Enhanced Financial Distress Prediction Model Using GWO-Adaboost Optimized Support Vector Machine

Xiaohua Ma
Anyang Normal University, Anyang 455000 China
E-mail: 02041@ayu.edu.cn

Keywords: ML, risk management, financial forecasting, listed company, support vector machine

Received: July 24, 2025

With the continuous changes in the global economy and increasingly fierce market competition, financial risks faced by municipal companies are becoming increasingly complex and diverse. To provide more accurate and timely risk warning for listed companies and support risk management, a financial distress prediction model is constructed using support vector machine (SVM) in machine learning algorithms to forecast financial data risks. Grey wolf optimization (GWO) algorithm is combined with adaptive boosting (Adaboost) algorithm to optimize parameters. The study is based on the CSMAR database of Guotai An and the data of listed companies on the official website of the National Bureau of Statistics from 2019 to 2024 as samples, and selects normal and special treatment (ST) company data in the A-share market for comprehensive analysis. The listed companies in the dataset include normal companies and ST companies, with a ratio of 1:1. The findings indicate that the prediction accuracy, recall rate, and F1 value of the research model reach 92.67%, 93.52%, and 93.09%, respectively. In the prediction of normal enterprises and ST enterprises, the prediction errors of the improved SVM model are 2.67% and 3.22%, respectively. In summary, research on financial distress prediction and risk management of listed companies based on machine learning can provide more accurate and timely risk warnings for enterprises, help identify potential financial distress in advance, and provide strong support for enterprise decision-making.

Povzetek: Izboljšani model strojnega učenja z visoko natančnostjo napoveduje finančno stisko podjetij ter omogoča pravočasno obvladovanje tveganj.

1 Introduction

As economic globalization rapidly develops, the uncertainty of the global economy continues to intensify, and the financial difficulties and risks faced by enterprises are increasing day by day. Economic globalization has made the economic connections between countries increasingly close, which not only brings more development opportunities for enterprises, but also comes with greater market risks [1]. Financial distress (FD) not only threaten the survival and development of enterprises, but if not handled properly, may also trigger systemic risks in the financial market, thereby affecting the stability of the entire industry and even the national economy [2]. Therefore, how to effectively predict and manage financial difficulties has become a key concern for enterprises and investors. By accurately predicting potential financial difficulties, companies can take measures in advance for risk management, avoiding or mitigating the negative impact of potential financial crises on the enterprise. Traditional financial distress prediction (FDP) methods often rely on static financial data, such as balance sheets and income statements. Although these data can reflect the financial health of a company, they lack the ability to dynamically predict long-term development trends and are difficult to cope with complex market changes [3]. As data mining and machine learning (ML) technologies rapidly evolve, traditional financial forecasting methods are no longer adequate for the needs of modern enterprise financial risk management. ML technology can extract potential patterns from complex data by learning and analyzing massive historical data, providing more accurate support for FDP. There have been earlier studies abroad on using ML for FDP, and empirical evidence shows that financial indicators such as pre tax return and current ratio can effectively predict a company's financial situation [4]. Although research in this field started relatively late in China, recent studies have shown that support vector machine (SVM) prediction models optimized by intelligent algorithms can greatly improve the accuracy of FDP [5]. Therefore, this study combines principal component analysis (PCA) algorithm to screen key financial indexes, and based on SVM prediction model, combines grey wolf optimization (GWO) algorithm and Adaptive Boosting (Adaboost) algorithm to optimize parameters, to further improve the accuracy of FDP of listed companies and provide effective decision support for risk management.

2 Related works

The FDP model can provide decision support for listed companies, which has attracted widespread attention from many scholars. Elhoseny M et al. suggested an FDP model based on deep learning's adaptive whale optimization algorithm (AWOA) to predict corporate bankruptcy and assess credit risk. This model combined a multi-layer perceptron with AWOA's high parameter tuning mechanism, and the findings indicated that its average prediction accuracy was as high as 95.80% [6]. To enhance the precision of FDP, Liu J. et al. conducted a research initiative that utilized gradient boosting decision trees, extreme gradient boosting, lightweight gradient boosting machines, and classification boosting models. The findings indicated that the tree-based model significantly outperformed traditional methods in terms of predictive performance [7]. In terms of text information fusion, Huang B's research team extracted text sentiment features from the annual reports of Chinese A-share listed companies to improve the accuracy of FDP. The findings indicated that the fusion of text sentiment scores significantly improved modeling performance [8]. Hajek P et al. further introduced semantic emotion recognition and text emotion analysis to capture the intentions and perspectives of company stakeholders, combining emotional information with traditional financial indicators. The findings demonstrated that this approach enhances the precision of predictions and substantiates the pivotal function of managerial emotions in predicting FD [9].

Among numerous ML methods, SVM is broadly utilized in the field of FDP. Kurani A et al. proposed to integrate artificial neural networks with SVM to construct a hybrid prediction model for comprehensive FDP. This method effectively solved the problems of financial data scarcity and cold start by integrating forward propagation algorithm and multi-layer feeding neural network [10]. Zhu B's research team developed a multi-objective least squares SVM model to improve prediction accuracy and trading performance. The model combined a mixed kernel function and used particle swarm optimization (PSO) for parameter tuning. The findings indicated that this method effectively improved the accuracy of asset price prediction and trading performance [11]. Yang L et al. proposed an SVM-based investment portfolio prediction model by combining LSTM and reinforcement learning to improve the accuracy of the investment portfolio model. The model was evaluated for its effectiveness through indicators such as Sharpe ratio, and others. The findings indicated that the model had good adaptability and accuracy in multivariate data environments [12]. Wasserbacher H et al. proposed a method of introducing causal inference combined with SVM model for accurate financial forecasting. The findings indicated that the method achieved a certain degree of interpretability and predictive effectiveness of the model, but also demonstrated a high dependence on quantitative analysis methods [13].

Table 1: Summary of related work

Method	Techniques	Limitations
Elhoseny M et al [6].	Deep learning-based AWOA combined with Multilayer Perceptron	Despite achieving an accuracy of 95.80%, it does not consider text sentiment information, potentially overlooking important unstructured data.
Liu J et al [7].	Four tree-based gradient boosting models (including GBDT, XGBoost, LightGBM, and CatBoost)	Lacks consideration of non-financial factors, such as management sentiment and market sentiment.
Huang B et al [8].	Extracting text sentiment features from annual reports	Solely relies on text data, neglecting the importance of financial data.
Hajek P et al [9].	Semantic emotion recognition and text sentiment analysis combined with traditional financial indicators	High method complexity and long model training time.
Kurani A et al [10].	Artificial Neural Network combined with SVM	Does not apply intelligent optimization algorithms for parameter tuning, potentially affecting prediction performance.
Zhu B et al [11].	Multi-objective least squares SVM combined with hybrid kernel functions and PSO	Limited improvement in prediction accuracy and high model complexity.
Yang L et al [12].	LSTM combined with Reinforcement Learning and SVM	Limited applicability and may not handle large-scale datasets effectively.
Wasserbacher H et al [13].	Causal inference combined with SVM	Requires a large amount of historical data for causal inference, making data acquisition difficult.
Proposed Method	PCA for dimensionality reduction combined with intelligent optimization algorithms for adaptive tuning of SVM models	Combines multiple optimization techniques to enhance the overall performance of the model.

The relevant work summary table is shown in Table 1.

In summary, despite significant progress in FDP models, existing research still faces a series of challenges when dealing with high-dimensional, redundant, and highly heterogeneous financial data. For example, although Elhoseny M et al.'s deep learning model [6] achieved high prediction accuracy, it ignored the importance of textual emotional information and may have overlooked important unstructured data. Although the tree model proposed by Liu J et al. [7] outperforms traditional methods in predictive performance, it does not take into account non-financial factors such as management sentiment and market sentiment. In addition, Huang B et al. [8] relied solely on textual data and ignored the importance of financial data, while Hajek P et al. [9]'s method was complex and time-consuming to train. Kurani A et al. [10] did not use intelligent optimization algorithms for parameter adjustment, which may affect prediction performance. Zhu B et al. [11]'s multi-objective least squares SVM improved prediction accuracy, but the model complexity was high. The method proposed by Yang L et al. [12] has limited applicability on large-scale datasets, while Wasserbacher H et al. [13]'s method relies on a large amount of historical data, making data acquisition difficult. Therefore, the study adopts PCA method to reduce the dimensionality of financial indicators, thereby extracting key feature variables, and then combines intelligent optimization algorithm to adaptively optimize the key parameters of SVM model. The ultimate goal is to improve the overall performance and practical application value of the FDP model.

