International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 6414-6422

https://doi.org/10.70517/ijhsa463551

Research on Intelligent Anomaly Recognition of Financial Data Based on Decision Tree Algorithm

Yunmin Zhu[1,*]
[1] School of Business, Sichuan University Jinjiang College, Pengshan, Sichuan, 620860, China
Corresponding authors: (e-mail: zhuyunmin502@163.com).

Abstract With the growth of financial data scale, the traditional financial anomaly identification method is difficult to meet the demand for efficient monitoring. This study constructs an intelligent anomaly identification model for financial data based on decision tree algorithm, which improves the accuracy and efficiency of financial anomaly detection. The study utilizes variance reduction and information gain as decision criteria, adopts post pruning strategy and advancement technique to optimize the model performance, and combines a variety of data preprocessing methods to process financial data containing 4616 listed companies in 26 industries. The experimental results show that the constructed model achieves an accuracy rate of 0.869, a precision rate of 0.830, a recall rate of 0.791, and an F1 value of 0.810 in financial data anomaly recognition, which is better than the comparative algorithms such as CNN, LSTM, and BP. In the financial fraud identification task, when both corporate governance indicators and financial control indicators are used, this model achieves an accuracy rate of 0.884 and an AUC value of 0.835, which is an improvement of 7.3% to 22% over the comparison models. The case study verifies the effectiveness of the model in identifying financial anomalies, and achieves accurate identification of anomalous signals such as abnormal net profit per capita and high equity pledge. The study shows that the decision tree-based intelligent anomaly identification model for financial data can accurately extract key features in financial time series data, improve the accuracy of anomaly identification, and provide support for the early warning of corporate financial risks.

Index Terms decision tree algorithm, financial data, anomaly identification, information gain, financial fraud, corporate governance

I. Introduction

With the development of China's economy, the number of listed companies is increasing, and the financial and accounting information provided by listed companies is an important basis for market participants to make investment decisions [1], [2]. However, in recent years, the competition among enterprises has been intensifying, and a few listed companies, for their own interests, try to commit financial fraud by forging accounting documents and altering financial statement data, and financial data fraud is common [3]-[5]. This has caused great harm to both the market and investors, and undermined the principle of fairness and justice in the capital market [6]. Therefore, accurately identifying abnormalities in the financial data of listed companies is of great significance in protecting the rights and interests of investors and maintaining the order of the capital market [7], [8].

However, the existing methods that rely solely on experts from regulatory agencies to analyze publicly disclosed financial data of listed companies often have shortcomings such as low data credibility, excessive time lag in data disclosure, and non-uniformity of identification data indicators [9]-[11]. With the continuous development of artificial intelligence, the application of decision tree algorithm brings new technical means and application scenarios to this field [12], [13]. Decision tree algorithm is a classification algorithm based on tree structure, by splitting and pruning the data to get a decision tree, the leaf nodes of the tree represent the classification results of the data [14], [15]. Decision trees can handle various types of data and work well for both discrete and continuous features [16]. Intelligent anomaly identification of financial data based on decision tree algorithm is a simple and effective anomaly identification method with the advantages of easy to understand and strong interpretability, which can significantly improve the efficiency and accuracy of anomaly identification by quantitatively analyzing the correlation between financial and non-financial data so as to build a dynamic detection model [17]-[20].

Financial data is an important embodiment of an enterprise's operating conditions, and its authenticity and reliability have an important impact on investor decision-making, supervision by regulatory authorities and the enterprise's own development. However, anomalies in financial data, especially financial fraud, not only harm the interests of investors, but also disturb the order of the capital market, and even lead to corporate bankruptcy and

