Research on Feature Selection Method for Enterprise Financial Crisis Early Warning Based on LRRF-MIC Integrated Screening Approach

Yuan Yan[1,2], Yujian Cui[3,*], Sheng Zhou[4], Ling Zhao[5]

[1]School of Business, Central South University of Forestry and Technology, Changsha, China
[2]Collaborative Innovation Center, Hunan Automotive Engineering Vocational University, Zhuzhou, China
[3]School of Management, Hunan Automotive Engineering Vocational University, Zhuzhou, China
[4]School of Continuing Education, Hunan Automotive Engineering Vocational University, Zhuzhou, China
[5]Changsha Commerce & Tourism College, Changsha, China
**Corresponding author: 13973365607@163.com*

Abstract: *In capital markets, Special Treatment (ST) designation for listed companies stems from multiple complex factors. To enable early ST risk identification, this study explores integrated machine learning methods combining feature selection and predictive models, focusing on comparative evaluation of feature selection techniques through empirical analysis. Using a sample of 430 Chinese A-share listed firms with financial/non-financial indicators, 35 key indicators distinguishing distressed vs. healthy firms are first identified via statistical screening.We propose an innovative "LRRF-MIC Integrated Screening Method" integrating Lasso regression, Recursive Feature Elimination (RFE), Random Forest (RF), and Maximal Information Coefficient (MIC). This hybrid framework generates multi-dimensional feature evaluations and visual analytics, systematically selecting 14 core predictive indicators based on established criteria. Empirical validation uses these features as inputs for MLP and gcForest models, compared with benchmark models using raw data. Results show the LRRF-MIC framework outperforms single methods (RF, RFE, MIC, Lasso) by over 8% in prediction accuracy on average and benchmark models by 13%, demonstrating the efficacy and innovation of the proposed integrated approach.*

Keywords: *Special Treatment (ST) Risk, Feature Selection, Machine Learning Models, LRRF-MIC Method*

1. Introduction

The rapid development of China's capital markets and the deepening of economic globalization have amplified the significance of listed companies within the global economy. However, the outbreak of the global financial crisis, the impact of the COVID-19 pandemic, and uncertainties in both domestic and international economic environments have substantially increased the financial risks faced by enterprises. Financial distress not only jeopardizes corporate survival and development but also inflicts significant economic losses on stakeholders, including investors, creditors, and employees. Consequently, establishing effective financial crisis early warning mechanisms has become a critical research focus for both academia and practitioners.

Recent advancements in machine learning have revolutionized financial crisis prediction. Domestic scholars have developed various models, such as the SMOTE-XGBoost combination model [14], CNN-based feature extraction frameworks [12], Logistic-Cox comparative analysis models [16], LightGBM ensemble algorithms [9], and Stacking-Bagging-Vote ensemble learning models [15], significantly enhancing the accuracy of financial distress warnings. Internationally, Dastkhan constructed systemic risk network indicators based on Conditional Value at Risk (COVaR) [3]; Tsai et al. demonstrated the critical role of data preprocessing through multi-algorithm comparisons [5]; Jemović et al. analyzed financial crises using panel data [4]. Abdalzaher et al. developed a novel financial distress prediction model (AWOA-DL) using deep learning techniques integrated with an Adaptive Whale Optimization Algorithm [1]; Yao et al. proposed a Sequential Backward Feature Selection algorithm based on Ranking Information (SBFS-RI) and an Ensemble Feature Selection method Fusing Multiple Ranking Information

(FS-MRI) [8]; and Wu et al. introduced a novel hybrid corporate crisis warning model combining the Z-score model with an MLP-ANN model [6].

Despite these advances, current research exhibits two key limitations. First, the singularity of feature selection methods. While existing techniques like Lasso and Recursive Feature Elimination (RFE) demonstrate utility in feature screening [7, 11], they fail to address biases arising from the coupling of linear and nonlinear features or algorithmic preferences, this undermines the robustness of feature subsets in capturing complex financial risk patterns. Second, the dimensional constraints of indicators. Existing models overly rely on financial metrics [2], neglecting the dynamic interrelationships of non-financial factors such as Corporate Social Responsibility (CSR) and management quality [13].