3 FDP model based on ML

The study conducts data preprocessing based on financial data of listed companies and uses PCA algorithm to screen key financial indicators. Subsequently, based on the SVM FDP model, the GWO algorithm is applied to optimize its parameters. Moreover, the Adaboost algorithm is further utilized to improve the optimized SVM model, to enhance the accuracy and robustness of FDP.

3.1 Financial data processing and selection of key indicators

In ML-based FDP models, the processing of financial data and the selection of key indicators are the foundation and core of model construction. For this purpose, the study selects the CSMAR database of Guotai An and the data of listed companies on the official website of the National Bureau of Statistics from 2019 to 2024 as samples, and selects data from normal companies and Special Treatment (ST) companies in the A-share market for comprehensive analysis. To ensure that the data can be effectively used for training ML models, it needs to first preprocess the raw financial data appropriately. The specific preprocessing steps include missing value processing, outlier detection and processing, data standardization and normalization, and processing of time series data [14]. For missing data, the study uses the mean of the column to fill in missing values, as shown in equation (1).

$$\hat{x}_i = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

In equation (1), $\hat{x}_i$ refers to the data after filling in missing values, x_i means the original data, and n means the total amount of i data. Data standardization is shown in equation (2).

$$z_i = \frac{x_i - \mu}{\sigma} \tag{2}$$

In equation (2), z_i refers to the standardized data, μ denotes the mean of the data, and σ denotes the data's standard deviation. Normalization operation is to scale data proportionally to a specified range, as shown in equation (3).

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{3}$$

In equation (3), x_i' denotes the normalized data, and $x_{\min}$ and $x_{\max}$ indicate the mini and max values of the feature. There are numerous factors that can affect a company's financial condition, and the degree of impact and criteria for determining financial risk for each factor may vary from company to company. Therefore, the study selects concise and representative financial warning indicators to comprehensively reflect the operation and development of enterprises. The preliminary financial warning indicators are shown in Figure 1.

In Figure 1, the selected financial warning indicators in the study include primary indicators such as profitability, solvency, development ability, and corporate governance ability. Each primary indicator includes several secondary indicators. For example, the secondary indicators of profitability include return on assets, operating profit margin, and earnings per share. The secondary indicators of solvency include equity ratio, asset liability ratio, and current ratio. However, facing massive financial indicators, it is a challenge to choose the key indicators that best reflect the financial difficulties of the enterprise. Therefore, to more effectively screen representative financial indicators, the PCA algorithm is adopted in the study. PCA is a frequently employed dimensionality reduction technique that facilitates the transformation of raw, high-dimensional data into a low-dimensional space through linear transformation, thereby extracting the most informative principal components (PCs) [15]. When studying the dimensionality reduction of financial indicators through PCA algorithm, the principal components that retain 95% of the total variance are selected. This choice not only ensures that the main

information of the data is not lost, but also effectively reduces the dimensionality of the data. The calculation expression for PCA is given in equation (4).

$$\sum = \frac{1}{n-1} X^T X \tag{4}$$

In equation (4), $\sum$ denotes the covariance matrix, and X denotes the standardized data matrix. The condition for the number of PCs is shown in equation (5).

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq \alpha \tag{5}$$

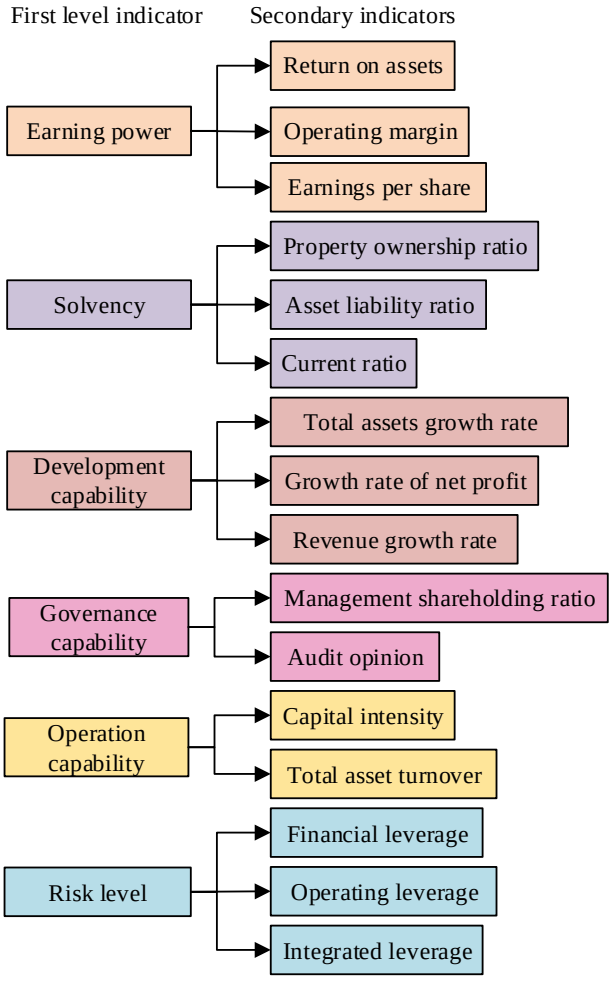


Figure1: Preliminary selection of financial warning indicators

In equation (5), λ_i denotes the variance size of each PC, p indicates the amount of dimensions in the data, α means the proportion of total variance required, and m denotes the amount of PCs. The calculation process of PCs is given in equation (6).

$$Z_i = \beta_{1i} X_1 + \beta_{2i} X_2 + ... + \beta_{pi} X_p \tag{6}$$

In equation (6), Z_i indicates the value of the i th PC, β_{ki} denotes the coefficient of the k th feature of the i th PC, and X_k denotes the X_k th feature of the original data, $k = 1, 2, ..., p$.

3.2 Construction of a distress prediction and risk management model based on improved SVM

By studying the use of ML methods to predict potential financial difficulties of listed companies, effective support can be provided for risk management of enterprises. By using FDP models, companies and investors can identify potential financial crises and take corresponding risk management measures in advance. With the continuous development of the field of FDP, traditional prediction methods, although able to provide accurate prediction results to a certain extent, are often limited by issues such as feature selection, non-linear data relationships, and high-dimensional data. In recent years, the SVM in ML has become a powerful tool for predicting FD due to its excellent classification ability and strong ability to handle high-dimensional data. However, traditional SVM's prediction accuracy is often greatly affected by kernel function parameters and penalty coefficients. Therefore, in practical applications, how to choose appropriate kernel functions and adjust model parameters has become a key issue affecting SVM performance. The GWO algorithm in intelligent optimization algorithms has advantages such as simplicity, ease of implementation, good convergence, and fewer parameters. Therefore, the study focuses on parameter optimization of SVM prediction models based on the GWO algorithm. The process of SVM distress prediction model based on the GWO algorithm improvement is illustrated in Figure 2.

