

An Enterprise Financial Management Technique Using Decision Tree with ID3 Algorithm

Chenghang Zhang

Instructor, School of Digital Intelligence and Trade Management, Lishui Vocational and Technical College, Lishui, 323000, China, E-mail: ranran612025@163.com

Project Management

Received January 20, 2026; revised March 18, 2026; accepted September 5, 2026
Available online October 5, 2026

Abstract: In the context of digitalization, enterprise financial management requires precise risk warning. Therefore, an improved generalized iterative binary decision tree model is proposed that enhances robustness by dynamically discretizing continuous attributes, using the information gain ratio criterion, and handling probabilistic missing values. Based on this, a financial risk decision tree model is constructed, which integrates financial indicator weighting, asymmetric misclassification cost-sensitive learning, and a dynamic feedback mechanism. Experiments on the Polish Companies Bankruptcy dataset showed that the accuracy increased by 3.40% compared to the original model, while reducing model complexity by 63.3%. The financial risk decision tree model achieved a high-risk enterprise detection rate of 89.4% and an F1 score of 85.6%, which are 6.6% to 13.6% higher than those of XGBoost, random forest, and support vector machines, while reducing misclassification costs by 43.2%. Managers can use models to accurately identify high-risk enterprises, optimize capital allocation strategies, and implement interventions in advance, thereby effectively reducing potential financial losses.

Keywords: Decision tree; ID3 algorithm; financial risk management; decision support system; enterprise management analysis.

Copyright © Journal of Engineering, Project, and Production Management (EPPM-Journal).
DOI 10.32738/JEPPM-2026-84

1. Introduction

The core goal of corporate financial management is to optimize resource allocation and maximize value, and it is now evolving into a data-driven strategic function, leveraging technologies such as AI (Hasanvand et al., 2023; Zhang and Li, 2022). Data-driven models are shifting finance from "post-event accounting" to proactive control (Hang et al., 2022a; Ang et al., 2022b). Recent studies highlight advanced predictive frameworks: Boyang et al. (2023) proposed a quantum optimization-LSTM model for financial crisis warning, outperforming baselines in accuracy and F1-score. Zeng et al. (2022) developed an IoT-based cash flow warning framework, achieving 91.6% and 75% accuracy in health and crisis states, respectively. Decision trees are widely used in finance for tasks such as credit assessment. Boruah et al. (2025) proposed a compressed hybrid decision tree to reduce rule redundancy, improving classification accuracy and interpretability. Wang et al. (2025) applied a decision-tree-based dynamic-game method to multi-round confrontation scenarios. Dou (2023) integrated cost-benefit trade-offs into a decision tree, achieving 93.2% accuracy for economic protection assessment. The ID3 algorithm is a classic method for building decision trees via information gain. Agarwal et al. (2023) improved ID3 by simplifying the entropy calculation for big medical data, demonstrating better performance than standard ID3 and Random Forest.

Although existing research has made significant progress in enterprise financial management, decision trees, and ID3 algorithms, there are still limitations, including insufficient ability to process real-world data, neglected business costs, limited accuracy, and a lack of continuous evolution capabilities. In response to these issues, the research first proposed an improved Generalized ID3 Decision Tree Model (G-ID3). By introducing a dynamic discretization algorithm for continuous attributes, adopting the information gain ratio as the splitting criterion, and integrating a cost-complexity pre-pruning strategy, this model provides a more robust framework for financial analysis. A Financial Risk Decision Tree Model (FR-DT) is constructed. The innovation of the research lies in proposing a progressive "base-domain" dual-model construction framework, thereby transforming the ID3 algorithm from a theoretical prototype into a robust tool suitable for enterprise financial management. The cost of financial misclassification is integrated into the decision tree growth

mechanism, thereby deeply coupling business logic and information-theoretic principles. A dynamic update mechanism based on confidence feedback is designed, thereby transforming a static prediction model into an adaptive decision system. The research preliminarily integrates financial risk management theory and algorithms at the theoretical level. Firstly, based on the theory of information asymmetry, there are significant differences in the importance of financial indicators in risk assessment. The study introduces prior weights for financial indicators to prioritize core indicators when selecting features for the model. Secondly, drawing on the asymmetric sensitivity of decision-makers to losses in prospect theory, an asymmetric misclassification cost matrix is designed. The business logic of "missing high-risk enterprises with much higher costs than mistakenly detecting low-risk enterprises" is quantified as the model optimization objective, making the model decisions more in line with the actual needs of enterprise risk control. Finally, through a confidence-driven dynamic feedback mechanism, the study expands enterprise risk management from static prediction to a dynamic decision-making process that adapts to environmental changes, providing a theoretical basis for building an intelligent decision-making framework for enterprises with continuous learning capabilities.

2. Methods and Materials

2.1. Construction of an Improved ID3 Model for General Data

Enterprise financial management activities are not purely technical operations, but management decision-making behavior embedded in a complex social and economic network. Its decision-making quality is deeply restricted by classic social science theories such as information asymmetry, managers' limited rationality, and environmental uncertainty. Therefore, the study systematically improves the ID3 algorithm to construct the G-ID3 model by introducing four key techniques: continuous-attribute discretization, gain-rate criterion, missing-value processing, and pre-pruning. While retaining the core advantages of the ID3 algorithm, this model improves its ability to process complex real-world data and provides a reliable basis for subsequent specialized applications in financial management. Continuous numerical attributes are common in real-world datasets, and the original ID3 algorithm cannot handle them. To address this challenge, a dynamic discretization method based on information-gain maximization is proposed. All possible split positions are systematically explored to generate candidate points between ranked values of consecutive attributes, as shown in Eq. (1).

$$T_A = \left\{ \frac{a_i + a_{i+1}}{2} \mid i = 1, 2, \dots, m-1 \right\} \quad (1)$$

In Eq. (1), T_A represents the set of candidate split points for attribute A . a_i represents the i -th value in the sorted sequence of values for attribute A . i represents the index variable. A represents the characteristic or attribute currently being evaluated. m represents the total number of sorted unique values of attribute A . To quantify the ability of each candidate split point to improve data purity, the information gain after bisection is calculated as a threshold, as shown in Eq. (2).

$$IG(S, A, t) = H(S) - \left(\frac{|S_L|}{|S|} H(S_L) + \frac{|S_R|}{|S|} H(S_R) \right) \quad (2)$$

In Eq. (2), $H(\cdot)$ represents the Gini impurity of the data set. $IG(S, A, t)$ represents information gain. S represents the entire training dataset for the current node. t represents the candidate split point being evaluated. To find the split point that best distinguishes categories, the study selects the candidate point with the largest information gain as the final split criterion, as shown in Eq. (3).

