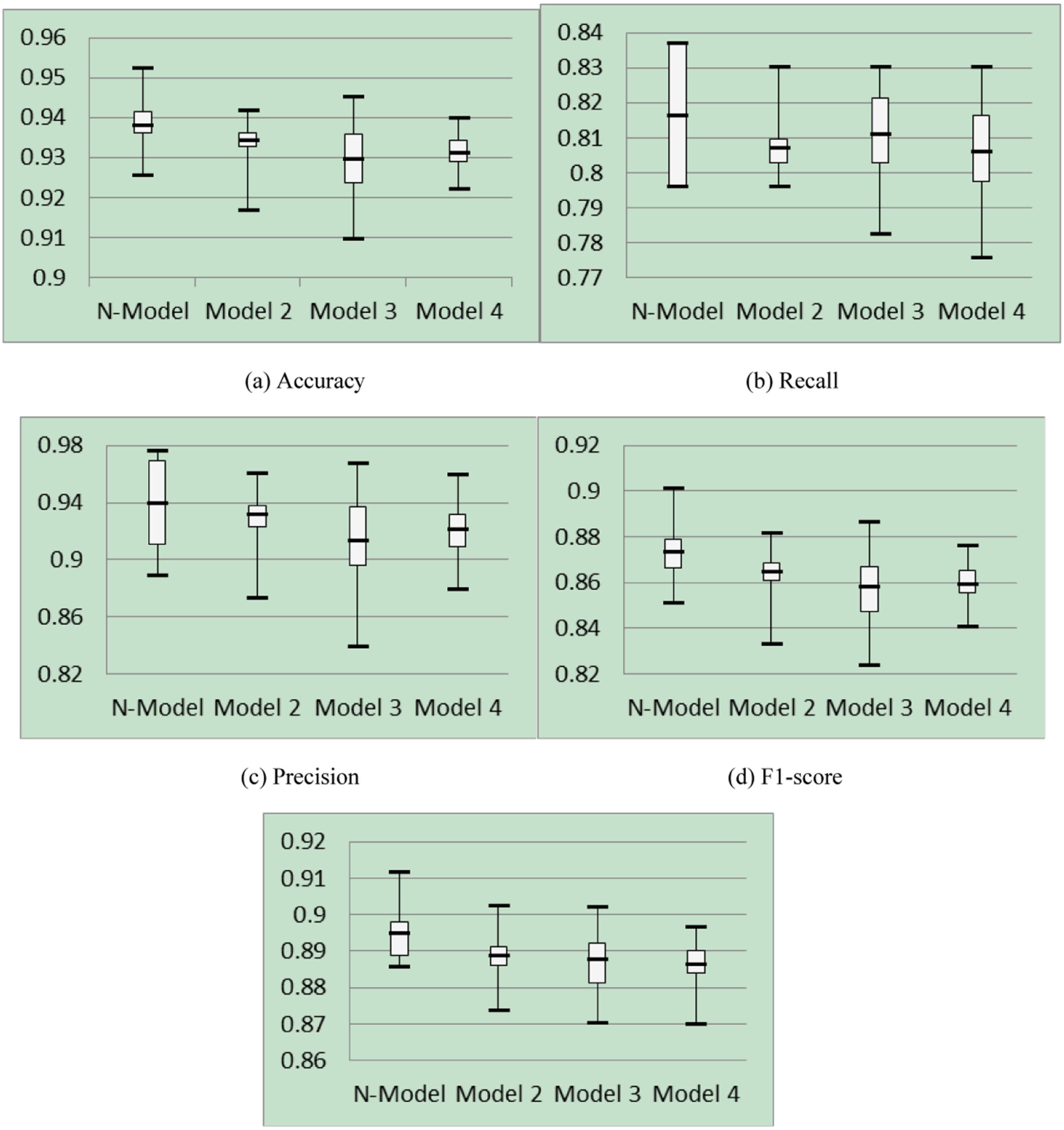
The new model still utilized three numbers - seven, eight and nine financial indicators as the sample data. Each comparative model was calculated 50 times to achieve mean values and standard deviations of evaluation indicators. As shown in Table 9, the underlined italics were the results of the ensemble model after Bagging, and the rest were the results of the model that has not been integrated. The class boxplot was constructed by the results after 50 times of calculation, and median value was replaced by mean value, as shown in Figs. 11 and 12.