economic crisis. With the explosive growth of data volume and the complexity of financial operations, traditional methods of identifying financial data anomalies have been difficult to meet the demand for accurate and efficient monitoring. Therefore, the development of intelligent financial data anomaly identification technology has become particularly important. In recent years, machine learning technology has made remarkable achievements in the field of data analysis, especially the decision tree algorithm provides new ideas and methods for financial data anomaly identification due to its advantages of strong interpretability, high computational efficiency, and applicability to mixed data. Based on the decision tree algorithm, this study constructs an intelligent anomaly identification model for financial data, and improves the accuracy and efficiency of financial anomaly identification through scientific feature engineering and model optimization. The study firstly analyzes the principle of decision tree algorithm in detail, explores the decision criteria based on variance reduction and information gain, and optimizes the model performance by adopting post pruning strategy, misclassification loss matrix and advancement technique. Then, it constructs a multi-level financial data anomaly identification model structure, which can not only discriminate whether the financial data is anomalous or not, but also identify specific anomalous variables and patterns. In terms of data processing, the study comprehensively processes 28,058 financial data containing 4,616 listed companies in 26 industries over the past 6 years, including feature analysis, missing value processing, data slicing and sampling, etc., to provide high-quality training data for the model. Aiming at the category imbalance problem of financial data, the study adopts the Borderline-SMOTE algorithm to sample a few categories of samples, which effectively improves the model's ability to recognize abnormal samples. In the experimental evaluation, the study compares the constructed model with classical models such as CNN, LSTM and BP to verify its superiority in the tasks of financial data anomaly identification and financial fraud identification. Further, the study verifies the effectiveness of the model in practical applications through the case study of Company A, especially the accurate identification of financial anomaly signals such as abnormal net profit per capita and high pledge of equity. The main innovations of this study are: applying the decision tree algorithm to the field of financial data anomaly identification to improve the accuracy of anomaly identification through scientific model construction and optimization strategies; integrating corporate governance indicators and financial control indicators to significantly improve the effect of financial fraud identification; and proposing an interpretable framework for financial anomaly identification, which not only identifies the anomalies, but also analyzes the causes of the anomalies.

II. Research on intelligent anomaly identification in financial data

II. A. Decision Tree Algorithm

The decision tree algorithm [21], [22] is a common machine learning algorithm used for prediction. The goal of this algorithm is to build a tree model that accurately predicts the outcome of a new data point based on the characteristics of that data point. The basic algorithmic steps to build a decision tree count are as follows:

(1) Select the best segmentation features: based on the criteria select the features that can best segment the data into subsets with the lowest variance or highest information gain. Here the highest variance reduction value is selected. The selection steps are as follows:

1) Calculate the target variable variance. The variance is calculated as in Eq:

$$\mathrm{var}\,iance(y)=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 \tag{1}$$

where n denotes the number of data points, y_i denotes the target variable, and $\bar{y}$ denotes the mean of the target variable.

2) Calculate the variance reduction of each feature. The variance reduction of the features is shown in Eq:

$$\begin{aligned}&\mathrm{var}\,iance\_reduction(x)=\mathrm{var}\,iance(y)\\&-weighted\_sum\left(\mathrm{var}\,iance(y_i)\right)\end{aligned} \tag{2}$$

where $\mathrm{var}\,iance(y)$ denotes the variance of the target variable, and $weighted\_sum\left(\mathrm{var}\,iance(y_i)\right)$ is the a weighted sum of the variance of the target variable for each subset resulting from partitioning the data according to the value of feature x, with the weight of each subset being proportional to the number of data points in that subset.

(3) Select the feature with the highest variance reduction as the best feature for segmenting the data.

(2) Create node: create a node for the selected feature and split the data into subsets based on the criteria for that node.

(3) Predictive value: at each leaf node, the algorithm assigns the mean or median value of the target variable for that subset as the predictive value for any new data point belonging to that subset.

(4) At each node of the decision tree, steps (1) to (3) are repeated, and the algorithm recursively splits the subset based on the selected criteria until the stopping criterion is satisfied. The stopping criterion can be based on

maximum tree depth, minimum number of leaf nodes, or other criteria. In addition, various pruning techniques can be used to prevent overfitting of the model to the training data.

Information gain [23] is also one of the more commonly used decision criteria, which measures the amount of information gained by segmenting the data based on specific features. The formula for information gain is:

$$\inf ormation\_gain(x) = entropy(y) - weight\_sum(entropy(y_i)) \tag{3}$$

where $entropy(y)$ is the entropy of the target variable, and $weight\_sum(entropy(y_i))$ is a weighted sum of the entropies of the target variable for each subset of the data produced by partitioning the data according to the value of feature x. The weight of each subset is proportional to the number of data points in that subset. The entropy of the target variable is calculated using the following equation:

$$entropy(y) = -\sum_{i=1}^{k} p_i \log_2(p_i) \tag{4}$$

where p_i is the proportion of data points in the subset with i values for the target variable, and $\log_2$ is a logarithm with base 2.

