Journal of Network Intelligence ©2025 ISSN 2414-8105
Taiwan Ubiquitous Information Volume **10**, Number **4**, November **2025**

Corporate Financial Crisis Early Warning Based on Integrated Learning and Genetic Algorithm in Big Data Environment

Hao Cheng*

School of Economics and Management
Shanghai Zhongqiao Vocational and Technical University
Shanghai 201514, P. R. China
tmmmmmb_12@163.com

Bo Tang

School of Economics and Management
Shanghai Zhongqiao Vocational and Technical University
Shanghai 201514, P. R. China
cheney202407@163.com

Zhao-Heng Jin

Sangji University
Jiangyuandao 26339, South Korea
luka0770731@163.com

*Corresponding author: Hao Cheng

Received August 8, 2024, revised February 25, 2025, accepted August 20, 2025.

Abstract. *With the gradual slowdown of economic growth, the pressure of economic downturn faced by enterprises is also increasing day by day, and it is of great importance to establish an efficient financial crisis early warning model for enterprises to warn the financial situation of companies in advance. Most of the existing studies focus on the optimization and prediction of a single model in the context of big data, and the prediction accuracy is not high. Therefore, this article suggests a corporate financial crunch early warning method relied on integrated learning and Genetic Algorithm (GA) in big data environment. Firstly, in view of the issues that the traditional integrated learning algorithm is prone to poor network generalization ability, genetic manipulation and species invasion of GA are used to obtain a globally optimal individual learning classifier. Then SMOTE algorithm was used to oversampling the enterprise sample to address the issue of sample imbalance. Secondly, financial early warning index was constructed, Principal Component Analysis (PCA) method was adopted to eliminate the remaining indicators to prevent the interference of noise on the forecast outcome. Finally, GA was used to optimize the parameters to be adjusted of each primary learner. The data learned by the primary learner is handed over to the secondary learner to output the final prediction outcome. The simulation outcome implies that the forecasting accuracy of the designed method can reach 94.08%. Compared with the forecasting accuracy of other methods and the class II error rate, the proposed model has a good early warning effect.*

Keywords: financial crisis warning; stacking integrated learning; genetic algorithm; principal component analysis; SMOTE algorithm

1. Introduction.

In current data-driven era, big data technology, with its powerful data processing and analyzing capabilities, is profoundly changing the development trajectory of various industries, among which, the field of enterprise financial management is no exception. In the face of the complex and changing market environment and increasingly fierce market competition, how enterprises can timely and accurately identify and warn of potential financial crises has become a key issue for their survival and development [1]. In recent years, some enterprises have fallen into financial crisis because of cash flow, debt and other problems, or exposed many risks, causing many negative impacts on society. Enterprise economic crunch early warning management is still a significant part of the operation that can not be ignored. Improvement of the financial crunch early warning model can help enterprises to identify potential financial risks and establish a more effective financial risk management mechanism, thus reducing the probability of enterprises falling into financial crisis [2].

1.1. Related work.

The choice of enterprise financial crisis early warning method is a critical factor affecting the performance of the initial warning model. The study of the existing financial crisis initial warning method has experienced the evolution from statistical method to artificial intelligence technology. The financial crunch initial warning method was first relied on statistical methods. Charalambakis and Garrett [3] predicted the financial distress of British companies and built the J-UK financial crisis early warning model. Sevim et al. [4] applied the Logistic model to the research of enterprise economic crunch warning and achieved good results. Yu and Zhang [5] built the financial crunch early warning method of agricultural listed companies through factor analysis and binary logistic regression. Shang et al. [6] introduced Benford's law into the Logistic model of economic crunch early warning. Huang et al. [7] proposed using GA to optimize the weight vector of case features and the nearest neighbor retrieval method relied on grey similarity, and carried out the prediction modeling of corporate financial distress. Statistical models, although simple in structure, are still affected by multiple nonlinear factors [8]. To solve this problem, researchers applied machine learning algorithms to financial crisis early warning research and achieved better results. Gholizadeh et al [9] used decision tree technology to establish a economic distress early warning system for Chinese listed companies. Ahn et al. [10] conducted financial crunch early warning research relied on Korean and Chinese data respectively, confirming that the support vector machine method achieves a Al-Assaf [11] conducted a comparative study using Linear Discriminant Analysis (LDA) [12], logistic regression analysis [13] and Artificial Neural Network (ANN) methods, and the empirical outcome indicated that the ANN had a better forecsting performance. Sun and Lei [14] constructed a BP neural network model on the basis financial indicators to predict the occurrence of corporate financial crisis. Ruan and Liu [15] combined Genetic Algorithm (GA) and BP neural network for early warning of enterprises in the manufacturing industry, and the results proved the superiority of GA-BP model over logistic and traditional BPNN. Zhang and Luo [16] used dual data mining techniques of rough set and GA to optimize the neural network-based financial crisis early warning system and improve the prediction accuracy. Single classifier is limited by the model itself, and its performance improvement has fallen into a bottleneck, so many scholars turn their attention to integrated learning algorithms. Wang et al. [17] selected a number of financial indicators and established a financial crunch early warning model relied on random forest classification technology. Budhidharma et al. [18] proposed a random forest financial crunch early warning method relied on Benford. Gholizadeh et al. [19] built a CFW-Boost model by integrating multiple CART trees, which is superior to other models in accuracy and early warning performance. Liang et al. [20] introduced a