For five models that have not been integrated, in the evaluation indicators of Recall, the optimal prediction results of the new model were all less than DT, MLP and LSTM, but mean values were only lower than DT and MLP. Furthermore, the new model showed very superior prediction results compared to evaluation indicators of the other four models. Especially for F1-score and G-mean, the new model's optimal values were 90.11% and 91.14%, respectively, slightly lower than 91.30% and 92.32% of LSTM and 93.39% of MLP. However, mean values (87.29% and 89.47%) and the lowest values (85.11 and 88.57%) of the new model were significantly higher than the corresponding values of other five models. Therefore, such comparative analysis was capable of confirming that the new model proposed had better prediction performance when dealing with data of financial indicators in time series. For the five models integrated by Bagging respectively, the comparison results obtained from Recall were the same with that from the models without being integrated. As to Accuracy and Precison, the mean value of the new model was slightly lower than that of LSTM. However, as to F1-score and G-mean, although the optimal value of the new model was slightly lower than that of MLP or LSTM, the mean and minimum values of the new model were not better than other comparison models, this is consistent with the comparison results of the unensemble model. In addition, compared with the models integrated before and after Bagging, Bagging had the ability to improve the performance of the model to some extent.

Additionally, T-test and nonparametric Wilcoxontest were still used for paired samples to examine if the results of six models were significantly different, as shown in Table 10. The underlined italics were the result of the model ensembled with Bagging, and the rest were the results of unensembled models. The values marked with * implied that there was no significant difference at the 1% level. Also, except for the Recall of the new model and MLP, the Recall and Precision of the new model and LSTM, the Recall of the new model and SVM with Bagging, the G-mean of the new model and MLP with Bagging, the Accuracy and F1-score of the new model and LSTM with Bagging, all the other indicators were considered to be significantly different from each other at the 1% level. Therefore, it was

Table 6
Prediction results of the improved and the original ensemble models.

Models	Accuracy		Recall		Precision		F1-score		G-mean	
	Mean%	S.D.	Mean%	S.D.	Mean%	S.D.	Mean%	S.D.	Mean%	S.D.
New Model	93.80	0.0053	81.63	0.0165	93.96	0.0310	87.29	0.0090	89.47	0.0062
Model2	93.41	0.0070	80.71	0.0129	93.15	0.0167	86.46	0.0143	88.88	0.0093
Model3	92.96	0.0104	81.09	0.0073	91.26	0.0350	85.78	0.0191	88.74	0.0090
Model4	93.10	0.0043	80.59	0.0145	92.08	0.0075	85.90	0.0096	88.63	0.0082



(a) Accuracy

(b) Recall

(c) Precision

(d) F1-score

(e) G-mean

Fig. 9. A class boxplot of 50-time calculation results of original ensemble model.

Table 7
The significance test of evaluation measures for common models.

Method	Measure	Model 2		Model 3		Model 4	
		T-test	Wilcoxon-test	T-test	Wilcoxon-test	T-test	Wilcoxon-test
New Model	Accuracy	0.000	0.000	0.000	0.000	0.000	0.000
	Recall	0.001	0.000	0.067*	0.044	0.000	0.001
	Precision	0.120*	0.106*	0.000	0.000	0.001	0.001
	F1-score	0.000	0.000	0.000	0.000	0.000	0.000
	G-mean	0.000	0.000	0.000	0.000	0.000	0.000

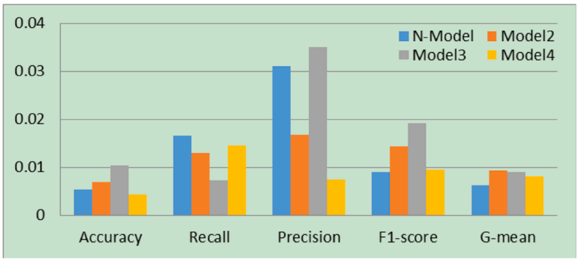


Fig. 10. The S.D. of evaluation measures for 4 ensemble models.

Table 8
KMO and Bartlett's test in SPSS.

KMO and Bartlett's test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.755
Bartlett's Test of Sphericity	Approx. Chi-Square	31,460.129
	df	820
	Sig.	.000

Table 9
Prediction results of different base classifiers.

Models	Accuracy		Recall		Precision		F1-score		G-mean	
	Mean%	S.D.	Mean%	S.D.	Mean%	S.D.	Mean%	S.D.	Mean%	S.D.
New Model	93.80	0.0053	81.63	0.0165	93.96	0.0310	87.29	0.0090	89.47	0.0062
SVM	88.30	0.0000	79.59	0.0000	76.47	0.0000	78.00	0.0000	85.28	0.0000
	91.49	*0.0000*	*81.63*	*0.0000*	*85.11*	*0.0000*	*83.33*	*0.0000*	*88.05*	*0.0000*
DT	87.14	0.0078	84.73	0.0103	71.35	0.0169	77.46	0.0117	86.34	0.0074
	90.31	*0.0143*	*85.92*	*0.0211*	*79.03*	*0.0416*	*82.26*	*0.0226*	*88.79*	*0.0119*
RF	92.02	0.0000	77.55	0.0000	90.48	0.0000	83.52	0.0000	86.79	0.0000
	92.55	*0.0000*	*79.59*	*0.0000*	*90.70*	*0.0000*	*84.78*	*0.0000*	*87.92*	*0.0000*
MLP	90.78	0.0217	83.14	0.0498	82.33	0.0649	82.51	0.0377	88.09	0.0259
	92.84	*0.0164*	*83.47*	*0.0317*	*89.00*	*0.0618*	*85.93*	*0.0263*	*89.48*	*0.0129*
LSTM	93.16	0.0152	79.31	0.0600	93.58	0.0277	85.69	0.0387	88.10	0.0343
	93.86	*0.0093*	*80.16*	*0.0303*	*95.62*	*0.0225*	*87.15*	*0.0203*	*88.93*	*0.0170*