The decision tree algorithm can handle both numeric and categorical (non-numeric) data, which makes it a generalized algorithm. However, there are some limitations. First, if the decision tree is too complex, the decision tree is prone to overfitting, which occurs when the tree is too specific to the training data, leading to poor application to new data. Also, the decision tree may be sensitive to small changes in the data, which may result in generating different trees. The following section details the decision tree generation strategy and parameter selection used and the financial data anomaly identification model constructed based on it.

II. B.Intelligent recognition model construction based on decision tree algorithm

II. B. 1) Decision Tree Generation Strategy and Parameter Selection

(1) Pruning strategy

In order to prevent the decision tree from being overfitted, the pruning rule of the decision tree should be adjusted first. A post pruning strategy is used in this study, where a fully grown tree is first formed and then pruned according to the error rate before and after pruning. Two adjustable parameters are usually included:

One is the minimum sample size that can be represented by any node, i.e., the minimum record of sub-branches, usually denoted by m. The subpopulation size can be used to limit the number of split points for any branch of the decision tree, which will only continue to split if two or more subsequent sub-branches include no less than m records from the training set. Increasing this parameter will help avoid overfitting to noisy data. The other is the pruning confidence (CF value), where an increase in the CF value will result in a higher level of model abstraction, so that the model is no longer overly focused on a single sample. The CF value is taken from 0% to 100% and the expected number of false positives is determined using the following formula:

$$N \times U_{CF}(E, N) \tag{5}$$

where N is the total number of training samples, E represents the number of samples that were misjudged among them, CF is the pruning confidence, and U_{CF} represents an upper bound on the error rate. The higher the pruning purity, the larger the value of CF and the smaller and more concise the generated decision tree. Lowering the purity value results in a more accurate decision tree, but may introduce the problem of overfitting.

In the next model, a pruning purity of 75 is used and the minimum number of records in the sub-branches is 2. In order to ensure the rationality of the pruning process, a global pruning pattern is used

(2) Type coding and misclassification loss matrix

Decision tree can deal with multi-categorization problems, assuming a total of k categories, then first of all, the k categories will be assigned a certain code, such as from 1 to k for each of the k categories assigned a natural number, this process is the type of coding. After coding the categories, the error rate of misclassification between different categories can be clearly distinguished.

When the decision tree misclassifies the i th sample as the j th sample, the error can be represented by (i, j), and so on, all error types can be represented by a two-dimensional vector, and all error types can form a matrix. It is up to the user to set the penalty values for different error types, with higher penalties meaning that the occurrence of that type of error is less likely to be tolerated. With the concept of penalty values, the user can build an error loss matrix to adjust the importance of a particular type of misjudgment.

For example, in process monitoring, the DT classifier responsible for monitoring categorizes the input samples into controlled and out-of-control, and in practice it is likely that the error rate at which out-of-control data is judged to be controlled will be tightly controlled due to the greater emphasis placed on whether or not the out-of-control occurs. Therefore, the error loss matrix established is shown in Table 1.

Table 1: Error loss matrix

Actual \ Forecast	Under control	Out of control
Under control	0.0	1.0
Out of control	3.0	0.0

(3) Advancement technique

In this paper, we adopt the following advancement strategy: let the set of training samples be S, the number of advancement trials be T, and the decision tree generated from the t th training be C_t. Then the final decision tree composite from these T decision trees is denoted as C^*, w_i^t is the weight of the samples, p_i^t is the normalization factor of w_i^t, and β^t is the adjusting factor of the weights.

Define a 0-1 function $\theta^t(i)$ that is 1 when sample i is misclassified and 0 otherwise.The training steps for the sample are described below:

1) Set the initial weight values: let $t=1$ and $w_i^t=1/n$.

2) Calculate $p_i^t=\dfrac{w_i^t}{\sum_{i=1}^{n} w_i^t}$ such that $\sum_{i=1}^{n} p_i^t=1$.

3) Assign a normalized weight value p_i^t to each sample and construct C^t based on this probability distribution.

4) Calculate the error rate $\varepsilon^t=\sum_{i=1}^{n} p_i^t\theta_i^t$ for the t th decision tree for the sample.

5) If $\varepsilon^t>0.5$, end the whole training process so that $T=T-1$. If $\varepsilon^t=0$, end the entire training process, making $t=T$. Otherwise, go to the next step.

6) Calculate $\beta^t=\varepsilon^t/(1-\varepsilon^t)$.

7) In the next iteration, multiply the weights of the samples by the adjustment factor β^t if they were misclassified in the previous modeling, otherwise leave them unchanged.

8) End the whole iteration process when $t=T$, otherwise make $t=t+1$ and go to step 2).