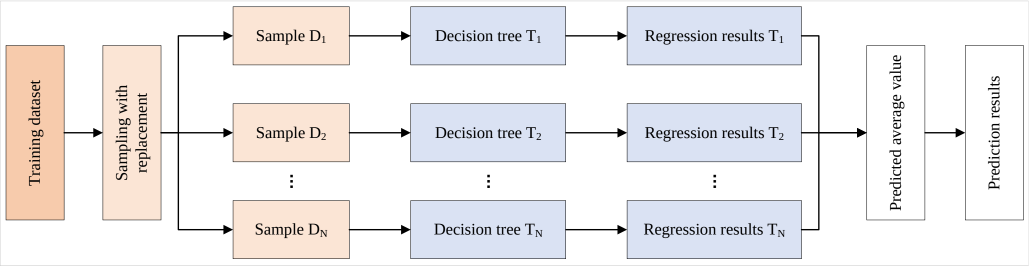

Figure 1. Schematic of stacking algorithm

new integrated algorithm Stacking. The algorithm does not restrict the types of learners and can combine the advantages of heterogeneous learners to form strong learners. Guo [21] used the integrated models of random forest, decision tree and BP neural network respectively to forecast the economic crunch of companies. The forecasting accuracy of the merged model was importantly better than that of a individual model.

1.2. **Contribution.** Throughout the study on financial early warning of enterprises in recent years, most scholars use a single variable model, which often has the problem of algorithm convergence and low prediction accuracy. To address the above issues, this paper designs an enterprise financial crunch early warning method relied on integrated studying and genetic algorithm in the big data environment, which has the following innovations:

(1) In view of the problems that existing integrated learning algorithms are prone to overfitting and poor network generalization ability, the continuous evolution of genetic algorithms to obtain the optimal integrated learning can accelerate the prediction speed, reduce the storage space requirements, and further improve the accuracy.

(2) Oversampleting the enterprise sample with SMOTE algorithm to address the issue of sample imbalance, construct economic early warning indicators, and use PCA method to remove redundant indicators to prevent noise interference on the forecast results.

(3) Primary learners with excellent performance, Random Forest and XGBoost, are stacked and merged, and the merged learning data is handed over to the secondary learner GA-BP neural network to output the final forecasting outcome.

(4) Experimental outcome implies that the designed method has higher prediction accuracy and lower Class I error rate and Class II error rate, which verifies the efficiency of the proposed model.

2. Theoretical analysis.

2.1. **Integrated learning algorithm.** By combining the prediction results of multiple classifiers, ensemble learning algorithms can make comprehensive use of their advantages to obtain more accurate predictions than a single learner [22]. Bagging, Boosting, and Stacking are all integration algorithms that use more [23]. Compared with Bagging and Boosting, Stacking has a higher efficiency in integrating multiple classifiers and can enhance the stability and generalization capability of the model.

The integrated learning algorithms in the Stacking process are generally divided into two level s. The first level is a primary learner and the second level is a secondary learner. The original data set is trained by the primary learner to produce a new data set, which is used to train the secondary learner and finally get the prediction result. In the process of training, different algorithm models are optimized to give complete play to their various benefits, for the goal of improving the forecasting accuracy of the whole model. Its structure is implied in Figure 1.

2.2. **Genetic algorithm.** GA is a casual seek optimization method which has been evolved recently. It is a casual seek method that draws on natural option and natural genetic system of biology, and is appropriate for addressing complicated improvement issues that are tough to be addressed by conventional seek algorithms, and can be broadly applied to combinatorial improvement, machine learning and other fields. The genetic operation of GA mainly includes election function, crossover function and mutation function [24].

(1) Election function. Choose the best peculiars from the population and eliminate the poor ones. Roulette selection method is the most normally adopted election method right now. Let the population size be m, where the fitness value of peculiar j is g_j, then the probability P_{s_j} of j being selected is

$$P_{s_j} = \frac{g_j}{\sum_{j=1}^{m} g_j}. \tag{1}$$

(2) Cross function. The crossover operator should guarantee that the traits of the prominent peculiars in the former generation can be awarded as much as possible in the novel peculiars in the later generation, and the design of the border operator should cooperate with the coding design.

(3) Variation function. The loci are randomly selected in each individual coding string of the population. Secondly, mutation probability P_m is used to mutate the gene values on the loci selected in the fitness ratio method.

3. Integrated learning algorithm optimization based on genetic algorithm.

Stacking integrated studying accomplishes studying tasks by building and connecting multiple learners. Therefore, the generation method of individual learners and the combination strategy of multiple individual learners are two key issues. Individual learners should not only ensure the accuracy of classification, but also fully consider the diversity of individual learners. Therefore, this article uses the natural law of GA "survival of the fittest and survival of the fittest" to get the optimal integrated learner through continuous evolution of selection, crossover, mutation, etc., and select the dominant individual learner to evolve, which can speed up the prediction speed, reduce its storage space requirements and further improve the accuracy. Meanwhile, species invasions increase the diversity of individual learners, reducing the risk of falling into local minima after multiple evolutions.

Suppose the training set is X, the input node of the individual learner x is a, the hidden layer node is b, the output level node is c, x_t^j is the j-th genetic individual of the t generation, the number of gene locations of the individual is m, the amount of individuals in the population is n, X_t is the t generation population, the individual fitness value is g_j, and the number of input samples of the individual learner is k. y_j is the expected output of the j-th sample of the individual learner, and O_j is the forecasted output of the i sample, then the steps of the integrated learning algorithm based on GA are as follows.