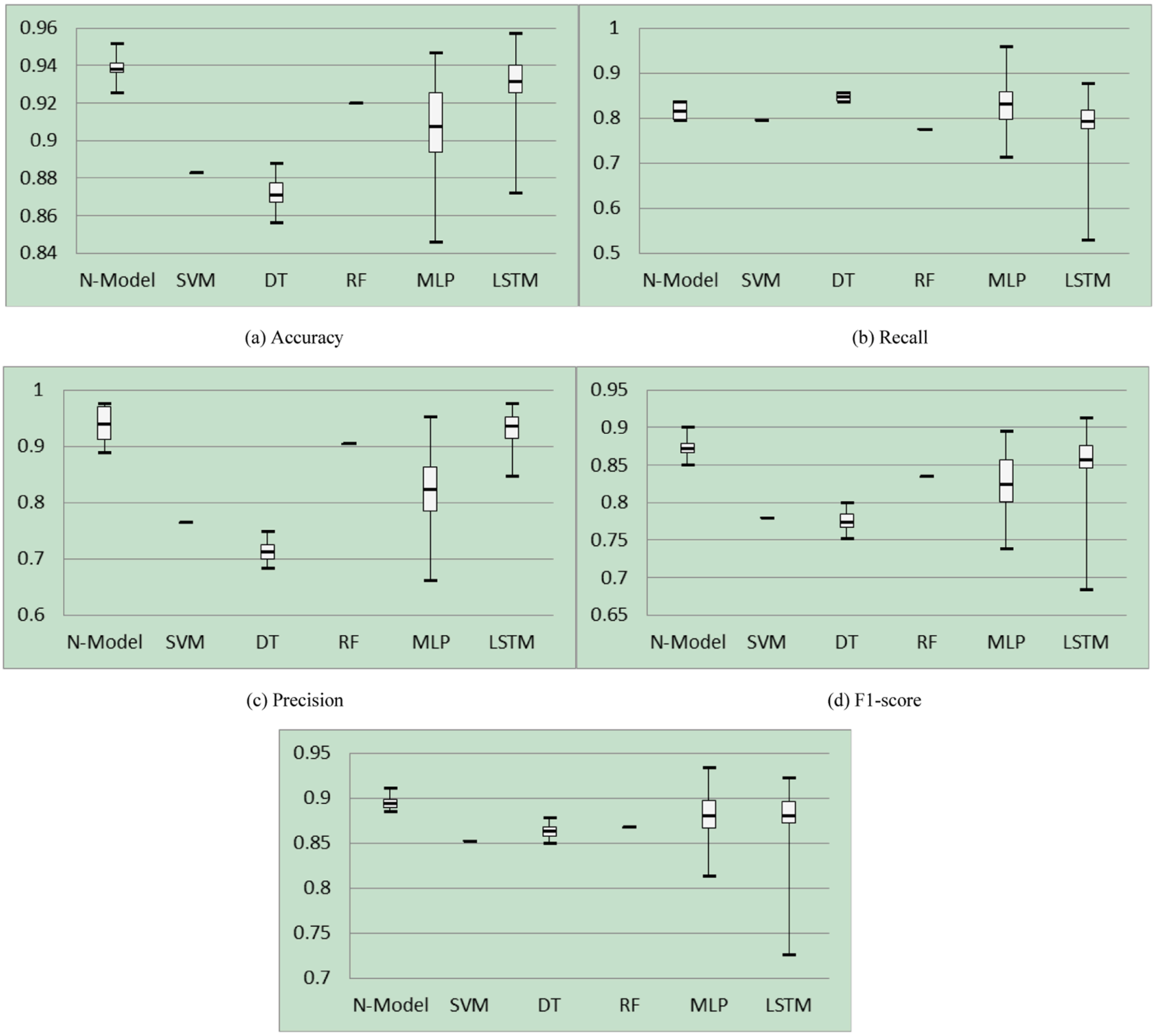
found that the new model's performance was better than the other five models indeed.

