Early Warning of Financial Crises in Manufacturing Using SMOTE-Tomek Random Forest and Sentiment-Enhanced Indicators

Yingchun Zhan
School of Economics and Management, Anyang University, Anyang,455000, China
E-mail: zhanychun@outlook.com

Keywords: financial crisis early warning, big data, sentiment analysis, machine learning, random forests

Received: August 3, 2025

In order to improve the accuracy of financial crisis warning in the manufacturing industry and solve the problems of single indicators and insufficient ability to handle imbalanced data in traditional models, a warning system integrating traditional financial indicators and text big data indicators has been studied and constructed. The synthetic minority oversampling technique Tomek link random forest (SMOTE-Tomek-RF) model for early warning is adopted. Moreover, using 21 manufacturing enterprises listed on the Shanghai Stock Exchange A-shares as samples and based on 22 warning indicators, core variables are selected through random forest (RF) feature selection to compare the warning performance of RF, SMOTE-RF, single decision tree (DT), and the proposed SMOTE-Tomek-RF model. The results showed that the importance scores of emotional inclination and popularity were 0.052 and 0.047, respectively. Both scores were higher than the threshold and were ranked high, effectively supplementing the information. The predictive model proposed by the research had a subject area under the working curve (AUC) of 0.968, an F1 score of 84.97%, and a G-Mean of 90.11%. The AUC of the traditional RF model, SMOTE-RF model, and DT model were only 0.934, 0.953, and 0.943, respectively. In addition, the prediction accuracy for healthy and crisis firms after combining text big data amounted to 100% and 92.86%, respectively. In summary, the prediction model can effectively deal with the data imbalance problem and improve the precision of early warning. This method provides a reliable method for financial crisis early warning in manufacturing industry, which is of great significance for enterprise risk control and investor decision-making.

Povzetek: Predstavljen je model za finančne krize v proizvodnji, ki z združevanjem finančnih kazalnikov in besedilnih podatkov izboljša napovedno uspešnost.

1 Background

In the context of global economic integration and increasingly fierce market competition, manufacturing industry as the pillar industry of the national economy. Its financial stability is directly related to industrial chain security and sustainable economic development. However, the financial risks faced by manufacturing enterprises have been intensifying due to the impact of multiple factors, such as fluctuations in raw material prices, changes in market demand, and business management. The suddenness and destructiveness of financial crises have also increased significantly [1-3]. Timely and accurate financial crisis early warning (FCEW) is of great significance for enterprises to optimize resource allocation and investors to avoid risks. Traditional FCEW research mostly relies on a single financial indicator or statistical model, including Z-Score model, logistic regression, etc. [4]. However, there are obvious limitations in such methods. On the one hand, the indicator system focuses on enterprises' internal financial data and ignores external information, such as market sentiment, that can impact their financial situation. This leads to an imbalanced early warning (EW) perspective. On the other hand, the financial crisis (FC) samples of manufacturing enterprises often show the unbalanced characteristics of more healthy enterprises and fewer crisis enterprises. The traditional model is prone to the problem of favoring the majority class and neglecting the minority class when dealing with such data, and the EW accuracy is insufficient [5-6]. In addition, It can be challenging to capture complicated nonlinear interactions in financial data since certain models are prone to overfitting due to basic structures or predetermined assumptions about data distribution. In recent years, the development of machine learning (ML) technology provides a new path for FCEW.

As a typical representative of integrated learning, random forest (RF) has been widely used for prediction tasks in unbalanced data scenarios due to its strong resistance to overfitting and ability to handle high-dimensional data. Aghware F O et al. proposed to combine synthetic minority oversampling technique (SMOTE) with various classification algorithms to construct a detection model for the problem of frequent and expanding credit card fraud. The study conducted comparative experiments of different algorithms in Python environment. The outcomes revealed that the prediction accuracy (PA) of RF was improved to 0.9919 after applying SMOTE, which was significantly better than the other models, verifying its fraud identification

advantage in unbalanced data scenarios [7]. Ghinaya H et al. proposed an integrated method combining SMOTE, recursive feature elimination (RFE) and RF to address the problem of software defect prediction where accuracy is affected by data imbalance. According to the results, the approach successfully addressed the issue of category imbalance and obtained the greatest accuracy of 0.9998 on the MC1 dataset [8]. Hasanah U et al. addressed the problem of class imbalance in cardiovascular disease prediction by proposing a combination of SMOTE with multiple ML algorithms for evaluation. The study was based on the UCI Heart Disease Database and comparative experiments were carried out using RF, support vector machines with different resampling techniques. The results showed that the combination of SMOTE and RF performed the best [9]. Rafrastara F A et al. proposed the use of random undersampling combined with RF for classification modeling to address the problem of unbalanced data leading to high risk of false negatives in malware detection. The study was conducted by balancing the sample categories and introducing multiple methods for comparative analysis. The outcomes showed that the accuracy, recall and specificity of RF reached 98.1%, 98.0%, and 98.2%, respectively [10].

A large number of researches have been carried out in various fields for FCEW. To solve the problem of early identification of FC, Reimann C proposed the method of combining multiple ML models to improve the prediction ability, and constructed a framework for comparing logistic regression with multiple advanced models. The results revealed that RF and extreme random tree significantly outperformed traditional models in area under the curve (AUC) assessment [11]. Chen S et al. proposed a temporal convolutional network model based on convolutional neural networks to address the problem of poor PA of FC. According to the results, the method's PA was noticeably higher than that of the comparative model. Through Shapley value decomposition, the study revealed that stock prices and real GDP growth had key impacts in crisis prediction [12]. To address the lack of harmonization of FC prediction indicators in the ASEAN region, Powell R J et al. used a multiple discriminant analysis model to construct an EW system for financial distress. The study-built country-level and regional holistic models based on an industry sample of 720 firms from six ASEAN countries. The results indicated that the holistic model performed robustly over multiple crisis periods and was applicable to support the risk EW needs of an integrated ASEAN banking system [13]. Barthélémy S et al. found that traditional models were difficult to accurately predict currency crises. The outcomes revealed that the method identified 91% of the crises within a two-year warning window. Moreover, the false alarm rate was significantly lower than that of logistic regression, indicating that it had higher practical value in policy making [14]. The summary of research on FC warning is shown in Table 1.

Table 1: Summary of research on FCEW

Research	Model/Method	Dataset features	Key performance indicators	Pros and cons
Aghware F O et al. [7]	SMOTE+Random Forest (RF)	Credit card fraud data is highly imbalanced	Accuracy: 0.9919	The effectiveness of SMOTE-RF in handling imbalanced data has been verified, but it has not been applied in the field of financial warning.
Reimann C [11]	Logistic regression, RF, extreme random tree	Financial institution data	AUC exceeds 0.9	Machine learning models are significantly better than traditional statistical models in terms of AUC metrics.
Chen S et al. [12]	Temporal convolutional network (TCN)	Macroeconomic and Financial Market Data	Accuracy rate of over 90%	Deep learning can capture time series features; The key role of stock prices and GDP growth has been revealed, but the model is more complex.
Barthélémy S et al. [14]	LSTM, GRU	Currency Crisis Data	Recall rate: 91%	Recurrent neural networks have a high recognition rate and low false alarm rate within the warning window, but are mainly targeted at macro currency crises.