The final composite decision tree C^* is obtained by weighted summation from C_1 to C_t. All models trained after this paper default to using $T=6$ to train the decision tree model.

II. B. 2) Model structure for identifying financial data anomalies

For process monitoring and anomaly identification of p variables, a single decision tree classifier DT_0 is trained to be used for process monitoring, and a total of p decision tree classifiers $DT_1, DT_2, \cdots, DT_p$ are trained to be used to perform anomaly identification of p variables, respectively. for anomaly identification. Among them, DT_0 classifies the input financial data as controlled or uncontrolled, and the data classified as uncontrolled will be input into DT_1 to DT_p in turn, and this group of classifiers will determine whether the mean values of the corresponding p variables are shifted upward, downward, or not shifted, and in total, derive the process occurrence of the The pattern of shifts. The structure of the decision tree-based financial data anomaly identification model is shown in Figure 1.

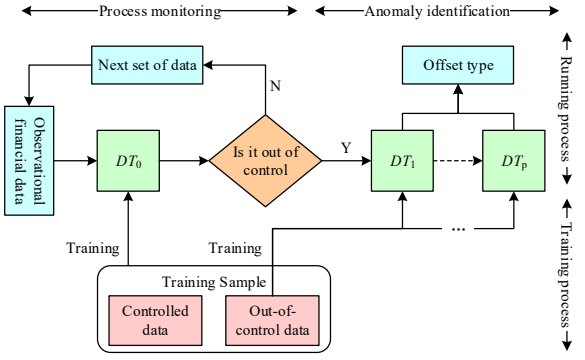

Figure 1: Financial data anomaly identification model structure

II. C.Pre-processing of financial data

II. C. 1) Data sources and data description

The data used in this paper is derived from the title data of the 9th Data Mining Challenge A competition. There are three data tables in the dataset, and the specific contents of each data table are as follows:

(1) The stock code of each company and its corresponding industry name.

(2) Financial statement data disclosed by each listed company.

(3) The Chinese interpretation of the characteristics corresponding to each field in the financial statement data.

The datasheet contains 28,058 financial data of 4616 listed companies in 26 different industries for the last six years. Among them, there are 20,167 financial data containing the feature field of whether or not financial fraud in the first five years, and 7,891 data not containing the feature field of whether or not financial fraud in the sixth year, and there are 337 financial features in each financial data. Among all the financial features, the FLAG feature is listed as a marker tag of whether or not the company is engaged in financial forgery, and FLAG takes the value of 0 if it is not engaged in forgery, and if it is engaged in forgery FLAG tag takes the value of 1.

II. C. 2) Data pre-processing

Data processing is a fundamental part of model building, and an effective data processing process can improve the performance of the model and the accuracy of the final results. Data processing is more like a potential investment, high-quality data is the key to output accurate results, when the data is processed comprehensively and carefully enough, the model will eventually return excellent results. Commonly used data preprocessing methods include missing value processing, duplicate value processing and outlier processing.

In this paper, we mainly process the data as follows:

(1) Analyze the actual meaning of data fields. The data used in this paper has a total of 337 characteristic fields, and after analyzing them, it is found that nine characteristic fields, such as stock code, actual disclosure time, release time, report deadline, report type, accounting interval, consolidation mark, accounting standard and currency code, do not have much impact on the actual financial data statement in essence. In order to avoid their impact on the final results, the above columns will be deleted from the operation.

(2) Missing value processing. A total of 553 data fields are missing, the proportion of missing more than 95%, the proportion of missing data can be seen is relatively high, so it is necessary to deal with the missing value of the data. For the missing proportion of 75% of the data below the missing value filling operation, this paper uses the KNN nearest neighbor method to fill the missing values.KNN nearest neighbor method is a common method of missing value data processing, its precision and accuracy is much higher than the use of the mean or 0 for filling. It confirms the final filling value after analyzing and estimating the proximity points in the sample, which is a more scientific and accurate method than using the mean or 0 to fill in the missing values.

(3) Outliers and duplicates processing. In terms of outlier processing, since the model constructed in this paper is to capture the few categories of financial fraud through the analysis of data, any outlier point that deviates from the distribution of data will become the primary object of observation focused on by the model, so this paper does not deal with outliers. After validation and analysis, the data does not contain duplicate values, which again do not need to be processed.

