

Research on the Application Effect of Artificial Intelligence Algorithms in the Detection of Financial Fraud in Listed Companies

Tianyue Lyu

Auditing Teaching and Research Section, School of Economics and Management, Hebei Oriental College, Langfang 065000, Hebei, China

Corresponding author: Tianyue Lyu (e-mail: fayelyu0102@163.com)

ABSTRACT Accurate detection of financial fraud remains a critical challenge due to information asymmetry, high-dimensional data complexity, and evolving fraudulent behaviors. This study develops an artificial-intelligence-based financial fraud identification framework for listed companies by integrating multiple machine-learning algorithms. A financial indicator database covering profitability, solvency, cash flow, and governance characteristics is first constructed. Principal Component Analysis is employed to reduce dimensionality and eliminate multicollinearity, followed by the training of Random Forest, Support Vector Machine, and Neural Network models. A fusion-learning strategy with cross-validation optimization is further introduced to enhance model robustness and classification performance. Experimental results demonstrate that the fusion model achieves an accuracy of 94.7%, outperforming individual RF (83.2%), SVM (85.6%), and NN (89.1%) models while maintaining a false-positive rate below 5%. The proposed framework provides effective support for intelligent risk warning and financial supervision and offers methodological references for pattern recognition, data fusion, and intelligent decision-making systems.

INDEX TERMS financial fraud detection, artificial intelligence, fusion learning, pattern recognition, intelligent decision-making

I. INTRODUCTION

FINANCIAL fraud within listed enterprises severely destabilizes the capital market and erodes investor confidence in the sector's overall manufacturing excellence. Such integrity breaches distort the perceived value of industrial material research and technological assets, undermining the high-dimensional data perception necessary for sustainable investment in modern engineering. It is hidden and complex, which makes it challenging to have it revealed through normal means of identification to give effective early warnings. The gap in risk identification is caused by incomplete information disclosure, financial statements manipulation, and governance malfunction. Traditional approaches to auditing are based on sampling and experience-driven audit, which cannot easily handle high-dimensional financial data of large-scale data environment [1,2]. Financial markets are too complicated and corporate behaviour too diversified to imply that the reasoning behind the assessment of fraudulent behavior is confined to a single financial indicator. This requires the application of smart algorithms in order to have deep pattern mining and automatic identification. The emergence of the Artificial Intelligent (AI) technology has offered additional avenues in the detection

of financial fraud. Multi-source data, which are nonlinear in feature relationships, can be modeled with machine learning, which increases the accuracy of detection and generalization ability. Self-learning and self-adaptation are the benefits of AI algorithms over the statistical models that are based on a static analysis, as they allow the algorithm to constantly optimize judgment criteria within constantly changing financial settings. It has become technically possible to construct many-dimensional fusion models with the introduction of algorithms like the Random Forest (RF), Support Vector Machine (SVM) and the Neural Network (NN).

According to this, the financial data of the listed companies is taken as the object of the analysis, a financial fraud identification model is built based on the combination of multiple AI algorithms, and principal component analysis (PCA) is used to mitigate the effects of multicollinearity on the model performance [3]. Cross validation (CV) is used to optimise the parameter settings to determine the identification effect and stability of the comprehensive model. The research aims to confirm the effectiveness of multi-algorithm fusion in detecting financial fraud within listed enterprises, thereby defining the possibilities of applying these models to audit risk warning and intelligent supervision of technical

industrial assets. By integrating material morphology data and industrial manufacturing parameters into the identification framework, the study offers robust technical solutions and empirical conclusions for maintaining manufacturing excellence and financial integrity in the modern sector.