In summary, previous studies have made some progress in the field of FCEW. However, there are still obvious deficiencies in the existing research. First, there are fewer specialized studies on FCEW in the manufacturing industry. Second, when addressing data imbalance, oversampling alone can lead to sample

overlap, while undersampling alone can lead to information loss. Both of these issues affect the accuracy with which the model recognizes a small number of classes. A manufacturing FC warning model has been proposed to address such issues. This model combines traditional FC indicators with text big data indicators, and proposes a prediction model based on SMOTE Tomek balanced random forest (SMOTE-Tomek-RF). The study aims to address the shortcomings of existing research in terms of indicator dimensions and data processing methods by providing clear answers to the following research questions and testing the corresponding hypotheses. Research Question 1: Compared to warning models that rely solely on traditional financial indicators, will the predictive performance of the model be significantly improved by integrating text big data indicators such as investor sentiment? Research Question 2: Regarding the problem of data imbalance caused by the scarcity of FC samples, will the proposed SMOTE-Tomek-RF model more effectively identify minority classes than traditional RF or single sampling methods? Therefore, the study proposes the following hypothesis: Hypothesis 1: Text-based big data indicators can provide forward-looking information that goes beyond traditional financial indicators. Since investor sentiment is often timelier than the release of financial statements, EW models that integrate text big data indicators will outperform models that only use traditional financial indicators in performance evaluations. Hypothesis 2: After using the SMOTE-Tomek mixed sampling technique to optimize imbalanced data, the SMOTE-Tomek-RF model will have significantly higher PA for FC enterprises than the benchmark model without data balancing treatment. The innovativeness of the study is mainly reflected in two aspects. First, the traditional financial indicators of manufacturing industry are deeply integrated with investor sentiment text big data, which breaks through the limitations of traditional internal financial data. Second, the the dataset of RF model is preprocessed by SMOTE-Tomek hybrid sampling technique, which solves the defects of the single sampling method and enhances the ability to deal with unbalanced data.

2 Methods

The study mainly elaborates the construction method and specific implementation steps of EW model construction of FC in manufacturing industry. It includes the selection criteria and scope of EW samples, the construction of EW indicator system, the dimensionality reduction processing of traditional indicators, and the whole process of EW model construction based on SMOTE-Tomek-RF.

2.1 Indicator construction and data processing for FCEW in manufacturing industry

In the context of the era of Internet technology, the business environment in which enterprises are located is experiencing unprecedented and drastic changes. The globalization of the market connection allows enterprises to quickly reach a broader user base. However, this high-speed development is also accompanied by multiple uncertainties, and the challenges faced by enterprises are increasing day by day. Therefore, it is necessary to improve the level of FC management. The traditional EW model of FC suffers from the defects of single index, overfitting model, and insufficient warning accuracy. Therefore, it is necessary to optimize the indicator system and prediction method of traditional EW model. The study mainly focuses on optimizing the indicators and model construction of EW. The specific implementation process is shown in Figure 1. Finding the research EW samples, screening FC and healthy manufacturing companies, and defining the sample pairing ratio are the first steps in the FCEW optimization design process for the manufacturing sector. Then the EW indicators are determined, including traditional financial indicators and big data indicators. Then the traditional EW indicators are downgraded, and the core variables are screened by RF feature selection. Then the big data indicators are quantified, which are realized by text acquisition, cleaning, and sentiment analysis. Finally, the construction method of EW model is determined.

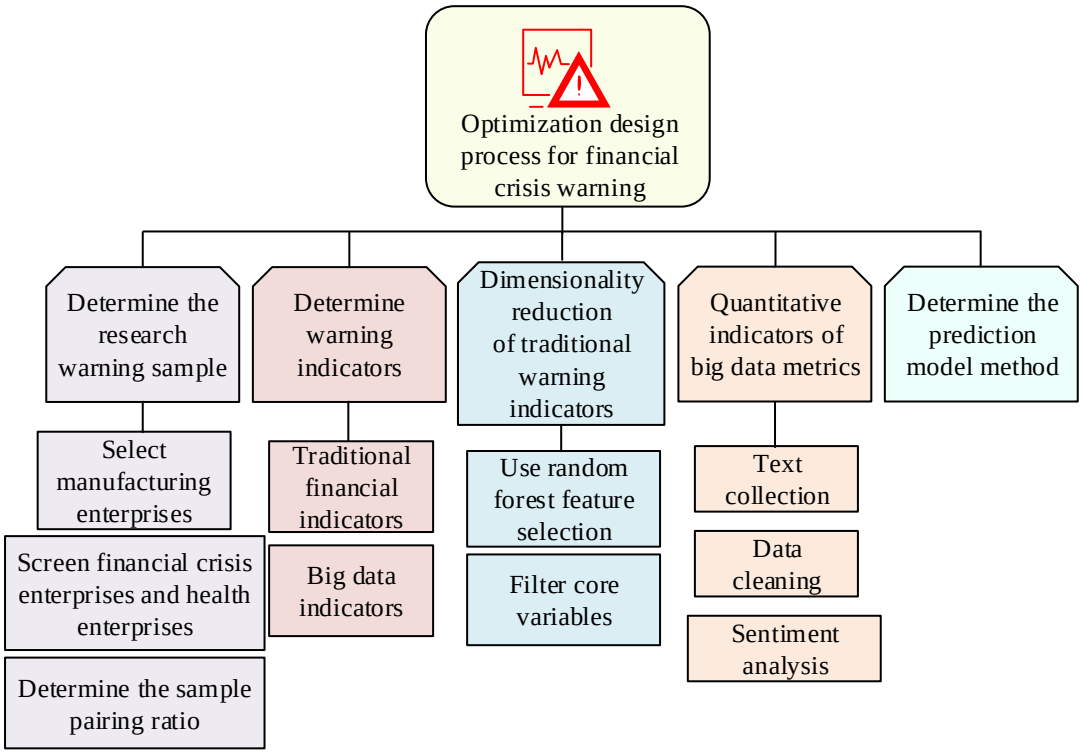


Figure 1: Optimization design process for FC warning

In the selection of EW samples, the study mainly focuses on the SEC A-share manufacturing enterprises. Due to two consecutive years of losses in 2021–2024, a total of seven businesses that have been listed for more than five years and have received special treatment are adopted by FC Enterprises. Healthy enterprises adopt enterprises with the same listing time and without special treatment, totaling 14. The year in which a listed company is subject to special treatment is defined as year T. Considering that there is a lag in the release of financial reports, both the annual report data and the special treatment in year T-1 occurs in year T. Using data from year T-1 can exaggerate the predictive ability of the model. Furthermore, firms are only treated exceptionally if they have lost money for two consecutive years. Therefore, the time range of traditional financial indicators is limited to T-3 years, so as to ensure that the EW is forward-looking. In the determination of EW indicators, text big data indicators and traditional financial indicators are mainly considered. All indicator data is sourced from publicly available data for the corresponding year. No cross-year data is reused to ensure there are no time leaks. The text-based big data indicator used to measure investor sentiment is sourced from multiple platforms to minimize bias from a single source. These include Eastmoney Stock Forum (https://guba.eastmoney.com/), Tonghuashun Finance (https://www.10jqka.com.cn/), and Xueqiu.com's Enterprise Column (https://xueqiu.com/). For each sample company, posts are captured from its dedicated section. The time horizon of the big data indicators covers T-3 years and T-2 years, avoiding sentiment polarization caused by staged events by integrating two years of data. Specifically, it includes two indicators, sentiment propensity value and heat value. The expression of emotional tendency value is shown in Equation (1).

$$A_1 = \frac{\sum_{i=1}^{n} P_i - \sum_{i=1}^{n} N_i}{M + N} \quad (1)$$

In Equation (1), A_1 represents the emotional tendency value, with a range of [-1,1]. The closer the value is to 1, the more positive the emotion. Moreover, the closer it is to -1, the more negative the emotion. $\sum_{i=1}^{n} P_i$ is the total emotional intensity of all positive comments. $\sum_{i=1}^{n} N_i$ is the total emotional intensity of all negative comments. M is the number of positive comments. N is the number of negative comments. The emotional heat value is expressed in Equation (2).

$$A_2 = \frac{\sum_{j=1}^{m} (L_j + C_j)}{T} \quad (2)$$

In Equation (2), A_2 is the emotional heat value, with a range of [0,+∞). The larger the value, the higher the heat of the text discussion. L_j is the number of likes for the j-th post. C_j is the number of comments on post j. m is the total number of valid posts, and T is the statistical time span. Traditional financial indicators are mainly derived from annual reports in the CSMAR database. Combined with previous research

results, the study constructs the indicator system from five core ability dimensions, including profitability, solvency, operating ability, growth ability, and cash flow analysis. A total of 22 financial EW indicators are constructed by combining traditional indicators and big data indicators, as shown in Figure 2.

