Research on the Financial Data Fraud Detection of Chinese Listed Enterprises by Integrating Audit Opinions

Leiruo Zhou[1], Yunlong Duan [2*], Wei Wei[3]
[1] International Business School, Yunnan University of Finance and Economics,
Kunming, Yunnan 650221 China
[e-mail: leiruozhou@gmail.com]
[2] Science and Technology Department, Yunnan University of Finance and Economics
Kunming, Yunnan 650221 China
[e-mail: lanjiangZ431@gmail.com]
[3] School of Computer Science and Engineering, Xi'an University of Technology
Xian, Shanxi 710048 China
[e-mail: weiwei@xaut.edu.cn]
[*]Corresponding author: Yunlong Duan

Received March 25, 2023; revised July 6, 2023; revised September 18, 2023; accepted November 26, 2023; published December 31, 2023

Abstract

Financial fraud undermines the sustainable development of financial markets. Financial statements can be regarded as the key source of information to obtain the operating conditions of listed companies. Current research focuses more on mining financial digital data instead of looking into text data. However, text data can reveal emotional information, which is an important basis for detecting financial fraud. The audit opinion of the financial statement is especially the fair opinion of a certified public accountant on the quality of enterprise financial reports. Therefore, this research was carried out by using the data features of 4,153 listed companies' financial annual reports and audits of text opinions in the past six years, and the paper puts forward a financial fraud detection model integrating audit opinions. First, the financial data index database and audit opinion text database were built. Second, digitized audit opinions with deep learning Bert model was employed. Finally, both the extracted audit numerical characteristics and the financial numerical indicators were used as the training data of the LightGBM model. What is worth paying attention to is that the imbalanced distribution of sample labels is also one of the focuses of financial fraud research. To solve this problem, data enhancement and Focal Loss feature learning functions were used in data processing and model training respectively. The experimental results show that compared with the conventional financial fraud detection model, the performance of the proposed model is improved greatly, with Area Under the Curve (AUC) and Accuracy reaching 81.42% and 78.15%, respectively.

Keywords: Audit opinion on financial statements, Bert model, data imbalance, LightGBM model, fraud financial data.

This job is supported by National Key R&D Program of China (No.2022YFE0138600). This job is also supported by Supported by Natural Science Foundation of Shaanxi Province of China(2021JM-344) and Open Fund for Chongqing Key Laboratory of Computational Intelligence (NO.2020FF02).

http://doi.org/10.3837/tiis.2023.12.001 **ISSN : 1976-7277**

1. Introduction

With the booming of the stock market, more and more enterprises raise funds by listing. However, to rapidly expand the scale of business, some enterprises often disclose financial information after whitewashing, causing huge losses to investors. According to the National Association of Certified Fraud Examiners (ACFE) [1], 10% of white-collar crimes involve falsifying financial statements. ACFE classified occupational fraud into three types: embezzlement, corruption, and financial statement fraud, with financial statement falsification causing the greatest damage. Financial statements are documents that describe the financial position, operating results, and cash flow of a company. On March 17th, 2023, the Chinese Ministry of Finance imposed a fine of 200 million yuan upon Deloitte Touche Tohmatsu Certified Public Accountants, and its Beijing branch was forced to suspend business for three months because Deloitte Touche Tohmatsu failed to audit the financial fraud of China Huarong Asset Management Co., Ltd. from 2014 to 2019. Chinese listed companies are mandated by the Securities Law and the China Securities Regulatory Commission to publish their financial statements quarterly and annually, which guarantees the feasibility of the study.

In the present environment of globalization, the detection of financial fraud is more important than ever. However, the traditional manual measurement method is inefficient because of the numerous financial indicators, which promotes the research boom of machine learning in the field of financial fraud detection. Financial fraud is generally done through the following ways [2]: (1) occupation of company assets, (2) false disclosure, (3) violation of guarantee regulations, (4) fraudulent listing, (5) unauthorized change in the use of funds, (6) improper general accounting, (7) falsified records, (8) delayed disclosure, (9) fictitious profits, (10) material omissions, and (11) fictitious assets. From the perspective of fraud, most of them can be identified through the digital indicators of financial statements, so the current research focuses more on the digital indicators modeling of listed enterprises [3, 4].

