Journal of Applied Mathematics and Physics, 2025, 13(5), 1629-1643
https://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Bankruptcy Prediction in the Polish Banking Industry Using Principal Component Analysis and BP Neural Network

Shiqing Li*, Qiancheng Tan#

Guangxi Colleges and Universities Key Laboratory of Data Analysis and Computation, College of Mathematics and Computing Science, Guilin University of Electronic Technology, Guilin, China
Email: #1907065810@qq.com

How to cite this paper: Li, S.Q. and Tan, Q.C. (2025) Bankruptcy Prediction in the Polish Banking Industry Using Principal Component Analysis and BP Neural Network. *Journal of Applied Mathematics and Physics*, **13**, 1629-1643.
https://doi.org/10.4236/jamp.2025.135089

Received: April 3, 2025
Accepted: May 5, 2025
Published: May 8, 2025

Abstract

With the rapid growth of the international banking industry, bank failures can lead to severe economic losses and social impacts. Although existing measures to address such failures are well-developed, timely prediction can significantly mitigate these effects. This study analyzes key indicators influencing bank failure through data analysis and correlation analysis, then develops a neural network-based risk prediction model to estimate failure probabilities. First, we extracted 64 indicators from the dataset, identified the most relevant indicators using the entropy weight method, and established a bank efficiency evaluation formula to determine the failure threshold. Next, we applied principal component analysis (PCA) to reduce dimensionality and derive a comprehensive scoring formula. Based on these findings, we constructed a machine learning model in MATLAB to predict bank failures. Finally, the model was used to predict the failure probabilities of all banks and identify 20 representative existing and failed banks. The developed models effectively predict bank failure risks and demonstrate strong applicability across different scenarios.

Keywords

BP Neural Network, Entropy Weight Method, Principal Component Analysis

1. Introduction

The stability and efficiency of banking institutions are critical to maintaining fi-

*First Author.
#Corresponding Author.

nancial market equilibrium and fostering economic development [1]. As financial systems become increasingly complex, the need for robust methods to analyze bank efficiency and predict potential risks has grown significantly. Understanding the operational health of financial institutions is essential for policymakers, investors, and regulators to mitigate financial crises and ensure sustainable economic growth [2].

In recent years, various methodologies have been developed to evaluate bank efficiency and predict financial risk. Traditional approaches, such as financial ratio analysis and expert evaluations, often suffer from subjectivity and limited predictive power. In contrast, data-driven methods, including machine learning [3] and statistical modeling [4], offer more objective and accurate assessments. Among these, the entropy weight method (EWM) [5] and principal component analysis (PCA) [6] have gained prominence due to their ability to handle complex financial data and extract meaningful features.

This paper proposes a two-stage framework for bank efficiency analysis and risk prediction. In the first stage, we employ the entropy weight method to construct an efficiency function using input-output indicators extracted from the Polish Companies Bankruptcy dataset. By computing efficiency scores, we effectively distinguish between operational and bankrupt banks. The efficiency evaluation graph illustrates a clear separation between these two groups, confirming the validity of our approach.

In the second stage, we apply principal component analysis to reduce the dimensionality of the dataset while preserving critical information. The extracted principal components serve as inputs to a BP neural network, which is optimized using a genetic algorithm (GA) to enhance predictive accuracy. Our proposed model achieves an accuracy of 83% in predicting bank failures, demonstrating its potential as a reliable tool for financial risk assessment.

2. Modeling of Bank Efficiency Analysis Based on Input-Output of Entropy Weight Approach

First, we used Python for preprocessing to calculate the correlation among the 64 indicators in the Polish Companies Bankruptcy dataset for the years 2017 and 2021, as shown in **Figure 1**.

The analysis results indicate that some indicators exhibit high correlation; however, certain indicators still contain missing values. To address this issue, we applied the mean imputation method, where the missing values were filled using the average of highly correlated indicators, thereby enhancing the completeness and usability of the data. Bank efficiency is an indicator to judge whether a bank is operating normally. The first question requires sorting out appropriate index input-output data from 64 indicators, selecting indicators that can measure bank efficiency, and providing the dividing line of bank failure. First, we standardize and normalize the existing data. Then the entropy weight method is used to assign the corresponding weights to the sorted indicators, and the efficiency function is