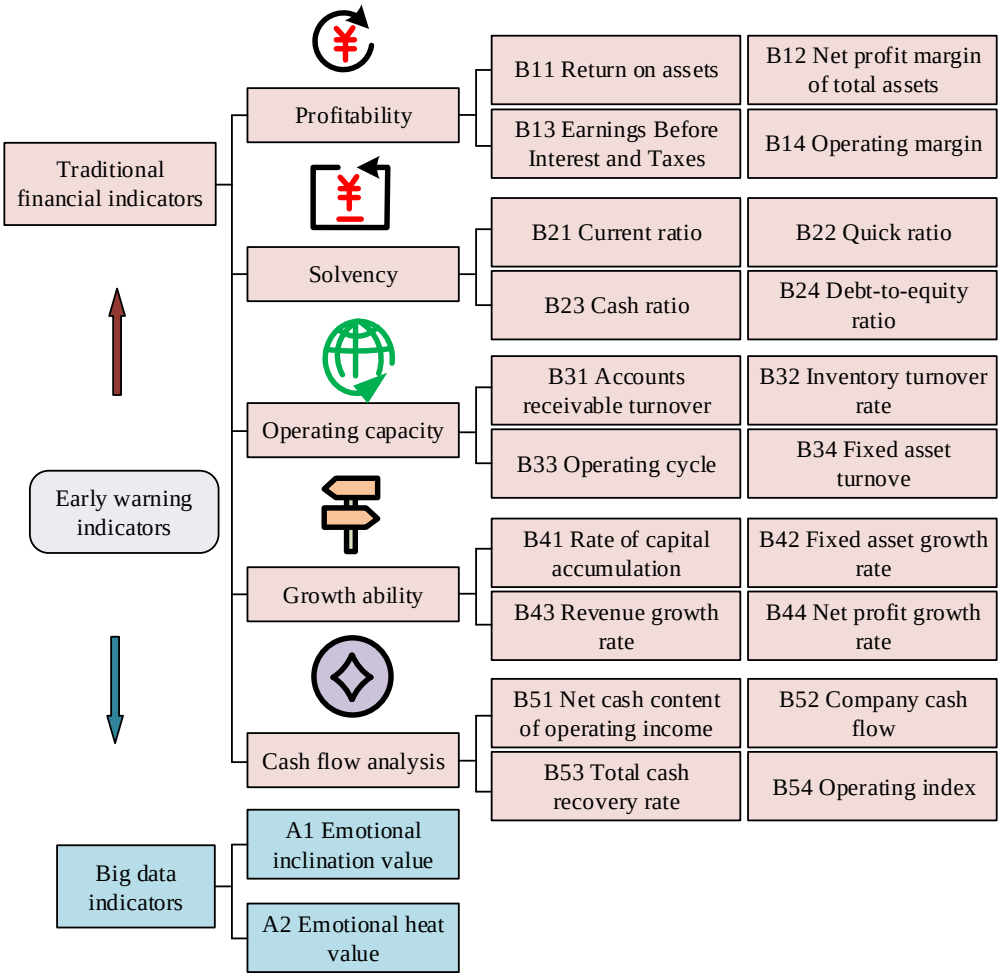


Figure 2: FCEW indicator system

The Pearson correlation coefficients of each indicator are all below 0.7, and the variance inflation factor (VIF) is below 3.2. This indicates that there is no strong correlation or collinearity risk. Therefore, there is no multicollinearity among the indicators. Considering that high-dimensional indicators in financial EW indicators may lead to redundancy and overfitting in EW models, for this reason, the study realizes indicator dimensionality reduction through RF feature selection. In RF model, Gini importance score is a common and efficient indicator to measure the importance of variables, and its calculation is derived from the Gini index of decision tree (DT). Taking a simple binary classification problem as an example, if the proportion of positive samples on the DT splitting node k is p, and the proportion of negative samples is $1-p$. The formula for calculating the Gini index of this node is shown in Equation (3) [15].

$$G_k = 2p(1-p) \quad (3)$$

In Equation (3), G_k denotes the Gini index of node k. The quantization process of text big data indicators is mainly divided into three parts: text collection, data cleaning and sentiment analysis. The quantization process is detailed in Figure 3.

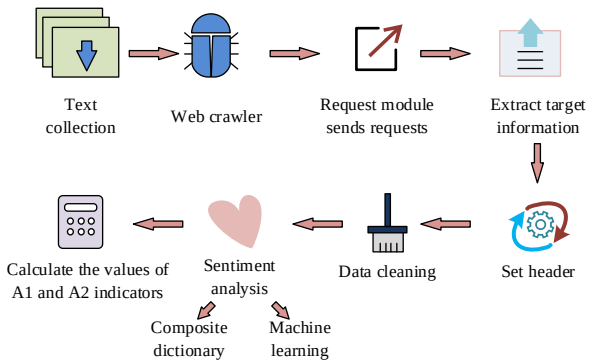


Figure 3: Quantitative process of text big data

In Figure 3, during the text collection stage, web crawler technology is used to obtain text data. The crawling fields include post title, body content, posting time, likes, comments, and replies. The study investigates

how to configure request headers to simulate browser login operations, incorporating a 5-second request delay mechanism and a pool of 30 dynamic IP proxies to avoid triggering the platform's anti-crawling mechanisms. When crawling public network data, researchers strictly adhere to website protocols, protect user privacy through anonymization, and ensure that all data is used solely for academic analyses that comply with research ethics. In text data processing, the data cleaning stage improves the purity of the text by using regular expressions to remove invalid information and redundant content, such as hypertext markup language (HTML) tags and single punctuation marks. Simultaneously, jieba segmentation is performed, Chinese stop words and meaningless words in the financial field are removed, and the Word2Vec to train word vectors are used to convert the text into numerical features. The sentiment analysis session integrates composite dictionaries and ML methods. On the one hand, it integrates basic sentiment dictionaries, online language dictionaries, and financial professional dictionaries to build a multi-dimensional sentiment vocabulary system. On the other hand, based on this system, each text is scored for sentiment, and A1 and A2 index values are finally calculated. Specifically, the study integrates the "CNKI Emotional Dictionary", "Professional Emotional Dictionary in the Financial Field", and Internet Finance Popular Language Dictionary, and annotates the emotional polarity and intensity of each vocabulary. In emotional polarity, positive=1, negative=-1, neutral=0. The range of emotional intensity is 1-5 points, with higher intensity indicating stronger emotions. Subsequently, a logistic regression classifier is trained using 1,000 manually annotated stock bar texts based on the dictionary scoring results. The fuzzy text is then subjected to secondary correction to address the limitations of the dictionary scoring method. Finally, a 5-fold cross validation is used to evaluate the classification performance.