(1) Divide the sample data set and conduct random out-of-order operations on each part;

(2) Standardization of training samples and PCA dimensionality reduction;

(3) Initialize the neural network structure and use machine learning algorithms to generate individual learners as genetic individuals;

(4) Judge whether the number of individual learners meets the number of population m, if not, return to step (3), otherwise enter step (5);

(5) The initial population individual is encoded by the individual weight and threshold, and the following formula is obtained.

$$\begin{cases} m = a{\cdot}b + b + b{\cdot}c + c, \\ x_t^j = \left[x_t^{j(1)}, x_t^{j(2)}, \ldots, x_t^{j(m)}\right] \in \mathbb{R}^m, \\ X_t = \left[x_t^1, x_t^2, \ldots, x_t^n\right]^T \end{cases} \tag{1}$$

(6) The prediction classification accuracy of the individual learner is taken as the individual fitness value g_j of the population, and the single selection operation is performed using the fitness value-based ranking strategy.

$$l_j = \begin{cases} 1, & y_j = O_j, \\ 0, & y_j \neq O_j, \end{cases} \tag{2}$$

$$g_j = \sum_{j=1}^{k} l_j \,/\, k \tag{3}$$

(7) Individual crossover operation was carried out using the individual fitness value ratio strategy.

$$x_{t+1}^j = \frac{g_l}{g_l + g_j}\, x_t^l + \frac{g_j}{g_l + g_j}\, x_t^j \tag{4}$$

(8) The heuristic strategy is used to carry out individual variation operations, and the following formula is obtained, where α is the mutation operator.

$$\begin{aligned} x_{t+1}^{j'} &= x_t^{j'} + \alpha \times \left(x_t^{j'} - x_t^{j+1}\right), \quad g_j > g_{j+1}, \\ x_{t+1}^{j'} &= x_t^{j'} + \alpha \times \left(x_t^{j+1} - x_t^{j'}\right), \quad g_j < g_{j+1} \end{aligned} \tag{5}$$

where α is the mutation operator.

(9) Species invasion, add a new individual learner as a genetic individual, calculate the fitness value of a new generation of population individuals, judge whether the termination condition is met, if not, return to step (8), otherwise, step (10);

(10) The weight and threshold of the optimal individual $f(x)$ as an integrated learner.

4. Enterprise financial crisis warning relied on integrated learning and GA in big data environment.

4.1. Processing of unbalanced enterprise data based on SMOTE algorithm.

The current financial early warning method has redundant indicators, which leads to low forecasting efficiency. Aiming at the above problems, this paper designs an company financial crunch warning method relied on ensemble studying and genetic algorithm. The SMOTE algorithm was used to process the unbalanced enterprise data and redundant indicators were removed by PCA. On this basis, random forest and XGBoost are stacked and fused, and the data learned by fusion is handed over to the secondary learner GA-BP neural network to output the final forecasting outcome. The overall process is implied in Figure 2.

At present, undersampling is a common pre-processing method for selecting normal company samples paired with crisis companies, but it is tough to entirely consider the real distribution features of samples in multi-class sample sets (financial normal companies), and the undersampling rate is large, which is easy to cause serious loss of classification information. Thus, this article adopts SMOTE algorithm [25] to sample the financial crisis samples to address the issue of sample imbalance.

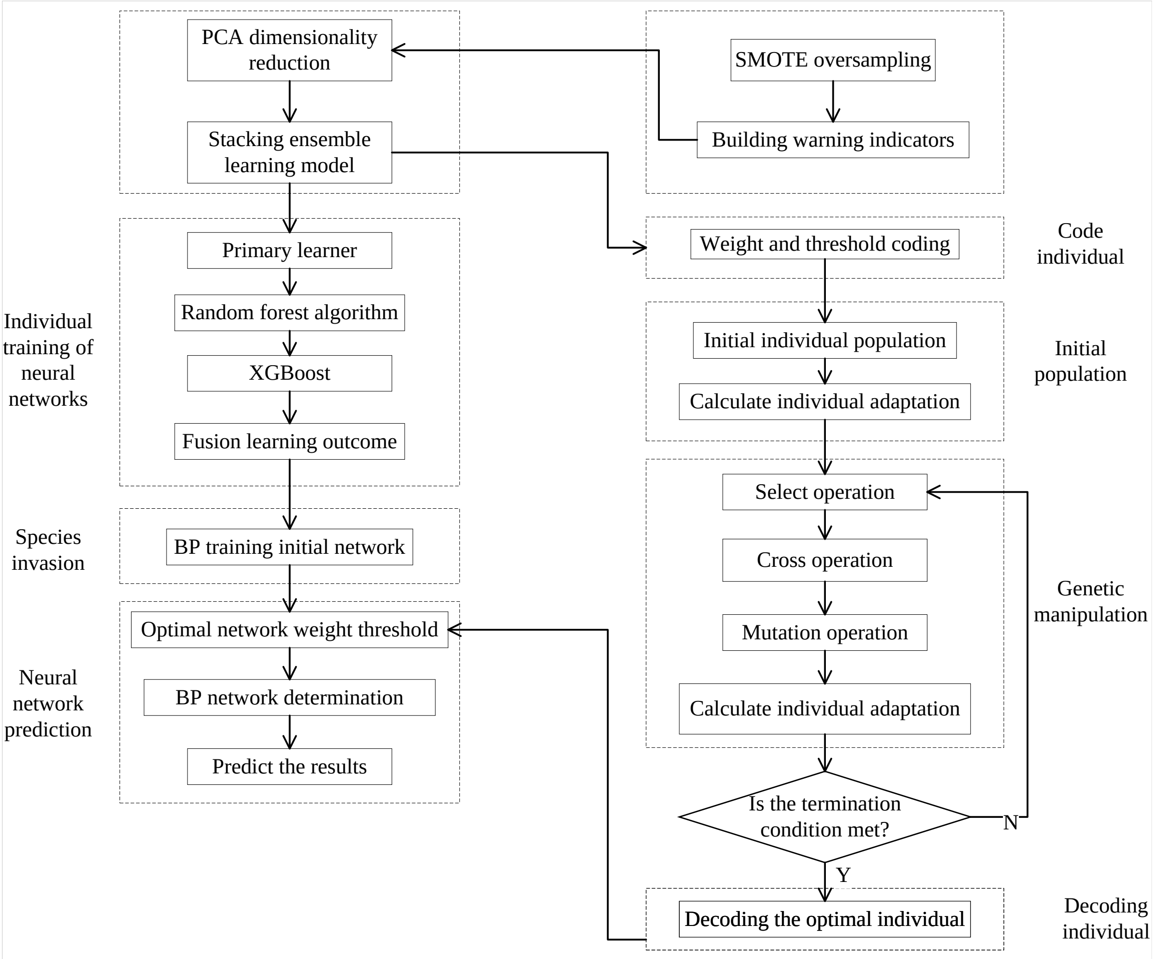
PCA dimensionality reduction
Stacking ensemble learning model
SMOTE oversampling
Building warning indicators
Primary learner
Random forest algorithm
XGBoost
Fusion learning outcome
Individual training of neural networks
Weight and threshold coding
Code individual
Initial individual population
Calculate individual adaptation
Initial population
Species invasion
BP training initial network
Select operation
Cross operation
Mutation operation
Calculate individual adaptation
Genetic manipulation
Neural network prediction
Optimal network weight threshold
BP network determination
Predict the results
Is the termination condition met?
N
Y
Decoding the optimal individual
Decoding individual

