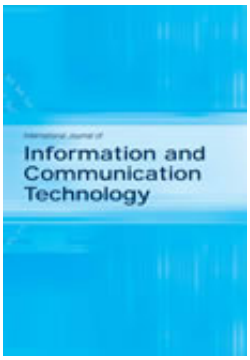
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Intelligent recognition of financial fraud based on CART decision tree

Guiyun Chen

DOI: 10.1504/IJICT.2025.10070828

Article History:

Received:	16 January 2025
Last revised:	26 February 2025
Accepted:	26 February 2025
Published online:	06 May 2025

Intelligent recognition of financial fraud based on CART decision tree

Guiyun Chen

School of Management,
Changsha Medical University,
Leifeng Avenue, Wangcheng District,
Changsha, Hunan, 410219, China
Email: 13467627901@163.com

Abstract: This paper proposes an intelligent recognition model of financial fraud based on classification and regression tree (CART) decision tree, which aims to improve the recognition rate of financial fraud and provide a preliminary reference for other industries to use non-financial information for fraud recognition. The decision tree model adopted is tuning Iterative Dichotomiser 3 (ID3) algorithm and CART algorithm, and optimises the decision tree parameters by particle swarm to avoid the occurrence of over-fitting. It is found that the area under curve (AUC) of CART tree recognition method is significantly higher than that of random forest (RF) and neural network recognition methods, reaching 70%, which has a good recognition effect. It can be seen that parameter combination search can make the accuracy of CART decision tree model achieve the best effect, and has a positive effect on improving the intelligent recognition effect of financial fraud behaviour.

Keywords: CART decision tree; finance; fraud; intelligent recognition.

Reference to this paper should be made as follows: Chen, G. (2025) 'Intelligent recognition of financial fraud based on CART decision tree', *Int. J. Information and Communication Technology*, Vol. 26, No. 11, pp.1–20.

Biographical notes: Guiyun Chen is an Associate Professor serving for Changsha Medical University. She has a Master's in Business Administration at Central South University. She honoured as Youth Backbone Teacher of Hunan Province, China and has 19 years of teaching and research experience, hosting five provincial-level research projects and publishing over 20 academic papers.

1 Introduction

With the rapid growth of the economy, the capital market is playing an increasingly important role in resource allocation, which provides financing channels for enterprises and a platform for investors to increase their wealth. Among the many elements of capital markets, financial reporting plays a crucial role. It is not only a bridge between investors and enterprises, but also directly affects the market value of companies and investors' decisions. Moreover, an accurate and transparent financial reporting system can improve

market efficiency, reduce transaction costs and enhance investors' confidence in the market (Gayam, 2021).

The business goal of enterprises is gradually changing from the traditional pursuit of profit maximisation to taking on corresponding social responsibilities while pursuing profit maximisation, and finally realising the overall value of enterprises. However, due to information asymmetry in the capital market, financial fraud, tax evasion, corruption, and infringement of employee health and safety have occurred frequently, and accounting information is no longer just a simple record of financial data, but has become an important tool for evaluating the overall performance of a company. At the same time, accurate and comprehensive accounting information can help stakeholders better understand the company's operating conditions, including its efforts in social responsibility and governance. Therefore, the transparency and accuracy of accounting information are directly related to the reputation and sustainable development ability of enterprises. Enterprises should pay attention to the quality of accounting information and ensure the comprehensiveness and transparency of information disclosure to meet the increasing concerns of stakeholders about corporate social responsibility and governance capabilities (Xu et al., 2022).

Financial fraud is an international and widespread economic behaviour, which not only destroys the fairness and transparency of the capital market, but also seriously damages the interests of investors and other stakeholders. Due to the complexity and concealment of financial fraud, there is currently no fully effective solution in the world (Mytnyk et al., 2023). The consequences of financial fraud are serious. For example, once the investment decisions made by financial report users based on false financial information lead to losses, these losses are usually irreversible, and the obstacles faced by victims in the process of seeking redress are often huge, including the long-term and uncertainty of legal proceedings. Under this background, the purpose of this paper is to explore and identify the patterns and characteristics of financial fraud through in-depth analysis of specific financial fraud cases. In this way, the research aims to improve the existing financial fraud recognition mechanism, and on this basis, provide constructive suggestions for improving the quality of corporate governance, strengthening the internal control system, improving the audit quality and preventing future financial fraud. This not only helps to protect the interests of investors, but also promotes the healthy development of the capital market and enhances the confidence of market participants (Kotagiri and Yada, 2024a).

The detection of financial fraud needs to overcome the complex financial data in the detection process, and the data characteristics are complex, which is prone to over fitting Aiming at the problem of feature selection and over fitting when the decision tree algorithm determines the split attribute, this paper innovatively proposes an improved decision tree model. By introducing pruning, feature selection and particle swarm optimisation (PSO) algorithm, the accuracy and robustness of the decision tree model in the detection of financial fraud are effectively improved, and the over fitting situation is overcome, and the robustness of the model is improved.

This paper proposes an intelligent recognition model of financial fraud based on CART decision tree, which aims to improve the recognition rate of financial fraud and provide a preliminary reference for other industries to use non-financial information for fraud recognition. In the model improvement, pruning mechanism and feature selection method are introduced to reduce the overfitting problem and improve the learning ability of the model. Moreover, the decision tree model adopted in this paper is tuning ID3

algorithm and CART algorithm, and this paper optimises the decision tree parameters by particle swarm to avoid the occurrence of over-fitting.

2 Related works

2.1 Financial fraud methods

In the academic discussion in the field of financial fraud, Kannagi et al. (2023) provided a comprehensive overview of financial fraud, including various techniques and strategies to identify and prevent fraud. The study focuses on the role of audit and internal control in detecting and preventing financial fraud, while also highlighting the potential of information technology to improve detection efficiency. Guo et al. (2024) studied the relationship between fraudulent financial statements and insider trading, and emphasised that insider trading may be the main means of financial fraud. Ali et al. (2022) studied the victims of financial fraud, made an in-depth study on the characteristics of the victims to predict the possible victims, and summarised the measures of financial fraud, mainly including inflated income and exaggerated accounts receivable.