Standard deviations of evaluation indicators of different models were compared in Fig. 13. For the comparison models ensembled, the output of results was quite stable in SVM and RF. Most changes in evaluation indicators of the new model and DT were less than 0.02, especially for F1-score and G-mean, all the changes were around 0.01. For MLP and LSTM, most changes were greater than 0.02, and the changes of F1-score and G-mean were around 0.03, but all were less than 0.07. The comparison models ensembled with Bagging showed the similar conclusions, but their standard deviations were generally lower than those of the corresponding unensembled models, indicating much more stable performance degree. By comparison, it can be seen that SVM and RF models had the most stable prediction performance, and the new model and DT had better stability. Even though MLP and LSTM were the least stable among them, prediction results were fluctuated within the allowable range.

6. Implications

Building complex ensemble models is a common practice used to effectively improve early warning effect in credit risk, because each model or method has its disadvantages in algorithms, but such a complex ensemble model is often difficult to analyze the source of risk for companies. Even if some studies take advantage of interpretability methods to analyze risks, their processes of interpretation are independent of early warning models, resulting in failure of recognizing what specific factors that will be related to predictive performance of the model based on predictive results. As a result, a new two-stage ensemble model is proposed on the basis of multiple machine learning to both effectively improve accuracy of credit risk early warning and qualitatively analyze sources of risk and more importantly,it can analyze the relevant factors that improve predictive performance based on predictive results. In this sense, this study has certain implications in terms of model construction method and practical application.

First, this paper enriches the studies that build complex ensemble models to analyze credit risk early warning effect by constructing the new two-stage ensemble model with the most superior predictive performance compared to traditional methods, by innovatively employing the improved grey relation analysis and convolutional neural network to improve the prediction performance, by adopting



(a) Accuracy

(b) Recall

(c) Precision

(d) F1-score

(e) G-mean

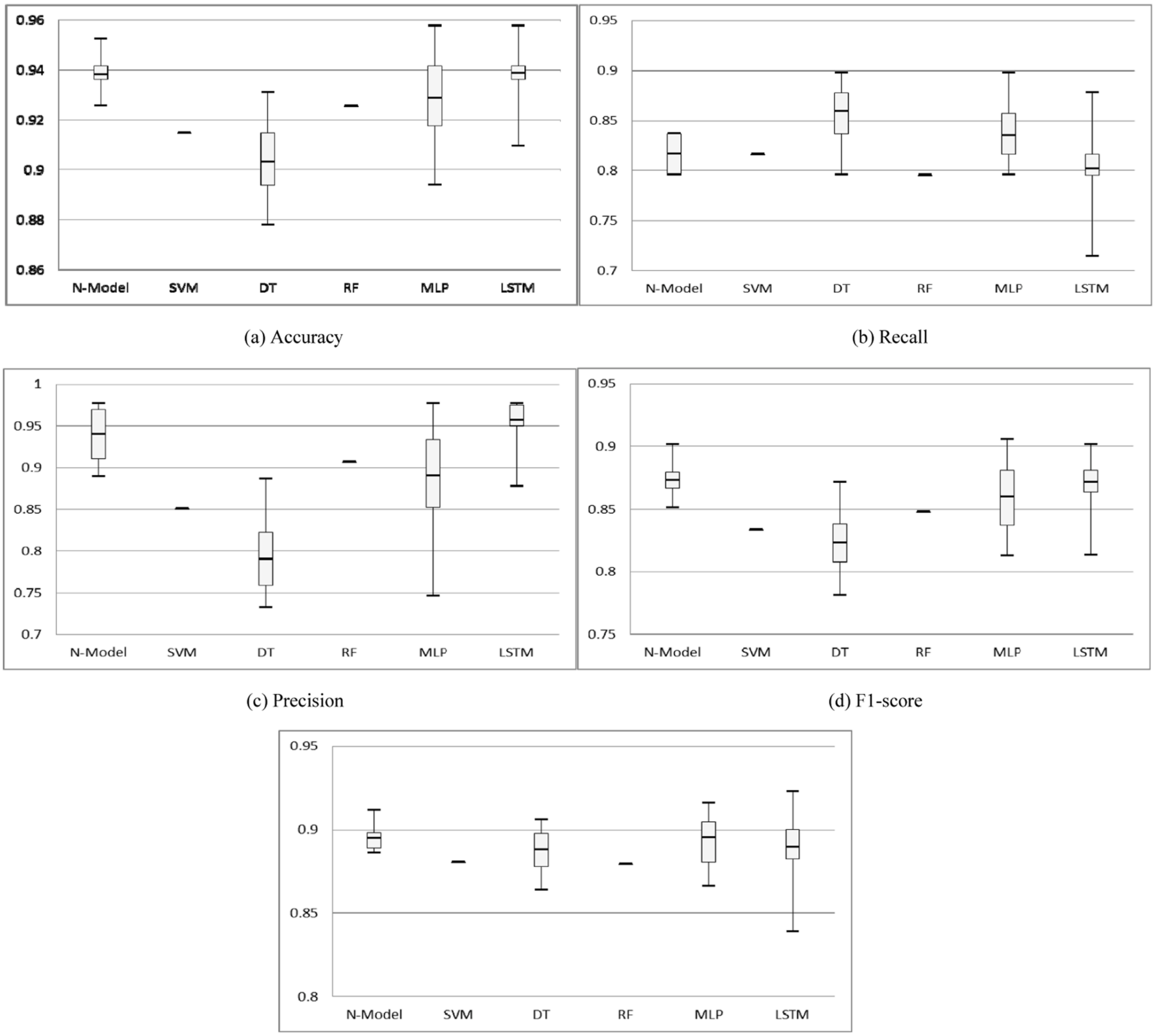
Fig. 11. A class boxplot of 50-time calculation results of comparative models.