(4) Data cut. Enterprises in different industries often choose appropriate management methods and business models according to the characteristics of their industries, which results in different industries having different emphasis in their financial statements. Therefore, it is considered to segment the data by industry. There are a total of 26 industry classifications in the financial sample data. Among all industries, the top three industries with the largest data volume are "Manufacturing", "Information Transmission, Software and Technical Services", and "Wholesale and Retail Trade". In order to complete the construction of the financial data anomaly identification model of this paper, the remaining 23 industries are uniformly categorized as the remaining industries.

(5) Data sampling processing. After the completion of data slicing, it is necessary to carry out feature engineering operations for each industry. After screening out the index system used in the final industries through feature engineering, the statistical analysis of the financial data found that the number of financial fraud is only about 1% of the number of financial fraud that did not occur, which belongs to the serious data imbalance, and it is necessary to carry out balancing processing of the data. Data sampling refers to the operation that needs to be carried out when the data are in the state of extremely biased category classification, where the classification bias refers to the large difference between the number of samples labeled as 1 and the number of samples labeled as 0. In order to improve the effectiveness and running speed of the final model, the Borderline-SMOTE algorithm [24], [25] is used to sample the few class samples that are at the sample boundary.

III. The effect of intelligent anomaly identification of financial data and case studies

III. A. Effectiveness of identifying anomalies in financial time series data

All the algorithms in the paper are realized by programming in Python language and standard libraries, and the experimental codes are run on the same physical machine, which is configured as follows: Intel(R)Core(TM) i5-7300HQ CPU, 2.50GHz, 16G RAM. The experiment uses the financial dataset collected above for evaluating the performance of the financial data anomaly recognition algorithm.

The time series anomaly identification results on the financial dataset are shown in Figure 2, with the light gray areas being the financial data identification anomalies and the black lines being the normalized financial feature data. It can be seen from the figure that the algorithm selected in this paper can accurately identify the anomalous data in the financial features.

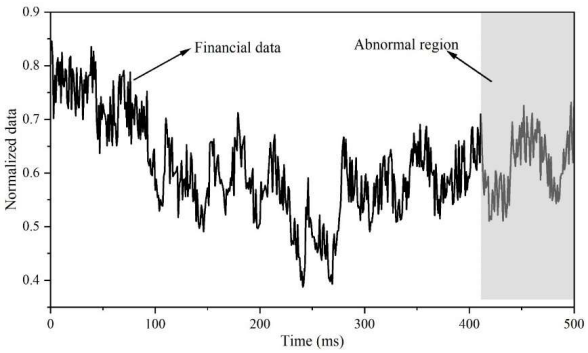

Figure 2: The time sequence exception recognition results on the financial data set

In addition, Convolutional Neural Network (CNN), Long Short-Term Memory Network (LSTM), and BP Neural Network (BP) models are selected and chosen for comparison. The comparison results of anomaly recognition on financial dataset between this paper's algorithm and the comparison algorithm are shown in Table 2. From the comparison data in the table, it can be seen that CNN, LSTM and BP algorithms are less effective in terms of recall R and F1 values when it comes to financial data anomaly recognition, and the variation is very obvious. While this paper's algorithm achieves better results in the accuracy, precision, recall and F1 values on financial data anomaly recognition, which are 0.869, 0.830, 0.791, 0.810 respectively, and the detection effect is more stable. In addition, the running time of this paper's algorithm is also similarly better than the comparison method, faster than the comparison method by 6.21s~12.93s.From the above experiments, it can be seen that the intelligent anomaly recognition model based on the decision tree algorithm is able to extract the temporal nature of the time-series data of the financial indexes very well. And it selects the more important features and patterns in the input data, reduces the computational cost and improves the recognition accuracy of abnormal data.

Table 2: Exception identification of the financial data set

Model	Accuracy	Precision	Recall	F1-score	Running duration
CNN	0.729	0.785	0.632	0.700	20.19s
LSTM	0.772	0.792	0.697	0.741	13.47s
BP	0.757	0.738	0.639	0.685	15.27s
Ours	0.869	0.830	0.791	0.810	7.26s

III. B. Financial fraud identification based on financial anomaly data

This section selects the corporate governance data of some A-share market listed companies from the above dataset for the years 2020-2024, and divides them into corporate governance indicators and financial control indicators based on previous studies.

Corporate governance indicators include: percentage of executives, size of board of directors, percentage of independent directors, two positions in one, percentage of shares held by the first largest shareholder, percentage of shares held by the top five shareholders, shareholding checks and balances, state-owned or privately owned, percentage of shares held by institutional investors, whether the directors and supervisors have financial backgrounds, and whether the directors and supervisors have overseas backgrounds.

