Journal of Engineering, Project, and Production Management
2026, 16(2), 2025-284

Identifying Financial Fraud in Listed Companies by Integrating Financial Text and Unstructured Data

Dong Peng[1] and Lu Yang[2]

[1] Lecturer, Pingdingshan Polytechnic College, Pingdingshan, 467001, China, E-mail: pengdongdp@outlook.com (corresponding author).

[2] Lecturer, Management School, Henan University of Urban Construction, Pingdingshan, 467036, China.

Project Management
Received November 22, 2025; revised February 13, 2026; accepted February 20, 2026
Available online March 7, 2026

Abstract: To effectively identify Financial Fraud (FF) in Listed Companies (LC), the study combined financial, non-financial, and unstructured data to develop an initial set of financial indicators. To further enhance the fraud detection model's efficacy, the indications were screened using chi-square tests and correlation coefficients. Convolutional Neural Networks (CNN) and bidirectional long short-term memory networks were merged in the study's model development, and an attention mechanism was added to emphasize important details. The outcome revealed that the average accuracy of the research design identification model was 97.48%, which was much higher than that of the comparison models (82.56%, 88.17%, 90.13%, and 91.17%). In addition, the maximum precision of this model was 98.49% and the average recall rate was 97.19%, both of which were superior to those of the comparison models. In summary, models that integrate multi-source data are better able to identify financial fraud in LC and provide strong technical support for maintaining the normal order of the securities market.

Keywords: Unstructured data, financial fraud, identification, BiLSTM, CNN, chi-square test.

DOI 10.32738/JEPPM-2025-284

1. Introduction

As key players in the market economy, listed firms are essential to maintaining the quality and transparency of financial information amid the explosive growth of the global capital market. This directly affects investor's decision-making processes, the efficiency of resource allocation, and the stable operation of financial markets (Liang et al., 2024; Sengupta and Das, 2023). However, as business operations become increasingly complex and market competition intensifies, some LCs resort to FF to achieve objectives such as "maintaining listing status," "raising capital," and "enhancing performance". These tactics include inflating revenue, fabricating transactions, reclassifying related parties as unrelated, and manipulating profits. Such actions severely infringe on investors legitimate rights and interests and exacerbate information asymmetry in the market (Filatova et al., 2023; Jin, 2025). Therefore, it is essential to identify FF in LCs. Currently, common methods for addressing this issue include traditional statistical methods and machine learning methods, such as Support Vector Machines (SVMs) and CNNs (Sowmiya et al., 2025). Many scholars have also researched this issue.

Liang and Lang (2023) addressed the problem of detecting corporate FF by adopting a backpropagation neural network for detection and by refining feature indicators using principal component analysis. In addition, the study also introduced the SVM algorithm for comparative verification. The results showed that principal component analysis could reduce the initial 30 indicators to 20. The detection model's performance was far superior to that of the decision tree and SVM algorithms. Li et al. (2023) proposed a data fusion-based FF anomaly detection framework to detect FF in companies listed on the growth enterprise market. The framework integrated structured, text, and multi-source heterogeneous data at the data layer to construct financial and non-financial information features. The study also used the Synthetic Minority Over-Sampling Technique (SMOTE) to handle imbalanced data. The results revealed that the model's F-β metric was 0.7738, indicating good performance. Li et al. (2023) addressed the issue of FF using data from 11,040 annual reports of non-financial LCs listed on the Shanghai and Shenzhen stock exchanges from 2009 to 2019. They proposed using machine learning to measure the tone of Management's Discussion and Analysis (MD&A) texts to investigate the correlation between text tone and corporate fraud. The results showed that a more positive MD&A tone was associated with a lower likelihood of non-compliance.

Li and Yu (2025) proposed a FF detection model based on stacked ensemble learning and accounting indicators. The model adopted a multi-level feature extraction method, combined multiple base learners with logistic regression, and also used accounting indicators. According to the findings, the model had a low false negative rate and 85.23% detection accuracy (Li and Yu, 2025). Gupta and Mehta (2024) proposed a correlation-based filtering method for feature selection to address feature correlation in FF detection. This method performed feature selection using an integrated model and tested the results by conducting an average-ratio analysis of the Financial Data (FD) of Indian companies. The results showed that this feature selection method identified the most important attributes or variable subsets that captured the original data. Karthikeyan et al. (2024) proposed a new hybrid method that integrates meta-heuristic optimization algorithms with neural network classifiers to achieve FF detection, combining the chimpanzee optimization algorithm with Long Short-Term Memory (LSTM). According to the findings, this approach had a 99.18% classification accuracy with an average absolute error of 25.7.

In summary, SVM cannot adequately represent long-term dependencies, and traditional statistical techniques cannot capture the intricate nonlinear correlations in FD. Furthermore, existing studies focus on FD or a single data type and do not delve deeply enough into the integration and application of multi-source data. To better identify FF in LCs, the study adopts FD, Non-Financial Data (NFD), and unstructured data, constructs and screens the corresponding indicator system, and combines Bidirectional LSTM (BiLSTM) and CNN.