Figure 2. The whole process of the designed financial early warning model

(1) For each positive sample X, find K nearest neighbor samples, and then select n samples from the nearest neighbor samples based on the oversampling rate n, denoted as $z_1, z_2, \ldots, z_n$, associate these n samples with sample X, and carry out random linear interpolation in its association formula to obtain a new sample q_i.

$$q_i = X + \text{rand}(0, 1)\,(z_i - X) \tag{6}$$

where X represents the positive sample, $\text{rand}(0, 1)$ is the casual amount in the interval $(0, 1)$, and z_i represents the i-th sample of the n nearest neighbor samples of sample X. Furthermore, the oversampling ratio n depends on the ratio IR of the negative sample to the positive sample.

$$n = \text{round}(\text{IR}) \tag{7}$$

(2) According to the proportion IR, the sampling rate N is decided. For every financial crisis instance x, several instances are casually chosen from its K-nearest neighbors, assuming that the chosen nearest neighbor is z_n.

(3) Each chosen neighbor z_n is built with the initial instance x in terms of the Equation (8) to obtain a new sample z_{new}.

$$z_{\text{new}} = x + \text{rand}(\text{IR}) \times (x - x_n) \tag{8}$$

4.2. Screening and extraction of enterprise financial crisis warning indicators. To fully reflect the financial status of the company, the election of financial indicators should ensure that the selected indicators can fully reveal the financial status and development trend of the enterprise. Based on the existing research results [26], financial indicators including the solvency, growth, profitability, operating capacity and cash capacity of the enterprise are selected. In addition, non-financial indicators including the types of audit reports and the proportion of independent directors are also selected. It can effectively combine the information of audit and macroeconomic indicators to build a comprehensive index system.

Assuming that the financial crisis early warning indicator variables selected in this article have m and a total of n evaluation objects, the value of the j-th indicator of the i evaluation object is x_{ij}, then the value of each indicator is converted to the standardized indicator $\tilde{x}_{ij}$.

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \tag{9}$$

where $\bar{x}_j = (1/n)\sum_{i=1}^{n} x_{ij}$, $s_j = \sqrt{(1/(n-1))\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$, that is, $\bar{x}_j$ and s_j represent the instance mean and instance standard deviation of the jTH indicator variable respectively.

Through the above method, the difference of orders of magnitude between the data of different dimensions is eliminated, and the influence of weights and measures is eliminated, and the pre-warning data of enterprise financial crisis after processing is obtained. Due to the large number of indicators initially screened, direct use of these indicators will cause dimensional disaster. Therefore, before early warning analysis, PCA will be adopted in this paper to extract the above selected enterprise financial crisis early warning indicators.

Before PCA, the KMO spherical test method was used to conduct PCA feasibility analysis on the original indicator variables. The KMO value essentially represents the ratio between the easy correlation coefficient and the partial correlation coefficient among the original indicators, and the calculation equation is as bellow.

$$\mathrm{KMO} = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} r_{ij,1\ldots k}^2} \tag{10}$$

where r_{ij} stands for the correlation coefficient between indicator i and indicator j, and

$$r_{ij} = \Big(\sum_{i \neq k=1}^{n} \tilde{x}_{ik} * \tilde{x}_{kj}\Big)/(n-1).$$

KMO can determine whether the original is suitable for PCA analysis. If the results are judged to be suitable, the correlation coefficient matrix $R = (r_{ij})_{m \times n}$ of each index is obtained, where $r_{ii} = 1$, $r_{ij} = r_{ji}$.

Then, the eigenvalue $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_m \geq 0$ of the correlation coefficient matrix R is calculated, and each eigenvalue corresponds to an eigenvector respectively, which is divided into $v_1, v_2, \ldots, v_m$, where $v_j = (v_{1j}, v_{2j}, \ldots, v_{nj})^T$, then m new index variables $y_1, y_2, \ldots, y_m$ can be formed from the eigenvector. To evaluate the ability of the extracted comprehensive index to include the original index, the information contribution rate $b_j = \mu_j / \sum_{k=1}^{m} \lambda_k$ and the cumulative contribution rate $a_p = \sum_{k=1}^{p} \lambda_k \;/\; \sum_{k=1}^{m} \lambda_k$ of the eigenvalue μ_j are calculated.