II. RELATED WORK

The study of financial fraud detection was started with statistical analysis techniques. Financial ratios and governance indicators were primarily taken as the key variables by which early scholars used discriminant analysis, logit regression, and other models to determine abnormal behavior. The approaches, however, presuppose that variables are related in a linear and independent manner, so it is hard to handle high-dimensional nonlinear data conditions. Other researchers tried to bring Bayesian networks and clustering algorithms in order to summarize the abnormal distribution features by using probabilistic modeling but the stability of the model was not adequate as the sample size grew [4,5]. The use of data mining technology has initiated a research shift and the scholars have started applying machine learning techniques to detect fraud, which includes decision trees, K nearest neighbor, and support vectors machines (SVM) so as to enhance the accuracy of classification. SVM performs well in terms of generalization in small, high-dimensional data, but is sensitive in terms of its parameters, and approaches to remedy this remain to be optimized so as to remain stable.

Due to the popularity of the artificial intelligence (AI) technology, the focus on multi-model fusion and deep learning has become a research focus over time. The current literature identifies that the single algorithms are vulnerable to the noise of the data and the imbalance of the samples whereas the use of ensemble learning can greatly enhance the recognition accuracy and resistance of risks due to multi-model fusion. Random Forest (RF) is a good example since it can be used in feature extraction and classifications, whereas Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) have the potential to be used in time series financial data analysis [6,7]. Nevertheless, the current research is largely limited to the comparison of algorithms and does not carry out systematic studies of the multi-dimensional fusion of financial indicators and validation developed using actual enterprise data. The results of many black-box models suffer from low interpretability, specifically regarding the ranking of feature importance and the marginal effect of individual financial ratios on the final fraud probability. Thus, there is an urgent need to create an ensemble AI recognition system that can be accurate and interpretable simultaneously.

III. METHODS

A. DATA CONSTRUCTION AND INDICATOR SELECTION

The financial data of a certain industry listed company in Shanghai and Shenzhen A-shares in the past five years were used as a sample. The data sources included Juchao Information Network and Wind Database. After missing

value detection and outlier correction, 1,560 sets of annual data from 312 companies were retained. The classification labels for ‘financial fraud events’ were defined based on official enforcement announcements from the China Securities Regulatory Commission (CSRC) and formal restatements of financial reports due to intentional irregularities, ensuring the legal and regulatory validity of the ground truth labels. In order to ensure data quality, companies that were delisted or suspended from listing were removed, and the indicators were standardized and the difference in dimensions was eliminated by range normalization.

The indicator system consists of four dimensions: profitability, solvency, cash flow and corporate governance [8]. Profitability is selected from return on net assets and return on total assets; solvency includes current ratio and debt-to-equity ratio; cash flow indicator uses operating cash flow and sales cash ratio; the corporate governance part includes the shareholding ratio of senior executives and the size of the board of directors to quantify the effectiveness of internal control. These indicators were prioritized due to their high data availability and consistent reporting across the sector; however, they serve as proxies for broader governance risks, such as executive turnover and related-party transactions, which are frequently correlated with internal control bypasses in fraudulent firms. The correlation test between variables uses Pearson correlation coefficient, and the results show that some financial indicators have strong linear correlation. To prevent multicollinearity from affecting the subsequent discrimination process, principal component analysis (PCA) is introduced to reduce the dimensionality of highly correlated variables, retaining only principal components with a cumulative variance contribution rate of 85% or more for modeling analysis. While 85% captures the primary structural variance, the remaining 15% was qualitatively screened to ensure that rare but significant ‘noise’—often representing the subtle outliers characteristic of sophisticated financial fraud—was not prematurely discarded, thereby balancing dimensionality reduction with signal retention.

Table 1 shows the indicator system and variable formulas used for AI model training, ensuring the quantification and comparability of financial fraud identification.

B. PRINCIPAL COMPONENT ANALYSIS

According to the developed financial indicator system, Principal Component Analysis (PCA) was conducted in order to reduce the linear dimension of each indicator to minimize the interference of high correlation of variables on the identification model. Once the data have been standardized, a covariance matrix was formed and the direction of distribution of the major information was obtained by calculating eigenvalues and eigenvectors. Principal components have been derived during the analysis process with the help of the statistical software and the cumulative variance contribution rate curve has been plotted with the variance explanation rate as the retention criterion. Once the