2.2 Construction of EW model of FC based on oversampling RF algorithm

The study has completed the first four steps in the optimization of EW model construction of FC in manufacturing industry. Then the last step, which is the construction of the EW model, needs to be completed. In the field of FCEW, sample imbalance is a common and difficult problem. Specifically, the sample of enterprises in FC is much smaller than the sample of enterprises in normal operation. ML can improve the PA by deeply analyzing the complex structure and nonlinear patterns hidden in the financial data. RF, as a typical representative of Bagging idea in integrated learning, not only has the generalized advantages of ML, but also has efficient learning speed [16-18]. Therefore, this algorithm is adopted to construct an EW model of FC.Currently, RFs are mainly classified into weighted and balanced RFs. The former optimizes the model by setting different misclassification costs for different categories. It usually assigns higher costs to a few categories of samples, which enhances the model's attention to those categories. The core idea of the latter is to adjust the original unbalanced data with the help of oversampling and undersampling, so as to realize the balanced processing of data. Since the misclassification cost of weighted RF is difficult to be set precisely in advance, the study mainly adopts balanced RF to deal with the unbalanced classification problem. The algorithm flow is shown in Figure 4.

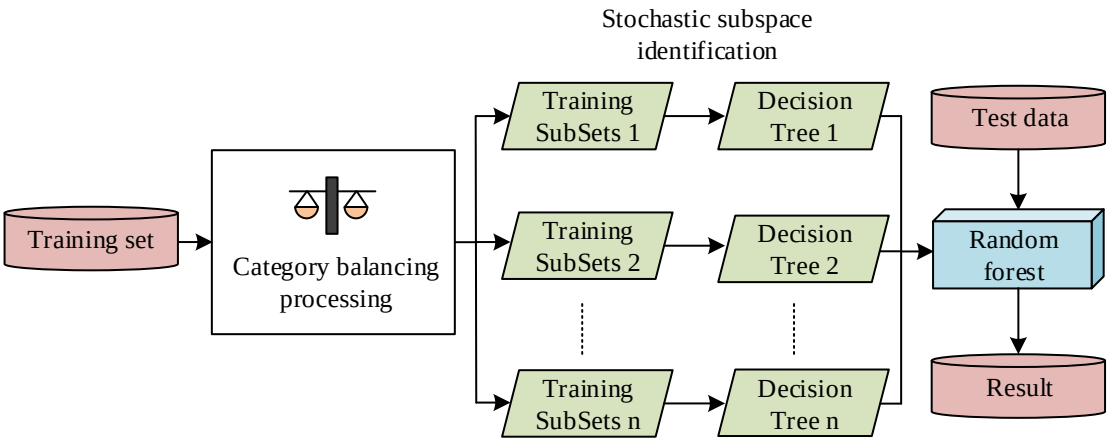


Figure 4: The process of balancing RF algorithm

In Figure 4, the training set is first processed with category balancing to solve the data imbalance problem. After that, Bootstrap sampling is carried out to generate multiple training subsets. Each subset constructs a DT on the basis of random subspace to form an RF. Finally, the test data is input and the prediction results are output from the RF. This reflects the idea of balanced RF, which borrows the category balanced processing and resampling, so that the data distribution better fits the conditions of traditional classification algorithms. However, in the balanced RF algorithm, the traditional resampling technique will lead to excessive replication of the minority class samples (CSs) and trigger model overfitting if simple oversampling is used. If undersampling is used, it will discard the valid information in the majority CSs, resulting in data waste.

Therefore, the study further introduces the SMOTE algorithm. This algorithm is specialized for solving the unbalanced data problem, and the core logic is to synthesize new samples for the minority CSs. Figure 5 displays the algorithm's schematic diagram.

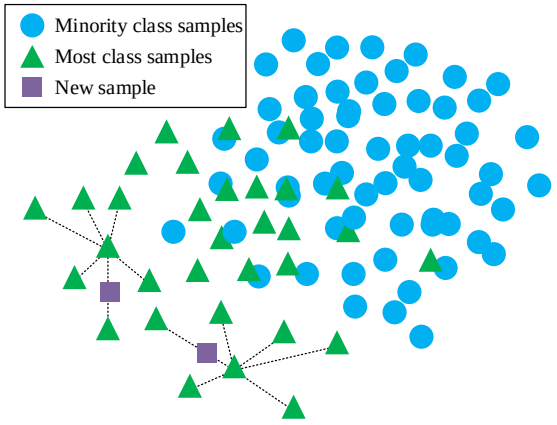


Figure 5: SMOTE algorithm schematic diagram

In Figure 5, the blue dots represent the majority CSs, the green triangles represent the minority CSs, and the purple squares are the synthesized new samples. For example, multiple nearest neighbors are first found for the green triangle in the minority CS. Then, the new synthetic sample of the purple square is generated by interpolating the feature space that connects this sample with its nearest neighbors. Without merely reproducing the minority CSs, this method can reduce data imbalance and increase the number of minorities CSs. It is considered that the new samples generated by SMOTE may overlap with most of the CSs, which leads to fuzzy classification boundaries [19-20]. To solve this problem, the study proposes the SMOTE-Tomek method. SMOTE-Tomek combines oversampling and undersampling techniques. First, SMOTE generates new samples. Then, Tomek links are used to remove boundary noise between the majority and minority classes, further optimizing the sample distribution. Based on this, the study selects SMOTE-Tomek as the resampling technique and constructs the SMOTE-Tomek-RF model. Its realization flow is shown in Figure 6.

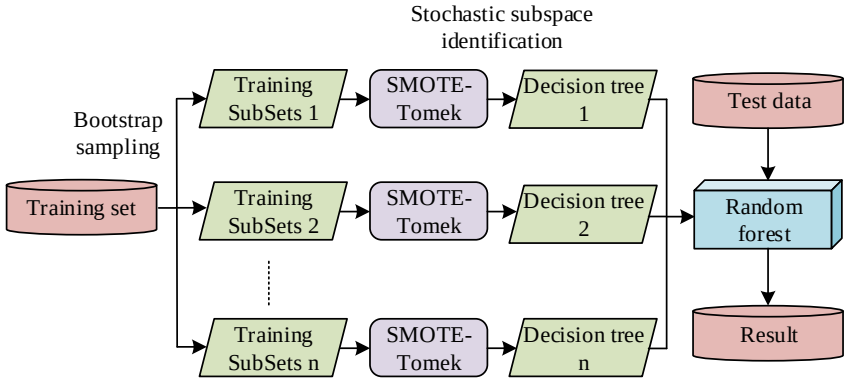


Figure 6: SMOTE-Tomek-RF implementation process

In Figure 6, the core idea of the SMOTE-Tomek-RF algorithm is that the SMOTE-Tomek resampling technique is embedded in the framework of RF to effectively deal with the category imbalance problem. For specific implementation, multiple subsets are first generated from the original training set by Bootstrap sampling. Then, the SMOTE-Tomek hybrid sampling method is applied to each subset independently. First, SMOTE oversampling synthesizes a few CSs, and then Tomek links remove the inter-class boundary noise to achieve data balancing and boundary optimization before training each DT. Finally, on these balanced and optimized subsets, multiple DTs are trained separately in conjunction with stochastic subspace methods. Meanwhile, they are integrated to form an RF model with stronger robustness to unbalanced data, which finally realizes accurate prediction of test data. All hyperparameters of the model are optimized and determined on the training set through grid search combined with 5-fold cross validation. The specific parameters are shown in Table 2.