$$t^* = \arg \max_{t \in T_A} IG(S, A, t) \quad (3)$$

In Eq. (3), t^* represents the optimal split threshold. This dynamic discretization mechanism enables the model to automatically identify key threshold points for financial indicators, thereby improving the scientific rigor and accuracy of financial indicator processing. The information gain criterion in the original ID3 favors attributes with many values, often leading the model to select irrelevant attributes and causing overfitting. The study uses the information gain rate in place of information gain. First, the split information equation is introduced to measure the dispersion of the attribute's own value, as shown in Eq. (4).

$$IV(S, A) = - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \log_2 \frac{|S_v|}{|S|} \quad (4)$$

In Eq. (4), $IV(S, A)$ represents split information. v represents a possible value of attribute A . S_v represents the sample subset whose attribute A takes the value v . Next, the information gain is divided by its intrinsic information to obtain the information gain rate, as shown in Eq. (5).

$$GR(S, A) = \frac{IG(S, A)}{IV(S, A)} \quad (5)$$

In Eq. (5), $GR(S, A)$ represents the information gain rate. Based on the new standardization criterion, the attribute with the largest gain rate is selected for splitting, thereby improving the robustness, as shown in Eq. (6).

$$A^* = \arg \max_{A \in \text{Attributes}} GR(S, A) \quad (6)$$

In Eq. (6), A^* represents the optimal splitting attribute. In financial analysis, this mechanism can effectively identify and suppress interference from irrelevant or redundant financial indicators, ensuring the model prioritizes core, truly differentiated financial indicators. In summary, the schematic of the gain-rate criterion is shown in Fig. 1.

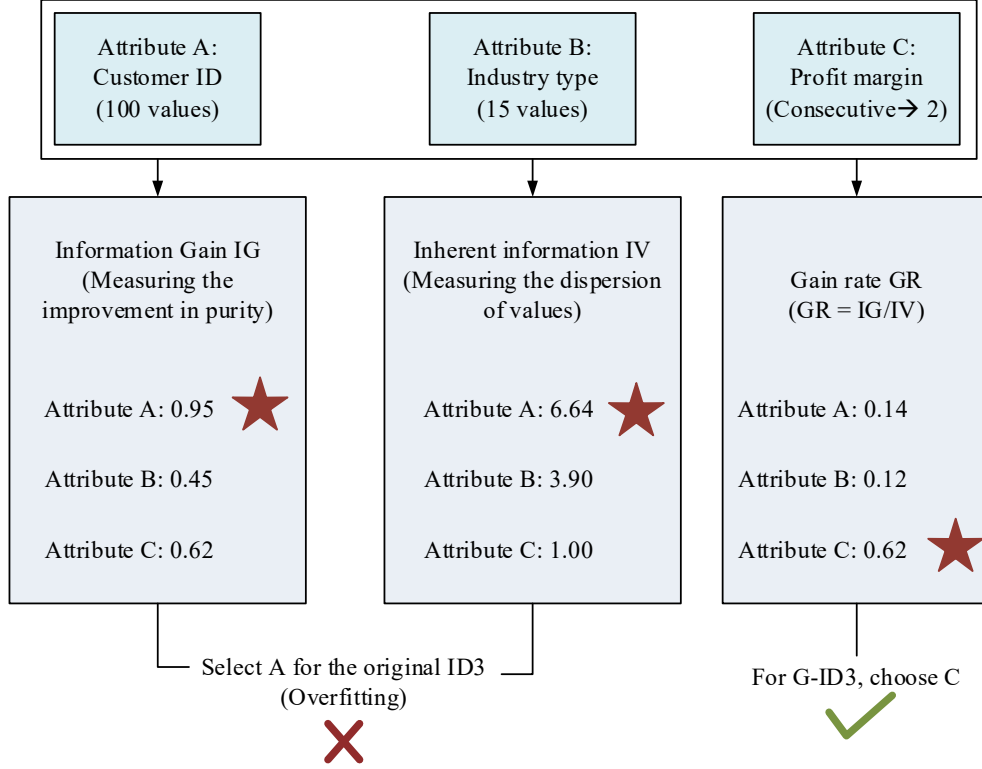


Fig. 1. Schematic diagram of the gain rate criterion

Fig. 1 illustrates feature selection using information gain and gain ratio in a decision tree. The original ID3 algorithm selects features solely based on information gain, which may prefer attributes like Customer ID (information gain=0.95) that have high inherent information (IV=6.64), leading to overfitting. In contrast, the gain ratio criterion (information gain = information gain/IV) introduces an IV penalty, making the Profit Rate attribute optimal (information gain = 0.62) due to its high information gain and low IV, thereby improving generalization. For samples missing an attribute, a weight vector is assigned to the sample based on the known value distribution of the attribute, so that all subsequent child nodes can be weighted, as shown in Eq. (7).

$$w_v = \frac{|\tilde{S}_v|}{|\tilde{S}|} \quad (7)$$

In Eq. (7), $|\tilde{S}_v|$ represents the number of samples whose value A is equal to v . $|\tilde{S}|$ represents the total number of samples with a value for attribute A . The greedy splitting mechanism of the ID3 algorithm can easily cause the model to become too complex and to overfit the training data. To automatically control complexity during tree growth, cost complexity pruning is introduced as a stopping condition. Based on the above improvements, the G-ID3 model systematically overcomes the inherent shortcomings of classic ID3 and becomes a powerful basic framework capable of processing complex general data. In summary, the study constructs a complete G-ID3 model, and its overall flow chart is shown in Fig. 2.

As shown in Fig. 2, the dynamic discretization of continuous attributes uses the information gain rate to select split attributes, probabilistically handles missing values, and controls overfitting through cost-complexity pre-pruning, ultimately generating a decision tree.

2.2. Construction of Domain Optimization Model for Enterprise Financial Management

The constructed G-ID3 model solves the technical bottleneck of the ID3 algorithm when processing general data. But it has not yet considered the knowledge in the field of enterprise financial risk management, and there is a gap between its

decision-making logic and the business objectives of financial risk control. Therefore, the research will conduct in-depth customization of financial management based on the G-ID3 model. A truly "business-savvy" FR-DT model is built by embedding domain knowledge into its core decision-making mechanism, thereby achieving a key transition from "universal classifier" to "domain expert system". The schematic diagram of building the FR-DT is shown in Fig. 3.