To address these gaps, this study focuses on China's A-share listed companies and innovatively proposes the LRRF-MIC integrated feature screening method. This approach synergistically combines the strengths of Lasso, RFE, Random Forest (RF), and Maximal Information Coefficient (MIC) to effectively overcome the limitations of single methods. The research first identifies 48 initial indicators across six dimensions: solvency, operational capability, cash flow capacity, profitability, growth potential,and social responsibility. Statistical analysis screens these down to 35 significant features. Subsequently, we apply the LRRF-MIC method to extract 14 core features exhibiting high discriminatory power. The refined feature set is then input into Multilayer Perceptron (MLP) and Deep Forest (gcForest) models for prediction.Empirical results demonstrate that LRRF-MIC significantly enhances the prediction accuracy of the warning models compared to single feature selection methods (RF, RFE), with an improvement up to 8%. The improvement is even more substantial (up to 13%) when benchmarked against models without feature selection. Statistical tests ($p < 0.05$) confirmed the significance of these differences.This study provides a novel, more robust feature selection method (LRRF-MIC) for financial crisis early warning. Its findings also offer direct practical value: optimizing dynamic risk assessment for the Special Treatment (ST) system; assisting investors in constructing risk-resistant portfolios; and guiding enterprises to enhance financial resilience by strengthening CSR practices.

2. Methods

Feature selection is a crucial step in the construction of machine learning models. Its objective is to extract the feature subset with the maximum amount of information and the minimum redundancy from high-dimensional data. The LRRF-MIC comprehensive screening method proposed in this paper aims to overcome the limitations of single methods by integrating the advantages of four types of methods: LASSO, Recursive Feature Elimination (RFE), Random Forest (RF), and Maximal Information Coefficient (MIC).

2.1 LRRF-MIC Comprehensive Screening Method

Feature selection aims to screen out a feature subset with high discriminability from the candidate features. Its significance lies not only in dimensionality reduction, alleviating overfitting, and enhancing the generalization ability of the model, but also in enhancing the interpretability of the model, accelerating the training efficiency, and ultimately optimizing the prediction performance. Currently, there are various feature selection methods in the field of machine learning, but each has its own applicable scenarios and limitations. Therefore, this paper proposes the LRRF-MIC comprehensive screening method, which integrates the advantages of the following four types of methods:

2.1.1 LASSO (Embedded Method)

The LASSO (Least Absolute Shrinkage and Selection Operator) method is a shrinkage estimation approach. By introducing an L1 - norm penalty term, LASSO imposes a sparsity constraint on the regression coefficients during the optimization process. It shrinks some coefficients to zero, thereby achieving automatic feature selection and model parsimony.

2.1.2 Recursive Feature Elimination (RFE, Wrapper Method)

Recursive Feature Elimination (RFE) was proposed by Guyon et al. It iteratively trains a base estimator (such as a support vector machine or logistic regression). In each iteration, it removes the feature that contributes the least to the model until the preset number of features is reached. This method requires updating the importance ranking metric at each step of the algorithm. Its core idea is to repeatedly build a model (such as a support vector machine model or a regression model), then select the best (or worst) features based on criteria like coefficients, and repeat this process for the remaining features until all features have been traversed, ultimately achieving feature selection.

2.1.3 Random Forest Feature Importance (Model - Based Method)

A random forest is a classifier composed of multiple decision trees. The output class is determined by the mode of the output classes of each decision tree. The principle of calculating feature importance in a random forest is based on the decrease in Gini impurity or Mean Decrease in Impurity (MDI). A random forest quantifies feature importance, which reflects the degree to which a feature reduces the classification uncertainty when splitting nodes.

2.1.4 Maximal Information Coefficient (MIC, Filter Method)

The Maximal Infor4mation Coefficient (MIC) can effectively measure the linear or nonlinear correlation strength between two variables. MIC evaluates the statistical dependence between a feature and the target variable by calculating the standardized value of the maximum mutual information between variables. Its value ranges from 0 to 1, with a larger value indicating a stronger correlation.Single feature selection methods may lead to biases due to algorithmic preferences. For example, LASSO tends to select linearly correlated features, while MIC is good at capturing nonlinear relationships. By integrating the screening results of the four types of methods, LRRF - MIC can balance stability and diversity and enhance the robustness of the feature subset.

2.2 Deep Forest Model

gcForest (Multi-Grained Cascade Forest) is a deep ensemble learning framework proposed by Professor Zhi-Hua Zhou's research team in 2017[10]. Its core mechanism simulates the hierarchical feature abstraction capability of deep learning through a cascaded multi-layer architecture. The model comprises stacked cascade forest layers, each containing multiple random forests and completely random forests. Data is iteratively processed across these layers, with dynamically enhanced feature representations generated at each stage. Specifically, the class probability vectors output by each layer are concatenated with the original features and propagated to subsequent layers, enabling multi-level feature interactions akin to neural networks. For instance, a 100-dimensional feature vector can expand to 104 dimensions after the first layer in a 4-class classification task. Prior to cascade processing, a multi-grained scanning mechanism mimics convolutional operations by partitioning the feature space with varying window sizes, extracting local patterns that are aggregated into enriched feature representations. The training process incorporates adaptive depth control—expansion terminates when the validation accuracy improvement from adding new layers falls below a predefined threshold (e.g., 0.5%), effectively mitigating overfitting (typically cascading 3–8 layers). This framework overcomes the limitations of conventional ensemble learning by integrating the interpretability of tree-based models with the representational capacity of deep architectures, demonstrating exceptional performance, particularly in small-sample scenarios.(Fig 1).