Finally, the first p index variable $y_1, y_2, \ldots, y_p$ is chosen as p principal constituents to substitute the initial m index variables, and a comprehensive analysis is carried out on the p principal components. So far, this paper has completed the dimensionality reduction of the original index system and extracted the characteristic indicators capable of enterprise financial crisis.

4.3. Enterprise financial crunch warning relied on ensemble learning and GA..

After extracting the characteristic index of enterprise financial crisis, the integrated learning algorithm based on GA optimization can give early warning to enterprise financial crisis. Models at the first layer in the Stacking process must have good and similar performance. Otherwise, the performance of modes formed by stack fusion deteriorates. Random Forest and XGBoost are both decision tree based algorithms with excellent performance. Thus, random forest and XGBoost are selected as the first layer model, and GA-optimized BP neural network is selected as the second level model. Each model in the first level generates a new sample set, which is input into the second layer learning GA-BP neural network to get the final forecasting outcome.

(1) Primary learners include Random Forest and XGBoost.

Random forest algorithm. Firstly, the training data set is extracted N times, and a new sub-training set $\{D_1, D_2, \ldots, D_N\}$ is obtained as the sample at the decision root node. Then the attributes are randomly selected as the node splitting attributes, and then a large number of decision trees are established to form a random forest. Finally, the predicted output value of each tree is averaged to get the final pre-test result.

XGBoost. XGBoost is an integrated learning algorithm with good prediction effect and high operation efficiency [26], and its base classifier is also a decision tree. Assuming that the 0-th tree predicts the sample value $y_i^{(0)} = 0$, the current root node is first determined by node splitting method, and the characteristic with the maximum score value is divided into two parts to get two sets of leaf nodes. Then, the depth of two leaf node sets reaches the depth of the set tree, and the entire tree is formed. Then, the forecasting value y_i^* of the leaf node is calculated by Equation (11), and the forecasting value of the first tree for instance x_i is shown in Equation (12).

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{11}$$

$$y_i^{(1)} = y_i^{(0)} + f_1(x_i) = 0 + w_{q1(x)} = w_{q1(x)} \tag{12}$$

where G_j represents the set of all leaf nodes, H_j represents the weight, $f(\cdot)$ represents the excitation function, and λ is the penalty coefficient.

Repeat the above steps until K trees, for the final forecasting value of intance x_i.

$$\hat{y}_i^{(K)} = y_i^{(0)} + \sum_{t=1}^{K} f_t(x_i) = \hat{y}_i^{(K-1)} + w_{q_k(x)} \tag{13}$$

For primary learners, it is necessary to set the maximum depth of the decision tree and the number of base classifiers. The method of relying on manual experience is often not effective, and the improved integrated learning algorithm in Chapter 3 can get better fitting effect. GA is used to obtain the optimal parameter combination (maximum depth and number of base learners) of each primary learner to optimize the model of the primary learner. The results of training set $\{y_{11}, y_{12}, \ldots, y_m^k\}$ and test set $\{P_1, P_2, \ldots, P_m^k\}$ are obtained by using primary learner training. The results obtained from the training of the primary learner are taken as a new data set, and the final outcome is obtained by training the secondary learner.

(2) The secondary learner is composed of a neural network model.

BP neural network is with strong studying ability and generalization ability, however its convergence speed is slow and it is simple to fall under partial minimum. In this article, GA is adopted to improve the weights and thresholds of BP [27] to further improve the early warning ability of the model. The specific steps are as bellow.

Step 1: Individual coding. GA is adopted to encode all the weights and thresholds in BP, and a complete BP neural network is constructed.

Step 2: Initialize the population and calculate the fitness function. The number of population evolution and population size were initialized, and the sum of absolute errors was adopted as fitness function F.

$$F = k\left(\sum_{i=1}^{m} \text{abs}(z_i - o_i)\right) \tag{14}$$

where m is the amount of output nodes of BP network; z_i is the planned output value of the i-th node; o_i is the forecasted output value of the i-th node; k is the coefficient.

Step 3: In the election function, the roulette algorithm is used to calculate the probability $p_i = f_i / \sum_{j=1}^{N} f_j$ of an individual being selected according to the ratio $f_i = k/F_i$ of fitness value F_i to the sum of fitness of N individuals.

Step 4: In the crossover operation, two randomly selected chromosomes a_k and a_j are crossed at the j position.

$$\begin{cases} a_{kj} = a_{kj}(1-b) + a_{ij}b, \\ a_{ij} = a_{ij}(1-b) + a_{kj}b \end{cases} \tag{15}$$

Step 5: Mutation operation. The j-th gene b_j of the i individual was selected for mutation.

Step 6: Calculate the individual fitness and judge whether the minimum error end condition is met. If it is not met, the operation of Step (2) to Step (5) is carried out again. If the end condition is met, the optimized weight and threshold c are distributed to the BP neural network.

Step 7: Using the optimized weight w_{ij}, the output y of the primary learner and the threshold c, the final financial crisis prediction results are calculated as follows.

$$O_j = f\left(\sum_{i=1}^{n} \omega_{ij} y + c_j\right) \tag{16}$$

5. Experiment and result analysis.

5.1. Analysis of early warning results of enterprise financial crisis.

The data in this article comes from the Guotai An CSMAR database, which contains 4,874 a-share listed enterprises in Shanghai and Shenzhen Stock Exchange, mainly distributed in 19 industries such as financial industry, real estate industry and manufacturing industry. Among them, there are 3,150 listed enterprises in manufacturing industry, accounting for about 65%. Therefore, this paper takes listed enterprises in manufacturing industry as the research object. In terms of the financial state of listed manufacturing enterprises from 2018 to 2020, a total of 2241 enterprises in T-2 annual report data were selected as samples, of which 2116 were common enterprises and the keeping 125 were enterprises in financial crisis. To avoid the influence of sample imbalance on the forecasting outcome of the model, SMOTE oversampling was adopted to adjust the proportion of instances in the dataset to 4:1, as implied in Table 1. Python programming tools were used to analyze the experimental results of the adjusted data set.