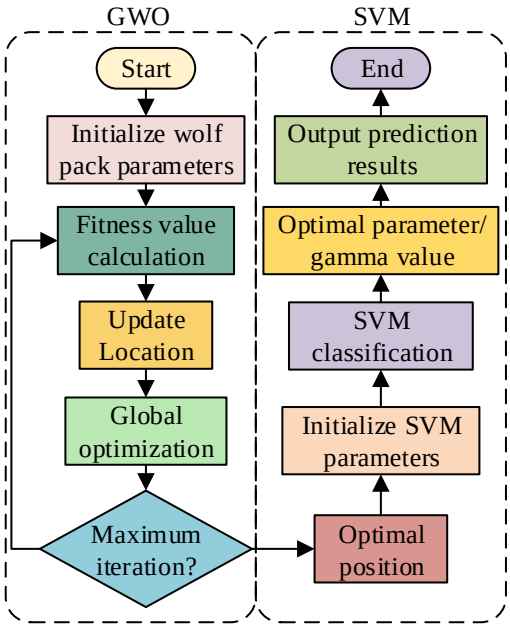


Figure 2: Process of SVM distress prediction model improved by GWO algorithm

As shown in Figure 2, in this process, the parameters of the wolf pack are initialized first, and a random gray wolf population is generated. Then, the fitness value of each individual is calculated and the location of the wolf pack is updated according to the fitness value. Subsequently, a global optimality search is conducted to ascertain whether the max number of iterations has been attained. In the event that the max number of iterations is reached, the optimal position is output and the parameters of the SVM are initialized according to this position. If the max number of iterations is not reached, it returns and recalculates the fitness value. After initializing the parameters of the SVM, the SVM classification model is built, the optimal parameters and gamma values are obtained, and these optimal parameters are used for the training and testing of the SVM model, and the final prediction outcomes are output. The location of the wolf pack is shown in equation (7).

$$X_j = l + r * rand(0,1) * (u - l) \tag{7}$$

In equation (7), X_j means the position of the j th gray wolf individual, j and u represent the lower and upper bounds of the search space, $rand(0,1)$ means a randomly generated value between 0 and 1, and r means an additional random factor. The position update formula is given in equation (8).

$$\begin{cases} D = \left| C * X_{P(t)} - X_{(t)} \right| \\ X(t+1) = X_{P(t)} - A * D \\ A = 2a \end{cases} \tag{8}$$

In equation (8), D denotes the distance between the current position of the grey wolf and the position of the prey, $X_{P(t)}$ indicates the position of the prey, $X_{(t)}$ means the current location of the grey wolf, A and C represent two adjustment parameters in the algorithm, which control the direction and step size of the search, respectively, and a denotes a constant that gradually decreases over time, used to control the search range. The mathematical expression for classification decision-making is given in equation (9).

$$f(x) = \omega x + b \tag{9}$$

In equation (9), $f(x)$ means the classification prediction value for input sample x, ω means the input feature vector, and b stands for the bias term. The objective function calculation is given in equation (10).

$$\varphi(w) = \frac{1}{2}\|w\|^2 \tag{10}$$

In equation (10), φ denotes minimizing the objective function, φ denotes the weight vector of the classification model, and φ denotes the square of its second norm. The research uses Lagrange functions to introduce constraints into optimization problems, and adjusts the contribution of each sample point to the optimization objective through these multipliers. The Lagrange function is given in equation (11).

$$L(w,b,\kappa) = \frac{1}{2}\|w\|^2 - \sum_{j=1}^{n} \kappa_j \left[y_j (w * x_j + b) - 1 \right] \tag{11}$$

In equation (11), L denotes the Lagrangian function, y_j denotes the label of sample j, x_j denotes the characteristics of sample j, and κ denotes the Lagrange multiplier, which is the degree of influence of each constraint. The final classification function expression formula is given in equation (12).

$$F(x) = sign\left(\sum_{j=1}^{n} \kappa_j^* y_j (x_j * x) + b^* \right) \tag{12}$$

In equation (12), $F(x)$ stands for the final classification function, $F(x)$ stands for the Lagrange multiplier obtained by optimizing the Lagrange function, and b^* stands for the optimized bias term. The pseudo code for GWO-SVM parameter optimization is shown in Table 2.

Table 2: Pseudo code for GWO-SVM parameter optimization

Input: Training Set Data_train, Grey Wolf Population N, Maximum Iteration Times T
alpha, beta, delta = initialize_wolves(N)
for t in range(T):
for each wolf in population:
fitness = cross_val_score(SVM(C, γ), Data_train)
update_leader_positions(alpha, beta, delta, fitness)
update_wolf_positions(population, alpha, beta, delta)
return alpha.position

However, predicting FD is not only about identifying potential financial issues, but more importantly, risk management is carried out through these predicted results. In this process, Adaboost algorithm, as a method of ensemble learning, can improve the model's prediction accuracy by combining multiple weak classifiers into a strong classifier [16]. However, a non-traditional combination approach is used in the study, which combines the Adaboost algorithm with SVM. SVM is usually considered a strong classifier, but this study combined multiple SVM models as weak classifiers and used Adaboost algorithm for weighted optimization. Although this combination method is not common, its purpose is to further improve the robustness and prediction accuracy of the dilemma prediction model. By iteratively optimizing the classifier weights multiple times, the overall model performance can be improved. Therefore, to further improve the robustness and prediction accuracy, on the basis of improving the SVM, the Adaboost algorithm is introduced to optimize the classifier weights through multiple iterations, thereby

improving the overall model performance. The principle of Adaboost algorithm is shown in Figure 3.

As shown in Figure 3, the Adaboost algorithm generates a strong learner by updating the weights of multiple weak classifiers and combining the weights of each sample through weighting. Finally, after all iterations are completed, the Adaboost algorithm weights and averages the prediction outcomes of each weak classifier to get the final prediction result of a strong classifier. The core of the Adaboost algorithm is to adjust the weight of each sample so that difficult to classify samples receive higher weights at each iteration, thereby allowing the model to focus more on these difficult to classify samples. The formula for updating sample weights is shown in equation (13).

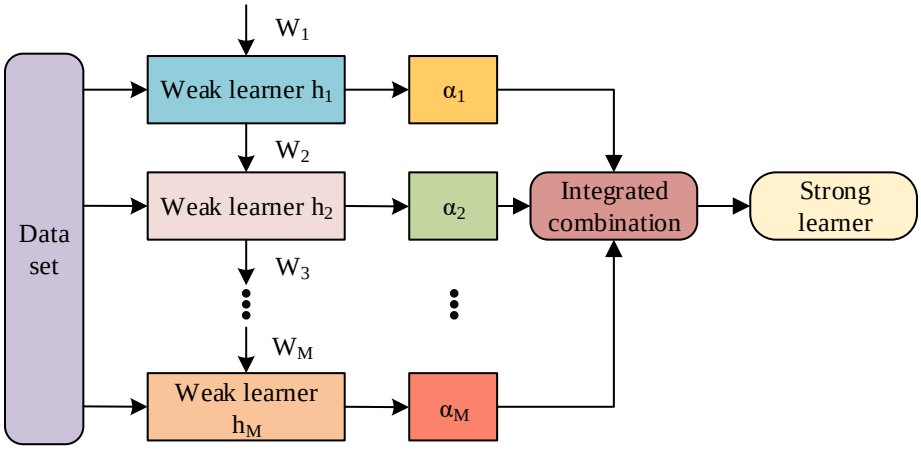


Figure 3: Schematic diagram of Adaboost algorithm

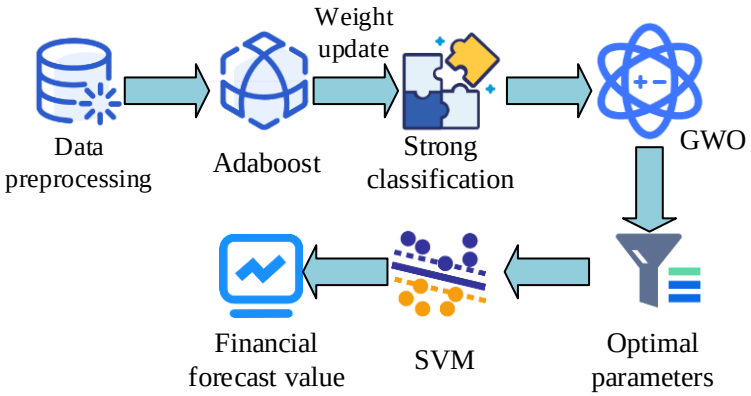


Figure 4: Structure of distress prediction model based on improved SVM

$$D_{t+1}(j)=\frac{D_t(j)*\exp\left[-\zeta_t y_j G_t(x_j)\right]}{Z_t} \tag{13}$$