The research in Narsimha et al. (2022) focuses on the specific means and behaviours of financial fraud of Chinese enterprises, such as income increase, expense decrease, improper capitalisation, etc. Their work focuses on real-world case analysis and uncovering fraud through audit reports and legal documents, and they also offer recommendations for improving accounting standards and strengthening legal supervision. Shoetan et al. (2024) focused on financial fraud in the stock market, analysed the influence of market supervision, corporate governance structure and cultural factors on the choice of fraud means, and discussed different types of fraud and the incentive mechanism behind it through quantitative and qualitative methods. By studying the financial fraud cases of listed companies, Agu et al. (2024) deeply analysed the evolution trend and characteristics of financial fraud of listed companies from the perspective of typical cases. From the perspective of listed companies, Hernandez Aros et al. (2024) studied the relationship between fraud in financial reports and audit disclosure, and further strengthened the important role of audit in revealing financial fraud based on the empirical analysis of the administrative penalty decision of the CSRC. Chang et al. (2022) focused on the means of financial fraud by using related party transactions, and summarised the following ways of fraud. The first is fraud between related companies, the second is fraud between associated companies and joint ventures, the third is concealing the relationship between related companies and conducting transactions, the fourth is using financial assets and inventories to conduct related-party transaction fraud, and the fifth is fraud by changing the depreciation provision model. Kotagiri and Yada (2024b) explored the evolution of financial fraud means over time and how the latest technologies, such as big data analytics and machine learning, can be used to detect and prevent these behaviours. The research highlights new tools and techniques for financial fraud detection and the effectiveness of these tools in improving the transparency and integrity of financial information.

2.2 Financial fraud recognition

In international academic circles, the research on financial fraud recognition adopted general analytical procedures to construct financial fraud recognition models earlier. The M-Score model in Akash et al. (2024) is one of the classic tools for identifying manipulative financial reports. This model is based on eight financial indicators, and aims to predict whether a company has financial fraud through statistical methods. These indicators include sales growth rate, changes in gross profit margin, asset quality, abnormal external financing, changes in costs, etc. Johora et al. (2024) studied more aspects of fraud detection, such as behavioural analysis and early warning signals, which may include abnormal financial reporting patterns, changes in management behaviour, or weaknesses in corporate governance. The research in Singh et al. (2022) focuses on specific financial reporting fraud, such as untrue statements on balance sheets and income statements. Furthermore, the reference proposes a variety of quantitative and qualitative methods for identifying fraud and explores how to improve the transparency of financial reporting and how to enhance the ability of regulators and auditors to identify fraud.

Odeyemi et al. (2024) discussed the methods and strategies of identifying financial fraud of listed companies. The work involves using machine learning and big data analysis methods to discover potential financial irregularities and analyse how market-specific regulatory environments and cultural factors affect the recognition of financial fraud. Hassan et al. (2023) focused on improving the accuracy of existing financial fraud detection models. Their work includes variable optimisation of traditional models, the introduction of new financial ratios or non-financial metrics, or the consideration of firm-specific environmental and industry factors. In addition, the research also discusses the applicability and effectiveness of the financial fraud model in listed companies. In the research field of financial fraud recognition, Bhaktiar and Setyorini (2021) focused on the applicability of financial fraud detection models in different cultures and regulatory environments. By comparing financial reporting standards and fraud behaviours in different countries and regions, it proposes model optimisation strategies tailored to specific environments. Such research helps multinational companies and their investors to conduct more effective risk management. Bello et al. (2024) focused on empirical analysis to examine the performance and limitations of different financial fraud models in actual situations. Their study evaluates the sensitivity, specificity, and stability of the model in multi-epoch data. Furthermore, the research also deals with how emerging regulatory policies and market trends can be incorporated into the model to improve its adaptability. Gautam (2023) discussed the impact of accounting information comparability on identifying financial fraud risks, and believed that improving information comparability helps to increase the difficulty of financial fraud and improve the effectiveness of recognition. Shoetan and Familoni (2024) discussed the structure of the bond market and the reform path of the private enterprise bond market. This study may indirectly affect the recognition of financial fraud because the transparency of the bond market and the financing conditions of private enterprises are related to the financial integrity of enterprises. Moreover, improvements in the bond market can improve the quality of financial disclosure, thereby helping to prevent and identify potential financial fraud.

3 Algorithm model

The intelligent financial fraud identification model based on cart decision tree constructs a binary tree classification structure by analysing the key characteristics of enterprise financial data. The core process is as follows: first, pre-process the historical financial data, standardise the indicators and fill in the missing values; Secondly, based on the Gini index minimisation principle, the optimal feature and segmentation threshold are selected recursively, and the dataset is divided into fraud/non fraud subsets to generate a binary tree; The pruning strategy is used to prevent over fitting and retain the critical decision path. The model judges the risk of new samples layer by layer according to the node rules, and outputs the classification results. Compared with traditional methods, cart model has both interpretability and nonlinear capture ability, and can effectively identify hidden abnormal financial patterns.

3.1 Decision tree algorithm

The construction process of decision tree model is based on the analysis of sample data and the generation of decision rules. The decision tree model can be used to identify potential fault situations and predict them. The process of building a decision tree mainly includes the following steps:

1 Feature selection: Before building a decision tree, it is necessary to first select features for decision making.

2 Tree establishment: After feature selection, the structure of the tree is constructed according to the selected features. The decision tree construction process is recursive, selecting one feature as the current split feature at a time, and dividing the dataset into different subsets according to this feature. This process continues until all samples are correctly classified or can no longer be divided.

3 Generation of decision rules: When the construction of the decision tree is completed, a series of decision rules can be generated according to the structure of the tree. Decision rules can be used to explain the decision process of the tree and make classification predictions for new samples.

Mainstream decision tree construction algorithms include ID3, C4.5 and CART. The main difference between these algorithms is the structure of the tree and the purity measurement method. ID3 algorithm constructs multi-fork tree, uses information entropy to measure purity, and judges segmentation by information gain.

3.1.1 ID3 algorithm

Starting from the root node, child nodes are formed according to different branches of attribute values. Subsequently, the algorithm recursively performs the same process for each branch node until all available attribute gains are too low to continue selecting features and finally a complete decision tree is formed.