TABLE 1. Indicator System and Variable Description

Indicator Categories	Indicator Name	Variable symbol	Calculation formula	Illustrate
Profitability	Return on equity	ROE	Net profit / average net assets	Measuring shareholder returns
Profitability	Return on total assets	ROA	Net profit / Total assets	Measuring asset utilization efficiency
Debt repayment ability	Current ratio	CR	Current assets/current liabilities	Reflects short-term solvency
Debt repayment ability	Debt-to-asset ratio	DAR	Total Liabilities / Total Assets	Measuring the robustness of the capital structure
Cash flow	Operating cash flow ratio	OCF	Operating cash flow / operating revenue	Reflects the quality of cash acquisition
Governance Structure	Shareholding ratio of senior executives	SHR	Management shareholding / total shares outstanding	Reflecting the level of governance incentives

rate of compilation of variance has exceeded 85 percent, the model will automatically take the first three principal components to be further analyzed. The outcome was that the initial three major components were capable of covering the original information like profitability, solvency, cash flow and governance structure [9,10]. The first principle component load was primarily focused on the variables of performance like the return on net assets, the return on total assets and operating cash flow that reflected the profitability specific to the enterprise; the second principle component was significantly correlated to the solvency indicators indicating the characteristics of the capital structure and the third principle component was significantly correlated with the shareholding ratio of high ranking executives and the magnitude of the board of directors indicating the differences in governance dimensions.

C. MODEL DESIGN AND FUSION STRATEGY

The fusion-based AI recognition framework is established on the principles of dimensionality-reduced features information through the application of the Random Forest (RF), Support Vector Machine (SVM) and Neural Network (NN) as the key learners. The training data is separated into training and test sets at a ratio of 8:2 and classification labels are identified by the availability of financial fraud. The RF model is a bootstrapping model that produces various decision trees and the classification results are produced by majority voting; grid search is used to optimize the tree depth and minimum leaf node sample size. The SVM uses radial basis function kernel as nonlinear mapping in order to have optimal boundary partitioning by tweaking the parameters of penalty factors and the width of the kernel. The NN part constructs a three-layer fully connected structure, with the number of input nodes determined by the number of principal components, and the hidden layers using the ReLU activation function. The learning rate weights are adjusted using the Adam optimizer. The prediction results of the three models are fused using a weighted average. The dynamic weight ω_i for each model m_i is determined by its validation accuracy Acc_i during cross-validation, expressed as

$$\omega_i = \frac{Acc_i}{\sum_{j=1}^3 Acc_j}.$$

To prevent bias reinforcement from a single learner, a smoothing penalty is applied to ensure no single model accounts for more than 50% of the total decision weight.

D. PARAMETER OPTIMIZATION

The parameter tuning is done with the help of cross-validation (CV) and cross-validation (CV). The dataset is partitioned into five equal segments, four of which are incorporated into the training set, and one of which is utilized in the validation set in every training round. The performance benchmark is calculated and is the average accuracy after iterative iteration. Grid Search was employed for the Random Forest model due to its discrete and smaller hyperparameter space, ensuring an exhaustive search for optimal tree depth. Conversely, Bayesian Optimization was selected for the Neural Network to efficiently navigate its continuous and high-dimensional parameter space (e.g., learning rate and batch size), reducing the computational overhead required for convergence. During the training, the loss function is watched in real-time to limit the convergence rate and the risk of overfitting. The fusion model weighting coefficients are dynamically set depending on the cross-validation performance of each single model in terms of classification accuracy and the optimum ratio is established with a weighted average error minimization criterion.

IV. RESULTS AND DISCUSSION

A. RECOGNITION ACCURACY

The experiment has been carried out in the environment of the Windows 11 system, and the hardware setup included an Intel Core i9 processor, 32GB of RAM, and an NVIDIA RTX 4090 graphics card. The Python 3.10 and TensorFlow 2.10 programming environment were used. The sample data included 5 years of financial data of A-share listed firms in Shanghai and Shenzhen with a total of 1560 groups. The training ratio of the model was 80 percent and the testing ratio was 20 percent. Control algorithms were chosen as the Random Forest (RF), Support Vector Machine (SVM), and the Neural Network (NN). Additionally, standard benchmarks for tabular financial data, specifically Logistic Regression (LR) and Extreme Gradient Boosting (XGBoost), were included to evaluate the fusion model against traditional linear and advanced gradient-boosting baselines. The fusion model was based on a weighted

average which was used in order to integrate the output results of the three algorithms. The experiments were done under the same initial parameters and data splitting. Each model was simulated ten times with the mean value being considered to minimize the fluctuations.