Table 2: Hyperparameter settings

Algorithm module	Hyperparameter Name	Value/Setting
SMOTE	k Neighborhood Number	5
Tomek Links	distance metric	Euclidean distance
RF	Number of decision trees	200
	Maximum tree depth	8
	Minimum sample size for node splitting	5
	Minimum sample size of leaf nodes	2

The study has set the ratio of the number of enterprises with FC to healthy enterprises as 1:2 in sample selection, which is unbalanced. Therefore, an oversampling method is used to control the ratio of the two at 1:1, thus increasing the total number of samples to 28. To prevent data leakage and guarantee the model's forward-looking warning capability, the study employs a forward validation strategy, dividing the training and testing sets. The training set data is limited to T-5 to T-4 years, and the testing set data is limited to T-3 to T-2 years, with the year of special treatment for the enterprise as the time node. During the verification process, it is strictly ensured that the training set precedes the testing set and that there is no overlap in any time dimension.

3 Results and analyses

The study mainly analyzes and discusses the experimental results related to EW model construction of FC in manufacturing industry, including two core parts. One is to analyze the importance of traditional financial indicators and the introduction of text-based big data indicators through feature selection validation in order to assess the practical value of the selected EW indicators. AUC, F1-score, G-Mean metrics, and statistical tests are used to assess the prediction impacts of various models and confirm the efficacy of the SMOTE-Tomek-RF model.

3.1 Importance analysis of FCEW indicators

The research subjects of the experiment are obtained from the CSMAR database. The FC firms employ seven firms, all of which have been listed for more than five years and have received special treatment for two years in a row of losses in 2021–2024. Healthy enterprises adopt enterprises with the same listing time and without special treatment, totaling 14 enterprises. Table 3 displays the particular sample companies.

Table 3: Sample company name

Enterprises with special treatment	Processing year	Health enterprise	Health enterprise
Lotus flower	2021	Hongdu Aviation Industry Group	Dongmu Corporation
Haihua	2022	Yangnong Chemical Industry	Jinxi Axle
source	2021	Daheng Technology	BRIGHT DAIRY
Coconut Island	2021	Guizhou Airlines Co., Ltd	Aerospace Information
Fugang	2021	Angel Yeast	Huafang Co., Ltd
Xinke	2022	Lingrui Pharmaceutical	Baoji Titanium Industry
Xiangdian	2022	Zhuolang Intelligent	Hualu Hengsheng

To validate the value of FCEW indicators used in the study, feature selection validation is performed. Considering that high-dimensional indicators in financial warning indicators may lead to redundancy and overfitting of warning models, traditional fixed threshold selection is subjective and arbitrary. Therefore, the study uses the Gini importance score of the RF model as the initial measurement standard and excludes indicators with extremely low contributions and importance scores below 0.01. Then, the robustness of the features is tested by arranging their importance, and the feature values of each indicator are randomly shuffled. The decrease in AUC of the shuffled model is calculated, and only the indicators with AUC decrease ≥0.02 are retained. The SHapley additive exPlans (SHAP) value is used to quantify the marginal contribution of each indicator to the prediction results, and indicators with SHAP mean ≥0.03 are selected. Finally, using RFE as the base classifier, RFE is used to gradually eliminate the indicators with the lowest importance. The verification results obtained by combining the optimized feature selection process are shown in Table 4. The importance scores in Table 4 show that B34 fixed asset turnover tops the list with an importance score of 0.125. This reflects the key impact of asset operation efficiency on FCEW under the asset-heavy attribute of manufacturing industry. B32 inventory turnover, B13 earnings before interest and taxes, and other indicators also score high, reflecting the importance of operating capacity and profitability indicators in FCEW. Although B23 cash ratio has relatively low importance, it still has some EW value.

Table 4: Verification results of feature selection for traditional indicators

Serial number	Symbol	Significance score	Serial number	Symbol	Significance score
1	B34	0.125	11	B14	0.065
2	B32	0.110	12	B22	0.062

3	B13	0.102	13	B24	0.058
4	B21	0.098	14	B53	0.055
5	B31	0.095	15	B12	0.048
6	B41	0.092	16	B42	0.045
7	B51	0.088	17	B44	0.042
8	B11	0.085	18	B52	0.038
9	B43	0.082	19	B54	0.035
10	B33	0.079	20	B23	0.032

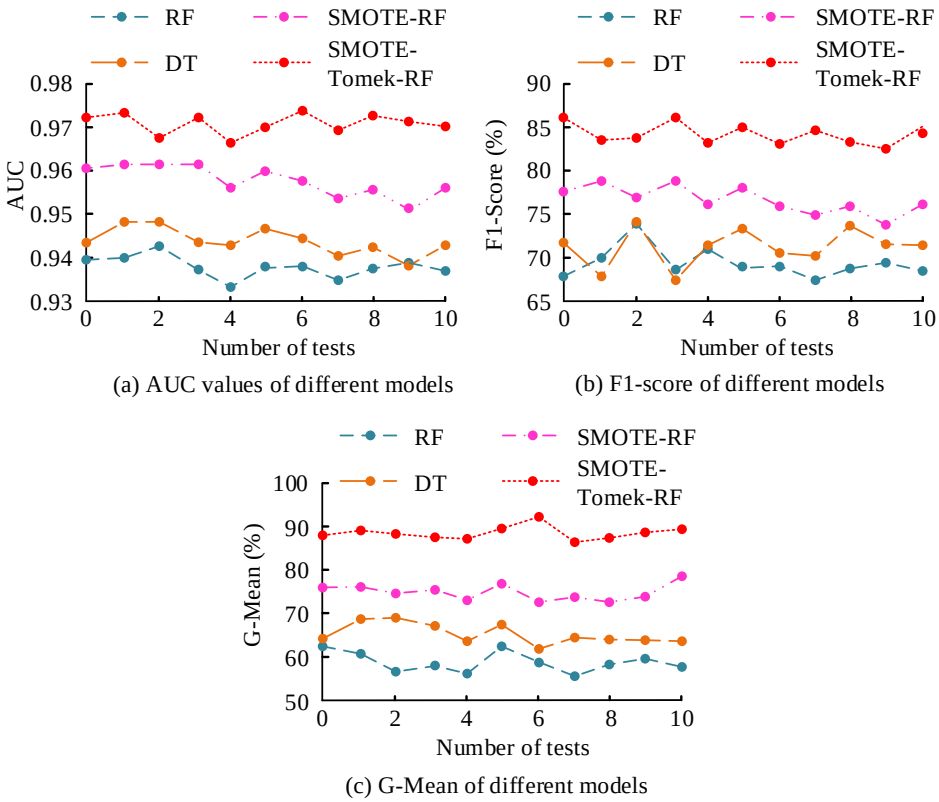
The study further introduces two text big data metrics that work together for feature selection. The threshold is set to 1/22. The feature selection importance scores are shown in Table 5. After the introduction of the text big data metrics, B34 fixed asset turnover (0.082), B32 inventory turnover (0.075), and B13 earnings before interest and taxes (0.071) still occupy the top three positions. A1 affective tendency value has an importance score of 0.052 and is ranked 9th. A2 affective hotness value has an importance score of 0.047 and is ranked 13th. The scores of A1 and A2 are higher than the thresholds, indicating that the investor sentiment indicator can provide effective information supplement for FCEW. Its importance is comparable to some traditional indicators, which verifies the practical value of text-based big data indicators in FCEW in manufacturing industry. In addition, the top five warning indicators primarily focus on operational efficiency, which reflects the manufacturing industry's characteristics of heavy assets and high inventory. The deterioration of its asset operational efficiency and core profitability is the direct trigger of the crisis. Second, the data clearly shows that crisis enterprises perform significantly worse than healthy enterprises in these key areas, providing quantitative references for risk monitoring. Unlike quarterly financial data, which lags, investor sentiment can capture negative market expectations about a company's prospects in real time. This provides a timelier warning signal before financial crises surface. Therefore, the importance score of A1 emotional tendency value is also relatively high.