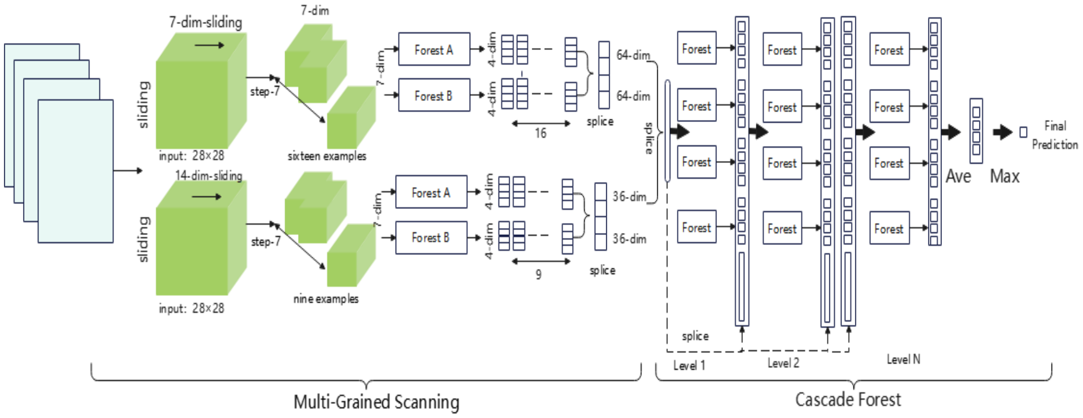

Fig.1 Deep Forest Model Flowchart

2.3 MLP Machine Learning Method

The Multi - Layer Perceptron (MLP) is a classic feed - forward neural network. Its structure consists of an input layer (which matches the feature dimensions), hidden layers (where non - linear feature extraction is achieved through activation functions such as ReLU or Sigmoid), and a task - oriented output layer (e.g., Sigmoid for binary classification or Softmax for multi - class classification).Its training is based on the error back - propagation algorithm: during forward propagation, the predicted values are calculated; the loss function (e.g., cross - entropy) quantifies the error; and during back - propagation, the weights are updated through the chain rule (e.g., using stochastic gradient descent) to minimize the prediction deviation.In this study, MLP is used as a benchmark model. By comparing the performance of different feature subsets with that of gcForest, the generalization ability of the feature selection method in traditional neural networks is verified, and the relationship between model robustness and feature quality is revealed.

2.4 Model Result Evaluation Method

Following model development, comprehensive evaluation is imperative. Given the characteristics of financial distress prediction as a binary classification task, this study establishes a multidimensional evaluation framework encompassing core discriminative dimensions. Listed companies under Special Treatment (ST) are designated as positive class samples (financial distress), while normally operating firms constitute the negative class. A confusion matrix framework is implemented, delineating four diagnostic categories: true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). The evaluation system incorporates multiple metrics: Classification Accuracy, Recall (Sensitivity), Precision (Positive Predictive Value), and Precision-Recall (P-R) curves.Their calculation formulas are as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \tag{1}$$

$$\text{Re}\,call = \frac{TP}{TP+FN} \times 100\% \tag{2}$$

$$\text{Precision} = \frac{TP}{TP+FP} \times 100\% \tag{3}$$

3. Results

3.1 Sample Selection and Preliminary Indicator Screening

This study employs the Special Treatment (ST) designation by the Shanghai and Shenzhen stock exchanges as the criterion for identifying financially distressed firms. Through rigorous data screening procedures, 215 ST-labeled companies from 2019 to 2021 were selected as the financial distress cohort. A matched sample of 215 healthy firms from the same industries with comparable total share capital was concurrently chosen to ensure comparability. The dataset sourced from the CSMAR Database and Wind Database. During the preliminary indicator selection phase, a comprehensive set of 41 indicators was initially identified—34 financial metrics and 7 corporate social responsibility (CSR) metrics—to holistically assess multifactorial influences on financial distress and validate dimensionality reduction efficacy. These indicators span seven critical dimensions: operational capacity, solvency, profitability, working capital efficiency, cash flow adequacy, growth potential, and social accountability (Table 1).