Financial indicators include: ratio of other receivables to total assets, financial expense ratio, gross profit margin index, tax profit ratio, fixed asset turnover ratio, shareholders' equity turnover ratio, growth rate of return on total assets, return on assets ratio, net cash flow from operating activities per share, and cash interest coverage multiple.

In this section, 128 fiscal years of companies involved in fraudulent behavior are selected as fraud samples, and financial fraud identification is carried out using an intelligent model through corporate governance index data (CGID) and financial control index data (FCID). Five random cross-validation is used, and the sample set is randomly divided by 8:2, and a sample of 0.2 is selected as the test set, and the accuracy of the model identification is verified out-of-sample. Meanwhile, Convolutional Neural Network (CNN), Long Short-Term Memory Network (LSTM), and BP Neural Network (BP) models are selected for comparison. Table 3 shows the financial fraud recognition rate of different models.

From the table, it can be seen that the accuracy, precision, recall, F1-score and AUC of different intelligent models for financial fraud recognition that only contains financial indicators are lower than the accuracy, precision, recall, F1-score and AUC of the recognition that contains both corporate governance indicators and financial indicators, which suggests that the corporate governance indicators can significantly improve the accuracy of the intelligent models in recognizing the financial fraud enterprises. . In addition, the comparison between different models reveals that this paper's model has the highest accuracy, precision, recall, F1-score and AUC for financial fraud recognition. Taking the example of including both corporate governance index data and financial control index data, this paper's model improves the accuracy rate by 7.3% to 22%, increases the precision rate by 32.3% to 36.3%, increases the recall rate by 5.8% to 12.3%, increases the F1-score by 20.2% to 25.2%, and improves the AUC by 10% to 16.1% compared with the comparison model. The study proves that the model in this paper is sensitive to the company's financial data, can accurately identify abnormal data and improve the identification of financial fraud. At the same time, corporate governance indicators have an obvious complementary effect on financial indicators to identify financial fraud.

Table 3: Different models financial fraud recognition

Model	Data source	Accuracy	Precision	Recall	F1-score	AUC
CNN	FCID	0.643	0.400	0.519	0.452	0.640
	CGID+ FCID	0.811	0.456	0.608	0.521	0.674
LSTM	FCID	0.735	0.364	0.567	0.443	0.567
	CGID+ FCID	0.788	0.496	0.673	0.571	0.701
BP	FCID	0.643	0.432	0.542	0.481	0.667
	CGID+ FCID	0.664	0.491	0.644	0.557	0.735
Ours	FCID	0.824	0.764	0.704	0.733	0.793
	CGID+ FCID	0.884	0.819	0.731	0.773	0.835

III. C. Case of abnormal financial data of Company A

This section takes Company A as an example for the case study of intelligent anomaly identification of financial data.Founded in 1992 and located in Nanyang City, Henan Province, Company A was successfully listed on the Shenzhen Stock Exchange in 2014, and is a leading enterprise in the commercial pig breeding industry. In terms of its main business, Company A mainly engages in the breeding, slaughtering and sales of live pigs, and its main products include commercial pigs, piglets, breeding pigs and pork products. In terms of business model, the Company currently adopts a “full self-raising, full chain, intelligent” business model.

III. C. 1) Industry net profit per capita

In the industry business dimension, common financial fraud identification signals are also net profit per capita and output per capita. In this section, the study uses the model constructed above to compare the changes in the financial indicators of net profit (million yuan), total number of employees (persons), and net profit per capita (million yuan per person) from 2022 to 2024 for four companies in the same commercial hog farming industry as Company A. Table 4 shows the net profit per capita of Company A compared with the industry from 2022 to 2024.

From the table, it can be seen that Company A has higher net profit per capita compared with comparable companies in the same industry.In 2023, except for Peer 4, where net profit per capita can reach 100,300 yuan due to the relatively low number of employees, Company A has a higher net profit per capita of more than 50,000 yuan compared with other companies.The gap between net profit per capita of Company A and the industry further widens in 2023, except for Industry 4, where net profit per capita is 38,300 yuan, and in the Company A reaches 114,800 yuan per capita when all other companies have negative net profit per capita.Such a large gap between Company A and the industry in terms of net profit per capita also conveys signals of financial anomalies.