In equation (13), $D_{t+1}(j)$ stands for the weight of sample j in the t th iteration, j stands for the weight of the weak classifier in the t th iteration, $G_t(x_j)$ stands for the predicted value of the weak classifier on sample j in the t th iteration, and Z_t refers to the normalization factor. The mathematical expression for classification error rate is shown in equation (14).

$$E_t=\sum_{j=1}^{N}\left|G_t(x_j)-y_j\right|D_t(j) \tag{14}$$

In equation (14), E_t refers to the classification error rate of the t th weak classifier, and N indicates the quantity of samples in the training set. The strong classifier is given in equation (15).

$$G(x)=sign\left(\sum_{t=1}^{T}\zeta_t G_t(x)\right) \tag{15}$$

In equation (15), $G(x)$ refers to a strong classifier, and the final prediction result is determined by the sign function *sign*, that is, if the weighted sum is positive, it is predicted as a positive class; If the weighted sum is negative, it is predicted as a negative class [17]. The structure of the improved SVM-based distress prediction model is illustrated in Figure 4.

As shown in Figure 4, the model first processes the preprocessed data and updates the weights of the samples using the Adaboost algorithm to generate a strong classifier prediction result. Next, the prediction results of

the strong classifier are input into the GWO algorithm to select the optimal parameters of the SVM model, thereby optimizing the performance of the model. Finally, the optimized parameters are passed to the SVM classifier for final classification, resulting in FDP values.

(i) The main objective of the research is to surpass the performance of existing RF and ANN models in FDP through an improved SVM model.

(ii) The reason for choosing the GWO algorithm for research is that it has strong global search capabilities and can effectively avoid local optimal solution problems. Moreover, the GWO algorithm has the characteristic of fast convergence when dealing with high-dimensional and complex optimization problems, which is suitable for optimizing the parameters of SVM. In addition, compared to GA and PSO, the GWO algorithm exhibits better adaptability and robustness in various practical applications. The reason for choosing the Adaboost algorithm is that by weighting and combining multiple weak classifiers, the Adaboost algorithm can significantly improve the overall classification performance of the model. The Adaboost algorithm can also pay more attention to misclassified samples, improving the model's generalization ability. In addition, compared with other integration methods such as Bagging, the Adaboost algorithm is simpler and more efficient in implementation and application, and is suitable for FDP in practical business scenarios.

(iii) The improved SVM model can detect FD risks 3 to 6 months in advance, providing sufficient time for enterprise management to take preventive measures. At the same time, the model also has fast computing speed and real-time update capability, which is suitable for real-time monitoring and analysis of enterprise financial status. In addition, the model can be used to regularly generate financial health reports, helping companies conduct periodic financial reviews and adjustments.

4 Validation of FDP model based on ML

The study first established an experimental environment, then validated the performance of the improved SVM distress prediction model, and finally conducted empirical analysis and verification on data from listed companies.

4.1 Experimental environment setup

To test the effect of the ML-based FDP model, an experimental environment was constructed. The experiment was carried out on the Ubuntu 16.04.7LTS operating system, and Matlab R2016b simulation software was used for model construction and experimentation. In terms of hardware configuration, the experimental equipment was equipped with an Intel Core i7-9700 processor, with 64 GB of memory and 11 GB of video memory, and used NVIDIA GeForce GTX 1080Ti graphics card to accelerate calculations. In addition, the LIBSVM 3.24 software package was used in the experiment to implement the SVM algorithm. To guarantee the accuracy and efficiency of the experiment, the listed companies in the dataset included normal companies and ST companies, with a ratio of 1:1. The preprocessed dataset was divided into 80% training data and 20% testing data. A normal company referred to a listed company with good financial condition and no significant financial difficulties or risks. ST companies referred to listed companies that have been specially treated due to poor financial conditions, consecutive losses for two years, or other reasons. These companies were subject to special regulation and labeling in the securities market, and faced significant delisting risks. The dataset in the study included normal companies and ST companies, with a ratio of 1:1. This balanced data distribution helped the model avoid being affected by class imbalance during the training process. However, to discuss the issue of category imbalance more formally, the number distribution of the two types of companies in the dataset was 500 normal companies and 500 ST companies each.

During the experiment, the GWO algorithm was used to optimize the hyperparameters of the SVM model. The search space and parameter range for hyperparameter optimization were as follows: the penalty parameter had a value range of $[2^{-5}, 2^{15}]$, the kernel function parameter had a value range of $[2^{-15}, 2^{3}]$, the GWO algorithm had 100 iterations, and the population size was set to 30. In addition, the study used PCA method to reduce the dimensionality of financial indicators. After dimensionality reduction, 95% of the variance was retained, resulting in a final feature count of 20. In order to evaluate the generalization ability of the model and reduce the risk of overfitting, a 5-fold cross validation was conducted during the training process. In addition, to address the issue of class imbalance, the SMOTE algorithm was used to generate synthesized minority class samples to ensure that the training data is more balanced in class distribution. Table 3 shows the experimental environment configuration.

Table 3: Experimental environment configuration

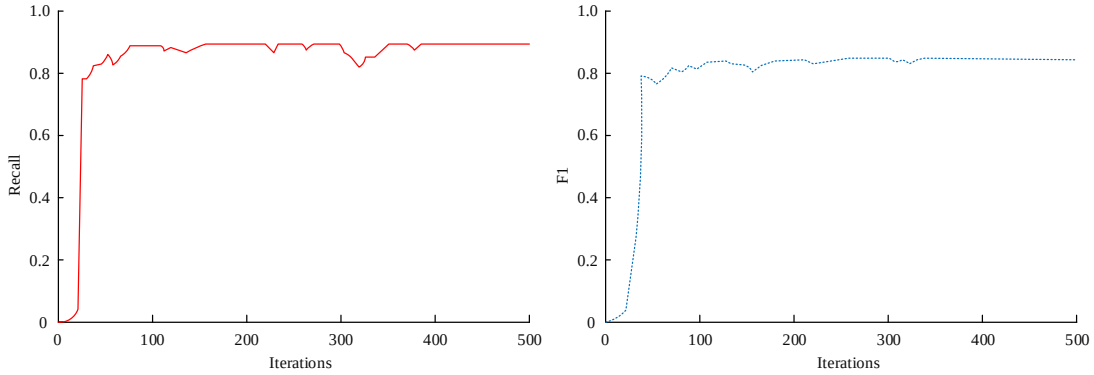
Equipment	Configuration
Operating system	Ubuntu 16.04.7 LTS
Simulation software	Matlab R2016b
Processor	Intel Core i7-9700
Memory	64 GB
Video memory	11 GB
Graphics card	NVIDIA GeForce GTX 1080Ti
Software package	LIBSVM 3.24

4.2 Performance verification of improved SVM based distress prediction model

To validate the effect of the FDP model based on improved SVM, the stability of the SVM model was first verified. The trend of recall rate and F1 value during the training of SVM model is shown in Figure 5. The horizontal axis in Figure 5 represents the number of