3.2 Statistical Validation of Initial Indicators

Data quality and discriminative power are foundational to reliable financial distress prediction models.To address potential data inconsistencies and enhance model robustness, a rigorous data-cleaning protocol was implemented. Central to this process is verifying significant intergroup differentiation between distressed and healthy firms across all candidate indicators. Statistical analyses were conducted using SPSS 24.0 to ensure methodological rigor.

The validation workflow comprised two stages:Normality Testing: A one-sample Kolmogorov-Smirnov (KS) test was applied to assess compliance with normal distribution assumptions—a prerequisite for parametric statistical methods. Indicators demonstrating normality ($p>0.05$) were retained for subsequent analysis.Inter group discrimination: For normally distributed indicators, independent samples t-tests were performed to evaluate mean differences between ST and non-ST groups.Indicators exhibiting statistically significant divergence($p<0.05$) were prioritized, as summarized in Table 1.This dual-phase approach ensures the selected indicators possess both distributional validity and discriminative power, forming a robust foundation for model development.

Table 1 Descriptive Statistics and Nonparametric Tests

Variable Code	Variable Name	Minimum	Maximum	Standard Deviation	p-value
X1	Accounts Receivable Turnover Ratio	0.076	336806.144	9352.735	0
X2	Inventory Turnover Ratio	0.005	411160.927	11608.965	0.026
X3	Accounts Payable Turnover Ratio	0.019	13744.347	386.056	0.036
X4	Cash and Cash Equivalents Turnover Ratio	0.008	313.148	11.255	0.056*
X5	Current Assets Turnover Ratio	0.004	20.634	1.167	0
X6	Capital Intensity	0.116	250.104	10.897	0
X7	Total Assets Turnover Ratio	0.004	8.455	0.507	0
X8	Current Ratio	0.093	29.080	2.293	0.037
X9	Quick Ratio	0.034	24.828	1.986	0.226*
X10	Cash Ratio	0.003	12.900	1.070	0.001
X11	Debt-to-Assets Ratio	0.019	1.352	0.226	0.042
X12	Equity Ratio	-340.171	231.651	13.284	0.710*
X13	Operating Current Liabilities Ratio	-2.105	2.642	0.328	0
X14	Net Cash Content of Net Profit	-84.114	677.465	32.760	0
X15	Net Cash Content of Operating Revenue	-24.884	10.251	1.167	0
X16	Net Cash Content of Operating Profit	-414.563	1095.301	40.053	0
X17	Total Cash Recovery Rate	-1.686	0.438	0.102	0
X18	Operating Index	-465.157	1171.529	37.272	0
X19	Cash Suitability Ratio	-1322.374	7758.487	218.293	0
X20	Cash Reinvestment Ratio	-18.673	16.186	1.200	0
X21	Cash Sufficiency Ratio for Investment	-272.221	63.282	9.776	0
X22	Return on Assets	-0.768	0.501	0.075	0
X23	Net Profit Rate on Total Assets	-0.775	0.431	0.070	0
X24	Net Profit Rate on Current Assets	-1.559	3.764	0.215	0
X25	Net Profit Rate on Fixed Assets	-12804.380	380.930	353.532	0
X26	Return on Equity	-14.819	8.670	0.522	0
X27	Return on Invested Capital	-1.548	14.284	0.415	0
X28	Return on Long-term Capital	-14.149	6.962	0.470	0
X29	Return on Investment	-42.839	1055.070	41.320	0.333*
X30	Capital Preservation and Appreciation Rate	-1.625	232.596	7.898	0.031
X31	Capital Accumulation Rate	-2.625	231.596	7.898	0.031
X32	Growth Rate of Total Operating Revenue	-0.873	96.024	4.153	0.001
X33	Sustainable Growth Rate	-16.852	1.486	0.532	0
X34	Growth Rate of Net Assets per Share	-6.983	32.856	1.763	0.005
C1	Shareholder Contribution Rate	0.000	0.169	0.019	0
C2	Interest Payment Rate	0.000	2.142	0.115	0.670*
C3	Employee Contribution Rat	0.000	5.787	0.203	0.928*
C4	Supplier Contribution Rate	0.000	10.217	0.556	0.007
C5	Consumer Contribution Rate	0.000	8.165	0.482	0
C6	Tax Contribution Rate	-0.132	3.787	0.181	0.003
C7	Social Donation Rate	-0.211	4.758	0.179	0

Note: * indicates that the indicator is not significant at the significance level of 0.05.