Table 2 shows that under full-sample testing, the fusion model achieved an accuracy of 94.7%, outperforming the single-algorithm models of 83.2% (RF), 85.6% (SVM), and 89.1% (NN). The model performed best in the manufacturing and large-cap enterprise samples, reaching over 95%, indicating that the model has higher applicability and stability under complex financial structures and heterogeneous samples, avoiding the overfitting and bias clustering problems of single algorithms.

TABLE 2. Recognition accuracy (%) of different algorithm models in different scenarios

Scene	RF	SVM	NN	Fusion Model
Full sample test	83.2	85.6	89.1	94.7
Non-manufacturing sample	82.4	84.1	88.5	93.9
Manufacturing Sample	84.6	86.8	89.7	95.1
Sample of highly indebted enterprises	81.9	83.2	87.9	94.2
Large-cap company sample	85.7	87.4	90.2	95.6

B. FALSE POSITIVE RATE

The false positive rate test used the same dataset and algorithm configuration as the accuracy test, and the misclassification rate was calculated using a confusion matrix. The experimental results were averaged from ten independent runs to eliminate random error. RF, SVM, NN, and a fusion model were used as control algorithms to analyze the model's misclassification performance across different enterprise types. Table 3 shows that the misclassification rate of the fusion model is generally controlled within 5%, demonstrating greater robustness in complex enterprise structure samples. It also maintains low error output in high-debt and mixed industry scenarios, indicating that the fusion strategy has better generalization ability and recognition accuracy under imbalanced sample conditions. The model's integrated weighting mechanism plays a significant role in reducing misclassification and balancing learning bias.

TABLE 3. False positive rate (%) of different algorithm models in different scenarios

Scene	RF	SVM	NN	Fusion Model
Full sample test	9.6	8.4	6.3	4.8
Non-manufacturing sample	10.2	8.9	6.9	4.2
Manufacturing Sample	9.1	7.8	6.1	4.5
Sample of highly indebted enterprises	10.8	9.6	7.2	4.1
Large-cap company sample	8.4	7.1	5.8	4.3
Mixed industry samples	9.7	8.2	6.5	4.7

C. FEASIBILITY OF PRACTICAL APPLICATION

The fusion model has a high practicality with regard to the accuracy of identification and false positive control. However, a 5% false positive rate implies that in a sample of 312 companies, approximately 15 firms may be

flagged erroneously. In a practical audit context, these 'false alarms' necessitate a tiered verification process to mitigate potential reputational damage to listed entities and minimize the unnecessary allocation of regulatory investigative resources. The input data of the model are the annual financial statements of companies and publicly available information pertaining to governance of companies that can be easily integrated with the current audit information systems without having to acquire separate data acquisition modules. The calculation algorithm in a high-performance server obtains the response time of minutes and the average time to complete a batch calculation of 300 companies is 42 seconds, which adheres to the real-time early warning needs of regulatory organizations. The system is capable of establishing risk levels, according to the output of the model, automatically detecting possible fraudulent firms and creating risk level reports, which may provide quantitative criteria of judgment by the audit departments. Repeated simulations tests reveal that the identification accuracy of the model varies by at most 2 percent with various samples of a year, which is a good result of stability to intertemporal variation of financial structure and industry difference.