1. Consideration of Financial Fraud Methods in Data Design and Model Construction

1.1. Data Design for Financial Fraud

To identify FF in LCs, the study constructs a relevant financial indicator system using financial, non-financial, and unstructured data. This system is then used to screen financial indicators and preprocess data. Next, based on the designed multi-source data, the study constructs an FF model combining BiLSTM and CNN. FD refers to the quantitative information in an LC's financial statements that directly reflects its financial condition. NFD, on the other hand, refers to information related to a company's operations that is not directly reflected in its financial statements. It plays an important supplementary role in FF (Raju et al., 2025; Kamyshanskyi et al., 2023). In the research sample, companies listed on the Beijing Stock Exchange in China from 2015 to 2024 are selected, and the fraud and non-fraud samples are marked as 1 and 0, respectively. The true labeling of fraudulent samples in the study is based on the regulatory penalty announcements of LCs (issued by the China Securities Regulatory Commission and exchanges) and non-standard opinions in audit reports. Moreover, the labeling process is cross-verified by two financial researchers to ensure its accuracy. In terms of data sources, FD primarily comes from the official database of the Beijing Stock Exchange (https://www.bse.cn/), covering regular reports, financial statements, and regulatory announcements for all LCs on the Beijing Stock Exchange from 2015 to 2024. NFD is primarily sourced from industry research reports by Wind and market analysis reports by Dongfang Wealth Network. It also comes from the information disclosure sections on the Beijing Stock Exchange's official website. This data covers dimensions such as equity structures, audit information, and corporate governance for LCs on the Beijing Stock Exchange. In unstructured data, MD&A text is a management discussion and analysis text in Chinese. The stock review text is sourced from three major platforms: Dongfang Wealth Network Stock Bar, Tonghuashun Financial Review Area, and Snowball Network. All data has been licensed without any commercial authorization restrictions. The research sampling process uses stratified random sampling to select fraudulent and non-fraudulent samples by industry and LC size. The potential sampling bias is mainly due to incomplete disclosure of unstructured data by some small-scale LCs. Samples with data-missing rates exceeding 30% were excluded during the sample screening stage to reduce bias. When constructing the financial indicator system, the study refers to existing research and makes adjustments accordingly. Indicator selection pipeline: First, a preliminary screening is performed; a threshold of mutual information ≥ 0.05 is applied to retain metrics associated with the fraud label. Second, a significant screening: after the chi-square test, significant indicators with *p*-values<0.05 are retained. Third, a redundancy screening: indicators with a Pearson Correlation Coefficient (PCC) ≥0.7 are removed. Finally, effective indicators are determined. The initial financial indicator system is shown in Fig. 1.

In Fig. 1, the initial financial indicator system contains both financial and non-financial indicators, with a total of nine major indicator dimensions. Since each major indicator dimension contains a large number of specific indicators, Chi-Square Tests (CST) and PCC are used to screen indicators to improve the ultimate accuracy of FF. The CST is used to determine whether two categorical variables are significantly related (Boldin, 2024). The calculation of the chi-square statistic χ^2 is shown in Eq. (1) (Zhao et al., 2025).

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{1}$$

In Eq. (1), O is the observed frequency, and E is the expected frequency. The calculation of E is shown in Eq. (2).

$$E = \frac{A \cdot B}{C} \tag{2}$$

In Eq. (2), A is the row sum, B is the column sum, and C is the total quantity of samples. The PCC is a measure of the degree of linear correlation between two variables (Noor et al., 2024). The calculation of the PCC r is shown in Eq. (3) (Yoon et al., 2023).

$$r = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2 \sum (Y-\bar{Y})^2}} \tag{3}$$

In Eq. (3), X an Y display the observed values of two variables. $\bar{X}$ and $\bar{Y}$ display the Average Values (AVs) of variables X and Y. The filtered indicators are shown in Table 1.

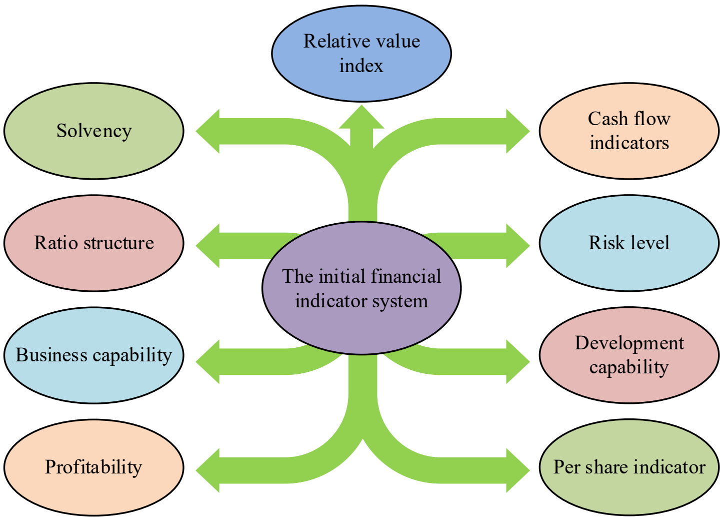


Fig. 1. The initial financial indicator system

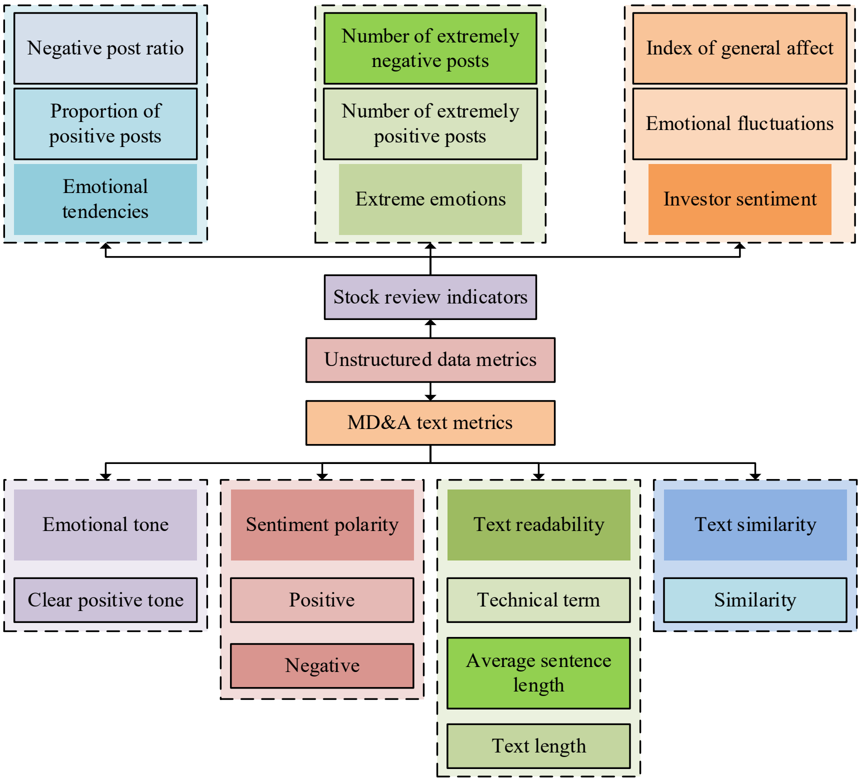


Fig. 2. MD&A text indicator and stock review indicator

In Fig. 2, there are four dimensions and seven indicators for the MD&A text. There are three dimensions and six indicators for stock comment indicators. Among them, the solution for net positive sentiment D is shown in Eq. (4).

$$D = \frac{F-G}{G} \tag{4}$$

In Eq. (4), F and G represent the number of occurrences of positive and negative words, respectively. The research

Table 1 shows that the screened indicators include interest coverage ratio, accounts receivable turnover ratio, and book-to-market ratio. For unstructured data, the study selects MD&A financial text data and stock commentary text (Liu et al., 2024; Musunuru, 2025). The MD&A text indicators and stock commentary indicators selected for the study are shown in Fig. 2.