As evidenced by the non-parametric test results in Table 1, the majority of indicators demonstrated statistically significant discriminative power between ST and non-ST firms at the 0.05 significance level. However, six metrics—Cash and Cash Equivalents Turnover Ratio (X4), Quick Ratio (X9), Equity Ratio (X12), Return on Investment (X29), Interest Coverage Ratio (C2), and Employee Contribution Rate (C3)—exhibited p-values exceeding 0.05, indicating insufficient differentiation capability between the two groups. In accordance with statistical theory, these metrics were subsequently excluded from further analysis. The remaining 35 indicators exhibited statistically significant intergroup disparities ($p < 0.05$), forming a refined feature set for subsequent modeling and validation.

3.3 Feature Selection Methodology

This study establishes an LRRF-MIC comprehensive evaluation framework integrating financial and corporate social responsibility (CSR) dimensions, implementing a multi-algorithm consensus feature selection protocol through the following workflow:

Phase I: Multidimensional Feature Importance Quantification Four distinct algorithms—Lasso regression (L), Recursive Feature Elimination (RFE), Random Forest (RF), and Maximal Information Coefficient (MIC)—were independently applied to quantify feature weights. This generated four sets of importance rankings, capturing algorithm-specific discriminative patterns.

Phase II: Dynamic Importance Stratification A tertile-based classification system was implemented:

Core Features: Weight values exceeding the upper tertile threshold (>66.7th percentile)

Auxiliary Features: Weights between lower and upper tertiles (33.3–66.7th percentile)

Candidate Exclusion Features: Weights below the lower tertile (<33.3th percentile)

Phase III: Consensus-Driven Feature Retention A cross-algorithm validation protocol was enforced using elimination thresholds:Elimination Criterion: Features labeled as "Candidate Exclusion" in ≥2 algorithms Retention Criteria (meeting either condition):Non-exclusion status (Core/Auxiliary) in ≥3 algorithms Single-exclusion status (Candidate in 1 algorithm) with Core/Auxiliary ratings in others. This triphase methodology ensures robust feature selection by harmonizing algorithmic diversity with statistical rigor, effectively mitigating single-algorithm bias while preserving critical predictive signals.

3.4 LRRF-MIC Integrated Screening Results

The LRRF-MIC methodology retained 14 critical features, categorized as follows (see Table 2).Social Responsibility: Shareholder Contribution Rate (C1), Tax Contribution Rate (C6); Operational Capacity: Accounts Receivable Turnover (X1), Accounts Payable Turnover (X3); Solvency: Current Ratio (X8); Cash Flow Adequacy: Total Cash Recovery Rate (X17), Operating Index (X18), Cash Reinvestment Ratio (X20), Cash-to-Investment Coverage Ratio (X21); Profitability: Fixed Asset Net Profit Margin (X25), Long-Term Capital Return Rate (X28); Growth Potential: Capital Accumulation Rate (X31), Sustainable Growth Rate (X33), Net Asset per Share Growth Rate (X34).

Table 2 Importance of Feature Combinations.

Variable Code	lasso	RFE	RF	MIC	Whether to retain
C1	0	3	0.026	0.131	1
C4	0	12	0.022	0.133	0
C5	0	13	0.022	0.122	0
C6	0	7	0.042	0.114	1
C7	0	29	0.031	0.145	0
X8	0	7	0.022	0.111	1
X3	-1.18E-05	1	0.023	0.121	1
X5	0	24	0.022	0.146	0
X6	0	32	0.022	0.151	0
X7	0	23	0.023	0.149	0
X22	0	22	0.026	0.167	0
X23	0	21	0.024	0.152	0
X24	0	20	0.024	0.140	0

X25	-6.33E-05	1	0.028	0.147	1
X26	0	19	0.020	0.166	0
X27	0	18	0.021	0.160	0
X28	-0.0002854	1	0.020	0.145	1
X14	-4.32E-05	1	0.019	0.140	0
X15	0	28	0.023	0.142	0
X16	-6.78E-05	1	0.014	0.141	0
X17	0	4	0.041	0.176	1
X18	0	1	0.026	0.150	1
X19	0	31	0.022	0.133	0
X10	0	27	0.026	0.131	0
X20	0	5	0.030	0.156	1
X21	0	6	0.024	0.152	1
X30	0	14	0.031	0.145	0
X31	0	8	0.031	0.145	1
X32	0	15	0.040	0.170	0
X33	0	9	0.031	0.158	1
X34	0	10	0.036	0.160	1
X11	0	26	0.024	0.117	0
X13	0	25	0.030	0.160	0
X1	-2.45E-07	1	0.030	0.136	1
X2	2.02E-06	30	0.022	0.111	0

Remarks: In the column of "Whether to Retain", 1 indicates retention, and 0 indicates elimination.