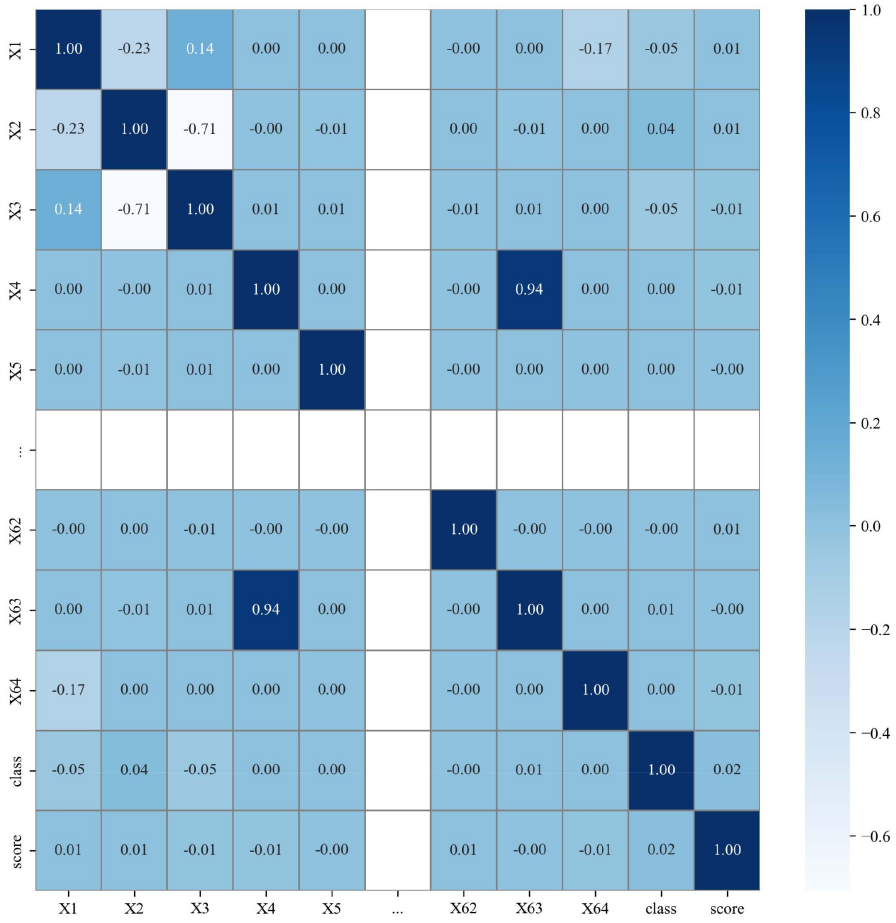


Figure 1. Heat maps related to the 64 indicators in the data for 2021 as well as for 2017.

constructed with these weights. The efficiency function is used to evaluate the efficiency of all data samples, and finally a bank efficiency graph is obtained. According to the different class values, we can distinguish the bankrupt banks from the banks that are still in operation.

According to the research of Wang Xinyang *et al.* [7], we extracted the input-output related indicators of the relative efficiency of the calculation of this question from 64 indicators, respectively (**Table 1**).

Table 1. Input-output index division table.

Index type	Index definition	Pointer symbol
Input index	Net profit	X1
Output indicator	Total liabilities	X2
	Working capital	X3
	Retained earnings	X6

To build a comprehensive evaluation system based on the input-output indicators determined above, the weight of each evaluation indicator is required first, and then quantified. The specific algorithm is as follows [8]:

First, the data of each indicator are standardized. There are three indicators, of which

$$X_i = x_1, x_2, \cdots, x_n \tag{1}$$

After standardization of the above indicators, the value is

$$Y_1, Y_2, \cdots, Y_n \tag{2}$$

The standardized formula can be obtained:

$$Y_{ij} = \frac{x_{ij} - \min(x_i)}{\max(x_i) - \min_{x_i}} \tag{3}$$

According to the definition of entropy in information theory, the information entropy of a set of data is:

$$E_j = -\frac{1}{\ln n} \sum_{i=1}^{n} P_{ij} \ln P_{ij} \tag{4}$$

where, $P_{ij} = \frac{Y_{ij}}{\sum_{i=1}^{n} Y_{ij}}$, If $P_{ij} = 0$, then $\lim_{P_{ij} \to 0} P_{ij} \ln P_{ij} = 0$.

According to the calculation formula of information entropy, the information entropy of each index is calculated as

$$E_1, E_2, \cdots, E_n \tag{5}$$

Calculate the weight of each index through information entropy

$$W_i = \frac{1 - E_i}{k - \sum E_i} \quad (i = 1, 2, \cdots, k) \tag{6}$$

According to the combination weight method, the efficiency index of the bank can be obtained as Z_i, and the calculation formula is

$$Z_i = \sum_{i=1}^{n} x_i w_i \tag{7}$$

where, x_i is i different indicators, w_i is i different indicator weights, and Z_i is the efficiency of different banks.

It can be known that all bank data can be divided, and the lower limit of $A \cap B$ represents the dividing line between the collapse of two different sets. Set A corresponds to class value 0, and set B corresponds to class value 1, based on mathematical principles.

According to the conclusion of the above model, we extracted X1, X2, X3, X6, class. We substituted these data into the above model and obtained the entropy weight value through the excel file as (Table 2).

Table 2. Weights of relevant indicators.

Correlation index	Weight W
X1	0.33
X2	0.32
X3	0.32
X6	0.02

Then, the weights of the above indicators are brought into formula (7) to obtain

$$Z_i = 0.33 * x_{1i} + 0.32 * x_{2i} + 0.32 * x_{3i} + 0.02 * x_{6i} \tag{8}$$

where, $x_{1i} \in \mathrm{X1}$, $x_{2i} \in \mathrm{X2}$, $x_{3i} \in \mathrm{X3}$, $x_{6i} \in \mathrm{X6}$.