Table 5: Importance score of feature selection after introducing text big data indicators

Serial number	Symbol	Significance score	Serial number	Symbol	Significance score
1	B34	0.082	12	B14	0.048
2	B32	0.075	13	A2	0.047
3	B13	0.071	14	B22	0.046
4	B21	0.068	15	B24	0.044
5	B31	0.065	16	B53	0.042
6	B41	0.062	17	B12	0.037
7	B51	0.059	18	B42	0.035
8	B11	0.056	19	B44	0.033
9	A1	0.052	20	B52	0.031
10	B43	0.053	21	B54	0.029
11	B33	0.051	22	B23	0.027

3.2 Evaluation of the predictive effectiveness of different models

To verify the validity of the SMOTE-Tomek-RF model, the study uses the traditional RF model, the SMOTE-RF model, and the single DT model to compare with it. AUC, F1-score, and G-Mean are used as evaluation metrics. Each model undergoes ten tests in total, with the average value serving as the ultimate outcome. In Figure 7(a), the AUC value of the SMOTE-Tomek-RF model is as high as 0.968, which is significantly higher than 0.934 for RF, 0.953 for SMOTE-RF, and 0.943 for DT. In Figure 7(b), the F1-score of the SMOTE-Tomek-RF model is as high as 84.97%, which is improved by 8.49% compared to SMOTE-RF. The G-Mean of the SMOTE-Tomek-RF model in Figure 7(c) is as high as 90.11%, which is substantially higher than that of the other comparison models. In summary, the SMOTE-Tomek-RF model has a better prediction effect in FCEW in manufacturing industry, and it can identify FC enterprises more accurately.



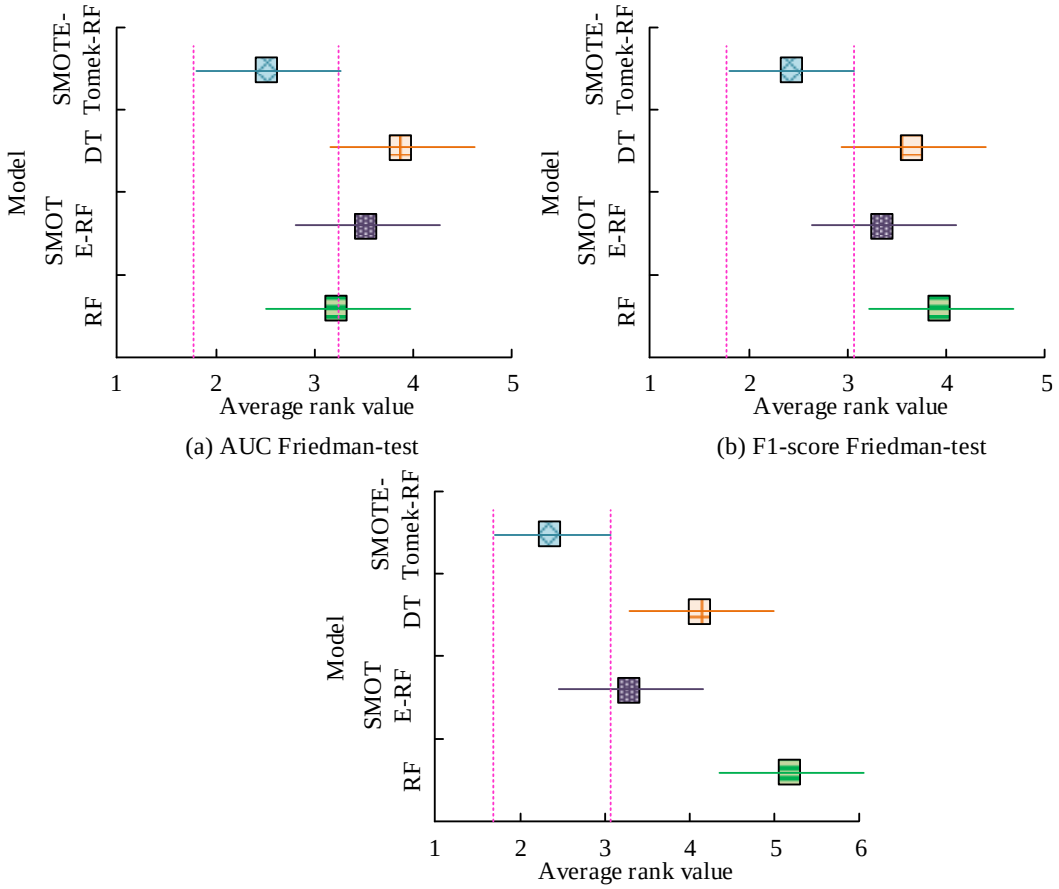
(a) AUC values of different models

(b) F1-score of different models

(c) G-Mean of different models

Figure 7: Accuracy, F1-score, and G-Mean of different models

The study continues with the Friedman test for each model, which is suitable for assessing differences between multiple related sample groups. The significance level is set at 0.05. If the test results show a *p*-value less than 0.05, the original hypothesis that there is "no significant difference in the performance of all models" is rejected. This indicates that at least some models substantially differ in terms of warning effectiveness. Conversely, if the *p*-value is less than 0.05, the original hypothesis can be rejected, suggesting that some models have substantial differences in EW effectiveness. Figure 8 displays the specific test findings. The crucial value domain of the difference in mean ordinal values is shown in the figure by the red dashed line. The SMOTE-Tomek-RF model's mean ordinal value is only 2.48 in Figure 8(a) for the AUC indicator at the significance level of 0.05, which is less than the values of the other models. Meanwhile, the differences in the mean ordinal values of the remaining comparison models do not exceed the critical value domain, representing insignificant differences. In Figure 8(b), for the F1-score metric, the mean ordinal value of the SMOTE-Tomek-RF model is only 2.24, which is significantly lower than the RF model. Meanwhile, the mean ordinal value of this model is still lower than that of the DT and SMOTE-RF models, and the difference is not significant. In Figure 8(c), for the G-Mean metric, the average ordinal value of the SMOTE-Tomek-RF model is only 2.19, which is significantly better than the DT and RF model. Moreover, it has a better prediction performance.