Table 1. The selected financial indicator system

Indicator dimension	Index	Symbol	Indicator dimension	Index	Symbol
Solvency	Interest coverage ratio	A4	Profitability	Net Profit Margin (NPM) of total assets	A33
	Tangible asset liability ratio	A6		NPM of fixed assets	A35
	Property rights ratio	A7		Return on equity	A36
	Equity multiplier	A8		Return on investment capital	A37
	Long term debt to equity ratio	A9		Long term capital return rate	A38
Ratio structure	Current asset ratio	A11		Operating NPM	A42
	Cash asset ratio	A12		Total operating cost rate	A43
	Accounts receivable asset ratio	A13		Cost expense profit margin	A44
	Operating capital ratio	A14		Investment return rate	A45
	Non-current asset ratio	A15		Period expense rate	A46
	Fixed asset ratio	A16		Net profit cash net content	A47
	Ratio of intangible assets	A17	Cash flow indicators	Financing activities creditor's net cash flow	A50
	Ratio of tangible assets	A18		Net cash flow from financing activities to shareholders	A51
Business capability	Accounts receivable turnover rate	A21	Development capability	Revenue growth rate	A67
	Inventory turnover rate	A22	Per share indicator	Comprehensive earnings per share	A74
	Business cycle	A23	Relative value index	Price-to-book ratio	A88
	Accounts payable turnover rate	A24		Book-to-market ratio	A90
	Fixed asset turnover rate	A27	Non-financial indicators	Ownership concentration	A93
	Noncurrent asset turnover ratio	A28		Tradable	A95
	Capital intensity	A29		Total audit fees	A99
Profitability	Return on assets	A32		Number of board meetings	A100

In Eq. (4), F and G represent the number of occurrences of positive and negative words, respectively. The research feature engineering process includes three steps. The first step is feature screening: the CST is used to screen for indicators. significantly correlated with cheating, and the Pearson PCC method is used to eliminate indicators with strong correlations. The second step is outlier handling: the box plot method is used to identify outliers across indicators, the median is used to replace outliers in continuous indicators, and outlier samples are removed from categorical indicators. The third step is feature normalization: all numerical features undergo Z-score normalization, mapping their values to a distribution with a mean of 0 and a variance of 1. Unstructured text features are vectorized before normalization to ensure features across different dimensions can participate in model training.

1.2. Building a Financial Fraud Model Combining BiLSTM and CNN

In constructing the FF model, the study uses BiLSTM and CNN. BiLSTM can simultaneously process the forward and backward information of sequence data (Chen et al., 2025). Fig. 3 depicts the BiLSTM model's structure.

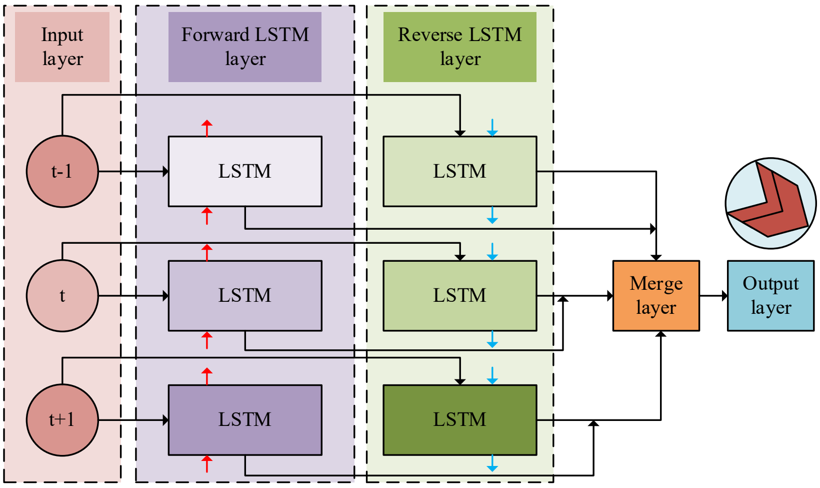


Fig. 3. The structure of BiLSTM model

In Fig. 3, the BiLSTM model consists of an Input Layer (IL), a forward LSTM Layer (LSTM-L), a backward LSTM-L, a merging layer, and an output layer. The expression of the merged vector h_t is shown in Eq. (5).

$$h_t = \left[h_t^{(f)}; h_t^{(b)} \right] \tag{5}$$

In Eq. (5) $h_t^{(f)}$ displays the hidden state of the forward LSTM. $h_t^{(b)}$ displays the hidden state of the backward LSTM. CNN is a deep learning model primarily used for processing data with grid structures, offering strong local perception capabilities. Therefore, research focuses on utilizing CNN to extract features from high-dimensional data. The convolution operation is shown in Eq. (6).

$$y_{i,j} = \sum_{m=1}^{M} \sum_{n=1}^{N} x_{i+m-1,j+n-1} \cdot w_{m,n} + R \tag{6}$$

In Eq. (6), $y_{i,j}$ displays the value of the output Feature Map (FM) at position (i,j). $x_{i+m-1,j+n-1}$ Displays the value of the input FM at the specified position $(i+m-1, j+n-1)$. $w_{m,n}$ denotes the weight of the Convolution Kernel (CK) at position (m,n). R displays the bias term. M and N denote the width and height dimensions of the CK, respectively. The structure combining a BiLSTM and a CNN-FF model is shown in Fig. 4.

A one-dimensional CL, a batch normalization layer, a ReLU Activation Function (AF), pooling, FMs, a BiLSTM layer, an AM, an FCL, the final output, and input data are all included in the FF model shown in Fig. 4. The research adopts feature concatenation to achieve the fusion of text and digital features, and the fusion process V_f is shown in Eq. (7).

$$V_f = Concat\left(V_n, V_t\right) \tag{7}$$

In Eq. (7), $Concat$ represents the feature concatenation function, V_n represents the numerical feature vector, and V_t is the textual feature vector. The fusion logic of text and numerical features is shown in Fig. 5. The fusion of text and digital features follows a multi-step process: First, multiple sources of raw data are input. Thus, independent preprocessing is performed on the digital features, and encoding is applied to the text features. Feature dimension adaptation is then verified, and cross-modal concatenation is used to complete the fusion. Finally, the fused features are fed into the model as input and output. The process for detecting FF in LCs by integrating financial text and unstructured data is shown in Fig. 6.