Based on the multi-algorithm fusion of LRRF-MIC, this study constructs a multidimensional financial distress early-warning system using 14 rigorously selected indicators. The corporate social responsibility indicators C1 and C6 overcome limitations of traditional models by identifying governance deficiencies and operational compliance risks. The working capital indicators X1, X3, and cash flow indicators X17–X21 establish a dual-monitoring mechanism for supply chain liquidity and cash flows, significantly enhancing anomaly detection rates compared to single-dimensional cash flow metrics. Solvency indicator X8 and profitability indicators X25, X28 cover critical nodes of short- and long-term risk transmission. Development capacity indicators X31, X33, and X34 dynamically track capital accumulation and sustainable growth parameters, enabling 6–12 month advance warnings of financial deterioration triggered by growth stagnation. Crucially, the interaction effect between X20 and X33 detects "illusory prosperity"-type crises.This integrated approach significantly enhances identification accuracy for major risks—including cash flow mismatches and earnings manipulation—while establishing a highly sensitive parametric framework for dynamic early-warning.

3.5 Results of the Empirical Analysis of the Model

Figure 2 demonstrates the comparison results of precision-recall (P-R) curves for five feature selection methods and the original data method on the MLP and gcForest models.

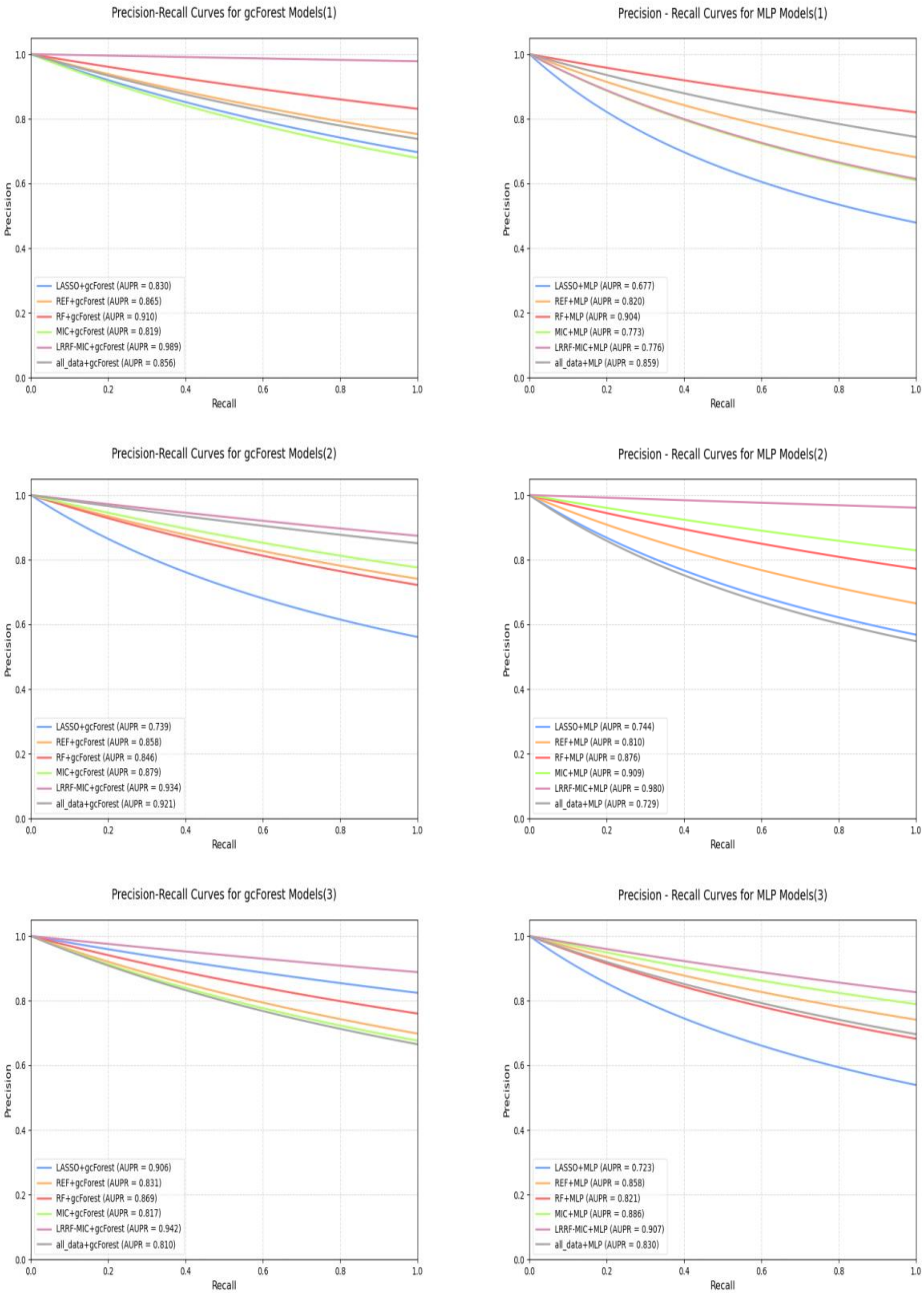

Fig.2 Precision-Recall (P-R) Graphs of Five Groups of Data in the gcForest and MLP Classifiers

The results demonstrate that LRRF-MIC exhibits remarkable superiority over other feature selection methods, notably achieving exceptional performance on gcForest with an Area Under the Precision-Recall Curve (AUPR) approaching 1.0. It also significantly outperforms the original data approach on MLP, validating its capabilities in multi-strategy evaluation and complex data adaptation. The necessity of model feature screening is further confirmed: all feature selection methods enhance model performance, and LRRF-MIC specifically maintains high AUPR after dimensionality reduction (from 35 to 14 features), proving the effectiveness of its "redundancy elimination and generalization enhancement" mechanism. Collectively, the precision-recall (P-R) curves visually demonstrate how feature screening optimizes both model accuracy and robustness.

To further present the average classification performance metrics (including accuracy, recall, precision, AUC, and AUPR) of each method under five-fold cross-validation, this paper's empirical analysis obtained the comparative values of various performances for the two classifiers in Table 3 and Table 4.

Table 3 Performance of Feature Selection Methods under the MLP Classifier (Average Values of the Data).

Reduction_Method +MODEL_NAME	Average ACC	Average AUC	Average Precision	Average Recall	Average AUPR
LASSO+MLP	0.570632911	0.61340137	0.594349206	0.499425558	0.606556907