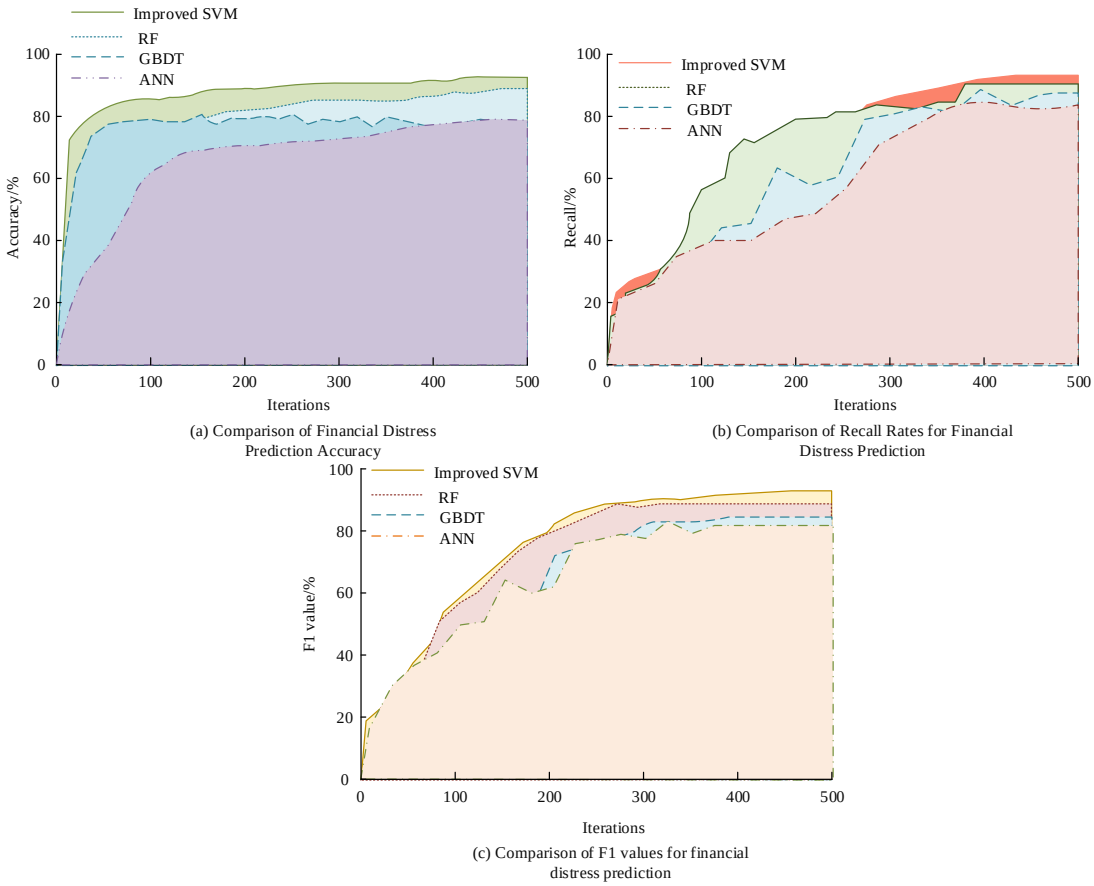
iterations, and the vertical axis represents the recall rate and F1 value, respectively. From Figure 5 (a), the recall rate increased rapidly in the early stages of training and gradually stabilized with the increasing number of iterations, eventually converging to 0.89. This indicated that in the initial stage of model training, SVM models could quickly capture key features of FD and gradually optimize them in subsequent iterations, ultimately reaching a stable state. The trend of F1 value variation shown in Figure 5 (b) is consistent with the trend of recall rate variation. During the training process, the F1 value gradually increased and eventually converged to 0.85, indicating that the model performed well in balancing accuracy and recall. In summary, the improved SVM model exhibits high stability and good predictive ability in FDP, and can effectively identify the risks of FD.

To further test the effectiveness of the improved SVM prediction model, a comparative analysis was conducted with other advanced prediction models. Other models include Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and Artificial Neural Network (ANN) [18-19]. The performance comparison of different prediction models is shown in Figure 6.



(a) The trend of recall rate changes (b) Trend of F1 value variation

Figure 5: The trend of recall and F1 value changes in SVM model during training



(a) Comparison of Financial Distress Prediction Accuracy

(b) Comparison of Recall Rates for Financial Distress Prediction

(c) Comparison of F1 values for financial distress prediction

Figure 6: Performance comparison of different prediction models

Table 4: The recognition performance of different prediction models on financial data

Data set	Model	Normal enterprise		ST enterprise	
		Accuracy/%	Error rate/%	Accuracy/%	Error rate/%
Training set	Improved SVM	98.11	1.89	96.22	3.78
	RF	83.01	16.99	86.79	13.21
	GBDT	92.45	7.55	90.56	9.44
	ANN	88.67	11.33	84.90	15.10
Test set	Improved SVM	97.33	2.67	96.78	3.22
	RF	82.51	17.49	83.24	16.76
	GBDT	89.60	10.40	87.68	12.32
	ANN	85.25	14.75	82.16	17.84

The horizontal axis in Figure 6 represents different models, and the vertical axes in Figures 6 (a), 6 (b), and 6 (c) represent the prediction accuracy, recall, and F1 value percentages, respectively. From Figure 6 (a), the improved SVM prediction model achieved the highest prediction accuracy of 92.67%, which was 4.13%, 12.44%, and 13.76% higher than the 88.54%, 80.23%, and 78.91% of the RF, GBDT, and ANN models, respectively. From Figure 6 (b), the improved SVM prediction model had a prediction recall rate of up to 93.52%, which was 2.90%, 5.18%, and 7.85% higher than the 90.62%, 88.34%, and 85.67% of the other three models, respectively. From Figure 6 (c), the improved SVM prediction model had a predicted F1 value of up to 93.09%, which was 3.84%, 8.47%, and 9.95% higher than the 89.25%, 84.62%, and 83.14% of the other three models, respectively. Overall, the research model has demonstrated higher prediction accuracy, stronger recognition ability, and better stability in FDP tasks.

To further validate the performance of each prediction model, the recognition performance of different prediction models on financial data was compared and studied, as shown in Table 4. From Table 4, in the training set, the improved SVM model had prediction errors of 1.89% and 3.78% for normal and ST enterprises, respectively, while the forecasting errors of the other three models were all higher than 7%. In the test set, the improved SVM model had prediction errors of 2.67% and 3.22% for normal and ST companies. The forecasting errors of the other three models were all higher than 10%. Overall, the improved SVM model exhibits high accuracy and low prediction error on both the training and testing sets, indicating its strong advantages and practicality in FDP tasks.

To verify the statistical significance of the performance improvement of the improved SVM model compared to the baseline model, paired sample significance tests were conducted on three key indicators: accuracy, recall, and F1 value. Wilcoxon signed rank test was used, with a significance level of $\alpha=0.05$. The test results are shown in Table 5. From Table 5, the performance differences of the improved SVM model compared to the RF, GBDT, and ANN models on all three indicators were highly statistically significant ($p<0.05$). This indicated that the improved SVM model observed in this article had a prediction accuracy 4.13% higher than RF, GBDT, and ANN, respectively. 12.44% and 13.76%. The recall rates increased by 2.90%, 5.18%, and 7.85% respectively. The F1 values increased by 3.84%, 8.47%, and 9.95% respectively. This indicated that the advantage in this aspect was not accidental, but significantly superior to each baseline model in statistical significance. Overall, it can be seen that the improved SVM model significantly outperforms the RF, GBDT, and ANN models in terms of accuracy, recall, F1 score, false negative rate, and specificity, confirming the robustness and effectiveness of the proposed improvement method.