four separate experimental methods to compare and analyze financial panel data to verify that sources of risk can be analyzed qualitatively, it is not in attribute dimension that the more financial indicators, the more accurate prediction performance, as well as it is not in time dimension that the larger the time span of the selected data, the more prediction accuracy improved.

Second, this paper complements the deficiencies of existing studies analysing credit risk early warning effect by improving the entropy method with grey correlation analysis as basis to make more reasonable judgment on correlation degree of each indicator because the entropy method has the problem of unreasonable weight distribution, and further by convolutional neural network(CNN) with better effect as base classifier and adopts the Bagging method to ensemble so as to further improve early warning effect, by borrowing the idea of N-fold cross-validation to divide training sets to differentiate base classifiers and thus increase prediction effect.

Third, this paper extends the extant studies focusing on credit risk early warning that stay in one method utilized to explain models or indicator data by integrating explainable methods into the prediction process, and perform qualitative analysis on how to improve the prediction results of models, and by according with the current developing trend of using qualitative analysis based on prediction results, meanwhile by satisfying the suggestions raised by existing scholars to analyze by building models and utilizing diverse interpretability and interpretable methods in the model.

Our results suggest professionals in the industy should clarify the way of how to choose the most appropriate years to build data sets in time sequence, especially in time selection, they should not choose the financial indicators in longer time span as the base for credit risk early warning but select the recent financial indicators. In other words, it should obtain the latest data of companies in a certain frequency and put them into the model to update the calculation results and realize the dynamic monitoring and warning of companies' debt risk. Meanwhile, in indicator dimension, it should lay more emphasis on observing earnings per share and operation profit per share as both of them are always highly correlated with net profit, signifying that these two indicators played a key role in judging



(a) Accuracy

(b) Recall

(c) Precision

(d) F1-score

(e) G-mean

Fig. 12. A class boxplot of 50-time calculation results of comparative models with Bagging.

the type of sample companies, instead of covering all financial indicators.

Professionals should pay more attention to the innovation and improvement of early warning methods by building complex ensemble models containing various machine learning methods, applying innovatively machine learning models, giving full play to the advantages of big data, mining massive, multidimensional and dynamic data to improve the accuracy, timeliness and foresight of monitoring and early warning. Professionals should re-examine the important value of financial indicators in credit risk early warning, especially the abnormal information financial data cannot reflect in the process of operation. Also, they should examine the profitability of companies from different aspects, combine financial indicators in multiple dimensions to corroborate each other, identify information distortion, and then evaluate credit risk.