We use all the Z_i in python, save them in a support file with the column name "score", and then we put the corresponding class values of Z_i and i into the drawing function to get the following Figure 2.

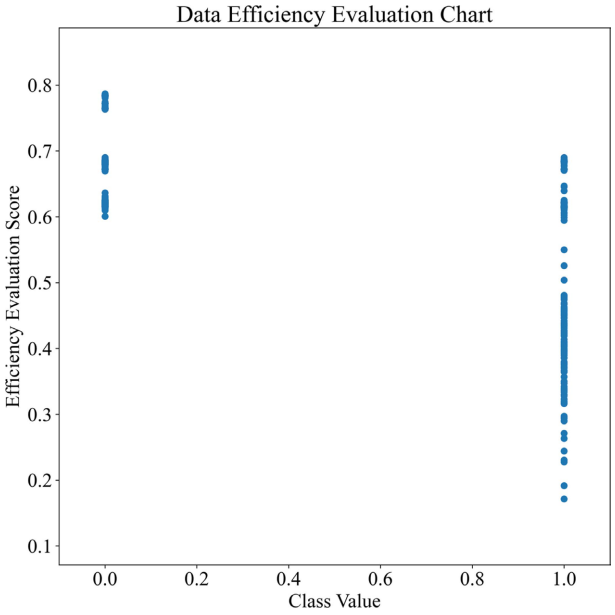


Figure 2. Data efficiency evaluation chart.

There is a significant difference in the image between the banks that went bankrupt during the observation period and the banks that did not go bankrupt during the observation period. I can take the lower bound of the group of data whose class value is 0 as the dividing line of bank bankruptcy efficiency is $Z = 0.60090761$.

3. The Bank Risk Prediction Model Based on Principal Component Analysis Dimensionality Reduction and BP Neural Network

We must first get the important factors affecting bank failure, and then we must determine the index that can comprehensively reflect the information reflected by the 64 indicators. If we use correlation analysis, we will find far more than 5 important indicators, so we choose principal component analysis.

The principle of principal component analysis is to try to recombine the original variables into a new set of several comprehensive variables that are unrelated to each other, and at the same time, according to the actual needs, several less sum variables can be extracted from the statistical method to reflect the information of

the original variables as much as possible, which is called principal component analysis or principal component analysis, and is also a method to deal with dimensionality reduction in mathematics.

3.1. Principal Component Analysis Model

The principal component is a linear combination of the original variables; The number of principal components is less relative to the original number; The principal component retains most of the information of the original variable; The principal components are independent of each other [9].

Let the data sample data X, in which there are m variables, a total of n samples

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \tag{9}$$

Calculate the average for each column $\mu_i = \frac{1}{n}\sum_{j=1}^{n} x_{ji}$

The variance of each column $\sigma_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_{ji} - \mu_i\right)^2$

Standardize the data, $z_{ji} = \frac{x_{ji} - \mu_i}{\sigma_i}$

We get matrix Z

$$Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nm} \end{bmatrix} \tag{10}$$

Calculated correlation coefficient,

$$r_{ij} = \frac{\sum_{k=1}^{n} z_{ki} * z_{kj}}{n-1}, \ \left(i, j = 1, 2, \cdots, m\right)$$

The correlation coefficient matrix R is obtained

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix} \tag{11}$$

where: r_{ij} is the correlation between the sample sequence in column i of the X matrix and the sample sequence in column j, whose value is between −1 and 1, and the R matrix should be a symmetric matrix, that is, $r_{ij} = r_{ji}$.

The difference in the degree of correlation coefficient is shown in the following Table 3 and Table 4.

The covariance matrix $\sum$ is a real symmetric matrix, knowing that its eigenvalue is non-negative, it may be useful to set its eigenvalue

$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_p \geq 0$, and their corresponding orthonormalized unit eigenvectors are as follows:

Table 3. Positive and negative correlation.

Correlation coefficient r	correlation
$r > 0$	Positive linear correlation
$r = 0$	Linearly independent
$r < 0$	Positive linear correlation

Table 4. Degree of correlation.