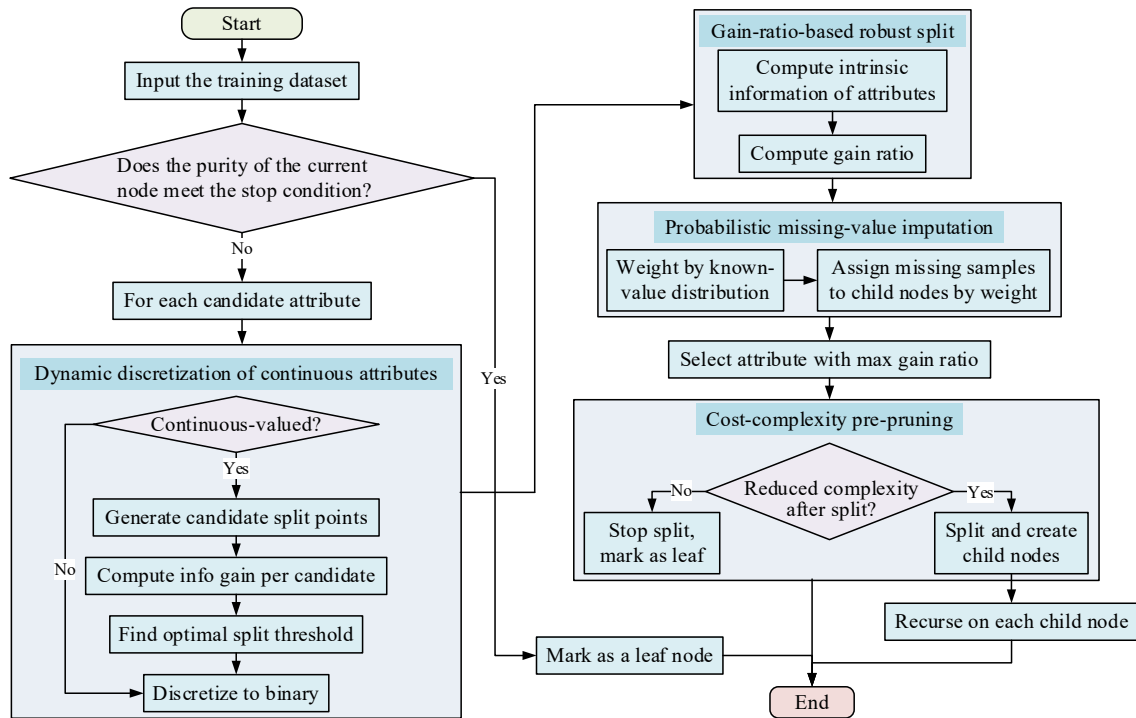


Fig. 2. Flowchart of the G-ID3 model

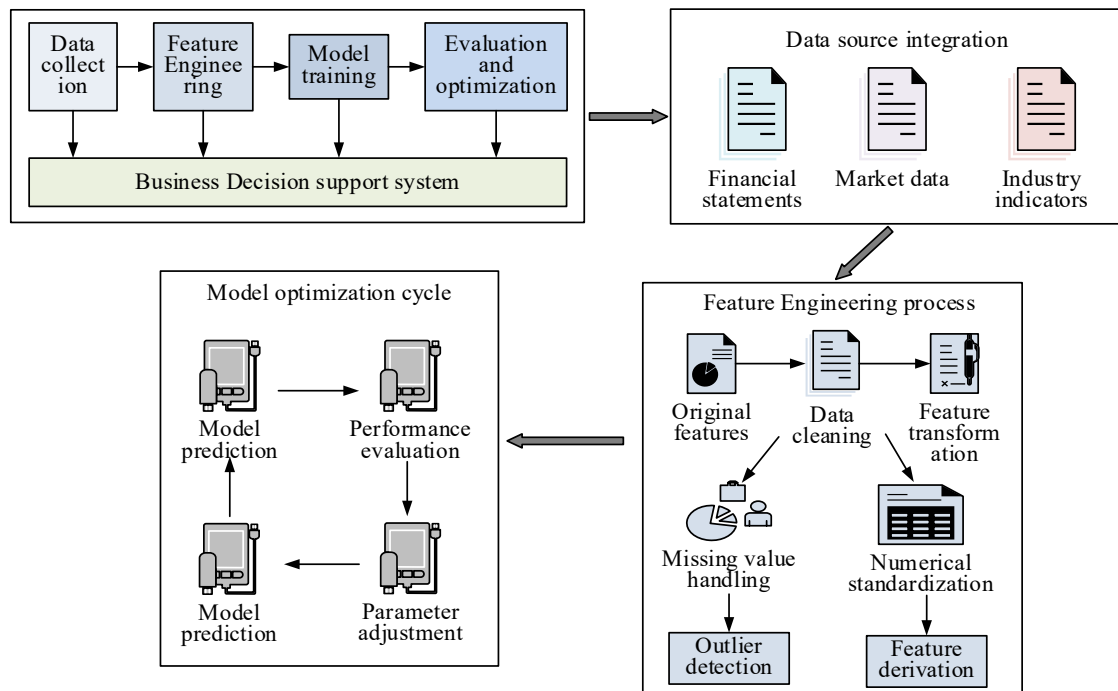


Fig. 3. Schematic diagram of the financial risk decision tree model

Fig. 3 outlines the systematic construction of the FR-DT. The process begins with multi-source data collection, followed by feature engineering, including data cleaning, missing-value handling, standardization, and outlier detection. The refined features are used to train the model, which then enters an iterative optimization cycle of prediction, evaluation, and parameter tuning. The final model is deployed into a decision support system, forming a closed-loop data-to-decision framework for financial risk management. In financial analysis, different metrics have different importance. To reflect the experience of domain experts in model decision-making, the importance of the basic splitting criteria is weighted, as shown

in Eq. (8).

$$WGR(S, A) = \omega_A \cdot GR(S, A) \quad (8)$$

In Eq. (8), ω_A represents the financial prior weight assigned to attribute A . $WGR(S, A)$ represents the weighted gain rate of attribute A . However, in risk control, the cost of failing to identify high-risk enterprises is much higher than misidentifying low-risk enterprises. This asymmetric misclassification cost is the core business logic of financial risk management, which is captured by defining the cost matrix C and reshaping the split criterion. To measure the lowest business risk of a node under the optimal classification, its minimum expected misclassification cost is calculated in Eq. (9).

$$EC(S) = \min_j \left(\sum_{i=1}^n p_i(S) \cdot C_{ij} \right) \quad (9)$$

In Eq. (9), n represents the number of categories. $p_i(S)$ represents the proportion of samples belonging to category i in node S . C_{ij} represents the cost of predicting a sample of true category i as category j . $EC(S)$ represents the expected cost of node S . To evaluate the extent to which an attribute split can reduce the overall business risk, the cost reduction gain is calculated, as shown in Eq. (10).