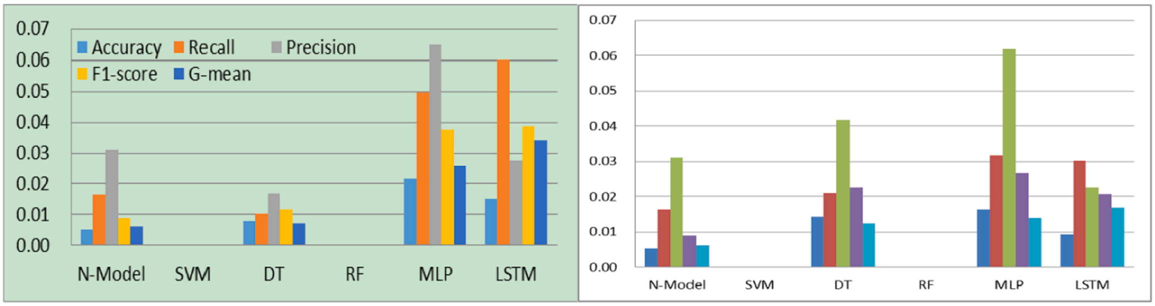
7. Conclusions and future works

Corporate credit risk warning is able to effectively supervise the risks that will be encountered in the process of operation and reduce the possibility of bankruptcy. Specific to indicator data in time series, this study builds a two-stage ensemble model for early warning of corporate credit risk based on the improved grey correlation analysis and CNN. This model can not only effectively improve the accuracy of corporate credit risk prediction, but also qualitatively analyzes prediction results from multiple perspectives. In empirical section, within the scope of listed companies in China’s manufacturing industry, this paper selects 121 sample companies that are especially treated for the first time and have been continuously listed for more than 15 years from 2012 to 2021 as the samples with credit risk, each of which contains the indicator data from the year of listing to the year of being specially treated; this study also takes 346 sample companies that have been listed for more than 15 consecutive years and have never been specially treated as the

Table 10
The significance test of evaluation measures for common models.

Method	Measure	SVM		DT		RF	
		T-test	Wilcoxon-test	T-test	Wilcoxon-test	T-test	Wilcoxon-test
New Model	Accuracy	0.000	0.000	0.000	0.000	0.000	0.000
		0.000	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*
	Rec	0.000	0.000	0.000	0.000	0.000	0.000
		*0.591**	*0.592**	*0.000*	*0.000*	*0.000*	*0.000*
	Pre	0.000	0.000	0.000	0.000	0.000	0.000
		0.000	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*
	F1-score	0.000	0.000	0.000	0.000	0.000	0.000
		0.000	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*
	G-mean	0.000	0.000	0.000	0.000	0.000	0.000
		0.000	*0.000*	*0.002*	*0.001*	*0.000*	*0.000*

Method	Measure	MLP		LSTM	
		T-test	Wilcoxon-test	T-test	Wilcoxon-test
New Model	Accuracy	0.000	0.000	0.006	0.009
		0.000	*0.001*	*0.708**	*0.437**
	Rec	0.047*	0.049*	0.012*	0.014*
		0.000	*0.001*	*0.003*	*0.006*
	Pre	0.000	0.000	0.575*	0.655*
		0.000	*0.000*	*0.006*	*0.018*
	F1-score	0.000	0.000	0.006	0.005
		0.002	*0.003*	*0.644**	*0.792**
	G-mean	0.000	0.001	0.000	0.000
		*0.755**	*0.842**	*0.005*	*0.007*



(a) without Bagging (b) with Bagging

Fig. 13. The S.D. of evaluation measures for comparative models.

samples with no credit risk, as well as chooses indicator data from the listing date to the year of 2021. Then, this paper utilizes four experiments to verify the superiority of this model and qualitatively analyzes based on prediction results from multiple angles.

The new findings are obtained from experiments in this article: First, the results obtained in financial indicator identify that earnings per share (X40), net profit margin on total assets (X15), net profit margin on current assets (X16), operation profit per share (X41), total profit cost ratio (X24) originating from profitability and stockholder's profitability, all maintain a high degree of correlation to net profit compared to other indicators. This correlation shows a downward trend in companies with credit risk: earnings per share (X40) > net profit margin on total assets (X15) > net profit margin on current assets (X16) > operation profit per share (X41); in companies with normal operation: earnings per share (X40) > operation profit per share (X41) > total profit cost ratio (X24). It is worth notifying that, no matter what type of sample companies, earnings per share (X40) and operation profit per share (X41) always maintain a high correlation with net profit, signifying that these two indicators played a key role in judging the type of sample companies. Also, it compares prediction results achieved after inputting different numbers of financial indicators into the model and the results demonstrate that it is not the more the number of indicators, the higher the prediction accuracy. The best prediction effect occurs when selecting seven to nine financial indicators.

Second, in time dimension, it compares prediction results from indicator data in different time ranges and the results indicate that it is not the longer the time range, the better the prediction effect. Further, the best prediction results can be achieved when selecting the data within three years.

Third, compared with the unimproved grey relational analysis and CNN model and the commonly used models for credit risk early warning (SVM, DT, RF, MLP, LSTM), the new two-stage ensemble model proposed has the optimal F1-score value (87.29%) and G-mean value (89.47%), implying its better performance in corporate credit risk early warning.

Additionally, this article has some limitations and future research directions. First, it only selects panel data of limited financial indicator categories from manufacturing-listed companies. Follow-up researches are encouraged to continuously expand the scope of categories to further verify this model. Second, this study only separate samples into two types, with risk and without risk. Future research is encouraged to utilize multi-class classification method to consider more about what other types of samples that can be achieved.

Data availability

Data will be made available on request.