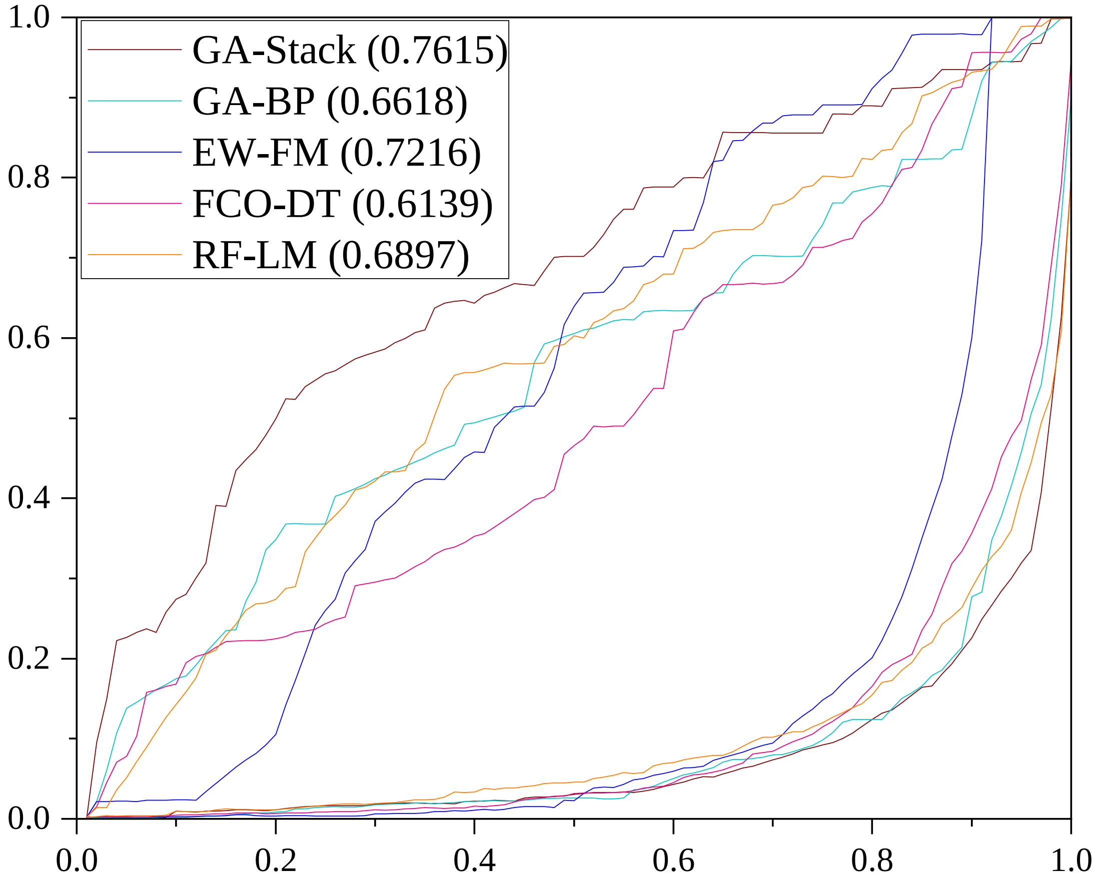

Figure 3. KS curves corresponding to different financial early warning models.

Table 1. Number of training and test set samples

	Sample number for training set	**Sample size of test set**
Non-financial crisis enterprises	1687	429
Financially troubled companies	385	90

To objectively evaluate the model results, In this paper, the classification accuracy rate (PA), Class I error rate (IER), Class II error rate (IIER) [28] and KS curve [29] were used to compare and analyze FCO-DT [9], GA-BP [15], RF-LM [18], EW-FM [21] and the suguseted model GA-Stack, respectively. The outcome are implied in Table 2.

Table 2. Comparison of the forecast accuracy of the T-2 and T-1 year models

Model	**T-2**			**T-1**		
	PA	IER	IIER	PA	IER	IIER
FCO-DT	0.729	0.5362	0.5215	0.7687	0.4087	0.4223
GA-BP	0.797	0.4711	0.3824	0.8144	0.3462	0.3811
RF-LM	0.837	0.4092	0.3071	0.8531	0.2896	0.3195
EW-FM	0.875	0.3184	0.2141	0.9056	0.1962	0.2473
GA-Stack	0.917	0.2516	0.1482	0.9423	0.1393	0.1854

As can be seen from Table 2, the prediction accuracy of GA-Stack for T-2 and T-1 years is 91.7% and 94.2% respectively, which is 18.8% and 17.4% higher than that of FCO-DT, 12% and 12.8% higher than that of GA-BP, and 8% and 8.9% higher than that of RF-LM. Compared with EW-FM, the prediction accuracy of GA-Stack model is the highest, which increases by 4.2% and 3.7% respectively. In addition, GA-Stack also had the lowest Class I and Class II forecast error rates in both years. FCO-DT does not reduce the dimension of financial early warning index, resulting in low forecasting efficiency. GA-BP uses GA-BP to forecast the type of financial early warning, which is a forecasting method of single neural network model. Although the forecasting efficiency is higher than that of traditional BP, it is far lower than that of integrated learning model. RF-LM is sensitive to noisy data, so the prediction efficiency needs to be improved. EW-FM does not optimize the fusion strategy of Stacking or the parameters of the learner, resulting in lower prediction performance than GA-Stack. In conclusion, GA-Stack has good prediction effect on samples.