REF+MLP	0.615632911	0.675728156	0.602569468	0.691818182	0.680439219
RF+MLP	0.693449367	0.739349954	0.711032873	0.660408505	0.752541829
MIC+MLP	0.678322785	0.705842936	0.691963576	0.641449275	0.705252156
LRRF-MIC+MLP	0.762797468	0.833717458	0.776496872	0.738171392	0.82038468
all_data+MLP	0.65585443	0.690300496	0.656603397	0.63087552	0.694946794

Table 4 Performance of Feature Selection Methods under the gcForest Classifier (Average Values of the Data).

Reduction_Method +MODEL_NAME	Average ACC	Average AUC	Average Precision	Average Recall	Average AUPR
LASSO+gcforest	0.668259494	0.708386515	0.670244613	0.667923968	0.673690258
REF+gcforest	0.685981013	0.738581973	0.699155811	0.649043062	0.726049799
RF+gcforest	0.708449367	0.786521191	0.721770626	0.707127749	0.767296104
MIC+gcforest	0.688386076	0.743575931	0.668435731	0.742694673	0.735120358
LRRF-MIC+gcforest	0.85021519	0.947501898	0.84638961	0.882392177	0.904861433
all_data+gcforest	0.723829114	0.808776253	0.726909637	0.738921396	0.768367815

Empirical analysis demonstrates that the LRRF-MIC method significantly outperforms the other four feature selection methods in overall performance, with its accuracy being on average more than 10% higher than those of other methods and showing a 13% improvement over the benchmark model without feature screening (taking gcForest as the base model). Additionally, the ensemble classifier gcForest exhibits remarkably better classification performance than the MLP classifier, verifying the adaptability of the LRRF-MIC method to different models. In summary, through streamlining variables and reusing key features, the LRRF-MIC comprehensive screening method effectively improves prediction performance while reducing computational complexity.

4. Conclusion

This study investigates Chinese A-share listed companies, initially identifying 48 indicators across six dimensions: solvency, operational efficiency, cash flow adequacy, profitability, growth potential, and social responsibility. Statistical analysis refined these to 37 discriminative features, which were further screened using Random Forest (RF), Recursive Feature Elimination (RFE), Maximal Information Coefficient (MIC), and Lasso regression. The proposed LRRF-MIC (Lasso-RF-RFE-MIC) Comprehensive Screening Method integrated these techniques to extract 18 optimal features, subsequently employed as inputs for financial crisis prediction using gcForest and Multilayer Perceptron (MLP) models. Key findings are summarized as follows:

(1) Methodological Innovation of LRRF-MIC. By synergizing four feature selection strategies—RF's nonlinear importance evaluation, RFE's iterative optimization, MIC's dependency measurement, and Lasso's sparsity constraints—the LRRF-MIC framework establishes a generalized approach for high-dimensional financial data optimization. Empirical results demonstrate that the integrated method achieves a 13% higher accuracy than the benchmark model without feature selection, validating its capacity to balance algorithmic biases and enhance predictive robustness.