(a) AUC Friedman-test

(b) F1-score Friedman-test

(c) G-mean Friedman-test

Figure 8: Friedman test chart

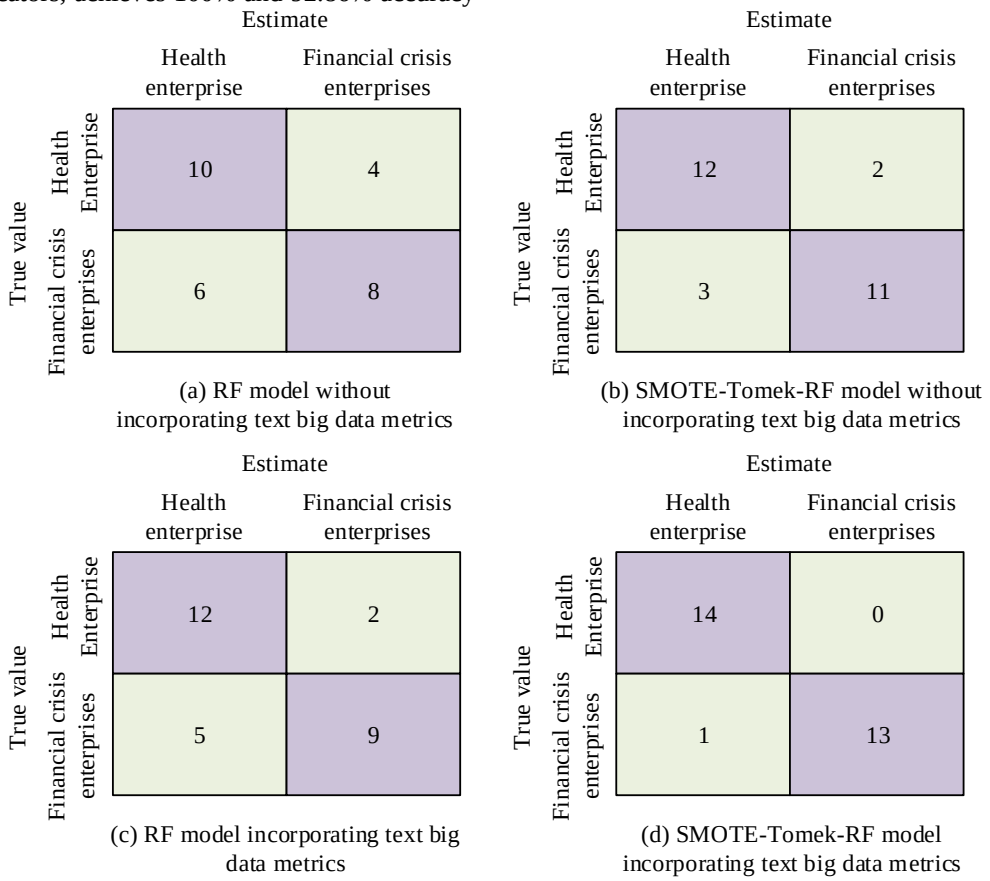
Considering the small sample size, in order to further verify the robustness of the model, a leave one cross validation is adopted to re evaluate the performance of each model. One enterprise should be left as the test set, and the remaining 20 should be used as the training set. This process should be repeated 21 times. The final results are shown in Table 6. The SMOTE-Tomek-RF model performs the best in LOOCV evaluation, with AUC, F1 score, and G-Mean improved by 4.2%, 8.66%, and 8.48%, respectively, compared to the traditional RF model. Compared with the SMOTE-RF model, there is an improvement of 1.9%, 3.89%, and 4.81%. This verifies the model's adaptability and effectiveness in providing warnings for imbalanced data under small sample constraints. A comparison of the results of 10 random tests shows that the core indicators of the LOOCV model have slightly decreased. The PA for healthy and crisis enterprises decreased by 4.76% and 7.15%, respectively. This reflects the fact that the model still exhibits overfitting tendencies in small sample scenarios and that its generalization ability is limited by the sample size. Further optimization is needed by expanding the sample size in the future.

Table 6: LOOCV test results

Model Type	AUC	F1 score (%)	G-Mean （%）	Health enterprise prediction accuracy (%)	Crisis enterprise prediction accuracy (%)
SMOTE-Tomek-RF	0.945	81.78	87.93	95.24	85.71
SMOTE-RF	0.926	77.89	83.12	90.48	78.57
RF	0.903	73.12	79.45	85.71	71.43

The study further validates the effect of incorporating text big data metrics on the predictive performance of the model, which is tested using the RF model without incorporating text big data metrics, the SMOTE-Tomek-RF model without incorporating text big data metrics, the RF model with incorporating text big data metrics, and the SMOTE-Tomek-RF model with incorporating text big data metrics. The prediction chaos matrix of each model for healthy enterprises and FC enterprises is shown in Figure 9. In Figure 9(a), the RF model without combining text big data indicators predicts healthy enterprises with an accuracy of 71.43%, and predicts FC enterprises with an accuracy of 57.14%, with a total PA of only 64.29%. In Figure 9(b), the

SMOTE-Tomek-RF model without combining text big data metrics predicts the accuracy of 85.71% and 78.57% for healthy and FC firms, respectively, and the total PA is 82.14%. It indicates that SMOTE-Tomek-RF is significantly better than RF model. The overall PA of the RF model with text big data indicators is 75% in Figure 9(c), which is greater than the PA of the RF model without text big data indicators. This confirms that text big data indications are logical. As shown in Figure 9(d), the SMOTE-Tomek-RF model, which incorporates text big data indicators, achieves 100% and 92.86% accuracy for healthy and financially crisis-stricken companies, respectively. This demonstrates the model's improved predictive performance and is noticeably better than alternative approaches. Considering that a small sample size can lead to randomness in the results, the study calculated its 95% confidence interval. The confidence intervals for predicting the accuracy of healthy enterprises are [98.2%, 100%] and for crisis enterprises are [84.0%, 99.8%]. This indicates a cautious attitude toward generalization ability.



(a) RF model without incorporating text big data metrics

(b) SMOTE-Tomek-RF model without incorporating text big data metrics

(c) RF model incorporating text big data metrics

(d) SMOTE-Tomek-RF model incorporating text big data metrics

Figure 9: Chaos matrix for predicting healthy enterprises and FC enterprises by various models

4 Discussion

The SMOTE-Tomek-RF model proposed in the study outperformed other models in both AUC and F1 scores. Compared with studies that also use ML methods, the performance of this model was at a comparable or even better level. There were two main reasons for this: firstly, the model innovatively integrated emotional information as a leading indicator. Unlike most models that relied solely on lagged financial indicators, the study captured negative market expectations that had not yet been reflected in financial statements, providing key incremental information. Second, a more refined data balancing strategy was adopted. Compared to studies that use single SMOTE oversampling, the SMOTE-Tomek-RF model removed noisy samples at class boundaries through the Tomek Links mechanism, making the recognition decision boundaries for minority classes clearer and improving the accuracy of EW.

However, there are certain limitations to the research. Firstly, the main limitation of the study is its small sample size. Due to the rarity of FC enterprises, the study only selects 7 eligible crisis enterprises and 14 matched healthy enterprises. Although SMOTE-Tomek technique is used to balance the data, a small base sample size may limit the model's generalization ability. Second, there is a potential risk of overfitting. Meanwhile, the construction dimensions of text big data indicators are relatively single. The study only constructed text indicators from two macro dimensions: overall emotional tendency and emotional heat, without delving into the deep structure and specific meaning of text information. Subsequent research will collect samples from more industries over longer time periods to enhance the statistical robustness of research conclusions. At the same time, techniques such as theme modeling, aspect-based sentiment analysis, and temporal dynamic

analysis can be used to create a more multidimensional text indicator system. This will improve the accuracy and depth of the warning model's predictions. In terms of practical deployment, the model must establish a robust process that incorporates automatic data collection, text sentiment quantification, feature engineering, and regular model retraining. At the same time, attention should be paid to hyperparameter optimization to ensure predictive stability in different market environments.