Absolute value of correlation coefficient	Degree of correlation
1	Perfect correlation
[0.8, 1)	Highly correlated
[0.5, 0.8)	Moderate correlation
[0.3, 0.5)	Low correlation
[0, 0.3)	uncorrelated

$$a_1 = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{bmatrix}; a_2 = \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{p2} \end{bmatrix}; \cdots; a_p = \begin{bmatrix} a_{1p} \\ a_{2p} \\ \vdots \\ a_{pp} \end{bmatrix} \tag{12}$$

If the index variable represented by the columns of the original X, the composite vector, is denoted as $Var = \left[Var_1, Var_2, \cdots, Var_p\right]^{\mathrm{T}}$, then the ith principal component of X is

$$F_i = \left(a_i\right)^{\mathrm{T}} Var = a_{1i} * Var_1 + a_{2i} * Var_2 + \cdots + a_{pi} * Var_p \tag{13}$$

The selection rule of the number of principal components is determined according to the cumulative contribution rate, which is generally required to reach more than 0.85, so as to ensure that the new variable can include most of the information of the original variable.

$$j = \frac{\lambda_j}{\sum_{k=1}^{p} \lambda_k} \left(j = 1, 2, \cdots, n\right), \quad \alpha_p = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=1}^{n} \lambda_k} \left(p \leq n\right) \tag{14}$$

Figure 3 and Figure 4 show the contributions of the indicators to the principal components. The first principal component is strongly influenced by indicators X13, X19, X20, X23, X30, X31, X39, X42, X43, X44, X49, X56, and X58, indicating that it primarily reflects the information from these variables. The second principal component is mainly driven by X4, X8, X12, X16, X17, X26, X33, X34, X40, X46, X50, and X63, suggesting it reflects the information contained in these indicators. For the third principal component, the indicators X1, X7, X11, X14, X18,

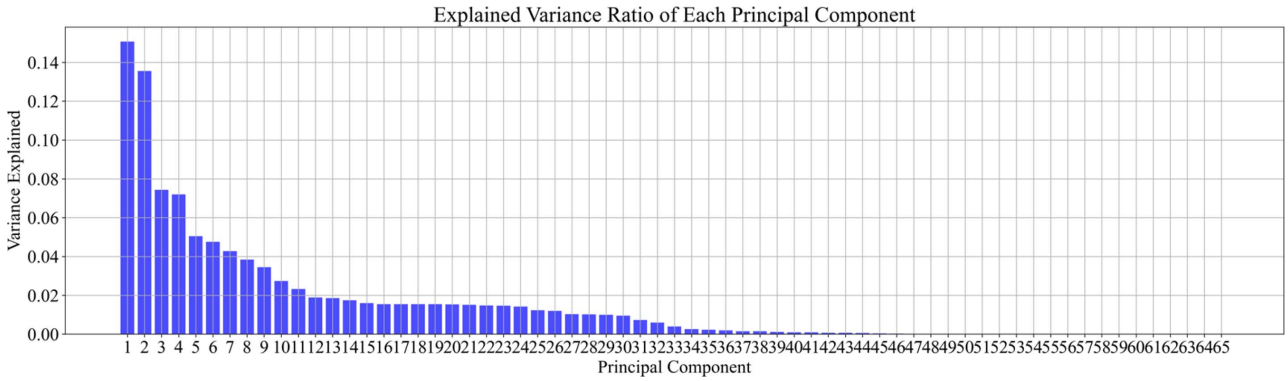


Figure 3. Explained variance ratio of each principal component.