The training set is labelled as D, and it is used to build the charging pile fault decision tree. C_i (i = 1, …, m) is m category attributes for judging the abnormality of financial data. $C_{i,d}$ is the set of tuples of class C_i in D, and $|D|$ and $|C_{i,d}|$ are the number of tuples in

training set D and training set $C_{i,d}$ respectively. First, the tuples in the training set D are divided according to attribute A. Attribute A has v different values $\{a_1, a_2, \ldots, a_v\}$, and attribute A divides v into $\{a_1, a_2, \ldots, a_v\}$ subsets $\{D_1, D_2, \ldots, D_v\}$, where D_j ($j = 1 - v$) should contain all tuples in D. Each split is based on the metric with the largest information gain until all the data in the training set is classified.

The training set contains fault data and normal data during the operation of the charging pile. The attribute A is the monitored charging pile parameters. This parameter can be the status parameter of the charging pile, including some electrical parameters and environmental parameters, such as voltage, current, temperature, etc. or it can be signal data, such as drive signal, locking signal, etc. The specific parameters shall be based on the collected data. C is the safety monitoring result of the charging pile. If the monitored data can only determine whether there is a fault, C only contains two parameters, 1 and 0. If the monitored data can determine the specific fault type, the fault type in C is not restricted and is subject to the specific collected data results.

1 The category information entropy of the training set D is calculated as:

$$Info(D) = -\sum_{i=1}^{m} \frac{C_{i,D}}{|D|} \log_2 \frac{|C_{i,D}|}{D} \tag{1}$$

2 The information entropy of attribute A is calculated as:

$$Info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} Info(D_j) \tag{2}$$

3 The information gain A of attribute is:

$$Dain(A) = Info(D) - Info_A(D) \tag{3}$$

3.1.2 CART algorithm

The CART algorithm uses the Gini index to represent the impurity of the sample (division basis) to generate a binary tree. The smaller the Gini index, the lower the impurity and the better the features. The classification method is similar to C4.5. Each splitting is based on the index with the smallest Gini coefficient until all the data of the training set are classified.

1 The Gini index of the training set D is calculated as:

$$Gini(D) = \sum_{i=1}^{m} \frac{|C_{i,D}|}{|D|} \left(1 - \frac{|C_{i,D}|}{|D|}\right) = 1 - \sum_{i=1}^{m} \left(\frac{|C_{i,D}|}{|D|}\right)^2 \tag{4}$$

2 The Gini index of attribute A is calculated as:

$$Gini_A(D) = \sum_{j=1}^{V} \frac{|D_j|}{|D|} Gini(D_j) \tag{5}$$

3.2 Improved CART decision tree algorithm

The current sample is divided into two samples, and a simple binary tree is constructed by CART algorithm. However, the biggest defect of dichotomy is local optimality. Each calculation can only find the optimal value of the current step, and it is easy to fall into local convergence. ID3 algorithm adopts the splitting method of information gain, and builds and generates a multi-fork tree. Compared with binary trees, it is less efficient, occupies more space, and cannot handle large amounts of data. Therefore, the decision tree model is used to optimise ID3 algorithm and CART algorithm, keep the purity measurement criterion of the original ID3 algorithm, and change the multi-tree into a binary tree, and introduce pruning to improve the running efficiency. The reconstructed ID3 algorithm is as follows: there are n features (attributes) in the sample set D, and the optimal features and optimal split points are found for n features. The information entropy of the training sample set is:

$$Info(D) = -\sum_{m=1}^{m} \frac{|C_{m,D}|}{|D|} \log_2 \frac{|C_{m,D}|}{D} \tag{6}$$

Among them, C is the label set, the label has M values, the set of values is $\{C_1, C_2, C_3, \ldots, C_M\}$ (financial data has only two categories, normal and abnormal, that is, the number 0 represents normal and 1 represents abnormal), $|D|$ is the number of tuples in D, and $C_{m,D}$ represents the number of tuples with label C_m in D.

If attribute A has v values, these v values divide the tuple into v subsets. Since the multi-fork tree is changed to a binary tree, one of the values is selected as the split point, and D is changed from v subsets to two subsets D_1 and D_2. If the data set D is divided into two subsets D_1 and D_2 based on attribute (or feature) A, the expected information required to classify the sample set D using attribute A is:

$$Info_A(D) = \frac{|D_1|}{|D|} Info(D_1) + \frac{|D_2|}{|D|} Info(D_2) \tag{7}$$

$$Info(D_i) = -\sum_{m=1}^{M} \frac{|D_{m,D_i}|}{|D_i|} \log_2 \frac{|C_{m,D_i}|}{|D_i|} (i = 1, 2) \tag{8}$$

Among them, $|D_i|$ is the number of tuples in D_i, and $|C_{m,D_i}|$ is the number of tuples with label C_m in subset D_i.

The information gain obtained by dividing on A is gain (A).

$$Dain(A) = Info(D) - Info_A(D) \tag{9}$$

CART decision trees are prone to overfitting, especially when the dataset contains noise or a high degree of variability. This results in the tree being overly complex and capturing noise instead of general patterns. The reconstructed algorithm has the same structure as CART algorithm. According to different feature data, the partition criteria can be selected to build a decision tree, which reduces the complexity of decision tree construction and increases the generalisation ability. However, the over-fitting phenomenon of decision tree has not been reduced, so the problem of over-fitting needs to be dealt with.

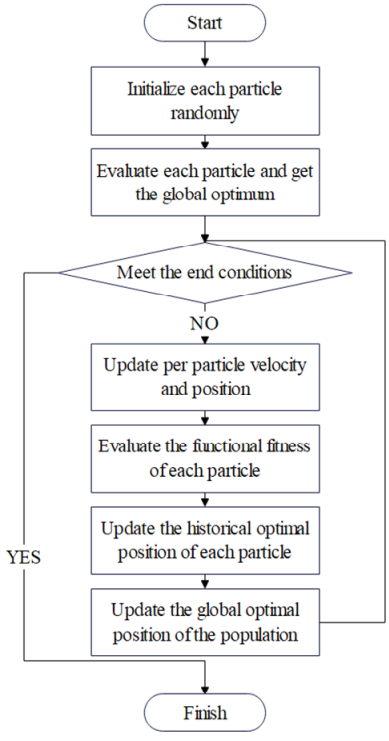
PSO algorithm is introduced to optimise the decision tree model, and the parameters of the decision tree are optimised by particle swarm, such as purity measurement criterion

(choose one of CART algorithm and reconstruction ID3 algorithm), maximum depth of the tree, maximum sample number of leaf nodes and other parameters. By using the method of optimised pruning, the training samples are classified correctly as much as possible, so as to prevent too many branches of the whole tree and avoid the occurrence of overfitting.