$$CRG(S, A) = EC(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} EC(S_v) \quad (10)$$

In Eq. (10), $CRG(S, A)$ represents the cost reduction gain caused by attribute A to split node S . To comprehensively consider the risk reduction ability and uncertainty reduction ability of attributes, the study normalizes them and then weights them to form the final splitting basis, as shown in Eq. (11).

$$S_{final} = \beta \cdot \frac{CRG(S, A)}{\max CRG} + (1 - \beta) \cdot \frac{WGR(S, A)}{\max WGR} \quad (11)$$

In Eq. (11), β represents the trade-off parameter. S_{final} represents the final score. This composite criterion drives the model to select attributes for splitting that can both reduce costly errors and improve data purity. As a result, the schematic diagram of domain knowledge injection and decision-making logic reshaping is shown in Fig. 4.

Fig. 4 illustrates how domain knowledge transforms the FR-DT model from a statistic-driven to a business-driven system. Key financial indicators are assigned business weights to prioritize risk signals, and an asymmetric cost matrix imposes a higher penalty for missing high-risk cases. The splitting criterion combines weighted gain rate with cost-reduction gain, balancing statistical purity with business impact. The resulting tree structure actively tilts toward high-risk branches, demonstrating sharper risk identification than the balanced yet potentially blind traditional model. To focus on and learn these critical minority classes, the data is synthetically oversampled before training. By linearly interpolating between the neighbors of high-risk enterprises in the feature space, new and reasonable high-risk samples are artificially synthesized, thereby addressing the class imbalance at the data level, as shown in Eq. (12).

$$x_{new} = x_i + \lambda(\hat{x} - x_i) \quad (12)$$

In Eq. (12), x_i represents a real high-risk enterprise sample. $\hat{x}$ represents a sample randomly selected by x_i from the K nearest neighbor high-risk samples. λ represents a random number uniformly distributed in the range $[0,1]$. x_{new} represents the newly generated synthetic high-risk enterprise sample. The data balancing strategy and cost-sensitive learning work together from both the data and processing logic levels to strongly guide the FR-DT model to focus on and accurately identify high-risk enterprises. To provide richer decision-making information and output the probability distribution of risk levels, the probability for each risk level is calculated from the category distribution of samples within leaf nodes, using smoothing techniques to avoid zero probabilities. To integrate the continuous value bankruptcy risk prediction theory and the data-driven model, it is integrated into FR-DT as an auxiliary continuous output. For the sample path that reaches the leaf node, the Z-Score meaning is calculated as the early warning indicator of the path, as shown in Eq. (13).

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5 \quad (13)$$

In Eq. (13), Z represents Altman Z-Score. X_1 represents working capital relative to total assets, measuring the proportion of a company's net current assets. X_2 represents retained earnings relative to total assets, which measures the company's cumulative profitability. X_3 represents profit before interest and tax relative to total assets, which measures the operating efficiency of corporate assets. X_4 represents the market value of shareholders' equity relative to total

liabilities, which measures the company's solvency. X_5 represents operating income relative to total assets, which measures the efficiency of corporate asset turnover. Integrating Z-Score allows FR-DT to provide not only classification results but also a widely recognized continuous measure of financial risk. To provide decision-makers with a unified and comprehensive risk view, a dynamic risk-scoring function is constructed to integrate the aforementioned probability output with the classic Z-Score, as shown in Eq. (14).

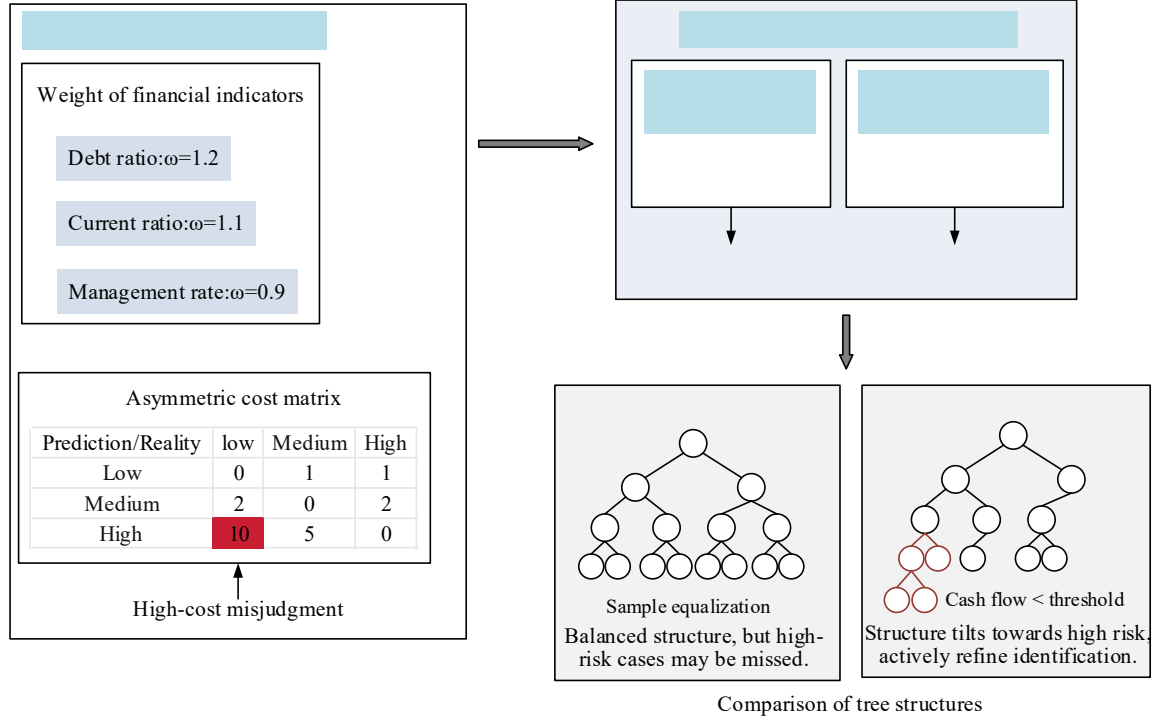


Fig. 4. Schematic diagram of domain knowledge injection and decision-making logic reshaping

$$RiskScore = \gamma \cdot p_{High} + (1 - \gamma) \cdot \Phi(Z) \quad (14)$$

In Eq. (14), p_{High} represents the high risk probability predicted by the FR-DT model. $\Phi(Z)$ represents mapping Z-Score to the [0,1] interval through the function Φ . γ represents the fusion weight. $RiskScore$ represents the final output comprehensive risk score. Dynamic risk scoring provides a concise, clear decision-making basis for sorting and prioritizing businesses. An excellent business model must be able to continuously evolve. For this reason, this study designs a feedback-driven online update mechanism for FR-DT. When the maximum predicted probability of a sample by the model is lower than the threshold τ , the sample is automatically marked for expert review, as shown in Eq. (15).