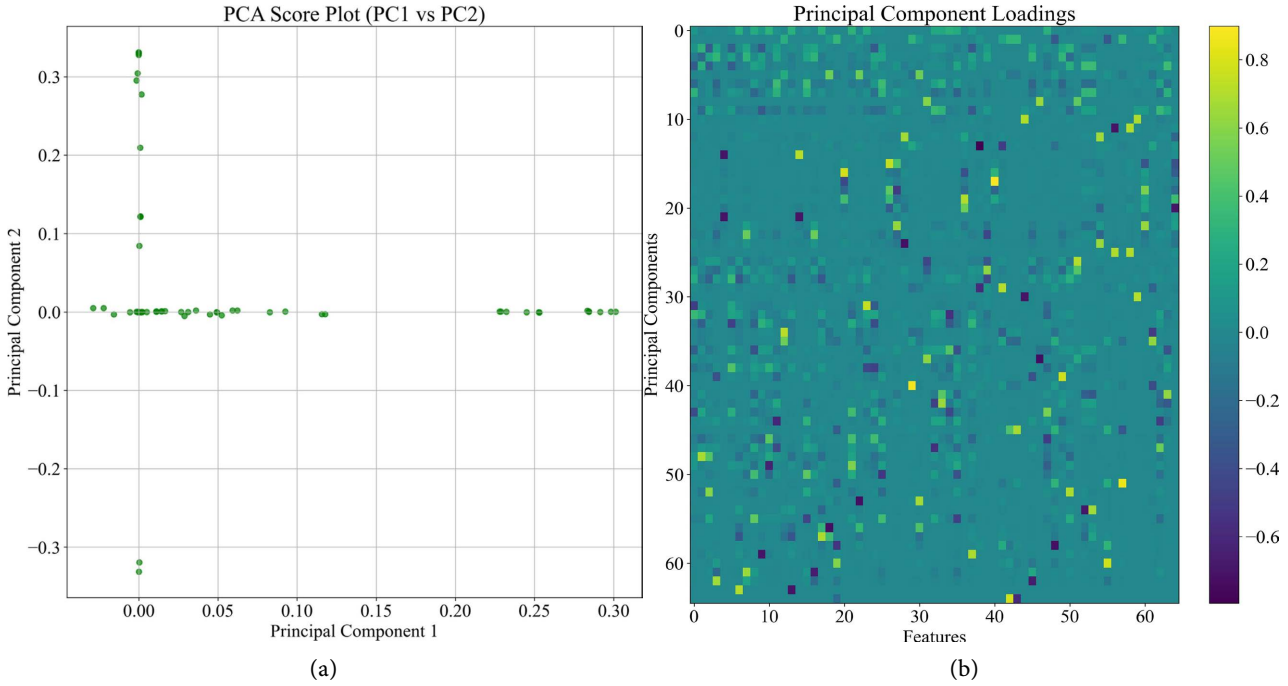


Figure 4. PCA analysis.

X22, X24, X35, X36, and X48 have the largest contributions, meaning this component is largely characterized by these variables. The fourth principal component is predominantly shaped by X2, X3, X6, X10, X25, X38, and X51, indicating it reflects the information from these indicators. Lastly, the fifth principal component is influenced mostly by X53 and X64, highlighting that it mainly captures the information of these two indicators.

By understanding which indicators each principal component represents, we can then calculate the contribution rate of each principal component (Table 5).

According to the component matrix diagram and the principal component contribution diagram, we can calculate the coefficients corresponding to each index in the five principal components:

The new eigenvectors W_1, W_2 and W_3 are calculated respectively

Table 5. Principal component contribution diagram.

Ingredient	Total	Percentage of variance (%)	Accumulate (%)
PC1	12,807	20.011	20.011
PC2	11,118	17.372	37.383
PC3	9447	14.761	52.144
PC4	5199	8.123	60.267
PC5	3481	5.439	65.706

$$W_1 = \mathrm{VAR00065}/\mathrm{SQR}(12.807)$$

$$W_2 = \mathrm{VAR00066}/\mathrm{SQR}(11.118)$$

$$W_3 = \mathrm{VAR00065}/\mathrm{SQR}(9.447)$$

$$W_4 = \mathrm{VAR00065}/\mathrm{SQR}(5.199)$$

$$W_5 = \mathrm{VAR00065}/\mathrm{SQR}(3.481)$$

Through the above, the formula after dimensionality reduction of principal component analysis can be calculated.

Set the component matrix to W_i, then

$$W_i = \begin{bmatrix} w_{1i} \\ w_{2i} \\ \vdots \\ w_{ji} \end{bmatrix} (i = 1, 2, \cdots, 5)(j = 1, 2, \cdots, 64) \tag{15}$$

Each index matrix is $A = \left[X_1, X_2, \cdots, X_j \right]$

Then $F_i = A * W_i$

$$F = (0.20011/0.65706) * F_1 + (0.17372/0.65706) * F_2$$

All available $$+ (0.14761/0.65706) * F_3 + (0.08123/0.65706) * F_4 .$$

$$+ (0.05439/0.65706) * F_5$$

The weight calculation results of principal component analysis show that the weight of principal component 1 is 0.34, the weight of principal component 2 is 0.26, the weight of principal component 3 is 0.22, the weight of principal component 4 is 0.12, and the weight of principal component 5 is 0.08.

3.2. Bank Risk Prediction Model Based on BP Neural Network

BP neural network is usually composed of three layers, namely input layer, hidden layer and output layer. In the case that the output vector has been defined, it is particularly important to choose the appropriate input vector. There are 38 indicators that mainly affect the comprehensive score, and the 38 indicators influence each other. When predicting the comprehensive score, a series of variables such as X1, X2 and X3 are selected as the input layer, and the comprehensive score is selected as the output layer, and the single hidden layer network structure is also selected.

BP neural network should be trained before the prediction, and through training, the network has the ability of associative memory and prediction. BP neural network training must first initialize the network parameters, including the initialization of the weight between the input layer and the hidden layer, the initialization of the weight between the hidden layer and the output layer, and the initialization of the threshold between the hidden layer and the output layer. The error between the network prediction result and the expected result is obtained through the calculation of the hidden layer and the output layer. The network then updates the initial weight and threshold according to the error result. Therefore, the final weight and threshold will be affected by the selection of its initial weight and threshold, which will further affect the convergence speed and prediction accuracy of the network model. Genetic algorithm can overcome the randomness of BP neural network in automatically generating initial weights and thresholds by iterating to find the optimal solution of initial weights and thresholds. Therefore, we combine the genetic algorithm with the BP neural network, and use the genetic algorithm to optimize the initial weight and threshold of the BP neural network, and get a more stable and reliable network structure [10] (**Figure 5**).