In Fig. 6, the FF in the LCs process comprises five core links: data preparation, indicator system construction, data preprocessing, model construction and application, and performance verification. This process allows the study to systematically integrate multi-source data from LCs. It also combines the BiLSTM-CNN model's ability to accurately capture time series and local features. This enhances the precision and effectiveness of FF detection in LCs.

2. Verification of the Performance of Financial Fraud Methods

2.1. Data Design and Analysis of Processing Results

The exact sequence of complete experimental operations for research is data collection, data preprocessing, feature engineering, dividing training and testing sets in chronological order, a SMOTE-balanced training set, model construction and hyperparameter optimization, model training, and model performance evaluation. To respect the temporal dependence of FD, the study redivided the dataset in chronological order: LC data from 2018 to 2022 were used as the training set (totaling 10020 samples). Moreover, LC data from 2023 to 2024 is used as the testing set (with a total of 1560 samples), ensuring that the training set is early-limited data and the testing set is later new data. Due to an imbalance between fraudulent and non-fraudulent samples in the dataset, the study uses SMOTE to generate new samples by synthesizing them from the feature distribution of the minority fraudulent samples. To avoid data leakage caused by SMOTE, the study applies SMOTE only to the training set (dividing the preprocessed samples into training and testing sets in chronological order), leaving the testing set with its original unbalanced distribution. After processing, the number of fraudulent samples increases to 4,234, while the number of non-fraudulent samples remains at 6,052, with the total number of samples adjusted to 10,286. To verify the effectiveness of SMOTE, SVM is used as the basic verification model. The CST results of the indicators are shown in Fig. 7.

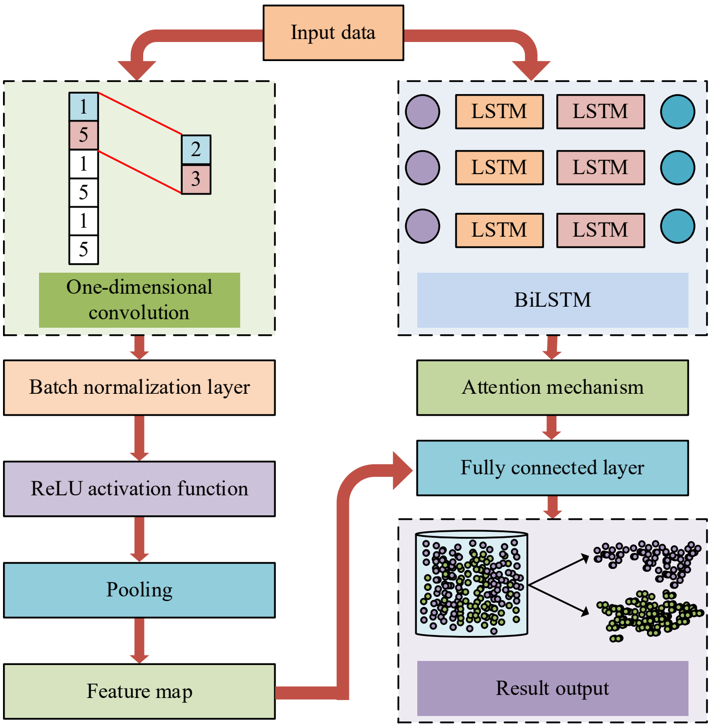


Fig. 4. Structure of financial fraud identification model combining BiLSTM and CNN

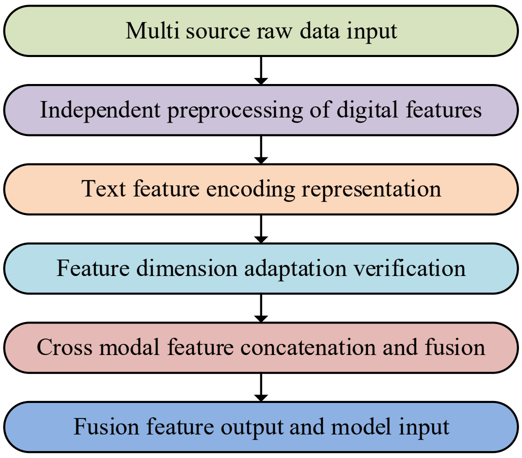

Multi source raw data input
Independent preprocessing of digital features
Text feature encoding representation
Feature dimension adaptation verification
Cross modal feature concatenation and fusion
Fusion feature output and model input

Fig. 5. Fusion logic of text and numerical features

In Fig. 7, due to the large number of preliminary indicators, the study selects only indicators with p-values that meet the analysis requirements. All indicators are statistically significant and appropriate for further study, as their *p*-values are less than 0.05. Furthermore, the Chi-Square Value (CSV) for indicator A18 is the highest at 84.21, indicating the strongest association with the target variable. Most indicators have CSVs within the range of (5, 30). These findings offer a strong basis for the ultimate indication selection. The results of the PCC screening for the indicators are shown in Table 2.

In Fig. 8(a), in the comparison of accuracy rates, the Maximum Value (MaxV) of the SVM base model after data balancing is 82.72%, and the Minimum Value (MinV) is 77.12%. Under unbalanced data conditions, the SVM base model achieves a maximum accuracy of 68.24%. This suggests that the SVM basic model's identification accuracy can be considerably increased by employing data balancing with SMOTE technology. In Fig. 8(b), in the time-consuming comparison, the maximum time consumption of the SVM base model after data balancing is 120ms, while the maximum time consumption of the SVM base model without data balancing is 157ms.