Acknowledgments

The authors acknowledge and appreciate all of the editors and reviewers who gave us high quality comments and suggestions. This work was supported by the National Natural Science Foundation of China (Grant #71801187, #72104115), Humanities and Social Sciences Research Project of the Ministry of Education of China (#20YJC630218), Project Supported by China Scholarship Council (CSC No. 202008370107).

References

Agarwal, P., Tamer, M., & Budman, H. (2021). Explainability: Relevance based dynamic deep learning algorithm for fault detection and diagnosis in chemical processes. *Computers & Chemical Engineering, 154*, Article 107467.

Bhattacharya, S., Ramalingam, D., Gunaseelan, D., Ramkrishna, S., & Sandhya, M. (2017). Improvement of e-polylysine production by marine bacterium Bacillus licheniformis using artificial neural network modeling and particle swarm optimization technique. *Biochemical Engineering Journal, 126*, 8–15.

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications, 39*(3), 3446–3453.

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics, 57*(1), 203–216.

Chen, F. L., & Li, F. C. (2010). Combination of feature selection approaches with SVM in credit scoring. *Expert Systems with Applications, 37*(7), 4902–4909.

Chen, Y. S., & Cheng, C. H. (2013). Hybrid models based rough set classifiers for setting credit rating decision rules in the global banking industry. *Knowledge-Based Systems, 39*, 224–239.

Chen, W., Wang, X., Wang, W., et al. (2021). A heterogeneous GRA-CBR-based multi-attribute emergency decision-making model considering weight optimization with dual information correlation. *Expert Systems with Applications, 182*, Article 115208.

Chi, B. W., & Hsu, C. C. (2012). A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model. *Expert Systems with Applications, 39*(3), 2650–2661.

Dai, W. (2022). Application of improved convolution neural network in financial forecasting. *Journal of Organizational and End User Computing, 34*(3), 1–16.

Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing, 91*, 1–21.

Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In. In *Proceedings of the IEEE symposium on security and privacy* (pp. 598–617).

Dumitrescu, E., Hue, S., Hurlin, C., & Tokpavi, S. (2021). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research, 297*(3), 1178–1192.

Elshawi, R., Sherif, Y., Al-Mallah, M., & Sakr, S. (2021). Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence, 37*, 1633–1650.

Fu, X. L., Ouyang, T. X., Chen, J. P., & Luo, X. (2020). Listening to the investors: A novel framework for online lending default prediction using deep learning neural networks. *Information Processing & Management, 57*(4), Article 102236.

Goerigk, M., Lendl, S., & Wulf, L. (2022). Two-stage robust optimization problems with two-stage uncertainty. *European Journal of Operational Research, 302*, 62–78.

Gramespacher, T., & Posth, J. A. (2021). Employing explainable AI to optimize the return target function of a loan portfolio. *Frontiers in Artificial Intelligence, 4*, Article 693022. Article.

Han, L., Han, L., & Zhao, H. (2013). Orthogonal support vector machine for credit scoring. *Engineering Applications of Artificial Intelligence, 26*, 848–862.

Henley, W. E., & Hand, D. J. (1996). A k-nearest-neighbour classifier for assessing consumer credit risk. *Journal of The Royal Statistical Society Series D-The Statistician, 45*(1), 77–95.

Hotchkiss, E., & Altman, E. I (2006). *Corporate financial distress and bankruptcy*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Huang, J., Chai, J., & Cho, S. (2020). Deep learning in finance and banking: A literature review and classification. *Frontiers of Business Research in China, 14*(1), 1–24.

Hu, Y. C. (2020). A multivariate grey prediction model with grey relational analysis for bankruptcy prediction problems. *Soft Computing, 24*, 4259–4268.

Jardin, P. (2016). A two-stage classification technique for bankruptcy prediction. *European Journal of Operational Research, 254*, 236–252.

Kadier, P., Abdeshahian, Y., Simayi, M., Ismail, A., Hamid, A., & Kalil, M. S. (2015). Grey relational analysis for comparative assessment of different cathode materials in microbial electrolysis cells. *Energy, 90*, 1556–1562.

Kliegr, T., Bahnik, S., & Furnkranz, J. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence, 295*, Article 103458. Article.

Kim, J. Y., & Cho, S. B. (2019). Towards repayment prediction in peer-to-peer social lending using deep learning. *Mathematics, 7*(11), 1–17.

Lecun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation, 1*(4), 541–551.

Liang, D., Tsai, C. F., & Wu, H. T. (2015). The effect of feature selection on financial distress prediction. *Knowledge-Based Systems, 73*, 289–297.

Mi, J. X., li, A. D., & Zhou, L. F. (2020). Review study of interpretation methods for future interpretable machine learning. *IEEE Access : Practical Innovations, Open Solutions, 8*, 191969–191985.