KS curves corresponding to different models are shown in Figure 3. The horizontal axis of KS curve is the threshold value, and the vertical axis is the TPR and FPR of the model prediction results at the level of the threshold value of the horizontal axis. The value of KS ranges from 0 to 1. The larger the KS value, the greater the degree to which the model separates the two types of labels, and the better the forecasting impact of the model. From Figure 3, KS values of GA-Stack, FCO-DT, GA-BP, RF-LM and EW-FM are 0.7615, 0.6139, 0.6618, 0.6897 and 0.7216, and KS values of GA-Stack are the largest, superior to other single models and integrated learning models. The effectiveness of the integrated learning algorithm of GA optimization in Stacking for enterprise financial crisis warning is further verified.

5.2. **Robustness test of enterprise financial crisis early warning model.** Although the above analysis shows that GA-Stack has a strong ability to forecast the financial crunch of enterprises, the evaluation of the performance of the model needs to compare its forecasting ability under different sample divisions. The robustness of the model is tested by the 5-fold cross-validation method. To objectively compare the results of 5-fold cross-validation of each model, the mean values of three evaluation indicators were used as the means to evaluate the model. As implied in Figure 4, AVR(PA), AVR(IER) and AVR(IIER) respectively represent the mean values of PA, Class I error rate and Class II error rate under the 5-fold cross-validation method.

The average prediction accuracy of GA-Stack five-fold cross-validation method in two different periods is the highest, which is 0.9281 and 0.9408 respectively, implying that the model constructed in this paper has relatively good prediction accuracy under the five-fold cross-validation method. From the mean value of class II error rates, GA-Stack has the lowest error rates of class I and class II in T-2 and T-1 phases, and the KS value of GA-Stack is the largest. It not only removes the redundancy of warning indicators, but also uses GA to optimize the fusion strategy of integrated learning, making the advantages of each learner fully apparent. The prediction efficiency is improved.

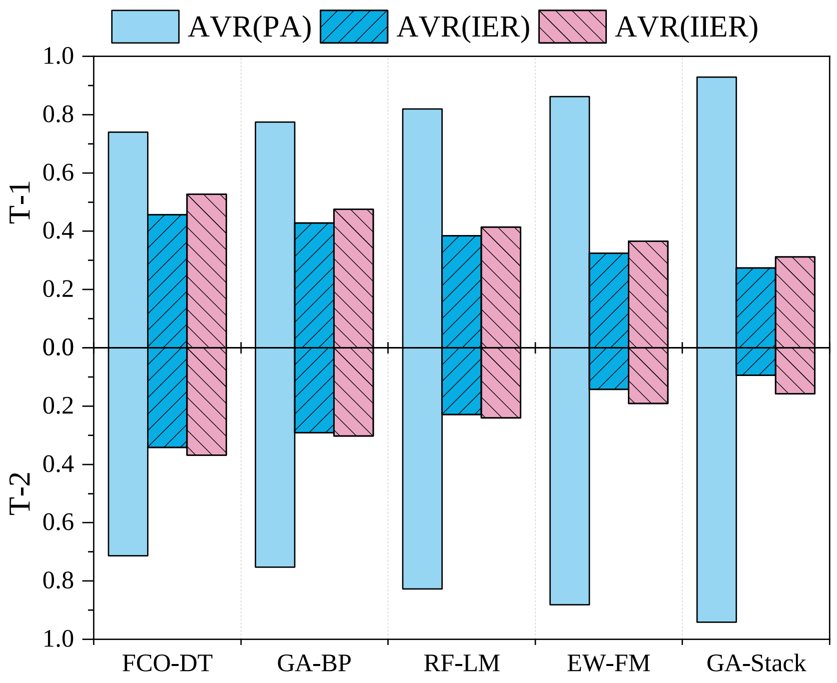

Figure 4. Comparison of the prediction performance of the 5-fold cross-validation method.

6. **Conclusion.**

Focusing on the current financial crunch early warning model using a single-variable model, which leads to poor prediction accuracy, an enterprise financial crisis early warning model relied on integrated studying and GA in big data environment is designed. Firstly, the GA is used to optimize the Stacking integrated learning algorithm to get the optimal integrated learner and speed up the prediction speed. Second, the SMOTE algorithm is used to oversample the samples to address the issue of enterprise sample imbalance, and establish the financial warning indicators, and use PCA to remove redundant indicators. Finally, the primary learners Random Forest and XGBoost are stacked and fused, and the fused data are given to the secondary learner GA-BP neural network to output the final prediction results. The simulation outcome implies that the forecasting accuracy of the designed model is improved by 0.037–0.188 compared with other models, indicating that the suggested model has a high prediction efficiency.

REFERENCES

[1] L. Xu, Q. Qi, and P. Sun, “Early-warning model of financial crisis: an empirical study based on listed companies of information technology industry in China,” Emerging Markets Finance and Trade, vol. 56, no. 7, pp. 1601-1614, 2020.

[2] R. Padhan, and K. Prabheesh, “Effectiveness of early warning models: A critical review and new agenda for future direction,” Bulletin of Monetary Economics and Banking, vol. 22, no. 4, pp. 457-484, 2019.

[3] E. C. Charalambakis, and I. Garrett, “On the prediction of financial distress in developed and emerging markets: Does the choice of accounting and market information matter? A comparison of UK and Indian firms,” Review of Quantitative Finance and Accounting, vol. 47, pp. 1-28, 2016.

[4] C. Sevim, A. Oztekin, O. Bali, S. Gumus, and E. Guresen, "Developing an early warning system to predict currency crises," European Journal of Operational Research, vol. 237, no. 3, pp. 1095-1104, 2014.