$$\text{If } \max(P) < \tau, \text{ then flag for review} \quad (15)$$

In Eq. (15), $\max(P)$ represents the model's prediction confidence. τ represents the confidence threshold. The results of expert review of labeled samples will be fed back into the model as new gold-standard data, driving incremental learning or local reconstruction, thereby forming an intelligent living system that can adapt to changes in the economic environment. In summary, through the confidence-driven feedback mechanism, the FR-DT model achieves a key transformation from static prediction to dynamic evolution. Finally, the complete flow chart of the enterprise financial management model is shown in Fig. 5.

The flowchart in Fig. 5 illustrates the entire process for constructing and applying the FR-DT model. The design of the FR-DT model establishes a clear mapping to the enterprise management decision-making process. In the risk identification stage, the model's high-risk probability and comprehensive risk score outputs can directly support managers in risk ranking and hierarchical management of the enterprise, thereby prioritizing the allocation of limited audit and risk control resources to high-risk enterprise groups. In the decision review stage, the confidence threshold mechanism built into the model can automatically screen out samples with high prediction uncertainty, triggering the expert manual review process. The review conclusion is fed back to the model as new training data, supporting incremental learning and dynamic optimization. Through the above mechanism, the FR-DT model achieves closed-loop coupling with the enterprise's "risk screening manual review model iteration" management process, providing intelligent decision support that can be embedded into existing business processes for financial risk management. The FR-DT model has a natural advantage in interpretability. As a decision tree model, the decision path of FR-DT can be fully traced: each path from the root node to a leaf node consists of specific financial indicator thresholds, and managers can clearly understand why the model identifies a given enterprise as high risk. In addition, the model's comprehensive risk score output integrates the leaf node probability

distribution and Z-Score, providing both classification results and continuous risk measurements, thereby enabling managers to rank and compare enterprises more finely. At the regulatory compliance level, traceable decision paths help meet the requirements of internal audit and external oversight for transparency in risk decision-making processes, thereby enhancing acceptability in practical enterprise applications.

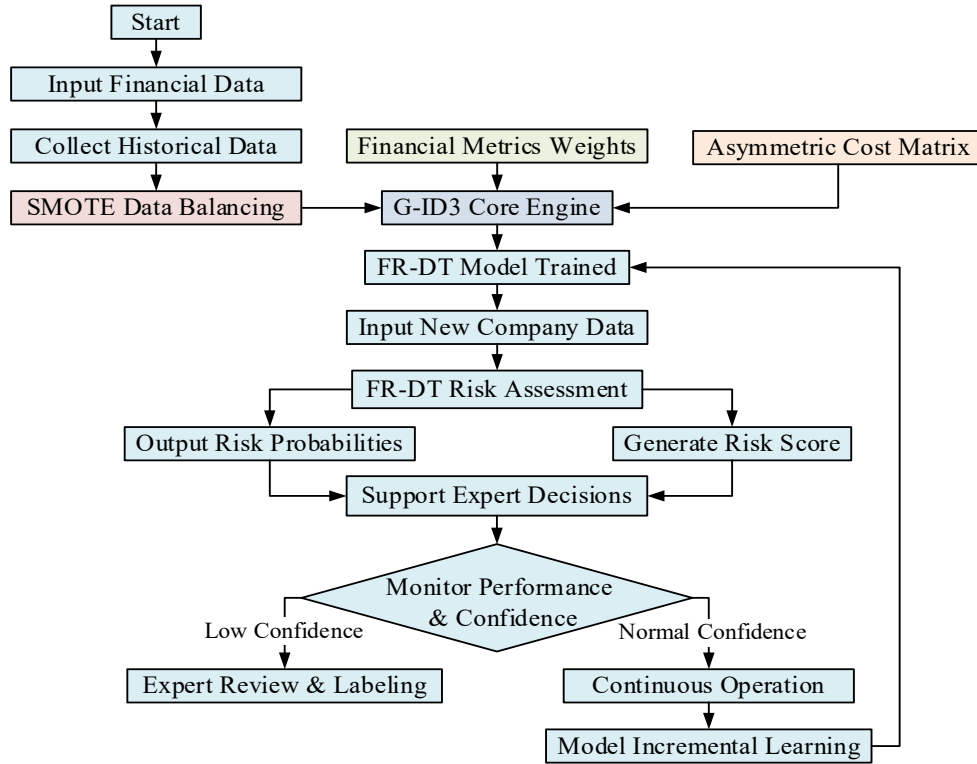


Fig. 5 Flowchart of the FR-DT model

3. Results

3.1. G-ID3 Model Performance Verification

Before the experiment, the researcher needs to prepare the required experimental environment and data. This includes the hardware and software configuration required for the experiment, as well as the experimental data set. An efficient hardware configuration and a stable software version are sufficient to support the smooth progress of the experiment. The specific parameters are shown in Table 1.

Table 1. Hardware and software configuration parameters

Category	Component	Model/Version
Hardware Configuration	CPU	Intel Core i7-12700K
	GPU	NVIDIA GeForce RTX 3060 (12GB)
	Memory	32 GB DDR4
	Storage	1 TB NVMe SSD
Software Configuration	Operating System	Ubuntu 20.04 LTS
	Programming Language	Python 3.9
	Key Libraries	Scikit-learn 1.2, Pandas 1.5, NumPy 1.23
	Development Environment	Jupyter Notebook

The study was conducted according to the parameters in Table 1. The Census Income (CI) data set in the UCI Machine Learning Repository was introduced. The CI data set contains 48,842 instances. Its 14 features include 8 categorical attributes (such as job type, education level, marital status) and 6 continuous attributes (such as age, years of education, capital gain/loss), with a small number of missing values. This mixed data type and the imperfections of the real world make it an ideal testing platform for verifying various improvement measures for G-ID3. The data set uses stratified random sampling to divide the complete data set into training, validation, and test sets in a 6:2:2 ratio. Given the ethical

considerations and the financial risk prediction's involvement of sensitive enterprise data, this approach is appropriate. The model deployment strictly complies with data privacy regulations to ensure proper data collection, storage, and processing. To ensure the reproducibility and comparability of experiments, the study adopted a unified hyperparameter-tuning process across all models, using the validation set to optimize key hyperparameters via grid search. The details are shown in Table 2.