The particle adjusts its own flight direction and speed by tracking its personal historical optimal position (individual extremum P_{best}) and the optimal position in the whole population (global extremum g_{best}). The update of the position depends on the current velocity of the particle, the gap between the individual historical optimal position and the current position, and the gap between the optimal position in the population and the current position. At the same time, randomness is introduced to increase the breadth and depth of exploration. In this way, the entire particle swarm searches in the solution space and gradually approaches the global optimal solution.

The algorithm flow is shown in Figure 1.

Figure 1 Flowchart of PSO algorithm



CART is sensitive to small changes in the training data, leading to significant differences in the tree structure. This lack of stability can make it less reliable for financial fraud detection, where consistent performance is critical. The methods to solve this problem are as follows:

The monitoring model performs predictive classification on the data through the decision tree technology, in which the abnormal data in the process of financial fraud analysis shows a nonlinear relationship with the prediction results, and the data shows

high volatility and random characteristics. This algorithm monitors potential problems based on financial characteristic parameters, and verifies the accuracy of the model through performance evaluation.

The specific steps are as follows:

1 Data pre-processing, processing missing and abnormal values of financial characteristic data.

2 Set the optimisation objective function and fitness function. The fitness function equation is as follows:

$$fitness(T_i) = acc(T_i) \tag{10}$$

T_i represents the decision tree constructed by each group of parameters, $acc(T_i)$ represents the classification accuracy of each decision tree.

3 Set relevant parameters of PSO algorithm and parameter interval of decision tree.

4 The pre-processed data is substituted into the decision tree for training, and all particles are substituted into the model to find the optimal fit. The velocity and position of particles are continuously updated according to the value of stress.

5 Repeat the iteration. After the iteration, the fitness value and the parameters of the optimal decision tree are output.

6 The optimised parameters are substituted into the decision tree prediction model.

7 The test set and cross validation set are input into the decision tree model for prediction, and the prediction results are output.

The combination of cart and PSO can partly alleviate its strong dependence on impurity measures such as Gini index. The specific mechanism and effect are as follows:

3.2.1 Enhanced global search capability

- Split point optimisation: CART's greedy strategy based on Gini coefficient is easy to fall into local optimisation, while PSO can traverse the multidimensional feature space through swarm intelligence search to explore a better combination of split points and reduce the dependence on a single impurity measurement.
- Dynamic weight adjustment: PSO can iteratively optimise the feature weight, assist cart in identifying high discrimination feature interactions, and reduce the interference of noise features on segmentation decisions.

3.2.2 Extensibility of segmentation criteria

- Multi-objective optimisation: PSO supports user-defined fitness functions (such as classification accuracy, model complexity, feature interaction strength), which can balance impurity measurement and other indicators in the segmentation process to avoid the limitation of relying only on Gini index.

- Nonlinear relationship capture: The global search ability of PSO can assist cart to find complex nonlinear segmentation boundaries, making up for the dependence of traditional single split on the linear separability assumption.

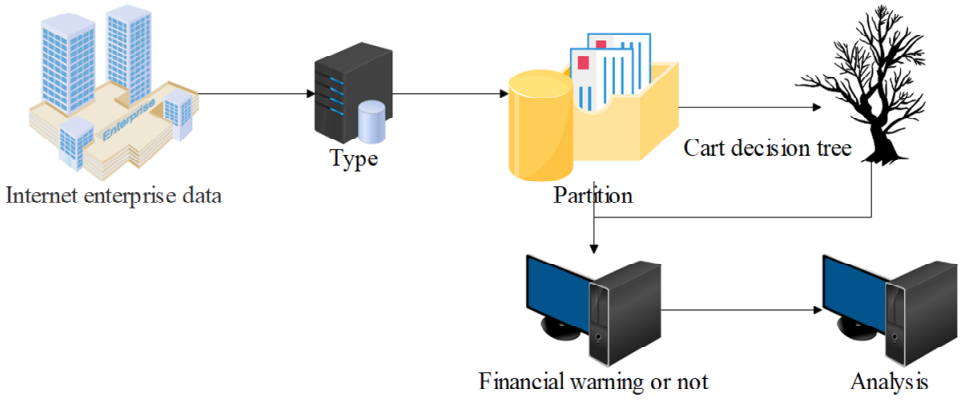
3.2.3 Anti-over fitting and stability improvement

- Regularisation effect: PSO avoids the over fitting of the model to the noise in the training data through population diversity, and generates a more robust tree structure. Dynamic pruning collaboration: combining pre pruning (limiting tree depth) and post pruning strategies, PSO can further optimise tree complexity and improve generalisation ability.

3.3 Model construction

The CART decision tree model is constructed by SPSS Modeler, and 70% of the sample data is set as the training set and 30% as the test set in the partition. This paper uses the five principal component factors previously screened as the input features of the enterprise financial risk warning indicators, and takes 'whether financial warning' as the target feature. The CART decision tree model construction flow chart is shown in Figure 2.

Figure 2 Modelling process of CART decision tree algorithm (see online version for colours)



Although pruning can effectively prevent data overfitting, after removing some nodes, the prediction results of the model also have certain deviations. Therefore, on this basis, a new method is proposed to solve this problem. Therefore, a variable α is selected as the equilibrium, and the loss function is defined as follows [formula (11)]:

$$C_\alpha(T) = C(T) + \alpha|T| \tag{11}$$

Among them, T represents any subtree, $|T|$ represents the number of leaf nodes on T, and α is a parameter that balances fitness and complexity. $C(T)$ represents the prediction error, which can be square error or Gini index, and α measures the fitness. On this basis, this paper proposes a new optimisation algorithm, which takes α as a positive value and searches for the optimal subtree $T(\alpha)$ in each given α to obtain the best subtree that

minimises $C_\alpha(T)$. T_0 is the best subtree when α is small. If the value of α is large, only a single root node needs to be selected as the best subtree.