Moraffah, R., Karami, M., Guo, R. C., Raglin, A., & Liu, H. (2020). Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter, 22*(1), 18–33.

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications, 165*(9), Article 113986.

Park, M. S., Son, H., Hyun, C., & Hwang, H. J. (2021). Explainability of machine learning models for bankruptcy prediction. *IEEE Access : Practical Innovations, Open Solutions, 9*, 124887–124899.

Peng, J. F., Zou, K. Q., Zhou, M., Teng, Y., Zhu, X. Y., Zhang, F. F., et al. (2021). An explainable artificial intelligence framework for the deterioration risk prediction of hepatitis patients. *Journal of Medical Systems, 45*(5), 61.

Oreski, S., & Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. *Expert Systems with Applications, 41*(4), 2052–2064.

Seo, Y., & Shin, K. (2019). Hierarchical convolutional neural networks for fashion image classification. *Expert Systems with Applications, 116*, 328–339.

Sezer, O. B., & Ozbayoglu, A. M. (2018). Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach. *Applied Soft Computing, 70*, 525–538.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies, 146*, Article 102551. Article.

Sinnl, M. (2022). Exact and heuristic algorithms for the maximum weighted submatrix coverage problem. *European Journal of Operational Research, 298*(3), 821–833.

So, Y. S., Dong, H. K., & Jin, H. Y. (2016). Technology credit scoring model with fuzzy logistic regression. *Applied Soft Computing, 43*, 150–158.

Sun, J., Fujita, H., Zheng, Y. J., & Ai, W. G. (2021). Multi-class financial distress prediction based on support vector machines integrated with the decomposition and fusion methods. *Information Sciences, 559*, 153–170.

Sun, J., Lang, J., Fujita, H., & Li, H. (2018). Imbalanced enterprise credit evaluation with dte-sbd: Decision tree ensemble based on smote and bagging with differentiated sampling rates. *Information Sciences, 425*, 76–91.

Wang, F., Ding, L., Yu, H., & Zhao, Y. (2020). Big data analytics on enterprise credit risk evaluation of e-business platform. *Information Systems and e-Business Management, 18*, 311–350.

Wang, J., Hedar, A. R., Wang, S., & Ma, J. (2012). Rough set and scatter search metaheuristic based feature selection for credit scoring. *Expert Systems with Applications, 39*(6), 6123–6128.

Wu, L., Liu, J., Zhou, J., Zhang, Q., Song, Y., Du, S., et al. (2022). Evaluation of tar from the microwave co-pyrolysis of low-rank coal and corncob using orthogonal-test-based grey relational analysis (GRA). *Journal of Cleaner Production, 337*, Article 130362.

Wu, W., Kou, G., & Peng, Y. (2016). Group decision-making using improved multi-criteria decision making methods for credit risk analysis. *Filomat, 30*(15), 4135–4150.

Xia, P., Ni, Z., Xiao, H., Zhu, X., & Peng, P. (2022). A novel spatiotemporal prediction approach based on graph convolution neural networks and long short-term memory for money laundering fraud. *Arabian Journal for Science and Engineering, 47*(2), 1921–1937.

Xia, Y., Liu, C., & Li, Y. Y (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications, 78*, 225–241.

Yang, Y. C., & Liu, H. P. (2016). Determinants of banking sector's credit granting policy for the yacht industry in Taiwan. *Maritime Business Review, 1*, 55–75.

Yin, L. L., Qin, Y. W., Hou, Y., & Ren, Z. J. (2022). A convolutional neural network-based model for supply Chain financial risk early warning. *Computational Intelligence and Neuroscience, 4*, Article 7825597.

Yoichi, H., & Naoki, T. (2020). One-dimensional convolutional neural networks with feature selection for highly concise rule extraction from credit scoring datasets with heterogeneous attributes. *Electronics, 9*(8), 1–15.

Youn, H., & Gu, Z. (2010). Predicting Korean lodging firm failures: An artificial neural network model along with a logistic regression model. *International Journal of Hospitality Management, 29*(1), 120–127.

Zhang, W., Yan, S., Li, J., Tian, X., & Taketoshi, Y. (2022a). Credit risk prediction of SMEs in supply chain finance by fusing demographic and behavioral data. *Transportation Research Part E, 158*, Article 102611.

Zhu, Y., Zhou, L., Xie, C., Wang, G. J., & Nguyen, T. V. (2019). Forecasting SMEs'credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach. *International Journal of Production Economics, 211*, 22–33.

Zhang, Z. J., Wu, C., Qu, S. Y, & Chen, X. F (2022b). An explainable artificial intelligence approach for financial distress prediction. *Information Processing and Management, 59*, Article 102988. Article.