Table 2. Model hyper-parameter configuration table

Model Category	Parameter Name	Set Value/Range
Baseline Models	criterion	ID3: 'information_gain' C4.5: 'gain_ratio' CART: 'gini'
	max_depth	[5, 10, 15, None]
	min_samples_split	[2, 5, 10]
	min_samples_leaf	[1, 2, 4]
G-ID3 Model	criterion	'gain_ratio'
	max_depth	[8, 12, 15]
	min_samples_split	[5, 10, 15]
	min_samples_leaf	[3, 5, 7]
	ccp_alpha	[0.01, 0.02, 0.05]
FR-DT Model	missing_value_handling	'probability_allocation'
	financial_weights	Key indicators: 1.5-2.0 Secondary indicators: 1.0
	cost_matrix	[[0, 1], [10, 0]]

Table 2 clearly shows that the success of FR-DT relies not only on algorithm optimization, but also on the systematic and quantitative integration of corporate financial management expertise into every key decision-making link of the model. To comprehensively evaluate the performance and generalization capabilities of different models as data scale increases, the study further compared the trends in accuracy, macro F1-Score, and time overhead, as shown in Fig. 6. C4.5 algorithm is an improved version of the ID3 algorithm that uses information gain as the splitting criterion. CART is a classification and regression tree that uses the Gini coefficient as the splitting criterion.

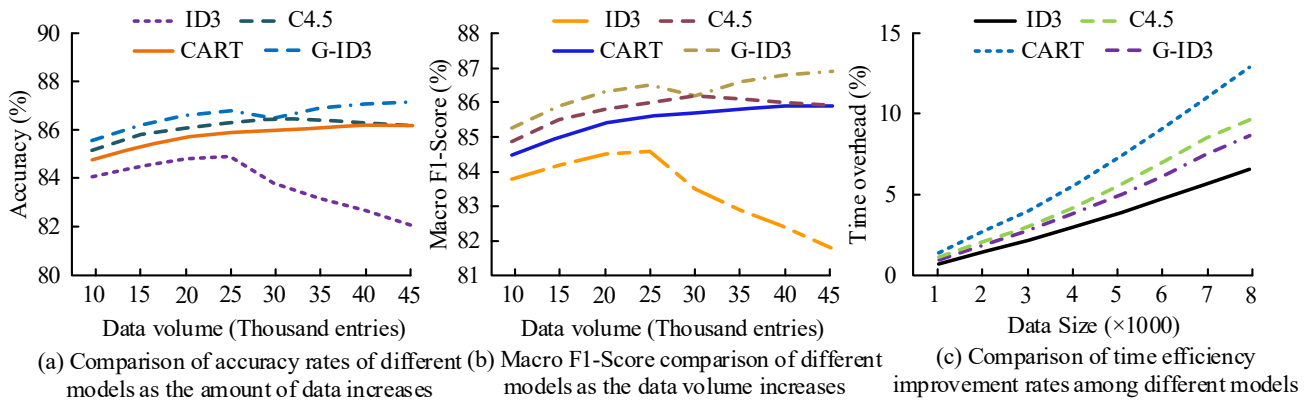


Fig. 6. Comparison of classification performance of different models

In terms of accuracy in Fig. 6(a), the average accuracy of G-ID3 on all data volumes has improved by about 0.60% and 0.98% respectively, compared to the second best performing C4.5 and Classification and Regression Trees (CART) models. Compared with the worst performing ID3 model, the improvement is as high as about 3.40%, which is better than the baseline model. In terms of macro F1 Score in Fig. 6(b), the average macro F1 Score of G-ID3 on all data volumes has improved by about 0.59% and 0.97% in accuracy compared to the second best performing C4.5 and CART models, respectively; Compared with the worst performing ID3 model, the improvement is as high as about 3.30%. In terms of total time cost, the G-ID3 model achieved 11.6% and 33.1% time efficiency improvement compared to the C4.5 and CART models, respectively, as shown in Fig. 6(c).

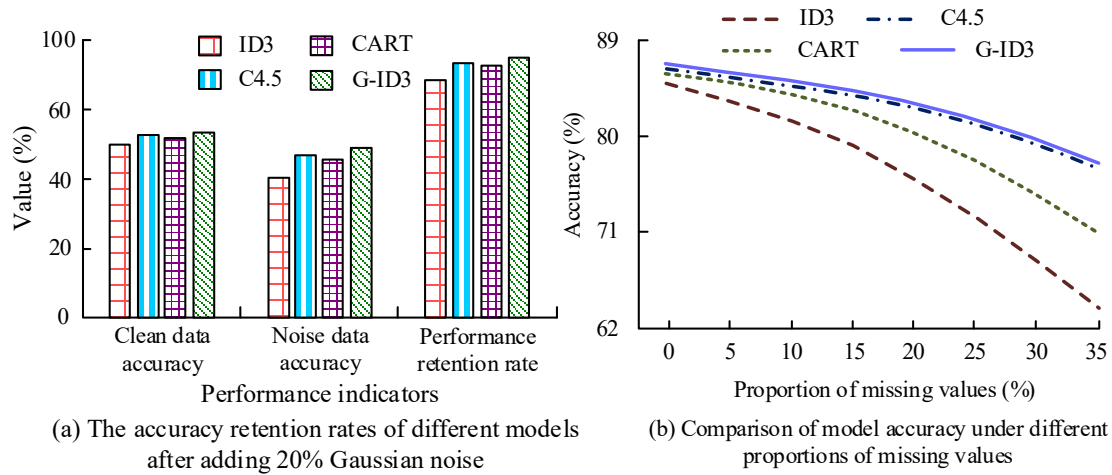


Fig. 7. Robustness analysis of different models

After adding 20% Gaussian noise in Fig. 7(a), the clean data accuracy of G-ID3 reached 86.8%, achieving relative improvements of 2.2%, 0.6%, and 1.0% compared to ID3, C4.5, and CART models, respectively; The noise data accuracy of G-ID3 reaches 84.7%, which is 5.7%, 1.4%, and 2.3% higher than ID3, C4.5, and CART models, respectively; The performance retention rate of G-ID3 reaches 97.6%, which is 3.5%, 0.8%, and 1.2% higher than ID3, C4.5, and CART models, respectively. In terms of accuracy across different proportions in Fig. 7(b), the average accuracy of G-ID3 during the missing stage was improved by 8.2%, 0.6%, and 3.5% compared to ID3, C4.5, and CART models, respectively. To comprehensively evaluate performance, the study compared two dimensions: computational efficiency and model complexity, as shown in Fig. 8.