Although the value of α can be infinite, the number of subtrees of T_0 is finite. T_n is the remaining subtree, which is generated based on the previous tree T_i. After removing certain internal nodes, the result T_{i+1} is obtained. On this basis, each subtree is interactively tested using test cases, and the best is selected to form a tree. The following is a subsequence [formula (12)].

$$T_0 > T_1 > T_2 > T_3 > \ldots > T_N \tag{12}$$

4 Experimental study

4.1 Data source

This paper takes manufacturing listed companies from 2017 to 2024 as the research object, and divides industries according to the industry standards of China Securities Regulatory Commission. The sample data required for the financial fraud recognition model mainly comes from Guotai'an database. The stocks of listed companies in the original sample mainly include three types: A shares, B shares and H shares. The difference of stock types may affect the final recognition result, so this paper selects A share samples with a large proportion of sample data. The report type of the sample is uniformly selected as the year-end report, and the listed companies marked as 'ST' and '*ST' are excluded. In addition, there are multiple companies in the sample that have violations in different years. In order to avoid duplication, only the data of the earliest year in which the company's violations occurred are screened. In order to avoid greater impact on the results, the outliers and missing values of related variables are deleted in this paper. Finally, this paper obtains 704 listed manufacturing companies and their 782 financial fraud index data, including 598 fraud data and 184 non-fraud data. In this paper, the train_test_split function is used to divide the sample data, and 70% of the sample data is used as the training set and 30% of the data is used as the training set.

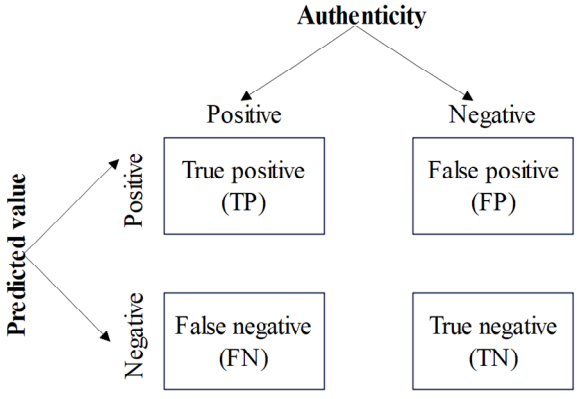
Confusion matrix is an important means to measure the performance of classification model, which can reflect the number of correct classifications and misclassifications of each class in the sample. For the recognition of two types of enterprise accounting fraud, '0' or '1' are usually selected to express it. In the process of fraud recognition, in order to complete the test of sample data, a proportion must be calculated for all sample data of the sample data test set, and the range of this proportion is [0, 1], which indicates the probability that the prediction of the sample by the fraud recognition mode is positive. If the probability is greater than 0.5, it means that the prediction of this sample is positive, that is, '1'. If it is the opposite, it is '0'. When the heat map is used for visual analysis, Figure 3 can be obtained.

For the construction and implementation of neural network, random forest (RF) model and CART decision tree recognition model, the effects of each model on financial fraud recognition will be further analysed. The recognition effects of specific models are shown in Tables 1 and 2.

For the processing of outliers, this paper uses adaptive threshold to identify sudden fluctuations in time series data, to avoid misjudgement caused by fixed threshold, especially in the scenario of severe market fluctuations Combined with financial

compliance constraints (build a rule engine to distinguish between real exceptions and normal business fluctuations.

Figure 3 Second-order confusion matrix



For the processing of missing values, the decision tree is constructed by associating external data, and the missing fields are inferred by using entity relationship (such as supplier customer link), which is better than the traditional mean interpolation.

4.2 Results

For the construction and implementation of neural network, RF model and CART decision tree recognition model, the effects of each model on financial fraud recognition will be further analysed. The recognition effects of specific models are shown in Tables 1 and 2.

Table 1 Financial fraud recognition results of each model-training set

	N_{ALL}	$N_{TP} + N_{TN}$	*Accuracy (%)*	*AUC (%)*
CART decision tree recognition model	542	423	77.22	69.30
RF recognition model	542	417	76.23	45.54
Neural network recognition model	542	417	76.23	56.43

Table 2 Financial fraud recognition results of each model-verification set

	N_{ALL}	$N_{TP} + N_{TN}$	*Accuracy (%)*	*AUC (%)*
CART decision tree recognition model	233	174	74.25	58.41
RF recognition model	233	176	75.24	50.49
Neural network recognition model	233	176	75.24	52.47

According to the results in Table 1 and Table above, this paper draws a visual image of the financial fraud recognition results. The confusion matrices of the validation set and the training set are plotted as shown in Figures 4 and 5.

In this paper, the ROC curve of the confusion matrix of the above models will be further drawn, and the results are in Figure 6.

Figure 4 Confusion matrix of each model training set (see online version for colours)

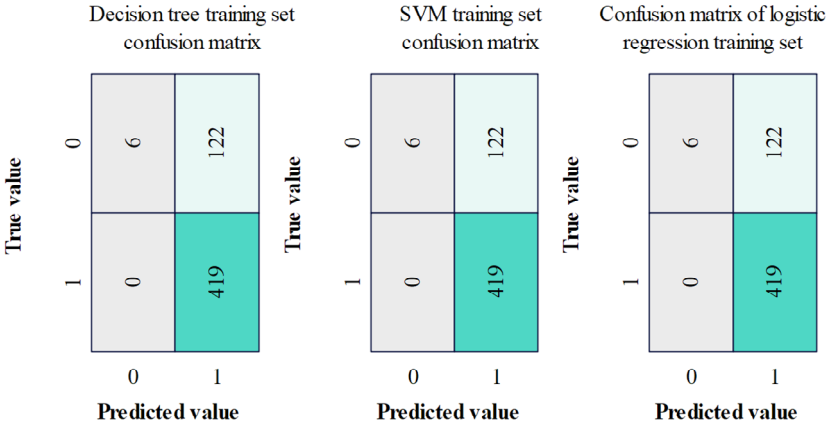
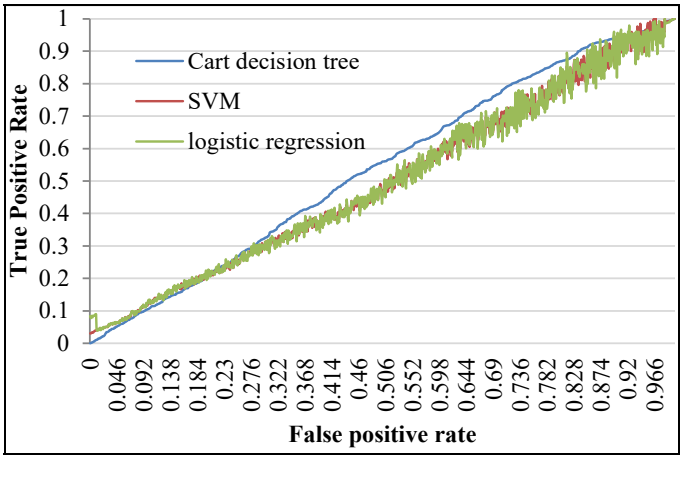