Table 4: A comparison between company and industry per capita net profit

Year	Index	Same industry 1	Same industry 2	Same industry 3	Same industry 4	A company
2022	Net profit	1179149	1716941	2146285	1841975	2764651
	Total number of employees	337016	371059	421609	294062	402607
	Per capita net profit	3.50	4.63	5.09	6.26	6.87
2023	Net profit	1605810	2179782	2499552	2620692	4718026
	Total number of employees	342454	380912	429788	261378	411377
	Per capita net profit	4.69	5.72	5.82	10.03	11.47
2024	Net profit	-1285936	-1815077	-1064506	1418485	4809669
	Total number of employees	349866	388171	436106	370141	419003
	Per capita net profit	-3.68	-4.68	-2.44	3.83	11.48

III. C. 2) Pledge of Company Equity

In the corporate governance dimension, it is necessary to pay attention to the equity pledges of listed companies, important personnel changes in management, and abnormal reductions in controlling shareholders' holdings. In this section, the statistics of controlling shareholders' pledges of Company A and its comparable companies from 2022 to 2024 are presented to identify abnormalities in the financial data of Company A's governance dimension.

Table 5 shows the controlling shareholder pledges of Company A and its peers. It can be seen that the controlling shareholders of Company A and peers2 are involved in high pledges of equity, with pledge ratios of 70% or more. Among them, Company A 2024 controlling shareholders' pledge ratio reaches 80.11%. Therefore, special attention needs to be paid to the characteristic of high pledge of controlling shareholders of Company A. Controlling shareholders' equity pledge will increase the financial risk of listed companies, and high equity pledge is one of the recognizable signals of financial fraud.

Table 5: A company and the controlling shareholder pledge in the same industry

	Year	Number of pledge stock	proportion
Same industry 1	2022	11789130	1.17%
	2023	11928774	1.94%
	2024	13327418	1.62%
Same industry 2	2022	343289135	78.73%
	2023	441307234	71.28%
	2024	419213072	72.73%
Same industry 3	2022	23276407	1.98%
	2023	11490972	3.77%
	2024	28169072	2.84%
Same industry 4	2022	102689201	35.24%
	2023	123098745	34.76%
	2024	92682698	27.19%
A company	2022	320359182	73.49%
	2023	419432075	75.28%
	2024	519357157	80.11%

IV. Conclusion

In this study, an intelligent anomaly recognition model for financial data based on decision tree algorithm is constructed, and the following conclusions are drawn through experimental validation and case study analysis: firstly, the constructed model performs excellently in financial time-series data anomaly recognition, with a running time of 7.26 s, which is 6.21 s to 12.93 s faster than the comparative method, and at the same time achieves an accuracy rate of 0.869. Second, in financial fraud identification, the model combines corporate governance indicators and financial control indicators to achieve an accuracy rate of 0.819 and an F1 value of 0.773, which is 20.2% to 25.2% higher than the comparison model. Third, corporate governance indicators have a significant complementary effect on financial fraud identification, and the model accuracy rate is improved from 0.824 to 0.884. Fourth, the case study of Company A shows that the model can effectively identify abnormal signals in the industry, such as the per capita net profit of Company A of $114,800 per capita in 2024, which is in stark contrast to the

negative value of the same industry, as well as the controlling shareholders' pledge ratio of 80.11%, which is much higher than the industry average. Finally, with its interpretability and computational efficiency, the decision tree algorithm successfully extracts the key features of the time-series data of financial indicators, which provides an effective tool for the early warning of corporate financial risks and the detection of abnormalities by regulatory authorities. Future research can explore the integration of decision tree and other algorithms to further improve the ability to identify anomalies in complex financial data.

References

[1] Huian, M. C. (2015). The usefulness of accounting information on financial instruments to investors assessing non-financial companies. An empirical analysis on the Bucharest Stock Exchange. Journal of Accounting and Management Information Systems (JAMIS), 14(4), 748-769.

[2] Hung, D. N., Ha, H. T. V., & Binh, D. T. (2018). Impact of accounting information on financial statements to the stock price of the energy enterprises listed on Vietnam's stock market. International Journal of Energy Economics and Policy, 8(2), 1-6.

[3] Sun, G., Li, T., Ai, Y., & Li, Q. (2023). Digital finance and corporate financial fraud. International Review of Financial Analysis, 87, 102566.

[4] Yin, L., Sun, G., & Kong, T. (2025). Regional big data development and corporate financial fraud. Pacific-Basin Finance Journal, 102693.