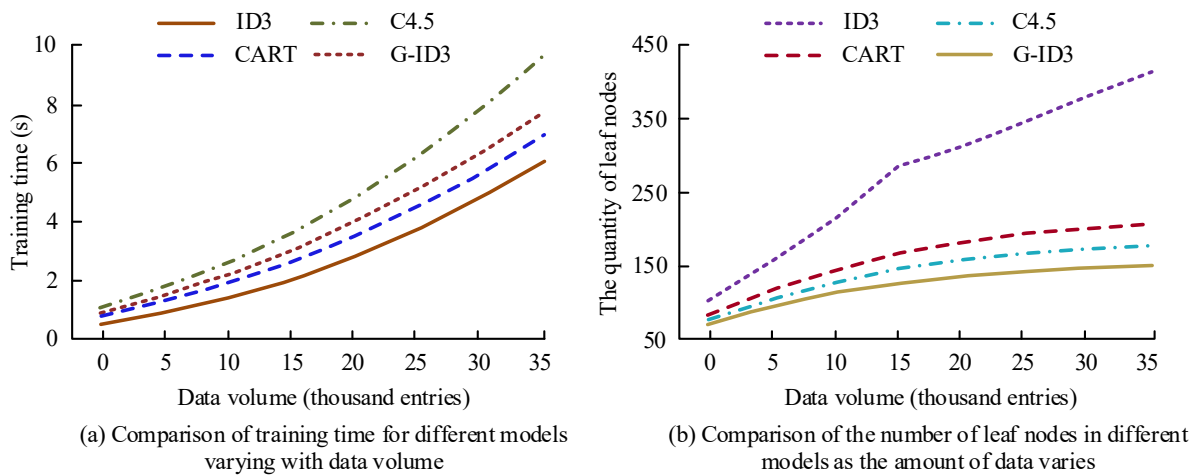


Fig. 8. Efficiency analysis of different models

Fig. 8(a) shows that, in terms of training time, when the data volume reaches 45000, G-ID3's training time (7.7 seconds) is reduced by 19.8% compared to C4.5 (9.6 seconds). Although it is 28.3% slower than ID3 (6.0 seconds), it demonstrates the potential for subsequent performance improvement within an acceptable time cost. Fig. 8(b) shows that in terms of training time, when the data volume reaches 45000, the number of leaf nodes in G-ID3 (151) decreased by 63.3%, 15.2%, and 27.1% compared to ID3 (412), C4.5 (178), and CART (207), respectively.

3.2. The Application Results of Enterprise Financial Management Based on FR-DT

The study further verified the comprehensive effectiveness of the FR-DT model in decision support for corporate financial risk management. To verify the superiority of the FR-DT model in the early warning management of corporate financial risks, the Polish Companies Bankruptcy Data (PCBD) dataset was selected. The dataset contains financial data of nearly 10,000 Polish companies, covering 64 key financial ratios (such as current ratio, return on assets and debt-to-equity ratio). After time series division, the data from 2000 to 2013 are used as a training set, the data from 2014 are used as a verification set to determine business parameters such as financial prior weights and cost matrix parameters, and the data from 2015 to 2017 are used as a test set to simulate the early warning capability for unknown future periods. To evaluate the business value of the model in real scenarios, the study compared the core effects of different models in enterprise risk control, as shown in Fig. 9.

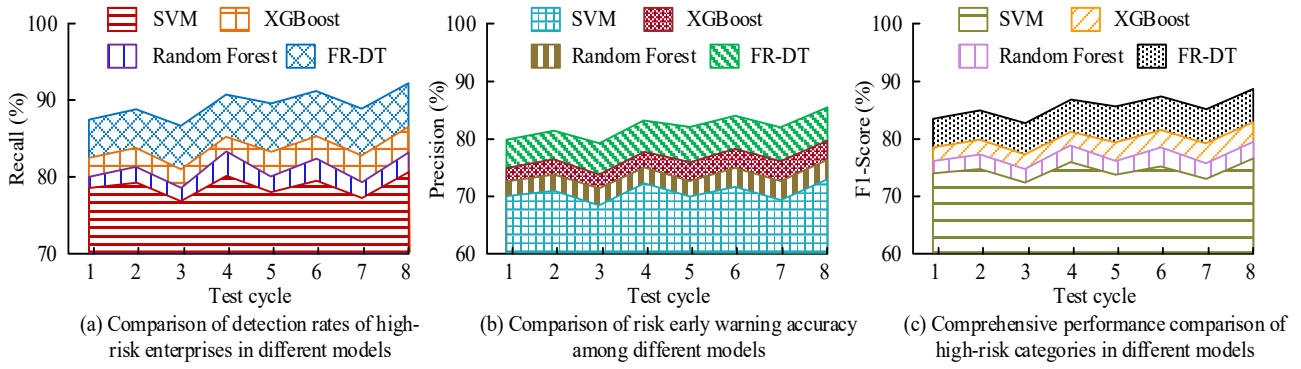


Fig. 9. Comparison of the core effectiveness of risk control in enterprises with different models

In the comparison of high-risk enterprise detection rates in Fig. 9(a), the FR-DT model performs exceptionally well, achieving detection rates consistently ranging from 86.7% to 92.1%, with an average of 89.4%. Compared to eXtreme Gradient Boosting (XGBoost), Random Forest, and Support Vector Machine (SVM), the FR-DT model achieved performance improvements of approximately 6.7%, 10.4%, and 13.6%, respectively. In terms of risk warning accuracy in Fig. 9(b), the FR-DT model also performs well, with an accuracy range of 79.3% to 85.4% and an average of 82.3%. Compared with XGBoost, Random Forest, and SVM, these models achieved performance gains of approximately 6.8%, 10.4%, and 14.0%, respectively. On F1 Score in Fig. 9(c), the FR-DT model significantly outperforms other models with an average score of 85.6%. Compared to XGBoost's 80.4%, Random Forest's 77.2%, and SVM's 74.5%, FR-DT achieved comprehensive performance improvements of approximately 6.6%, 9.9%, and 13.0%, respectively. To evaluate practicality, the study compared resource overhead and stability under incomplete data, as shown in Fig. 10.