Figure 5 Confusion matrix of each model verification set (see online version for colours)

Figure 6 ROC curve of each model (see online version for colours)



Under this premise, the financial fraud recognition and analysis are carried out based on the data of listed companies in manufacturing industry from 2019 to 2024. The set-aside method is used to divide the data set into 70% training set and 30% validation set. Based on these data, programming calculations are performed respectively, and the results are in Figures 7 and 8.

Figure 7 Visualisation of prediction recognition of CART decision tree training set (see online version for colours)

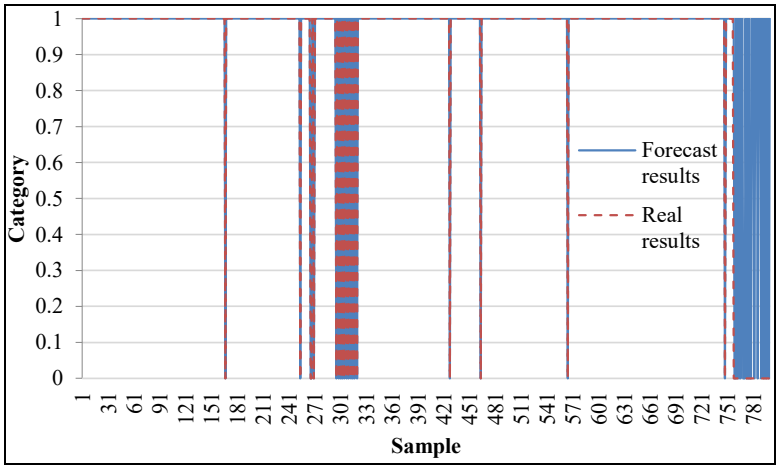
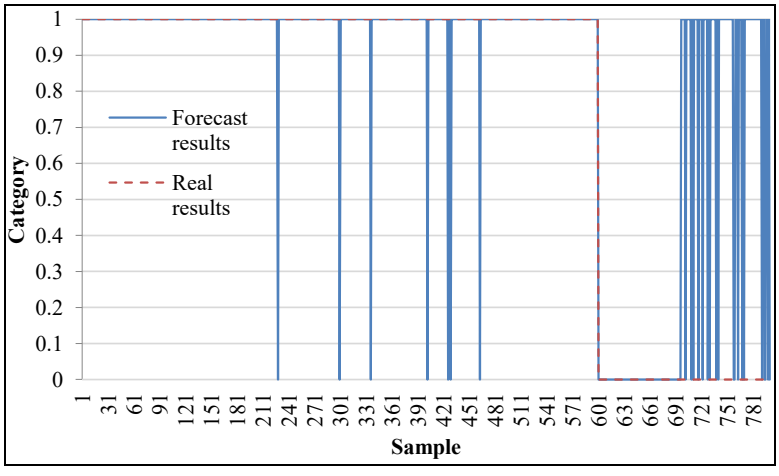


Figure 8 Visualisation of prediction recognition of CART decision tree validation set (see online version for colours)



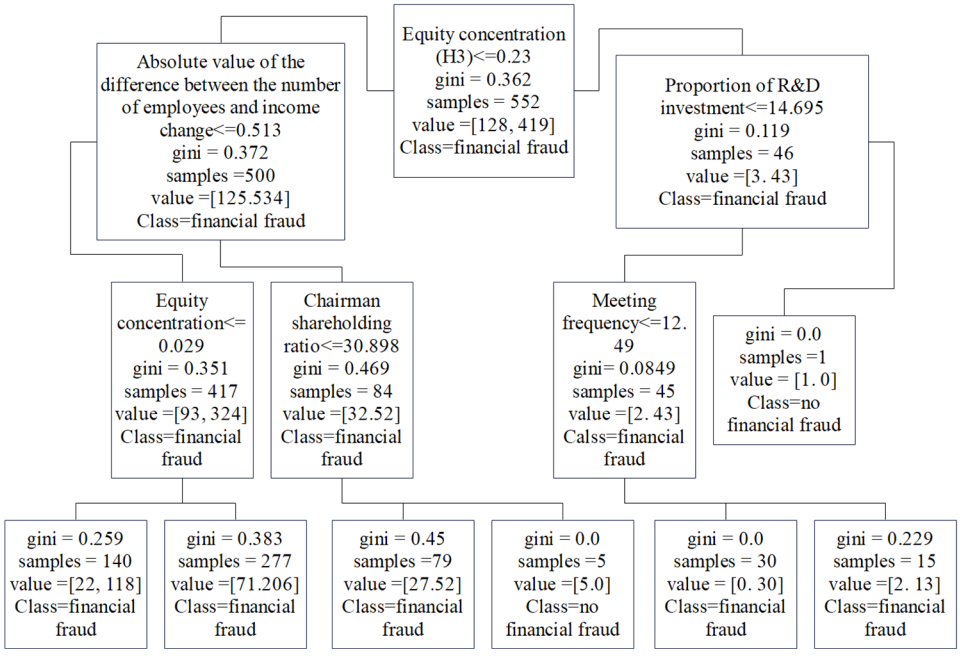
The generated decision tree is shown in Figure 9.

4.3 Analysis and discussion

Through the comparative analysis of the accuracy and AUC of the above methods, it can be found that AUC is a general evaluation method, and the AUC of the three recognition

methods is quite different. The AUC of CART tree recognition method is significantly higher than that of RF and neural network recognition methods, reaching 70%, which has a better recognition effect, while the AUC values of RF and neural network recognition models are only 57% and 46%, slightly higher than the random results, which indicates that the recognition effect of this model is general.