5 Summary and future work

To realize effective manufacturing FCEW, the study constructed an EW system integrating traditional financial indicators and text big data indicators, and used the SMOTE-Tomek-RF model for manufacturing FCEW. The results indicated that in terms of the importance of indicators, the Gini importance scores of fixed asset turnover, inventory turnover, and earnings before interest and taxes among traditional financial indicators were at the top of the list, which were 0.125, 0.110, and 0.102, respectively. This value reflected the key role of operating capacity and profitability in the manufacturing industry under the attribute of heavy assets on the FCEW. Under the category of heavy assets of the manufacturing industry, the value demonstrated the critical role that operational capacity and profitability play in FCEW. After the introduction of text-based big data indicators, the scores of emotional tendency value and emotional hotness value were 0.052 and 0.047, respectively, which were higher than the thresholds. It indicated that the investor sentiment indicators could provide an effective complement to the EW, and their importance was comparable to some traditional indicators. In terms of model performance, the AUC value of the SMOTE-Tomek-RF model reached 0.968, which was significantly higher than the 0.934 of traditional RF and the 0.953 of SMOTE-RF. Meanwhile, the F1-score of this model was as high as 84.97%, which was an improvement of 8.49% over SMOTE-RF. In addition, the G-Mean of this model was as high as 90.11%, which was significantly better than the rest of the compared models. Text big data indicators, represented by investor sentiment, can serve as key leading indicators. They provide significant, incremental information for traditional financial models, effectively improving the accuracy of EW. At the same time, the SMOTE-Tomek-RF model proposed in the study effectively addresses the issue of imbalanced data in FC samples using refined mixed sampling techniques. Its ability to identify minority classes is notably superior to that of the benchmark model. For business managers, research provides a more sensitive tool for self-inspection of risk, warning them that monitoring public opinion on the network must be incorporated into their daily risk management system. For investors and creditors, this model provides a quantitative auxiliary decision-making basis, which helps to identify and avoid potential investment risks earlier. This framework can also serve as a supplementary tool for regulatory agencies to use in their macroprudential supervision. It can be used to dynamically screen and monitor high-risk enterprises within specific industries.

References

[1] Sahiner M. Volatility spillovers and contagion during major crises: an early warning approach based on a deep learning model. Computational Economics, 2024, 63(6): 2435-2499. DOI: 10.1007/s10614-023-10412-4.

[2] Cao Y. Financial risk early warning method for modern manufacturing enterprises based on RBF neural network. International Journal of Manufacturing Technology and Management, 2025, 39(3-5): 183-194. DOI: 10.1504/IJMTM.2025.145936.

[3] Li H. Research on financial risk early warning system model based on second-order blockchain differential equation. Intelligent Decision Technologies, 2024, 18(1): 327-342. DOI: 10.3233/IDT-230318.

[4] Arifin M A S, Stiawan D, Yudho Suprapto B, Susanto, S., Salim, T., Idris, M. Y., & Budiarto, R. Oversampling and undersampling for intrusion detection system in the supervisory control and data acquisition IEC 60870-5-104. IET Cyber-Physical Systems: Theory & Applications, 2024, 9(3): 282-292. DOI: 10.1049/cps2.12085.

[5] Elhoseny M, Metawa N, Sztano G, El-Hasnony, I. M. Deep learning-based model for financial distress prediction. Annals of operations research, 2025, 345(2): 885-907. DOI: 10.1007/s10479-022-04766-5.

[6] Yonghong L, Jie S, Ge Z, Ru, Z. The impact of enterprise digital transformation on financial performance — Evidence from Mainland China manufacturing firms. Managerial and decision economics, 2023, 44(4): 2110-2124. DOI: 10.1002/mde.3805.

[7] Aghware F O, Ojugo A A, Adigwe W, Odiakaose, C. C., Ojei, E. O., Ashioba, N. C., ... & Geteloma, V. O. Enhancing the random forest model via synthetic minority oversampling technique for credit-card fraud detection. Journal of Computing Theories and Applications, 2024, 1(4): 407-420. DOI: 10.62411/jcta.10323.

[8] Ghinaya H, Herteno R, Faisal M R, Farmadi, A., & Indriani, F. Analysis of Important Features in Software Defect Prediction Using Synthetic Minority Oversampling Techniques (SMOTE), Recursive Feature Elimination (RFE) and Random Forest. Journal of Electronics, Electromedical Engineering, and Medical Informatics, 2024, 6(3): 276-288. DOI: 10.35882/jeeemi.v6i3.453.

[9] Hasanah U, Soleh A M, Sadik K. Effect of Random Under sampling, Oversampling, and SMOTE on the Performance of Cardiovascular Disease Prediction Models. Jurnal Matematika, Statistika dan Komputasi, 2024, 21(1): 88-102. DOI: 10.20956/j.v21i1.35552.

[10] Rafrastara F A, Supriyanto C, Paramita C, Astuti, Y.

P., & Ahmed, F. Performance Improvement of Random Forest Algorithm for Malware Detection on Imbalanced Dataset using Random Under-Sampling Method. Jurnal Informatika: Jurnal Pengembangan IT, 2023, 8(2): 113-118. DOI: 10.30591/jpit.v8i2.5207.

[11] Reimann C. Predicting financial crises: an evaluation of machine learning algorithms and model explainability for early warning systems. Review of Evolutionary Political Economy, 2024, 5(1): 51-83. DOI: 10.1007/s43253-024-00114-4.

[12] Chen S, Huang Y, Ge L. An early warning system for financial crises: A temporal convolutional network approach. Technological and Economic Development of Economy, 2024, 30(3): 688-711. DOI: 10.3846/tede.2024.20555.

[13] Powell R J, Dinh D V, Vu N T, Vo, D. H. Accounting-based variables as an early warning indicator of financial distress in crisis and non-crisis periods. International Journal of Finance & Economics, 2024, 29(4): 4105-4124. DOI: 10.1002/ijfe.2864.

[14] Barthélémy S, Gautier V, Rondeau F. Early warning system for currency crises using long short-term memory and gated recurrent unit neural networks. Journal of Forecasting, 2024, 43(5): 1235-1262. DOI: 10.1002/for.3069.

[15] Korkoman M J, Abdullah M. Evolutionary algorithms based on oversampling techniques for enhancing the imbalanced credit card fraud detection. Journal of Intelligent & Fuzzy Systems, 2023, 44(6): 10311-10323. DOI: 10.3233/JIFS-222344.

[16] Kaope C, Pristyanto Y. The effect of class imbalance handling on datasets toward classification algorithm performance. MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, 2023, 22(2): 227-238. DOI: 10.30812/matrik.v22i2.2515.

[17] Datta S, Ghosh C, Choudhury J P. Classification of imbalanced datasets utilizing the synthetic minority oversampling method in conjunction with several machine learning techniques. Iran Journal of Computer Science, 2025, 8(1): 51-68. DOI: 10.1007/s42044-024-00207-7.

[18] Nugroho A, Harini D. Teknik Random Forest untuk Meningkatan Akurasi Data Tidak Seimbang. JSITIK: Jurnal Sistem Informasi Dan Teknologi Informasi Komputer, 2024, 2(2): 128-140. DOI: 10.53624/jsitik.v2i2.379.

[19] Anusha Y, Visalakshi R, Srinivas K. Imbalanced data classification using improved synthetic minority over-sampling technique. Multiagent and Grid Systems, 2023, 19(2): 117-131. DOI: 10.3233/MGS-230007.

[20] Bakri R, Astuti N P, Ahmar A S. Evaluating random forest algorithm in educational data mining: Optimizing graduation on-time prediction using imbalance methods. ARRUS Journal of Social Sciences and Humanities, 2024, 4(1): 108-116. DOI: 10.35877/soshum2449.