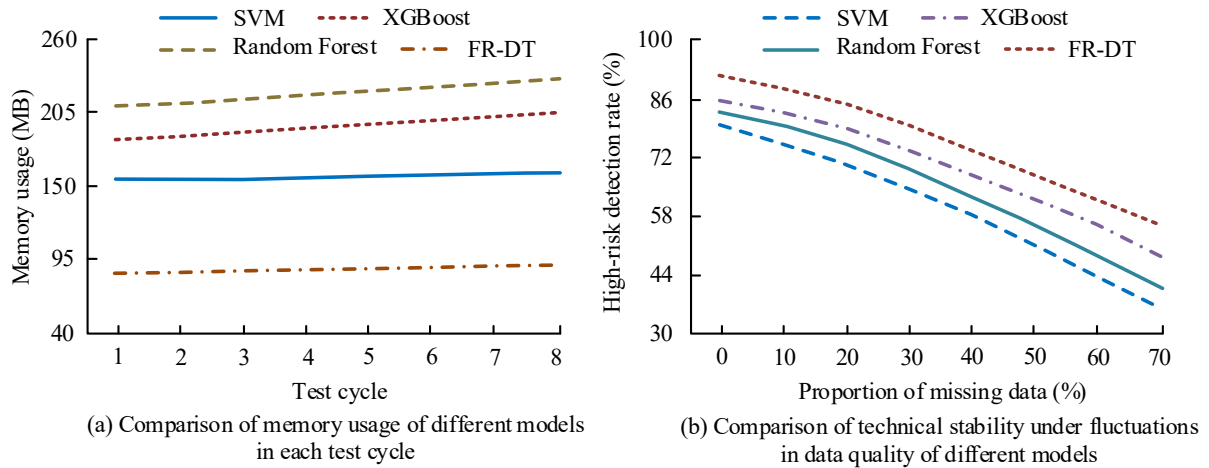


Fig. 10. Analysis of model ablation experiments

Note: Financial Risk Decision Tree (FR-DT), Support Vector Machine (SVM), Random Forest (RF), Xtreme Gradient Boosting (XGBoost).

In terms of memory efficiency in Fig. 10(a), the average memory usage of FR-DT is 88.5 MB, which is about 43%, 60%, and 54% higher than those of the SVM, Random Forest, and XGBoost models, respectively. In the technical stability test shown in Fig. 10(b), when the data missing ratio reached 50%, the high-risk detection rate of the FR-DT model remained at 67.6%. Even with a high missing rate of 70%, the performance advantage of FR-DT is further expanded, with a detection rate about 15.3% higher than that of XGBoost.

4. Conclusion

To address the challenges in enterprise financial risk warning, this paper developed a dedicated decision-tree model. By integrating four techniques, including dynamic discretization, information gain ratio, missing value handling, and pre-pruning, and combining cost-sensitive learning and data balancing strategies, it achieves the upgrade from a general classifier to an intelligent business decision system. Compared with the baseline model, the G-ID3 model reduced the model's structure complexity by 15.2% to 63.3% while maintaining an acceptable computational overhead (training time: 7.7 seconds). The recall rate for high-risk enterprises in the FR-DT model reached a peak of 92.1%, the comprehensive macro F1-Score remained stable at 85.6%, and the total business cost due to misjudgment was reduced by 21.0% to 36.6%. In extreme scenarios where the data missing rate was 50%, the risk identification performance was more than 9.4% lower than that of the comparison model, demonstrating its strong robustness in real-world bad-data environments. Managers can shift their financial risk decisions from relying on empirical judgment based on a single indicator threshold to quantitative grading management based on model output of high-risk probability and comprehensive scores. Meanwhile, the model's

confidence feedback mechanism can be used to establish a dynamic closed-loop decision-making process, "model pre-screening expert review," that precisely allocates and continuously optimizes risk control resources. However, there are still limitations to the research. The experimental data come from a single country's publicly available dataset and do not include external variables such as macroeconomic indicators, which may affect the cross-regional universality. Future research will incorporate data from enterprises across multiple countries, along with macroeconomic indicators, to verify the findings.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence (AI) Tools

The author confirms that no AI tools were used in the preparation of this manuscript.

References

- Agarwal, A., Jain, K., and Yadav, R. K. (2023). A mathematical model based on modified ID3 algorithm for healthcare diagnostics model. *International Journal of System Assurance Engineering and Management*, 14(6), 2376-2386.
- Boruah, A. N. and Biswas, S. K. (2025). A condensed hybrid decision tree for decision rules (CHDTDR). *International Journal of System Assurance Engineering and Management*, 16(8), 2887-2900.
- Dou, Q., Zhang, J., and Jing, B. (2023). A ML-based economic protection development level using Decision Tree and Ensemble Algorithms. *Soft Computing*, 27(24), 18929-18947.
- Hang, P., Lv, C., Huang, C., Xing, Y., and Hu, Z. (2022a). Cooperative Decision Making of Connected Automated Vehicles at Multi-Lane Merging Zone: A Coalitional Game Approach. *IEEE Transactions on Intelligent Transportation Systems*, 23(4), 3829-3841.
- Hang, P., Huang, C., Hu, Z., and Lv, C. (2022b). Decision Making for Connected Automated Vehicles at Urban Intersections Considering Social and Individual Benefits. *IEEE Transactions on Intelligent Transportation Systems*, 23(11), 22549-22562.
- Hasanvand, M., Nooshyar, M., Moharamkhani, E., and Selyari, A. (2023). Machine Learning Methodology for Identifying Vehicles Using Image Processing. *Artificial Intelligence and Applications*, 1(3), 170-178.
- Tyagi, S. K. S. and Boyang, Q. (2023). An Intelligent Internet-of-Things-Aided Financial Crisis Prediction Model in FinTech. *IEEE Internet of Things Journal*, 10(3), 2183-2193.
- Wang, L., Wang, Y., Chen, M., and Ding, S. (2025). Multi-Round Dynamic Game Decision-Making of UAVs Based on Decision Tree. *Journal of Systems Engineering and Electronics*, 36(4), 1006-1016.
- Zeng, H. (2022). Influences of mobile edge computing-based service preloading on the early-warning of financial risks. *The Journal of Supercomputing*, 78(9), 11621-11639.
- Zhang, Q. and Li, F. (2022). Financial Resilience and Financial Reliability for Systemic Risk Assessment of Electricity Markets With High-Penetration Renewables. *IEEE Transactions on Power Systems*, 37(3), 2312-2321.



Chenghang Zhang obtained her bachelor's degree in accounting from Zhejiang Agriculture and Forestry University in 2011. She then earned her master's degree in accounting from Nanjing University of Finance and Economics in 2015. She is currently a teacher at the School of Digital Intelligence and Trade Management at Lishui Vocational and Technical College. Her research interests include financial management, financial accounting, and tax administration.