(2) Significance of Social Responsibility Indicators. The retention of nonfinancial metrics, including shareholder contribution rate and tax contribution rate, underscores the critical role of Corporate Social Responsibility (CSR) in financial risk dynamics. Statistical and empirical analyses reveal that CSR

indicators exhibit a strong correlation ($p < 0.01$) with future financial distress, emphasizing the need for enterprises to align profit-driven objectives with sustainable governance practices. Enhanced CSR performance not only mitigates financial risks but also fosters long-term stakeholder trust and capital cost reduction.

(3) Cross-Method Integration and Extensibility. The LRRF-MIC framework exemplifies the value of methodological pluralism in financial analytics. By harmonizing diverse feature selection techniques, this study provides a multidimensional perspective that improves both the depth and reliability of early warning systems. This integrative paradigm is extensible to related domains, such as credit risk assessment and market volatility prediction, where hybrid methodologies can overcome the limitations of singular approaches. Future research may further refine this framework by incorporating adaptive weighting mechanisms or domain-specific constraints.

Acknowledgements

Scientific Research Projects of the Education Department of Hunan Province (24C0978). Natural Science Foundation of Hunan Province (2023JJ50183).

References

[1] Abdalzaher M.S., Soliman M.S., El-Hady S.M., Benslimane A., Elwekeil M. A deep learning model for earthquake parameters observation in IoT system-based earthquake early warning. IEEE Internet of Things Journal, 2021,9(11):8412-8424.

[2] Altman E.I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. Journal of Finance, 1968,23(4):589-609.

[3] Dastkhan H. Network-based early warning system to predict financial crisis. International Journal of Finance & Economics, 2021,26(1):594-616.

[4] Jemović M., Marinković S. Determinants of financial crises—An early warning system based on panel logit regression. International Journal of Finance & Economics, 2021,26(1):103-117.

[5] Tsai C.F., Sue K.L., Hu Y.H., Chiu A. Combining feature selection, instance selection, and ensemble classification techniques for improved financial distress prediction. Journal of Business Research, 2021,130:200-209.

[6] Wu D., Ma X., Olson D.L. Financial distress prediction using integrated Z-score and multilayer perceptron neural networks. Decision Support Systems, 2022,159:113814.

[7] Yanli Z., Xiaolin Y., Wei Q., Yuhan Z., Juntao L. LASSO-based early warning of enterprise's financial crisis and the selection of key indicators. Journal of Henan Normal University (Natural Science Edition), 2016,44(3):1-8.

[8] Yao G., Hu X., Wang G. A novel ensemble feature selection method by integrating multiple ranking information combined with an SVM ensemble model for enterprise credit risk prediction in the supply chain. Expert Systems with Applications, 2022,200:117002.

[9] Yim J., Mitchell H. A comparison of corporate failure models in Australia: Hybrid neural networks, logit models and discriminant analysis. Paper presented at: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, 2003.

[10] Zhou Z.H., Feng J. Deep forest: Towards an alternative to deep neural networks. Paper presented at: Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017.

[11] Huang H., Xu Q.G., Zhang Y.Q., Shi H.W. Financial crisis warning model based on KPCA dimension reduction and weight-LSSVM. Statistics & Decision, 2020,(20):157-161.

[12] Tan Y.Y., Chen J.Y., Sun J. Research on financial crisis warning of listed companies based on CNN. Journal of Southwest China Normal University (Natural Science Edition), 2021,46(5):88-95.

[13] Wu C., Chen X.F., Miao B.W. Research on financial crisis prediction of IT listed companies considering financial network indicators. Operations Research and Management Science, 2023,32(8):159-165.

[14] Wu Z.Y., Jin L.M., Han X.L., Wang Z.L., Wu B. Financial crisis warning model for foreign trade enterprises based on SMOTE-XGBoost. Computer Engineering and Applications, 2024,60(11):281-289.

[15] Zhang L., Liu J.P., Tian D.M. Financial early warning application based on Stacking-Bagging-Vote multi-source information fusion model. Journal of Computer Applications, 2022,42(1):280-286.

[16] Zhao J. Research on financial crisis warning model of listed companies based on logistic and Cox regression models. China Collective Economy, 2023,(10):95-98.