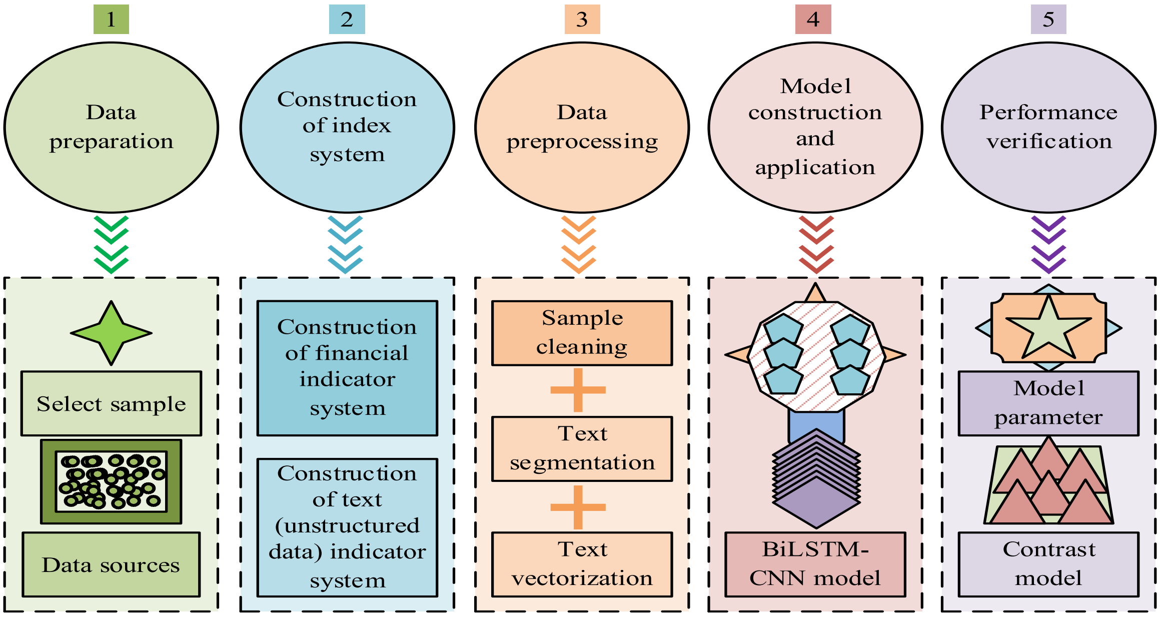


Fig. 6. Detecting financial fraud in listed companies via text and unstructured data

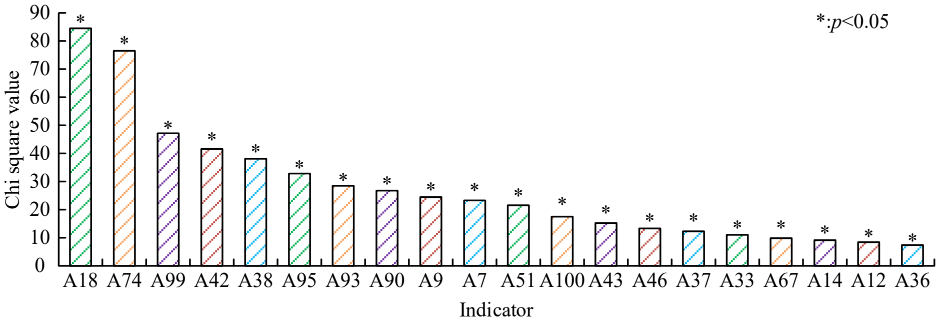


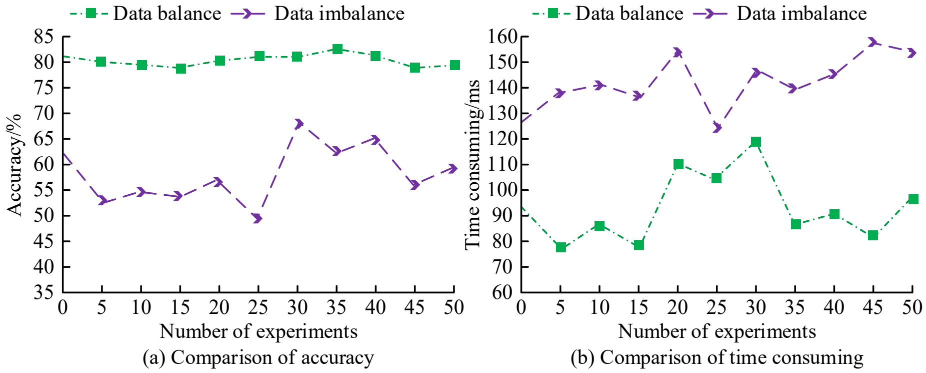
Fig. 7. Chi-square test results of indicators

Table 2. Selection results of correlation coefficients for indicators

Indicator pair	Correlation coefficient	Indicator pair	Correlation coefficient	Indicator pair	Correlation coefficient
A8-A13	0.65	A8-A16	0.58	A8-A17	0.60
A8-A21	0.55	A8-A22	0.51	A8-A24	0.48
A8-A27	0.45	A8-A35	0.52	A8-A44	0.56
A8-A45	0.61	A8-A47	0.59	A8-A11	0.42
A8-A15	0.58	A8-A50	0.48	A8-A29	0.51
A8-A32	0.46	A8-A23	0.49	A8-A6	0.53
A8-A4	0.50	A8-A28	0.54	A8-A88	0.62
A8-A17	0.47	A8-A7	0.58	A8-A9	0.49
A8-A90	0.43	A8-A33	0.42	A8-A42	0.57

The screening rule for PCC is to remove strongly correlated factors, i.e., indicators with a PCC greater than 0.7. In Table 2, after indicator screening, the indicators presented are all based on A8. Under the combination of A8 indicators and other indicators, the Pearson PCCs are all in the range of 0.42-0.65, which are all below 0.7, providing a good indicator basis for the subsequent FF model. In addition, due to the overlap between the indicators selected by the PCC screening and the CST screening, the final evaluation indicator system (a total of 42 indicators) is obtained after removing duplicate indicators.

The effect of data balancing is shown in Fig. 8.