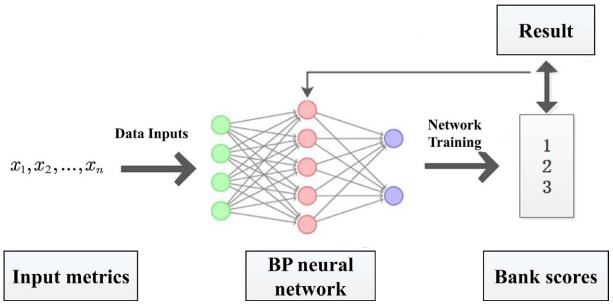


Figure 5. Network structure.

According to the above two questions, we can get the formula about the comprehensive score of the bank:

$$Q_j = \sum_{i=1}^{64} w_i x_{ij},\ j = 0,1,2,\cdots,N \tag{16}$$

Among them, Q_i is the comprehensive score of the bank, w_i is the weight of 64 indicators, x_{ij} is the data set.

The following weights are our conclusions based on the first and second models. The values of w_i in the above equation are all from the following table.

Through the above conclusions and formulas, the relevant indicators of the input layer and the output layer can be obtained, and the neural network prediction model can be established as follows:

The calculation process is as follows:

Through the above neural network model, the indicators of 38 banks are trained with the comprehensive score of banks, so that the comprehensive score value of each bank is Z, and the neural network trained by the above model predicts the comprehensive score is Z, then the relative error is δ:

$$\delta = \frac{|Z - Z'|}{Z} \times 100\% \tag{17}$$

4. Experiments

We did the following experiments with the above model, which used MATLAB 2023RA and Python 3.9.7, The experiments were conducted in Guilin, China, on an ASUS computer with the following The experiments were conducted in Guilin, China, on an ASUS computer with the following specifications: an AMD Ryzen 7 4800 H processor with Radeon graphics, operating at 2.90 GHz, 16 GB of RAM, and an NVIDIA GeForce GTX 1660 Ti graphics card. The operating system used was Windows 11. Through the above neural network model to predict the comprehensive score of each bank, we use MATLAB program to write the program, the detailed program can be found in the attachment and supporting materials, we set the parameters as follows (Table 6):

Table 6. Weights of relevant indicators.

Correlation index	Weight W	Correlation index	Weight W
X1	0.03	X33	0.00
X2	0.03	X34	0.01
X3	0.03	X35	0.01
X4	0.00	X36	0.00
X5	0.00	X37	0.02
X6	0.01	X38	0.00
X7	0.00	X39	0.00
X8	0.01	X40	0.01
X9	0.00	X41	0.01
X10	0.00	X42	0.01
X11	0.00	X43	0.16
X12	0.01	X44	0.00
X13	0.01	X45	0.00
X14	0.00	X46	0.01
X15	0.01	X47	0.00
X16	0.01	X48	0.00
X17	0.01	X49	0.00
X18	0.00	X50	0.01
X19	0.00	X51	0.01
X20	0.00	X52	0.01
X21	0.06	X53	0.02

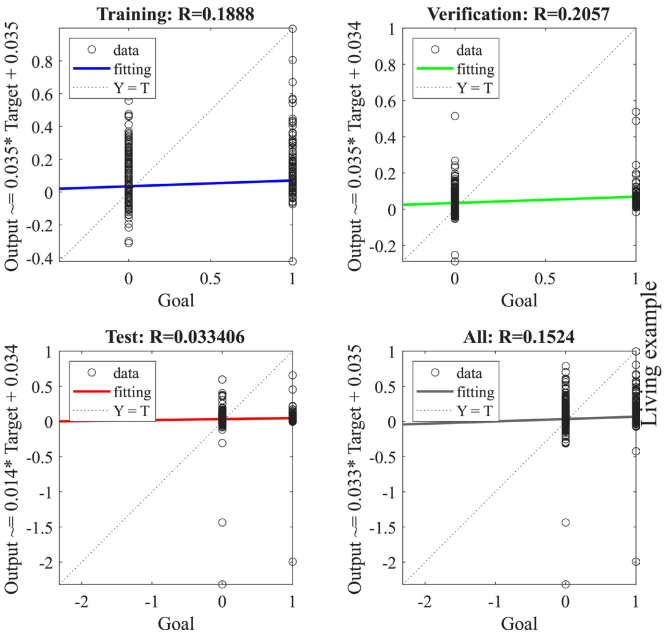
Continued

X22	0.01	X54	0.02
X23	0.00	X55	0.01
X24	0.01	X56	0.00
X25	0.01	X57	0.00
X26	0.01	X58	−0.01
X27	0.01	X59	0.02
X28	0.02	X60	0.00
X29	0.03	X61	0.00
X30	0.01	X62	0.00
X31	0.00	X63	0.01
X32	0.01	X64	0.02