[5] Dong, W., Liao, S., & Zhang, Z. (2018). Leveraging financial social media data for corporate fraud detection. Journal of Management Information Systems, 35(2), 461-487.

[6] Dyck, A., Morse, A., & Zingales, L. (2024). How pervasive is corporate fraud?. Review of Accounting Studies, 29(1), 736-769.

[7] Ogunmokun, A. S., Balogun, E. D., & Ogunsola, K. O. (2022). A strategic fraud risk mitigation framework for corporate finance cost optimization and loss prevention. International Journal of Multidisciplinary Research and Growth Evaluation, 3(1), 783-790.

[8] Ahmed, M., Choudhury, N., & Uddin, S. (2017, July). Anomaly detection on big data in financial markets. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (pp. 998-1001).

[9] Lopata, A., Gudas, S., Butleris, R., Rudžionis, V., Žioba, L., Veitaitė, I., ... & Zwitserloot, M. (2022). Financial data anomaly discovery using behavioral change indicators. Electronics, 11(10), 1598.

[10] Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: a review of anomaly detection techniques and recent advances. Expert systems With applications, 193, 116429.

[11] Huang, D., Mu, D., Yang, L., & Cai, X. (2018). CoDetect: Financial fraud detection with anomaly feature detection. Ieee Access, 6, 19161-19174.

[12] Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. Journal of applied science and technology trends, 2(01), 20-28.

[13] Rana, K. K. (2014). A survey on decision tree algorithm for classification. International journal of Engineering development and research, 2(1), 1-5.

[14] Alsagheer, R. H., Alharan, A. F., & Al-Haboobi, A. S. (2017). Popular decision tree algorithms of data mining techniques: a review. International Journal of Computer Science and Mobile Computing, 6(6), 133-142.

[15] Zhang, Q. (2022). Financial data anomaly detection method based on decision tree and random forest algorithm. Journal of Mathematics, 2022(1), 9135117.

[16] Bakumenko, A., & Elragal, A. (2022). Detecting anomalies in financial data using machine learning algorithms. Systems, 10(5), 130.

[17] Lokanan, M., Tran, V., & Vuong, N. H. (2019). Detecting anomalies in financial statements using machine learning algorithm: The case of Vietnamese listed firms. Asian Journal of Accounting Research, 4(2), 181-201.

[18] Hong, S., Wu, H., Xu, X., & Xiong, W. (2022). Early warning of enterprise financial risk based on decision tree algorithm. Computational Intelligence and Neuroscience, 2022(1), 9182099.

[19] Krishna, M. H., Nithin, K., Charmitha, G., Vignesh, T., Ch, V., & Kuchibhotla, S. (2023, February). Studies on Anomaly Detection Techniques. In 2023 7th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 813-817). IEEE.

[20] Malashin, I., Masich, I., Borodulin, A., Bukhtoyarov, V., Nelyub, V., & Tynchenko, V. (2024, December). Anomaly Detection in Financial Data Using GA-Optimized MLP Models. In 2024 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 278-285). IEEE.

[21] Yutao He,Jiamin Guo,Huan Ping,MingLiang Zhu & Sujith Mangalathu. (2025). A robust force-finding framework for tensegrity structures using gradient-boosting decision trees and Latin hypercube sampling. Structures,76,108863-108863.

[22] Bingzhi Zhong,Hui Xie,Zhengkai Zhang & Yan Wen. (2025). Non-linear effects of ecoacoustic indices on urban soundscape assessments based on gradient boosting decision trees in summer Chongqing, China. Building and Environment,278,112984-112984.

[23] David Howard Foster & Kinjiro Amano. (2025). Contributed Talks III: Surface color information gained in searching natural scenes.. Journal of vision,25(5),58.

[24] David Riañ,o,Eva Onaindia,Miguel Cazorla,P. Toribio,R. Alejo... & J.H. Pacheco-Sanchez. (2012). Using Gabriel graphs in Borderline-SMOTE to deal with severe two-class imbalance problems on neural networks. Frontiers in Artificial Intelligence and Applications,248,29-36.

[25] Yongchao Li,Jianping Chen,Chun Tan,Yang Li,Feifan Gu,Yiwei Zhang & Qaiser Mehmood. (2020). Application of the borderline-SMOTE method in susceptibility assessments of debris flows in Pinggu District, Beijing, China. Natural Hazards,105(3),1-24.