Table 5: Significance test results

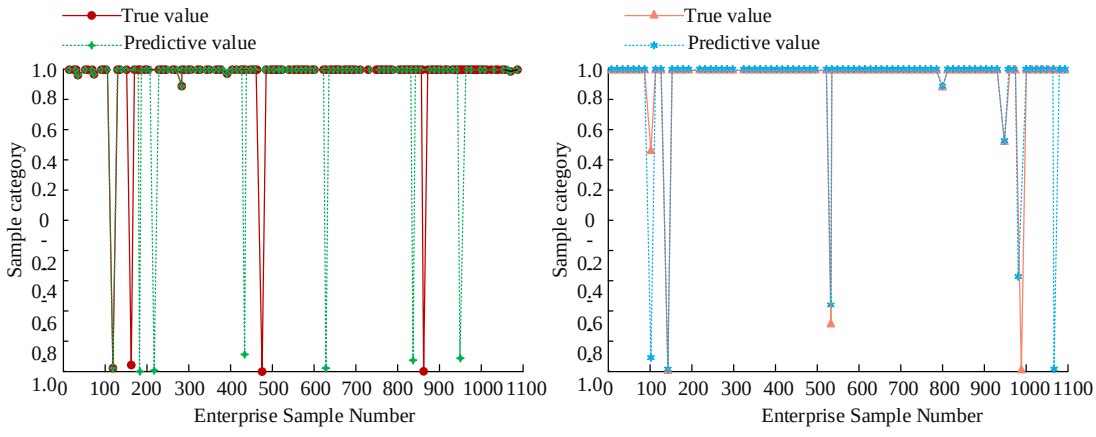
Performance metrics	Model	Average difference/%	Test statistic/W	*p* value
Accuracy	RF	4.13	3521	0.023
	GBDT	12.44	4186	0.001
	ANN	13.76	4265	0.016
Recall	RF	2.90	3382	0.048
	GBDT	5.18	3897	0.027
	ANN	7.85	4013	0.031
F1 value	RF	3.84	3465	0.040
	GBDT	8.47	4024	0.043
	ANN	9.95	4158	0.006
False Negative Rate	RF	-2.50	3291	0.052
	GBDT	-4.75	3710	0.039

	ANN	-6.12	3845	0.028
Specificity	RF	3.21	3415	0.045
	GBDT	6.44	3952	0.018
	ANN	7.88	4090	0.012

4.3 Empirical analysis

To prove the effectiveness of the FDP model in practical applications, the study utilized financial data from the Ruisi database to screen data from 1231 listed companies, including 179 ST companies and the remaining 1052 normal companies. Based on these data, the study analyzed the prediction results of the improved SVM model under normal and ST enterprise samples, as shown in Figure 7. The horizontal axis in Figure 7 represents the number of samples, and the vertical axis represents the percentage of prediction accuracy and error rate. From Figure 7 (a), in the prediction findings of normal enterprises, the improved SVM prediction model only had 6 samples whose prediction results are inconsistent with the real samples, with an error of only 0.57% and an accuracy rate of 99.43%. Figure 7 (b) shows the forecast results of ST companies, among which only three companies had forecast results that did not match the actual situation, with an error rate of 1.68% and an accuracy rate of 98.32%. In summary, the improved SVM model demonstrates high accuracy in both normal and ST enterprise samples, validating the effectiveness of the model in predicting FD.

To assess the effect of key indicators in the FDP model, the SHAP results of the selected indicators were studied and analyzed, as shown in Figure 8. The horizontal axis in Figure 8 (a) represents SHAP values, and the vertical axis represents different financial indicators. Figure 8 (b) shows the SHAP waterfall plot, where each point represents a sample and the color changes from red to blue. The SHAP method provides a means of elucidating the prediction results of ML models. The method is based on the concept of the Shapley value in game theory, which is utilized to quantify the contribution of each feature to the final prediction result [20]. From Figure 8 (a), the top nine indicators in the FDP model that have the greatest influence on the prediction outcomes were all secondary indicators related to the profitability and solvency of enterprises selected through research. Among them, the most important indicator was the asset net profit margin, with a SHAP value of 0.21. The asset liability ratio was the lowest indicator, with a SHAP value of 0.05. The indicator of income growth rate was originally expected to have a higher importance, but the results showed that its SHAP value was only 0.07, which unexpectedly lowered the importance of this feature. This may be due to other profitability indicators already providing sufficient information in the model, making the marginal contribution of operating profit margin relatively small. Figure 8 (b) showcases the SHAP waterfall plot, where each point refers to a sample and the color changes from red to blue. Red showcases a higher value for the feature, while blue showcases a lower value. A positive SHAP value indicates that the feature has increased the predicted value. From the figure, the SHAP values of the selected indicators were heavily clustered to the right of the 0.0 value, indicating that these indicators have a great positive influence on the model's prediction outcomes. In summary, the SHAP results validate the effectiveness of the key indicators selected in the FDP model.



(a) The predicted results of normal enterprises (b) ST Enterprise's Forecast Results

Figure 7: The prediction results of improved SVM model

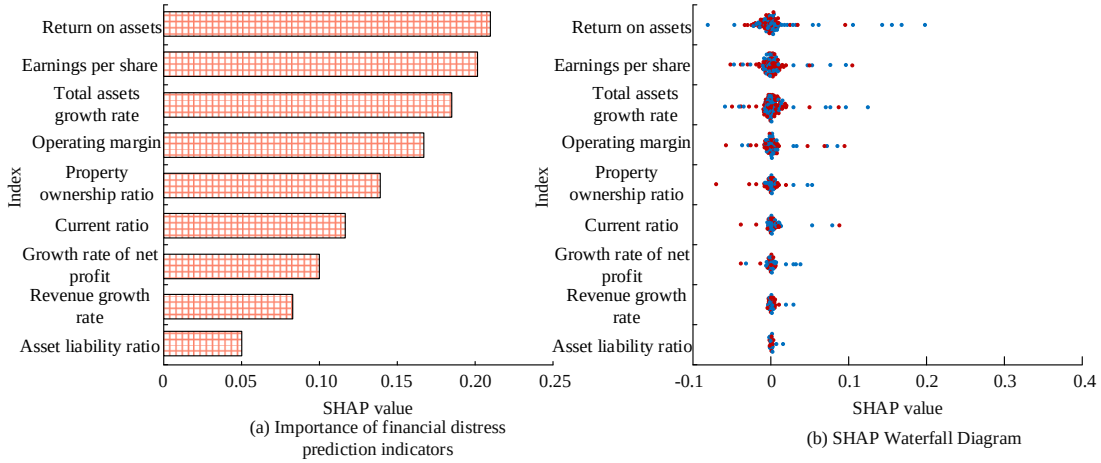


Figure 8: SHAP results for selecting indicators

To further validate the effectiveness of key feature selection, a visual analysis was conducted on the two most important features, as shown in Figure 9. The horizontal and vertical axes of Figure 9 represent the asset net profit margin and earnings per share, respectively. From Figure 9, the horizontal and vertical axes represent asset net profit margin and earnings per share, respectively. Based on these two most important indicators, it is possible to clearly distinguish between normal enterprise samples and ST enterprise samples. In summary, the selection of key financial indicators in FDP models is crucial. By selecting these indicators reasonably, the accuracy of predictions can be significantly improved, which further verifies the effectiveness of the selected indicators in the study.

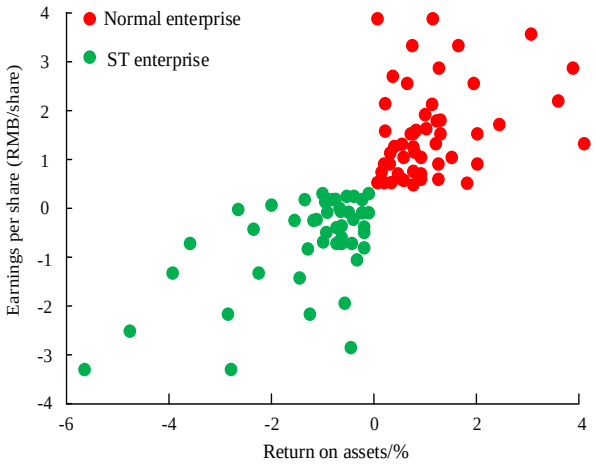


Figure 9: Visualization analysis results

The study conducted Panel logit regression analysis, with ST enterprises as the dependent variable and key financial indicators as the independent variable. The table below shows the results of the regression analysis. In Table 6, the results of the panel logit regression confirmed how well the chosen key financial indicators can be used to predict FD. The *p*-values of each indicator were all less than 0.05, showing that they have a considerable effect on the FDP of ST enterprises. These results further demonstrate the scientific and rational nature of the research in the feature selection stage, and provide a solid theoretical basis for the accuracy of FDP models. In addition, it can be seen that features such as return on total assets (ROA) and earnings per share (EPS) dominate the model prediction, as they directly reflect the profitability and return on capital of the enterprise. The return on total assets (ROA) is an indicator of the efficiency with which a company generates profits using its assets. A higher ROA indicates excellent performance in managing assets, while a lower ROA may indicate an increased risk of financial distress for the company. Earnings per share (EPS) represents the value created by a company for its shareholders and is also an important indicator of a company's profitability. A higher EPS usually means that a company has strong profitability in market competition and can better withstand market fluctuations and potential financial risks. Through SHAP (SHapley Additive exPlanations) analysis, it is possible to gain a deeper understanding of the contribution size of each feature to the model output, which has important practical significance for managers and investors. Managers can identify the most important factors for the financial health of the company based on SHAP values, making resource allocation and strategic decisions more precise. For example, if SHAP analysis shows that the return on total assets (ROA) has the greatest impact on predicting financial distress, managers can strengthen asset management and operational efficiency to increase ROA and reduce the risk of financial distress. For investors, SHAP analysis results can help them identify high-risk and low-risk enterprises more clearly in the investment decision-making process. Investors can focus on financial indicators with high SHAP values and evaluate the financial health of the company through changes in these indicators. For example, if a company's earnings per share (EPS) SHAP value is high and showing an upward trend, investors can consider the company to have good profitability and lower financial

distress risk, and make more informed investment decisions.