V. CONCLUSION

The fusion identification model, developed upon artificial intelligence algorithms, achieves high accuracy and stability in detecting financial fraud within listed enterprises by incorporating high-dimensional material morphology data and industrial manufacturing parameters. This integrated approach ensures that the data-driven precision necessary for manufacturing excellence is mirrored in the financial oversight and technical audit mechanisms of the modern sector. With the creation of a system of multi-dimensional financial indicators and dimensionality reduction with the help of the principal component analysis, the interference amongst the indicators is removed to a large extent, which increases the efficiency in feature expression. The algorithms of random forest, support vector machine, and neural network are used complementary, and they play a huge part of the fusion structure in such a manner that they substantially minimize the model bias and the risk of overfitting. Tests demonstrate that the overall model has an accuracy rate of 94.7, and false positive rate suppressed to below 5 per cent, which is very high compared to individual algorithm models. This process allows quickly and effectively screening of anomalies in large volumes of financial data, having a practical application value of risk discovery and internal auditing of listed companies. This framework provides regulatory authorities with intelligent analysis tools to facilitate financial transparency and enhance market control over listed textile enterprises by integrating fiber morphology data and industrial manufacturing parameters into the oversight process. This multidimensional approach ensures that the high-dimensional data perception necessary for manufacturing excellence is mirrored in the financial integrity and technical supervision of the modern textile sector.

AUTHOR CONTRIBUTIONS

Tianyue Lyu designed, collected and analyzed the data, and drafted the manuscript. Tianyue Lyu conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Tianyue Lyu participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] Q. Ye, F. Ye, and S. Huang, "Construction of a framework for identifying financial fraud-based on accounting information system theory and big data perspective," *Accounting Research*, no. 3, pp. 3-16, 2022, doi: 10.3969/j.issn.1003-2886.2022.03.001.
- [2] S. Tao, H. Zhong, F. Liu, et al., "Research on intelligent financial fraud identification based on three-dimensional integrated index system and residual neural network," *Operations Research and Management*, vol. 33, no. 11, pp. 145-151, 2024, doi: 10.12005/orms.2024.0366.
- [3] C. Liang, M. Zhang, and S. Gu, "Machine learning-driven financial fraud identification model," *Statistics and Applications*, vol. 14, no. 11, pp. 236-248, 2025, doi: 10.12677/sa.2025.1411326.
- [4] X. Zhang and Y. Shi, "Research on Financial Fraud Detection Based on Meta-learning," *Journal of Management Science*, vol. 26, no. 10, pp. 95-113, 2023, doi: 10.19920/j.cnki.jmsc.2023.10.006.
- [5] F. Ye, Q. Ye, and S. Huang, "Identification and Response to Financial Fraud in the Environmental Governance Industry-A Case Study Based on Botian Environment," *Accounting Monthly*, vol. 46, no. 20, pp. 96-101, 2025, doi: 10.19641/j.cnki.42-1290/f.2025.20.015.
- [6] L. Hernandez Aros, L. X. Bustamante Molano, F. Gutierrez-Portela, et al., "Financial fraud detection through the application of machine learning techniques: a literature review," *Humanities and Social Sciences Communications*, vol. 11, no. 1, pp. 1-22, 2024, doi: 10.1057/s41599-024-03606-0.
- [7] P. Kamuangu, "A review on financial fraud detection using AI and machine learning," *Journal of Economics, Finance and Accounting Studies*, vol. 6, no. 1, pp. 67-77, 2024, doi: 10.32996/jefas.2024.6.1.7.
- [8] R. Udayakumar, A. Joshi, S. S. Boomiga, et al., "Deep fraud Net: A deep learning approach for cyber security and financial fraud detection and classification," *Journal of Internet Services and Information Security*, vol. 13, no. 3, pp. 138-157, 2023, doi: 10.58346/JISIS.2023.14.010.
- [9] Y. Alghofaili, A. Albattah, and M. A. Rassam, "A financial fraud detection model based on LSTM deep learning technique," *Journal of Applied Security Research*, vol. 15, no. 4, pp. 498-516, 2020, doi: 10.1080/19361610.2020.1815491.
- [10] R. Li, Z. Liu, Y. Ma, et al., "Internet financial fraud detection based on graph learning," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 3, pp. 1394-1401, 2022, doi: 10.1109/TCSS.2022.3189368.