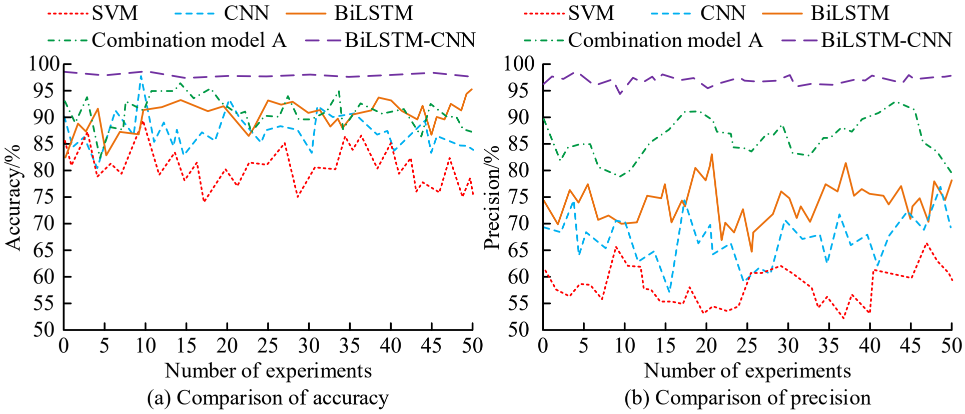


(a) Comparison of accuracy (b) Comparison of time consuming

Fig. 8. The effect of data balancing processing

2.2. BiLSTM-CNN Model Performance Verification

To validate the performance of the BiLSTM-CNN model design, the study also uses Windows 10 and an Intel Core i5-13500 central processing unit. In the comparison model, the study selects SVM, CNN, BiLSTM, and a combination model A that combines CNN and LSTM. In terms of model parameter settings, the BiLSTM-CNN model uses Adam as the optimizer, a learning rate of 0.01, a BiLSTM hidden layer dimension of 64, and a CNN kernel size of 3×32. Furthermore, the SVM kernel function is a radial basis function, and the LSTM hidden layer dimension is 64. The batch size is 32, the sequence length is 12, and the number of features is 42. The dataset consists of 10286 training samples and 1560 test samples. In terms of evaluation metrics, the study selects classification performance indicators and model efficiency indicators. Among them, the classification performance indicators include accuracy, precision, recall rate, and F1 score. Area Under the Curve (AUC). confusion matrices, Receiver Operating Characteristic (ROC)/Precision-Recall (PP) curves, AUC PR, and calibration plots. The model efficiency indicator includes time consumption. All performance evaluation indicators for the model are calculated using the original unbalanced test set. This ensures that the evaluation results align with real-world FF sample distributions, enhancing the model's practical application value. A comparison of the accuracy and precision of different models is shown in Fig. 9.

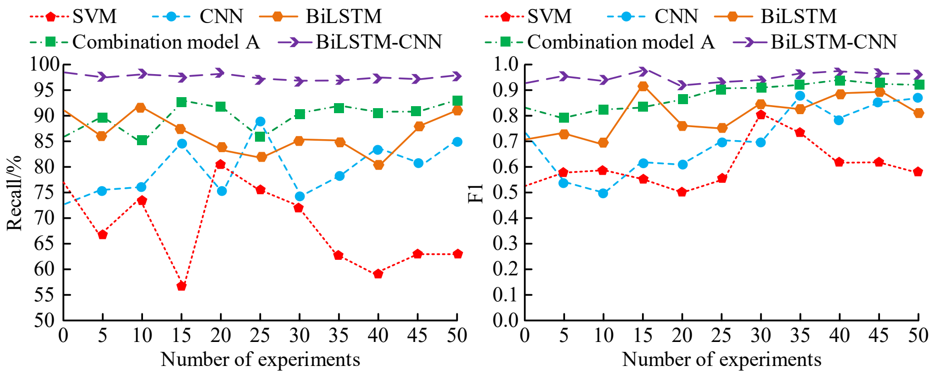


(a) Comparison of accuracy (b) Comparison of precision

Fig. 9. Comparison of accuracy and precision of different models

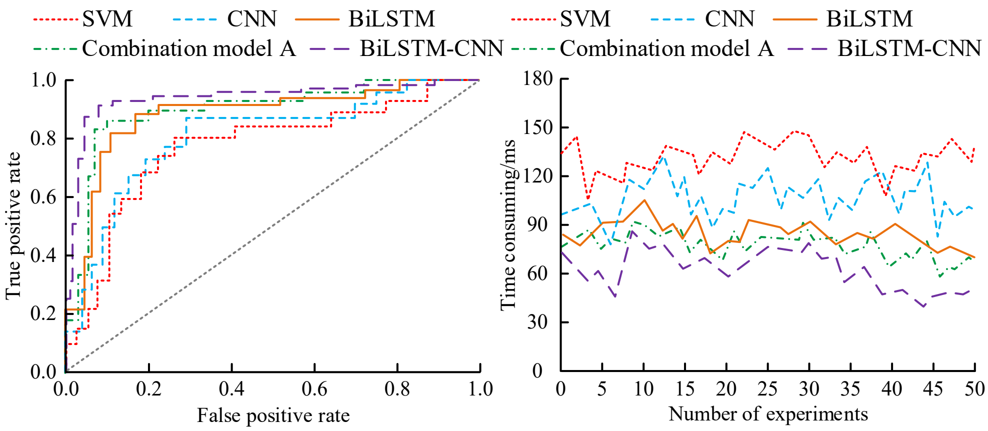
In Fig. 9(a), the proposed BiLSTM-CNN model achieves the best accuracy. The average accuracy rate of the BiLSTM-CNN model is 97.48% across multiple experiments. In Fig. 9(b), in terms of precision comparison, the BiLSTM-CNN model designed in this study performs better, followed by the combination model A, with maximum precision values of 98.49% and 93.22%, respectively. In summary, the BiLSTM-CNN model designed in this study performs better. This could be a result of the BiLSTM-CNN model's ability to capture the long-term dependencies of financial time series data with BiLSTM and efficiently extract local features from the data with CNN. A comparison of the recall and F1 values of different models is shown in Fig. 10.

In Fig. 10(a), the BiLSTM-CNN model still ranks first in terms of recall rate, with a MaxV of 98.23% and an AV of 97.19%. This indicates that the model can more comprehensively identify fraudulent samples and reduce false negatives. In Fig. 10(b), the BiLSTM-CNN model also performs exceptionally well in terms of F1 value, with MaxVs of 0.9867, 0.8092, 0.8876, 0.9011, and 0.9123 compared to the four comparison models. The BiLSTM-CNN model achieves a better balance between accurately identifying fraudulent samples and comprehensively covering fraudulent samples. The AUC and time consumption of different models are compared in Fig. 11