[5] Q. Yu, and L. Zhang, "Financial crisis early-warning model for listed company in China energy industry based on logistic regression," Bulgarian Chemical Communications, vol. 49, pp. 31-36, 2017.

[6] H. Shang, D. Lu, and Q. Zhou, "Early warning of enterprise finance risk of big data mining in internet of things based on fuzzy association rules," Neural Computing and Applications, vol. 33, no. 9, pp. 3901-3909, 2021.

[7] Y. Huang, G. Kou, and Y. Peng, "Nonlinear manifold learning for early warnings in financial markets," European Journal of Operational Research, vol. 258, no. 2, pp. 692-702, 2017.

[8] C. Filippopoulou, E. Galariotis, and S. Spyrou, "An early warning system for predicting systemic banking crises in the Eurozone: A logit regression approach," Journal of Economic Behavior & Organization, vol. 172, pp. 344-363, 2020.

[9] A. Gholizadeh, M. Fallahshams, and M. A. Afsharkazemi, "The designing early warning system of financial crisis outbreak in Tehran stock exchange by decision tree," Journal of Investment Knowledge, vol. 10, no. 40, pp. 35-55, 2021.

[10] J. J. Ahn, K. J. Oh, T. Y. Kim, and D. H. Kim, "Usefulness of support vector machine to develop an early warning system for financial crisis," Expert Systems with Applications, vol. 38, no. 4, pp. 2966-2973, 2011.

[11] G. Al-Assaf, "An early warning system for currency crisis: a comparative study for the case of Jordan and Egypt," International Journal of Economics and Financial Issues, vol. 7, no. 3, pp. 43-50, 2017.

[12] P. Xanthopoulos, P. M. Pardalos, T. B. Trafalis, P. Xanthopoulos, P. M. Pardalos, and T. B. Trafalis, "Linear discriminant analysis," Robust Data Mining, pp. 27-33, 2013.

[13] G. Tripepi, K. Jager, F. Dekker, and C. Zoccali, "Linear and logistic regression analysis," Kidney International, vol. 73, no. 7, pp. 806-810, 2008.

[14] X. Sun, and Y. Lei, "Research on financial early warning of mining listed companies based on BP neural network model," Resources Policy, vol. 73, 102223, 2021.

[15] L. Ruan, and H. Liu, "Financial distress prediction using GA-BP neural network model," International Journal of Economics and Finance, vol. 13, no. 3, pp. 1-1, 2021.

[16] H. Zhang, and Y. Luo, "Enterprise financial risk early warning using BP neural network under internet of things and rough set theory," Journal of Interconnection Networks, vol. 22, no. 03, 2145019, 2022.

[17] T. Wang, S. Zhao, G. Zhu, and H. Zheng, "A machine learning-based early warning system for systemic banking crises," Applied Economics, vol. 53, no. 26, pp. 2974-2992, 2021.

[18] V. Budhidharma, R. Sembel, E. Hulu, and G. Ugut, "Early warning signs of financial distress using random forest and logit model," Corporate and Business Strategy Review, vol. 4, no. 4, pp. 69-88, 2023.

[19] A. Gholizadeh, M. Fallahshams, and M. A. Afsharkazemi, "The designing early warning system of financial crisis outbreak in Tehran stock exchange by decision tree," Journal of Investment Knowledge, vol. 10, no. 40, pp. 35-55, 2021.

[20] M. Liang, T. Chang, B. An, X. Duan, L. Du, X. Wang, J. Miao, L. Xu, X. Gao, and L. Zhang, "A stacking ensemble learning framework for genomic prediction," Frontiers in Genetics, vol. 12, 600040, 2021.

[21] H. Guo, "The design of early warning software systems for financial crises in high-tech businesses using fusion models in the context of sustainable economic growth," Peerj Computer Science, vol. 9, e1326, 2023.

[22] T.-Y. Wu, H. Li, S. Kumari, and C.-M. Chen, "A Spectral Convolutional Neural Network Model Based on Adaptive Fick's Law for Hyperspectral Image Classification," Computers, Materials & Continua, vol. 79, no. 1, pp. 19-46, 2024.

[23] T.-Y. Wu, A. Shao, and J.-S. Pan, "CTOA: Toward a Chaotic-Based Tumbleweed Optimization Algorithm," Mathematics, vol. 11, no. 10. 2339, 2023.

[24] T.-Y. Wu, H. Li, and S.-C. Chu, "CPPE: An Improved Phasmatodea Population Evolution Algorithm with Chaotic Maps," Mathematics, vol. 11, no. 9. 1977, 2023.

[25] S. Wang, Y. Dai, J. Shen, and J. Xuan, "Research on expansion and classification of imbalanced data based on SMOTE algorithm," Scientific Reports, vol. 11, no. 1, pp. 24039, 2021.

[26] X. Shi, Y. D. Wong, M. Z.-F. Li, C. Palanisamy, and C. Chai, “A feature learning approach based on XGBoost for driving assessment and risk prediction,” Accident Analysis & Prevention, vol. 129, pp. 170-179, 2019.

[27] K. Cui, and X. Jing, “Research on prediction model of geotechnical parameters based on BP neural network,” Neural Computing and Applications, vol. 31, pp. 8205-8215, 2019.

[28] C. J. Mecklin, and D. J. Mundfrom, “A Monte Carlo comparison of the Type I and Type II error rates of tests of multivariate normality,” Journal of Statistical Computation and Simulation, vol. 75, no. 2, pp. 93-107, 2005.

[29] D. J. Lee, and J. H. Yoon, “The New Keynesian Phillips Curve in multiple quantiles and the asymmetry of monetary policy,” Economic Modelling, vol. 55, pp. 102-114, 2016.