In order to conduct a more comprehensive evaluation, the study conducted a comprehensive analysis using indicators such as ROC-AUC, PR-AUC, and Matthew's correlation coefficient (MCC). The comprehensive performance evaluation results of the test set model are shown in Table 7.

Table 6: Regression analysis results

Variable	Coefficient	Standard deviation	t-statistic	*p* value
Return on assets	-0.825	0.213	-3.874	<0.001**
Earnings per share	0.317	0.149	-2.128	0.035*
Total assets growth rate	-0.276	0.139	-1.988	0.048*
Operating margin	-0.563	0.162	-3.478	<0.001**
Property ownership ratio	0.224	0.090	2.489	0.013*
Current ratio	-0.154	0.117	-1.321	<0.001**
Growth rate of net profit	-0.601	0.145	-4.145	<0.001**
Revenue growth rate	-0.198	0.088	-2.262	0.024*
Asset liability ratio	0.274	0.134	2.023	0.043*

Note: * is $p<0.05$, ** is $p<0.001$.

Table 7: Comprehensive performance evaluation results of the test set model

Model	Confusing matrix elements	Precision（%）	MCC	ROC-AUC
Improve SVM	TP=141, FP=23 FN=5, TN=831	97.20	0.895	0.988
RF	TP=122, FP=149 FN=24, TN=705	82.76	0.382	0.842
GBDT	TP=128, FP=89 FN=18, TN=765	89.35	0.648	0.918
ANN	TP=120, FP=126 FN=26, TN=728	84.81	0.498	0.892

From Table 7, it can be seen that the accuracy of the research model is as high as 97.20%, while the accuracy of other models is over 90%. From the perspective of MCC, the improved SVM model has an MCC value of 0.895, which is 134.56%, 38.12%, and 79.72% higher than RF, GBDT, and ANN, respectively. In addition, the improved SVM model has a ROC-AUC value of up to 0.988, which is also significantly better than other models, indicating its excellent performance in distinguishing positive and negative samples. In terms of PR-AUC, the PR-AUC value of the research model is 0.975, which is 33.20%, 12.97%, and 24.02% higher than the other three models, respectively. From the perspective of the confusion matrix, the specific elements of the confusion matrix also demonstrate the superiority of the improved SVM model in various types of misclassifications. The TP of the improved SVM model is 141, FP is 23, FN is 5, and TN is 831. In contrast, the FP and FN values of other models were significantly higher, indicating that the improved SVM model can more accurately identify positive and negative samples and reduce misclassification. In summary, the improved SVM model has higher accuracy and robustness.

5 Discussion

With the development of economic globalization, the risk linkage between companies has gradually increased, and financial risk management has received widespread attention. To provide more accurate and timely risk warning for listed companies, the research proposed an enhanced FDP model based on GWO-Adaboost optimized support vector machine. The results showed that the improved SVM model exhibited high prediction accuracy and stability in predicting FD. During the training process, the recall rate and F1 value of the model rapidly increased and eventually stabilized, reaching 0.89 and 0.85, respectively. This indicated that the improved SVM model could quickly capture key features of FD and gradually optimize performance through iteration. In

addition, compared with other advanced prediction models, the improved SVM model showed significant advantages in prediction accuracy, recall rate, and F1 value, reaching 92.67%, 93.52%, and 93.09%, respectively. The main reasons for this performance improvement were as follows: 1. Feature selection: In the study, the effectiveness of the selected financial indicators was verified through SHAP value analysis, especially key indicators such as asset net profit margin and earnings per share, which had a significant impact on the prediction results. Reasonable feature selection enabled the model to better capture the essential characteristics of FD, thereby improving prediction accuracy. 2. Model optimization: The improved SVM model included parameter adjustment and algorithm optimization, such as kernel function selection and hyperparameter tuning, which further enhanced the performance of the model. 3. Data processing: In the data preprocessing stage, the study standardized the data to eliminate noise and outliers, making the model training process more stable and efficient.

Although the improved SVM model performed well in terms of performance, its complexity and interpretability also need to be considered. The SVM model itself is a black box model with relatively weak interpretability. However, the study effectively improved the interpretability of the model through SHAP value analysis. SHAP values could quantify the contribution of each feature to the prediction results, making the decision-making process of the model more transparent and interpretable. In addition, visual analysis was conducted on the most important features in the study, further verifying the effectiveness of these features. This model design that combined explanatory methods not only improved the predictive performance of the model, but also enhanced its interpretability and credibility in practical applications. Compared to other models in relevant literature, the improved SVM model significantly improved prediction accuracy and stability. For example, the prediction accuracy of the random forest, gradient boosting decision tree, and artificial neural network models in the literature were 88.54%, 80.23%, and 78.91%, respectively, while the improved SVM model achieved 92.67%. This significant performance improvement was mainly attributed to the rationality of feature selection and the effectiveness of model optimization.

Although the improved SVM model showed a high accuracy of 92.67% in experimental results, it is more important to evaluate its value in practical applications. Compared with current FDP methods, the improved SVM model not only significantly improved prediction accuracy, but more importantly, it could detect potential financial risks earlier. For example, the improved SVM model could detect FD risk 3 to 6 months earlier than traditional methods. The ability to detect in advance is extremely important for business management and investors, as they can have more time to take preventive measures and avoid greater losses. In addition, early detection of FD risks can help companies make financial adjustments and optimize strategies in the early stages, thereby improving their long-term stability and competitiveness. Therefore, the improved SVM model not only had high prediction accuracy in theory, but also demonstrated high practical value in practical applications. In addition, in practical applications, the computational complexity and cost of the model were important considerations. Especially in the field of risk management, the efficiency of models was directly related to their popularization and application. Although the improved SVM model based on the GWO algorithm and the Adaboost algorithm integration mechanism proposed in the study performed well in prediction accuracy, its computational cost and training time were relatively high. Therefore, in practical applications, enterprises need to choose appropriate models for risk management based on their own computing resources and time constraints. In order to enhance the universality and persuasiveness of the research, future studies can consider the following points: 1. Cross market validation, verifying the predictive performance of the improved SVM model in stock markets of different countries or regions, to evaluate its applicability and effectiveness under different economic environments and market conditions. 2. Data diversity, introducing more diverse datasets, including companies from different markets, industries, and scales, to improve the robustness and generalization ability of the model. 3. Market characteristic analysis, studying the impact of different market characteristics, such as market volatility, regulatory environment, and investor behavior, further optimizing the model to meet the needs of different markets. Through these extensions and improvements, the practicality and value of the improved SVM model can be more comprehensively evaluated on a global scale, providing a wider range of financial risk management tools for businesses and investors.

In summary, the study achieved high prediction accuracy and stability in FDP by improving the SVM model, and improved the interpretability of the model through SHAP value analysis. These results not only validated the effectiveness of the improved SVM model, but also provided new ideas and methods for predicting FD. Future research can further explore other optimization algorithms and feature selection methods to further improve model performance.