Figure 9 CART decision tree diagram



By analysing the financial fraud recognition model under the non-financial information in Figure 6, the recognition ability of CART decision tree model is obviously better than that of the other two models, and it should be used as the optimal recognition model in this paper.

Analysing Figures 7 and 8, it can be found that the parameter combination search can make the accuracy of CART decision tree model achieve the best effect.

According to CART decision tree theory, feature selection plays an important role in model construction. At the root node of the decision tree, the selected features have the greatest influence on sample segmentation, that is, the closer the index is to the root node, the more important it is. In the CART decision tree fraud recognition model constructed in this paper, X9 (ownership concentration) is regarded as the most critical feature, followed by X20 (proportion of R&D investment) and X21 (difference between the number of employees and income change). Through the drawn CART decision tree image, it can be clearly seen that the six indicators selected in this paper play an important role in predicting whether financial fraud occurs in listed companies. Among them, equity concentration, proportion of R&D investment, and difference between the number of employees and income are regarded as indicators that play a decisive role in identifying financial fraud.

A single cart decision tree is vulnerable to data noise, feature redundancy and over fitting problems in financial fraud detection. The introduction of PSO can improve the robustness of the model from the following aspects PSO uses swarm intelligence iteration to screen high discrimination feature subset, reduce the interference of redundant financial indicators on model decision-making, and improve the ability to capture the characteristics of hidden fraud; PSO allocates dynamic weights for different financial indicators, and strengthens the contribution of key features (such as abnormal cash flow and related party transaction frequency) to classification decisions; the PSO fitness function can simultaneously optimise the classification accuracy, false positive rate and characteristic dimension, and adapt to the balance between high accuracy and low false positive in financial fraud detection.

Because there are too many indexes and some of them are correlated, substituting all of them into the model will reduce the accuracy of the model, so this paper chooses to use stepwise linear regression method to screen the initial index system. After screening, most of the indicators are filtered out, and six significant variables are finally obtained, which further improved the accuracy of the fraud recognition model.

In the complex interactive feature extraction task, cart decision tree has the following differentiation advantages over RF, gradient lifting tree (such as XGBoost) or neural network by introducing pruning, feature selection and PSO:

4.3.1 Collaborative optimisation of feature selection and pruning

- Accurate feature screening: CART screens key features based on information gain or Gini coefficient in the feature selection stage, and combines with PSO algorithm to dynamically optimise the weight of feature subset, which can more efficiently capture the nonlinear interaction between features.
- Anti overfitting ability: Through pre pruning (limiting tree depth, minimum number of samples at nodes) and post pruning (pruning at cost complexity), avoid excessive dependence of redundant branches on noise features, and improve the generalisation expression ability of core interaction features.
- Comparison of RFs: RFs rely on random feature subsets to build multiple trees, which may reduce variance but may ignore key feature combinations; CART+PSO can focus on high discrimination interactive features through directional optimisation.

4.3.2 Model interpretability and computational efficiency

- Rule traceability: The tree structure generated by cart directly displays the feature splitting logic, which is convenient for analysing the action path of interactive features, and is superior to the black box feature of neural network.
- Lightweight computing: The complexity of cart model after pruning is significantly reduced. Combined with the parallel search mechanism of PSO, feature interaction optimisation can be quickly completed under limited computing resources, which is suitable for scenes with high real-time requirements.
- Contrast gradient lifting tree: Gradient lifting needs to generate multiple trees iteratively and superimpose regularisation constraints, which has high computational

cost; CART+PSO can effectively extract interactive features in a single tree framework.

4.3.3 Algorithm flexibility adapts to complex scenes

- Dynamic adaptation to data distribution: PSO algorithm adjusts feature weights through swarm intelligence iteration, which can adapt to the change of interactive feature importance under different data distribution, and is better than RF with fixed feature sampling.
- Multi-objective optimisation capability: PSO can simultaneously optimise classification accuracy, model complexity and feature dimensions, and balance the comprehensiveness and simplicity of interactive feature extraction, while traditional integration methods usually only optimise a single loss function.

Based on the above model analysis results, the following suggestions are put forward:

1 Investors should always be sensitive to the stock market, pay more attention to the capital market, and pay attention to the use of non-financial information published by the company. During this process, investors should fully understand the various types of information provided by the company, especially the company's non-financial information, and pay full attention to and make use of it.

2 For auditors, after accepting an audit project, they should conduct a detailed investigation and understanding of the auditee and his environment, and carry out the audit according to relevant requirements, so as to better identify and evaluate major errors. Moreover, they need to conduct comprehensive analysis and evaluation of non-financial information such as various situations in the same industry, relevant laws and regulations and other external factors, the specific circumstances of the audited unit, and the choice of accounting policies.

3 For enterprises, it is necessary to comprehensively improve the security risk recognition ability of enterprises by analysing their own personnel and external environment information, establish a regular assessment and evaluation mechanism and attach importance to the role of internal audit.

4 For the regulatory authorities, it is necessary to improve the punishment and corresponding joint and several liability for enterprises with false accounting information.

5 Conclusions

Financial fraud not only poses a great challenge to the authority of information disclosure system, but also seriously damages the integrity foundation of the market. It weakens market confidence and causes serious damage to investor interests. This paper studies the intelligent recognition model of financial fraud behaviour based on CART decision tree, and draws the following conclusions:

1 The AUC of CART tree recognition method is significantly higher than that of RF and neural network recognition methods, reaching 70%, which has a better

recognition effect, while the AUC values of RF and neural network recognition model are only 57% and 46%, which are slightly higher than the random results, indicating that the recognition effect of this model is general.

2 Parameter combination search can make the accuracy of CART decision tree model achieve the best effect.

3 The stepwise linear regression method is selected to screen the initial index system. After screening, most of the indicators are filtered out, and six significant variables are finally obtained, which further improves the accuracy of the fraud recognition model.

In practical application, how to convey early warning information to users in time is a key problem. Therefore, future research can explore how to establish an efficient information transmission mechanism to improve the timeliness and effectiveness of early warning. Based on previous studies, this paper selects financial indicators and non-financial indicators to build the index system from a comprehensive perspective, but does not use unstructured data, such as media reports, community reviews and other text data. These text data may contain some information to help the model more accurately identify or predict. Future research can use natural language processing methods to include text data into the index system after processing.