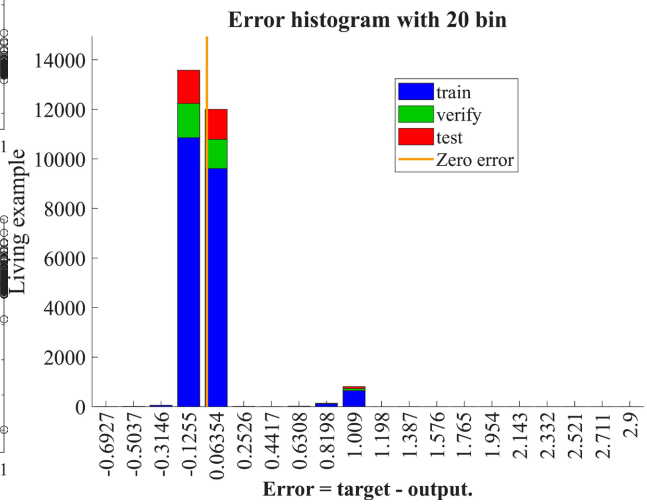
After a period of training, the results shown in the figure are obtained. We can see the changes of the relevant parameters of the neural network, as follows (Table 7):

Table 7. Machine learning related parameters table.

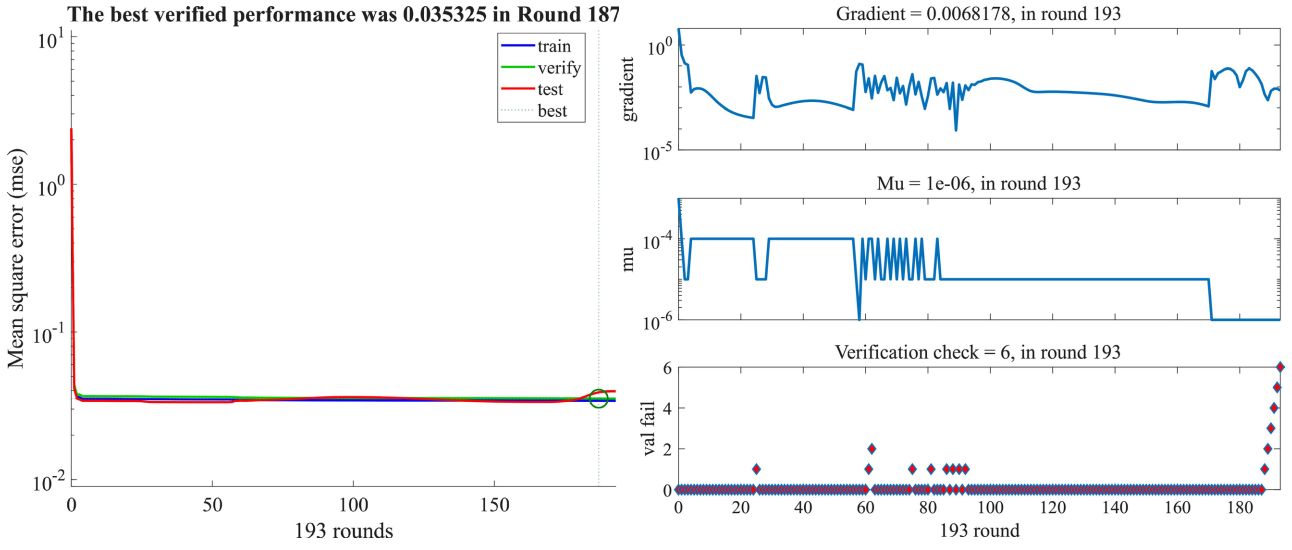
Correlation parameter	Set value
Net.trainParam.show	10,000
Net.trainParam.Lr	0.05
Net.trainParam.epochs	50,000
Net.trainParam.goal	$0.78*10^{-3}$



(a) Neural network regression state

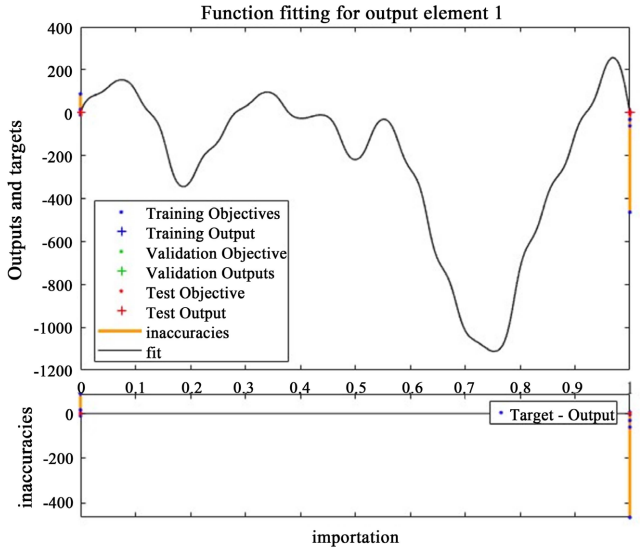


(b) Error histogram



(c) performances

(d) training state



(e) BP neural network fit

Figure 6. BP Neural network structure.