(a) Comparison of recall rates (b) Comparison of F1

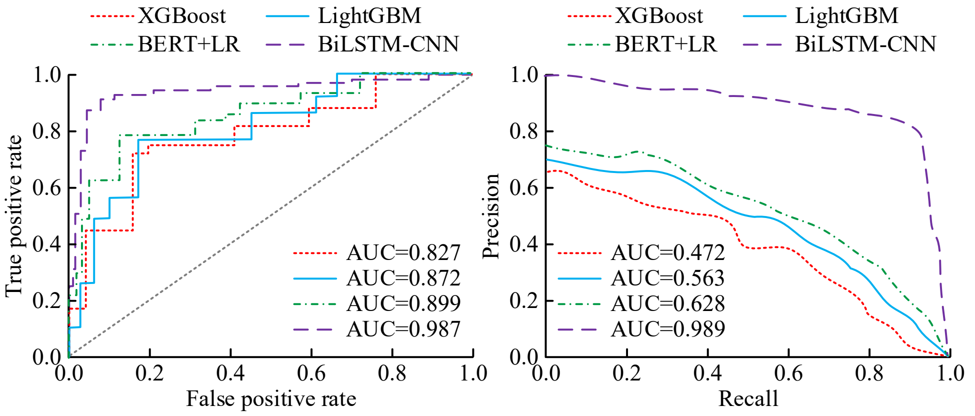
Fig. 10. Comparison of recall rates and F1 values for different models



(a) Comparison of AUC (b) Comparison of time consuming

Fig. 11. Comparison of AUC and time consumption of different models

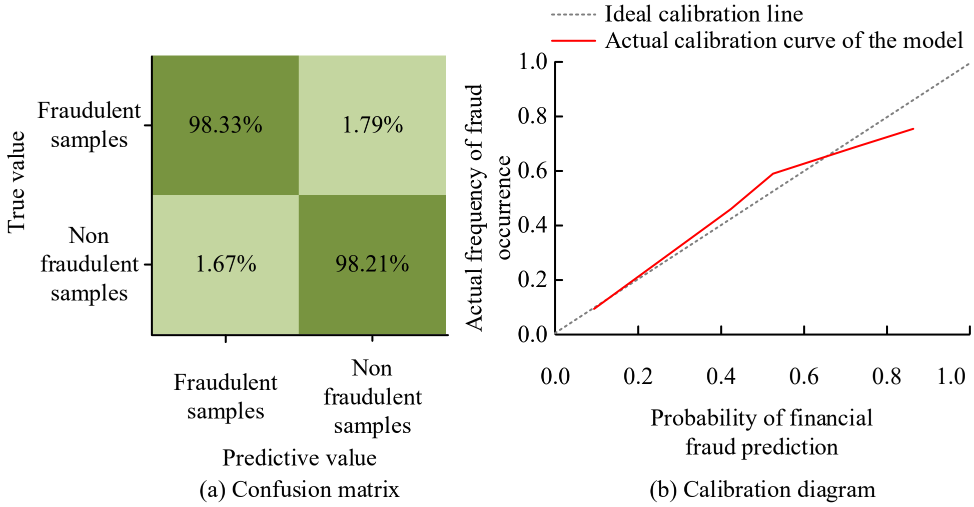
In Fig. 11(a), the AUC value for the BiLSTM-CNN model is 0.987. The values for SVM, CNN, BiLSTM, and the combined model A are 0.825, 0.869, 0.903, and 0.916, respectively. This suggests that the BiLSTM-CNN model developed for this investigation performs better overall at distinguishing between fake and non-fake samples. In Fig. 11(b), in terms of time consumption, the BiLSTM-CNN model designed in this study performs better, with a MaxV of 82ms and a MinV of 43ms. The MaxVs for the four comparison models are 148ms, 123ms, 108ms, and 95ms, respectively, all of which exceed 82ms. In summary, the BiLSTM-CNN model designed in this study performs better in terms of AUC and computation time. The study introduces the latest baseline model: eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Bidirectional Encoder Representations from Transformers (BERT)+Logistic Regression (LR). The comparison between the ROC/PR curve and the AUC PR is shown in Fig. 12.



(a) Comparison of ROC (b) Comparison of PR

Fig. 12. ROC/PR curve and AUC PR comparison

In Fig. 12, the ROC and PR curves of four models, XGBoost, LightGBM, BERT+LR, and BiLSTM-CNN, are compared. In the ROC curve, BiLSTM-CNN has the highest AUC (0.987). In the PR curve, the AUC-PR of BiLSTM-CNN is also the highest (0.989), significantly better than other models. The BiLSTM-CNN model performs the best in both classification performance and probabilistic prediction reliability. The confusion matrix and calibration plot of the BiLSTM CNN model are shown in Fig. 13.



(a) Confusion matrix

(b) Calibration diagram

Fig. 13. Confusion matrix and calibration graph of BiLSTM-CNN model

In Fig. 13(a), the BiLSTM CNN model achieves a recognition accuracy of 98.33% for fraudulent samples, with only 1.79% misclassified as normal. The recognition accuracy of normal samples is 98.21%, and the misjudgment rate is 1.67%. This indicates that the model achieves high classification accuracy across both sample types. In Fig. 13(b), the model's actual calibration curve closely aligns with the ideal calibration line (dashed line). This indicates a good match between the predicted probability of FF and the observed frequency of fraud.

3. Conclusion

To address FF in LCs, the study used multiple data sources and constructed a recognition model combining BiLSTM and CNN. The outcomes revealed that after SMOTE balancing, the maximum accuracy of the basic SVM model was 82.72%, which was 14.48% higher than the maximum accuracy without balancing. This suggested that the model's performance could be enhanced by balancing the data. The maximum precision of the proposed BiLSTM-CNN model was 98.49%, whereas the other four comparison models achieved precisions below 94%. This indicated that the BiLSTM-CNN model could filter out genuine cheating cases from complex data and reduce the misclassification of non-cheating samples as cheating samples. In addition, in terms of processing time, the BiLSTM-CNN model had a MaxV of 82ms and a MinV of 43ms, which was also significantly better than the comparison models. In summary, the BiLSTM-CNN model that integrates multiple sources of data achieves better overall performance in FF detection. However, the study's non-structured data categories are limited to stock comments and MD&A texts, among other limitations. Future research can introduce more non-structured data, such as news reports and social media, to further improve the model's recognition efficiency and accuracy.

Author Contributions

[Editor removed]

Funding

The author(s) declare none.

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence (AI) Tools

The authors confirmed that no AI tools were used in the preparation of this manuscript.