Declarations

All authors in this paper declare that they don't have any conflict of interest.

Acknowledgements

This paper is supported by Study on Coordination, Coupling and collaborative Promotion between Ecological Environmental Protection and High-quality Economic Development in Dongting Lack Ecological Economic Zone (Xiang Jiao Tong No. 2023, 361-23C0445).

References

Agu, E.E., Abhulimen, A.O., Obiki-Osafiele, A.N., Osundare, O.S., Adeniran, I.A. and Efunniyi, C.P. (2024) 'Utilizing AI-driven predictive analytics to reduce credit risk and enhance financial inclusion', *International Journal of Frontline Research in Multidisciplinary Studies*, Vol. 3, No. 2, pp.20–29.

Akash, T.R., Islam, M.S. and Sourav, M.S.A. (2024) 'Enhancing business security through fraud detection in financial transactions', *Global Journal of Engineering and Technology Advances*, Vol. 21, No. 2, pp.79–87.

Ali, A., Abd Razak, S., Othman, S.H., Eisa, T.A.E., Al-Dhaqm, A., Nasser, M. and Saif, A. (2022) 'Financial fraud detection based on machine learning: a systematic literature review', *Applied Sciences*, Vol. 12, No. 19, pp.9637–9650.

Bello, H.O., Idemudia, C. and Iyelolu, T.V. (2024) 'Integrating machine learning and blockchain: conceptual frameworks for real-time fraud detection and prevention', *World Journal of Advanced Research and Reviews*, Vol. 23, No. 1, pp.56–68.

Bhaktiar, R.E. and Setyorini, A. (2021) 'The effect of the fraud triangle on fraud financial statements (case study on manufacturing companies in the food and beverage subsector)', *Jurnal Mantik*, Vol. 5, No. 2, pp.841–847.

Chang, J.W., Yen, N. and Hung, J.C. (2022) 'Design of a NLP-empowered finance fraud awareness model: the anti-fraud chatbot for fraud detection and fraud classification as an instance', *Journal of Ambient Intelligence and Humanized Computing*, Vol. 13, No. 10, pp.4663–4679.

Gautam, A. (2023) 'The evaluating the impact of artificial intelligence on risk management and fraud detection in the banking sector', *AI, IoT and the Fourth Industrial Revolution Review*, Vol. 13, No. 11, pp.9–18.

Gayam, S.R. (2021) 'Artificial intelligence for financial fraud detection: advanced techniques for anomaly detection, pattern recognition, and risk mitigation', *African Journal of Artificial Intelligence and Sustainable Development*, Vol. 1, No. 2, pp.377–412.

Guo, L., Song, R., Wu, J., Xu, Z. and Zhao, F. (2024) 'Integrating a machine learning-driven fraud detection system based on a risk management framework', *Applied and Computational Engineering*, Vol. 87, No. 1, pp.80–86.

Hassan, M., Aziz, L.A.R. and Andriansyah, Y. (2023) 'The role artificial intelligence in modern banking: an exploration of AI-driven approaches for enhanced fraud prevention, risk management, and regulatory compliance', *Reviews of Contemporary Business Analytics*, Vol. 6, No. 1, pp.110–132.

Hernandez Aros, L., Bustamante Molano, L.X., Gutierrez-Portela, F., Moreno Hernandez, J.J. and Rodríguez Barrero, M.S. (2024) '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.

Johora, F.T., Hasan, R., Farabi, S.F., Akter, J. and Al Mahmud, M.A. (2024) 'AI-powered fraud detection in banking: safeguarding financial transactions', *The American Journal of Management and Economics Innovations*, Vol. 6, No. 6, pp.8–22.

Kannagi, A., Mohammed, J.G., Murugan, S.S.G. and Varsha, M. (2023) 'Intelligent mechanical systems and its applications on online fraud detection analysis using pattern recognition K-nearest neighbor algorithm for cloud security applications', *Materials Today: Proceedings*, Vol. 81, No. 2, pp.745–749.

Kotagiri, A. and Yada, A. (2024a) 'Improving fraud detection in banking systems: RPA and advanced analytics strategies', *International Journal of Machine Learning for Sustainable Development*, Vol. 6, No. 1, pp.1–20.

Kotagiri, A. and Yada, A. (2024b) 'Crafting a strong anti-fraud defense: RPA, ML, and NLP collaboration for resilience in US finances', *International Journal of Management Education for Sustainable Development*, Vol. 7, No. 7, pp.1–15.

Mytnyk, B., Tkachyk, O., Shakhovska, N., Fedushko, S. and Syerov, Y. (2023) 'Application of artificial intelligence for fraudulent banking operations recognition', *Big Data and Cognitive Computing*, Vol. 7, No. 2, pp.93–102.

Narsimha, B., Raghavendran, C.V., Rajyalakshmi, P., Reddy, G.K., Bhargavi, M. and Naresh, P. (2022) 'Cyber defense in the age of artificial intelligence and machine learning for financial fraud detection application', *IJEER*, Vol. 10, No. 2, pp.87–92.

Odeyemi, O., Ibeh, C.V., Mhlongo, N.Z., Asuzu, O.F., Awonuga, K.F. and Olatoye, F.O. (2024) 'Forensic accounting and fraud detection: a review of techniques in the digital age', *Finance & Accounting Research Journal*, Vol. 6, No. 2, pp.202–214.

Shoetan, P.O. and Familoni, B.T. (2024) 'Transforming fintech fraud detection with advanced artificial intelligence algorithms', *Finance & Accounting Research Journal*, Vol. 6, No. 4, pp.602–625.

Shoetan, P.O., Oyewole, A.T., Okoye, C.C. and Ofodile, O.C. (2024) 'Reviewing the role of big data analytics in financial fraud detection', *Finance & Accounting Research Journal*, Vol. 6, No. 3, pp.384–394.

Singh, A., Jain, A. and Biable, S.E. (2022) 'Financial fraud detection approach based on firefly optimization algorithm and support vector machine', *Applied Computational Intelligence and Soft Computing*, No. 1, pp.1468015–1468027.

Xu, L., Wang, J., Xu, D. and Xu, L. (2022) 'Integrating individual factors to construct recognition models of consumer fraud victimization', *International Journal of Environmental Research and Public Health*, Vol. 19, No. 1, pp.461–473.