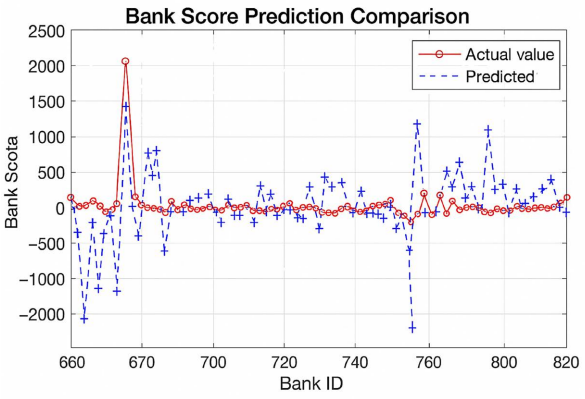


Figure 7. Image of the function between the predicted value and the true value.

Figure 6 illustrates both the state and performance of our trained BP neural network. Using this predictive model, we can assess a bank's risk of failure based on the designated cut-off score. When a bank's predicted value exceeds the cut-off line provided in Problem I, the model effectively identifies potential risks. Furthermore, as shown in Figure 7, our model demonstrates strong performance, achieving an accuracy of 83% when tested on over 800 banks, making it a valuable tool for risk detection across various financial institutions.

We also compared our data with other methods, and the results are shown in Table 8 below:

Table 8. This model is compared with other models.

Model	Overall accuracy	Bankruptcy prediction accuracy
LDA and SVM methods [10]	63.5%	66.3%
Bankruptcy Prediction Models [11]	88.0%	92.0%
Bp Neural Network	66.5%	68.5%
Our model	81.5%	83.5%

5. Conclusion

The experimental results demonstrate the effectiveness of our proposed bank efficiency analysis and risk prediction models. By applying the entropy weight method to the selected input-output indicators, we successfully constructed an efficiency function that distinguishes between operational and bankrupt banks. The generated efficiency evaluation graph clearly shows a significant difference between these two groups, validating the feasibility of our approach. Furthermore, the principal component analysis effectively reduced the dimensionality of the dataset while retaining essential information, enabling a more concise representation of key factors influencing bank stability. The BP neural network, optimized using a genetic algorithm, achieved an accuracy of 83.5% in predicting bank failure, as illustrated in the correlation analysis between predicted and actual values. The model effectively identifies high-risk banks, providing a reliable tool for financial risk assessment.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Beck, T., Demirgüç-Kunt, A. and Pería, M.S.M. (2010) Banking Sector Stability, Efficiency, and Outreach in Kenya. World Bank Policy Research Working Paper, 5442. https://documents1.worldbank.org/curated/en/428671468048252880/pdf/WPS5442.pdf

[2] Bayrakçeken, A. (2024) Economic Security in Times of Crisis: A Legal Framework for Sustainable Financial Resilience. *Mayo RC Journal of Communication for Sustaina-*

ble World, **1**, 65-74.

[3] Petropoulos, A., Siakoulis, V., Stavroulakis, E. and Vlachogiannakis, N.E. (2020) Predicting Bank Insolvencies Using Machine Learning Techniques. *International Journal of Forecasting*, **36**, 1092-1113.

[4] Stengel, M. and Glennon, D. (1999) Evaluating Statistical Models of Mortgage Lending Discrimination: A Bank-Specific Analysis. *Real Estate Economics*, **27**, 299-334. https://doi.org/10.1111/1540-6229.00775

[5] Sun, Q., Xia, J., Deng, S. and Zhou, M. (2022) A Bank Erosion Risk Evaluation Procedure Based on the Analytic Hierarchy Process and Entropy Weight Method. *Arabian Journal of Geosciences*, **15**, Article No. 1772. https://doi.org/10.1007/s12517-022-11065-7

[6] Bruce Ho, C. and Dash Wu, D. (2009) Online Banking Performance Evaluation Using Data Envelopment Analysis and Principal Component Analysis. *Computers & Operations Research*, **36**, 1835-1842. https://doi.org/10.1016/j.cor.2008.05.008

[7] Wang, X. (2016) Efficiency Evaluation of Commercial Banks in China. Ph.D. Thesis, Tianjin University of Finance and Economics.

[8] Zhu, Y., Tian, D. and Yan, F. (2020) Effectiveness of Entropy Weight Method in Decision-making. *Mathematical Problems in Engineering*, **2020**, Article ID: 3564835. https://doi.org/10.1155/2020/3564835

[9] Abdi, H. and Williams, L.J. (2010) Principal Component Analysis. *WIREs Computational Statistics*, **2**, 433-459. https://doi.org/10.1002/wics.101

[10] Li, J., Cheng, J.H., Shi, J.Y. and Huang, F. (2012) Brief Introduction of Back Propagation (BP) Neural Network Algorithm and Its Improvement. In: Jin, D. and Lin, S., Eds., *Advances in Computer Science and Information Engineering*, Springer, 553-558.

[11] Ptak-Chmielewska, A. (2021) Bankruptcy Prediction of Small- and Medium-Sized Enterprises in Poland Based on the LDA and SVM Methods. *Statistics in Transition New Series*, **22**, 179-195. https://doi.org/10.21307/stattrans-2021-010