6 Conclusion

With the continuous development of the economy, accurately predicting the financial difficulties of listed companies is greatly significant for risk management. To improve the accuracy of FDP, an improved SVM model based on the GWO algorithm and the Adaboost algorithm integration mechanism was proposed. The outcomes indicated that the improved SVM model exhibited good performance during the training process, with a rapid increase in recall and eventual convergence to 0.89. The F1 value also gradually increased and stabilized at 0.85, reflecting a good balance between accuracy and recall of the model. In prediction accuracy, the improved SVM

model achieved prediction accuracy, recall rate, and F1 value of 92.67%, 93.52%, and 93.09%, respectively, all significantly better than other compared models. Meanwhile, in empirical analysis, for normal and ST enterprises, the improved SVM model had prediction errors of 2.67% and 3.22% in the training and testing sets, while the prediction errors of other models were all above 10%. This result indicated that the improved SVM model had significant advantages in accuracy and reliability. Through SHAP value analysis, the key financial indicators that affect the prediction results of FD were further revealed. These indicators were mainly concentrated in the secondary indicators related to the profitability and solvency of the enterprise, especially the impact of asset net profit margin and earnings per share on the prediction results was particularly significant. In summary, the study has successfully improved the accuracy of FDP and risk management in listed companies. Through accurate FDP, companies can identify financial risks in advance, take necessary preventive measures, and achieve more efficient risk management. However, the study only focuses on A-share listed companies, which may result in its conclusions not having broad applicability. Therefore, future research can be extended to other types of enterprises and attempts can be made to establish FDP models applicable to different fields of enterprises, to further verify and optimize the universality of the models.

To further enhance the universality of the model, future research should consider expanding to other markets and types of enterprises, and explore in depth the impact of different market structures on the applicability of the model. In addition, in high-risk financial environments, false positives and false negatives may have significant impacts on corporate stakeholders. Therefore, it is necessary to include discussions on these ethical and practical impacts in research to ensure that models are more reliable and responsible in practical applications. When analyzing the results, it can be found that the regression coefficients of certain financial indicators are consistent with the expected economic effects. For example, the positive coefficient of the asset liability ratio indicated that the higher the asset liability ratio, the greater the probability of a company falling into financial difficulties, which was in line with economic theory. However, the regression coefficients of certain indicators had opposite signs to the expected effect. This was because there might be biases or outliers in the data sample that affected the direction of the regression coefficients. There may be complex interactions between certain financial indicators, and their significance when analyzed separately may differ from when considered comprehensively. Therefore, in future research, the model can be further refined to consider more variables and interactions, or other more complex models can be used to more accurately predict corporate FD.

References

[1] Zhou J, Zhu S, Qiu Y, Armaghani DJ, Zhou A, Yong W. Predicting tunnel squeezing using support vector machine optimized by whale optimization algorithm. Acta Geotechnica. 2022, 17(4):1343-1366. https://doi.org/10.1007/s11440-021-01239-1

[2] Zhu J, Li S, Song J. Magnitude estimation for earthquake early warning with multiple parameter inputs and a support vector machine. Seismological Society of America. 2022, 93(1):126-136. https://doi.org/10.1785/0220210144

[3] Youssef AM, Pradhan B, Dikshit A, Mahdi AM. Comparative study of convolutional neural network (CNN) and support vector machine (SVM) for flood susceptibility map**: a case study at Ras Gharib, Red Sea, Egypt. Geocarto International. 2022, 37(26):11088-11115. https://doi.org/10.1080/10106049.2022.2046866

[4] Tanveer M, Rajani T, Rastogi R, Shao YH, Ganaie MA. Comprehensive review on twin support vector machines. Annals of Operations Research. 2024, 339(3):1223-1268. https://doi.org/10.1007/s10479-023-05015-3

[5] Kim H, Cho H, Ryu D. Corporate bankruptcy prediction using machine learning methodologies with a focus on sequential data. Computational Economics. 2022, 59(3):1231-1249. https://doi.org/10.1007/s10614-021-10126-5

[6] Elhoseny M, Metawa N, Sztano G, El-Hasnony IM. Deep learning-based model for financial distress prediction. Annals of operations research. 2025, 345(2):885-907. https://doi.org/10.1007/s10479-023-05532-3

[7] Liu J, Li C, Ouyang P, Liu J, Wu C. Interpreting the prediction results of the tree-based gradient boosting models for financial distress prediction with an explainable machine learning approach. Journal of Forecasting. 2023, 42(5):1112-1137. https://doi.org/10.1002/for.3042

[8] Huang B, Yao X, Luo Y, Li J. Improving financial distress prediction using textual sentiment of annual reports. Annals of Operations Research. 2023, 330(1):457-484. https://doi.org/10.1007/s10479-022-04655-3

[9] Hajek P, Munk M. Speech emotion recognition and text sentiment analysis for financial distress prediction. Neural Computing and Applications. 2023, 35(29):21463-21477. https://doi.org/10.1007/s00521-023-08470-8

[10] Kurani A, Doshi P, Vakharia A, Shah M. A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting. Annals of Data Science. 2023, 10(1):183-208. https://doi.org/10.1007/s40745-022-00556-8

[11] Zhu B, Ye S, Wang P, Chevallier J, Wei YM. Forecasting carbon price using a multi-objective least

squares support vector machine with mixture kernels. Journal of Forecasting. 2022, 41(1):100-117. https://doi.org/10.1002/for.2796

[12] Yang L, Li J, Dong R, Zhang, Y., & Smyth, B. Numhtml: Numeric-oriented hierarchical transformer model for multi-task financial forecasting. Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(10): 11604-11612. https://doi.org/10.1609/aaai.v36i10.20833

[13] Wasserbacher H, Spindler M. Machine learning for financial forecasting, planning and analysis: Recent developments and pitfalls. Digital Finance. 2022, 4(1):63-88. https://doi.org/10.1007/s42485-022-00082-7

[14] Dash RK, Nguyen TN, Cengiz K, Sharma A. Fine-tuned support vector regression model for stock predictions. Neural Computing and Applications. 2023, 35(32):23295-23309. https://doi.org/10.1007/s00521-022-07437-5

[15] Zhang H, Zou Q, Ju Y, Song C, Chen D. Distance-based support vector machine to predict DNA N6-methyladenine modification. Current Bioinformatics. 2022, 17(5):473-482. https://doi.org/10.2174/1573402x17666220609162102

[16] Nguyen V G, Sharma P, Ağbulut Ü, Le, H. S., Cao, D. N., Dzida, M., ... & Tran, V. D. Improving the prediction of biochar production from various biomass sources through the implementation of eXplainable machine learning approaches. International Journal of Green Energy, 2024, 21(12): 2771-2798. https://doi.org/10.1080/15435075.2024.2312379

[17] Owolabi O S, Uche P C, Adeniken N T, Ihejirika, C., Islam, R. B., Chhetri, B. J. T., & Jung, B. Ethical implication of artificial intelligence (AI) adoption in financial decision making. Computer and Information Science, 2024, 17(1): 1-49. https://doi.org/10.5539/cis.v17n1p49

[18] Okeke N I, Bakare O A, Achumie G O. Forecasting financial stability in SMEs: A comprehensive analysis of strategic budgeting and revenue management. Open Access Research Journal of Multidisciplinary Studies, 2024, 8(1): 139-149. https://doi.org/10.53022/oarjms.2024.8.1.0055

[19] Ullah M, Shaikh M, Channar P, & Shaikh, S. Financial forecasting: an individual perspective. International Journal of Management (IJM), 2021, 12(3): 60-69. https://doi.org/10.34218/IJM.12.3.2021.005

[20] Bowden R, Haddadin S. Enhancing financial market risk forecasting through hybrid k-means and support vector machine models. Social Sciences Spectrum, 2024, 3(1): 189-201. https://sss.org.pk/index.php/sss/article/view/69?articlesBySimilarityPage=3