Reference

Boldin, V. (2024). On symmetrized chi-square tests in autoregression with outliers in data. *Theory of Probability and Its Applications*, 68(4), 559–569. doi: 10.1137/S0040585X97T991623

Chen, Y., Cao, H., Xiang, Z., Chen, B., Ma, Y., and Zhang, Y. (2025). Vehicle lane-change intention recognition based on BiLSTM attention model for the Internet of Vehicles. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 239(7), 2551–2564. doi: 10.1177/09544070241240225

Filatova, H., Tumpach, M., Reshetniak, Y., Lyeonov, S., and Vynnychenko, N. (2023). Public policy and financial regulation in preventing and combating financial fraud: A bibliometric analysis. *Public and Municipal Finance*, 12(1), 48–61. doi: 10.21511/pmf.12(1).2023.05.

Gupta, S., and Mehta, K. (2024). Feature selection for dimension reduction of financial data for detection of financial statement frauds in context to Indian companies. *Global Business Review*, 25(2), 323–348. doi: 10.1177/0972150920928663

Jin, Y. (2025). Study on digital financial fraud risk identification based on heterogeneous graph convolutional attention network. *International Journal of Networking and Virtual Organisations*, 32(1), 203–218. doi: 10.1504/IJNVO.2025.145376

Kamyshanskyi, O., Dykyi, A., Kryzhanovskyi, A., Baranovska, T., and Kovalchuk, O. (2023). Conduct of search actions in the investigation of fraud with financial resources. *Cuestiones Políticas*, 41(77), 830–852. doi: 10.46398/cuestpol.4177.54

Karthikeyan, T., Govindarajan, M., and Vijayakumar, V. (2024). Enhancing financial fraud detection through chimp-optimized long short-term memory networks. *Traitement du Signal*, 41(2), 835–845. doi: 10.18280/ts.410224

Li, A., Wang, D., Xu, W., Li, Z., and Yao, S. (2023). Financial fraud detection for growth enterprise market listed companies based on data fusion. *Data Analysis and Knowledge Discovery*, 7(5), 33–47. doi: 10.11925/infotech.2096-3467.2022.0585questio

Li, H., and Yu, X. (2025). Construction of financial fraud identification model based on stacking and accounting indicators. *Journal of Computational Methods in Sciences and Engineering*, 25(4), 3369–3383. doi: 10.1177/14727978251316402

Li, S., Jiang, L., and Bian, S. (2023). Annual reports' tone and violation behavior identification of listed companies: Evidence from textual analysis based on machine learning. *Modern Economic Science*, 45(6), 97–109. doi: 10.20069/j.cnki.DJKX.202306008

Liang, X., Xie, Q., Luo, C., Tang, L., and Sun, Y. (2024). Principal model analysis based on bagging PLS and PCA and its application in financial statement fraud. *Journal of Systems Science and Information*, 12(2), 212–228. doi: 10.21078/JSSI-2023-0144

Liang, Z., and Liang, Y. (2023). A study of identification of corporate financial fraud using neural network algorithms in an information-based environment. *Informatica*, 47(9), 165–171. doi: 10.31449/inf.v47i9.5220

Liu, H., Xu, J., Cao, S., Liu, Y., and Qiao, L. (2024). Research progress of digital financial fraud detection oriented to social relations network. *Journal of Zhejiang University, Science Edition*, 51(1), 41–54. doi: 10.3785/j.issn.1008-9497.2024.01.006

Musunuru, K. (2025). Big data analytics for financial auditing practices: Identification of conceptual patterns, implications and challenges using text mining. *Contaduríay Administración*, 70(2), 184–219. doi: 10.22201/fca.24488410e.2025.5283

Noor, M., Nugroho, R. A., Saputro, S. W., Herteno, R., and Abadi, F. (2024). Optimization of backward elimination for software defect prediction with correlation coefficient filter method. *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, 6(4), 397–404. doi: 10.35882/jeeemi.v6i4.466

Raju, P., Babu, T., Selvakumar, V., Thakur, S., Srinu, M., and Srivastava, P. (2025). Supervised learning models for enhancing financial fraud detection systems. *Journal of Information Systems Engineering and Management*, 10(17), 644–652. doi: 10.52783/jisem.v10i17s.2786

Sengupta, K., and Das, P. K. (2023). Detection of financial fraud: Comparisons of some tree-based machine learning approaches. *Journal of Data, Information and Management*, 5(1), 23–37. doi: 10.1007/s42488-023-00086-w

Sowmiya, B., Ida Seraphim, B., Fancy, C., Abirami, R., and Hussain, A. (2025). Harnessing artificial intelligence in financial fraud detection and prevention systems. *International Journal of Innovative Research and Scientific Studies*, 8(3), 1449–1459. doi: 10.53894/ijirss.v8i3.6821

Yoon, J. H., Kim, J., and Koo, Y. (2023). Novel fuzzy correlation coefficient and variable selection method for fuzzy regression analysis based on distance approach. *International Journal of Fuzzy Systems*, 25(8), 2969–2985. doi: 10.1007/s40815-023-01546-6

Zhao, J., Ge, X., Cheng, Y., Li, J., and Zhou, H. (2025). A robust Kalman filter with bias estimation based on variational Bayesian inference and chi-square test. *Circuits, Systems, and Signal Processing*, 44(4), 2922–2935. doi: 10.1007/s00034-024-02943-4

Dong Peng graduated from Liaoning Institute of Technology in 2006 with a master's degree in Economics, majoring in International Economy and Trade. He currently works as a lecturer at Pingdingshan Industrial College of Technology and has been repeatedly honored as an "Outstanding Teacher." He has published articles in journals such as Economic Research Guide and Modern Business and has presided over or participated in multiple research projects, including those organized by the Henan Federation of Social Sciences and the National Teaching Resource Database. Additionally, he has authored two textbooks and has frequently guided students to award-winning achievements in competitions. His research interests include Management and Economics.

Lu Yang graduated from the International Economics and Trade major at Liaoning Institute of Technology in 2006, earning a Master’s of Economics degree. She currently works as a lecturer at the School of Economics and Management, Henan University of Urban Construction. She has been awarded the honors of “Advanced Educator” and “Civilized Teacher” on multiple occasions. She has published papers in journals such as China Collective Economy, China Foreign Investment and Economic Research Guide, presided over and participated in several research projects sponsored by the Henan Provincial Government, Henan Federation of Social Sciences, and national-level teaching resource databases, compiled one textbook, and guided students to win awards in various competitions many times. Her research interests span economics and management.