In fact, an experienced investor's assessment of financial fraud relies not only on the numerical data of the financial statements but also on other textual information, such as audit opinions of the statements, textual modifiers [5], and Management Discussion and Analysis [6]. Audit statements are the opinions of certified public accountants on the quality of listed enterprises' financial reports, so audit opinions for investors to distinguish financial fraud has important significance. Take a Chinese fraud enterprise annual audit opinion for example, "We remind the users of financial statements: The operating income of enterprise A declined significantly for three consecutive years. The net loss after deducting non-recurring gains and losses in YYYY year was XX million yuan. By the end of YYYY year, current liabilities were XX million yuan higher than current assets. Company A has disclosed the proposed improvement measures in the notes to the financial statements, but significant uncertainty remains regarding its ability to go as a going concern." The above audit opinions conveyed pessimistic views on the enterprise operation, which no doubt can help improve the detection of financial fraud. At present, no authority has publicly published any research on the audit opinion of financial statements in the detection of financial fraud. Consequently, how to integrate the audit opinion of financial statements into the model becomes the main motivation of this study.

In recent years, the vigorous development of Natural Language Processing technology (NLP) in artificial intelligence has brought feasibility to this study. NLP is a subject that uses computer technology to analyze, understand, and process language as its object. This paper presents a method of financial fraud detection based on audit opinions. First, NLP's Bert (Bidirectional Encoder Representations from Transformer) technology was used to analyze

financial audit opinions and transform text information into digital features with emotional information [7, 8, 9]. Second, the digital audit opinions and financial digital features were integrated as the training data of LightGBM (Light Gradient Boosting Machine) [10] model. Finally, LightGBM after completion of training was able to conduct fraud discrimination on new enterprise samples. Due to the large difference between the number of fake and non-fake enterprises (1:85 sample ratio of text collection), when the model encounters extremely imbalanced data samples, it often leads to over-fitting or under-fitting of training [11]. To solve this problem, the study first used the combined technology of over-sampling and under-sampling to enhance the sample data and then used Focal Loss function in Bert model to force the model to pay more attention to the fake enterprise samples in the process of feature learning.

2. Related Work

In recent years, to protect the rights and interests of investors, corporate financial fraud detection research has become a hot topic in the financial field. These studies mainly focus on feature selection in financial statements and the selection of classification models. The core focus of our research in our research will also be on these two areas. This section will introduce the current research status in these fields.

2.1 Financial Fraud Detection Indicators Selection

Financial statements are an indispensable component of the reports submitted by publicly traded companies. They reflect the recent and future financial conditions of a company and serve as a vital source of data for financial fraud detection. Currently, feature selection in financial statements is primarily categorized into the following two types (**Table 1**).

The first category is the utilization of financial numerical data. W. T. Mongwe and K. M. Malan [3] used the income statement, balance sheet, and cash flow statement in 1,560 annual financial reports of enterprises in South Africa to calculate 13 financial ratios to measure the financial status of enterprises. The results showed that the ratio of debt to total operating income of fraudulent entities is very high, while non-fraudulent entities have high liquidity ratios. E. Hytis et al. [4] used the financial numerical indicators of listed enterprises in the Athens Stock Exchange to calculate 22 financial ratios as characteristic factors, which were used to measure profitability ratio, capital structure ratio, current ratio, and operating ratio respectively. The experimental results demonstrated the effectiveness of feature factors. X. Z. Yuan et al. [2] used Gibbs random search to extract 8 feature factors from 183 financial digital features under the framework of big data to describe the financial fraud behavior of listed companies, and their AUC value reached 0.76, These eight feature factors are: non-deductible net return on equity, growth rate of construction in progress, growth rate of prepayments, interest expense/total operating revenue, net investment income rate/total operating revenue, other income/total operating revenue, other receivables/total assets, and long-term borrowings/total assets. Their research provides us with a reference basis for utilizing financial digital data, such as how to calculate higher-dimensional financial ratios, how to filter a large number of financial ratio factors and the financial attributes represented by each financial ratio. For example, Formula (1) can calculate the cash ratio, which represents the solvency:

$$\text{Cash Ratio} = (\text{Cash} + \text{Marketable Securities}) / \text{Current Liabilities} \quad (1)$$

In this study, we will draw from the above-mentioned feature factors to construct a large-scale financial digital database as our research data source. However, in research on the Identification of Financial Fraud, reliance is not solely on financial digital data; reference is also made to company news, statements by management, or evaluations by other impartial and authoritative organizations regarding the company.

The second category is the use of non-numerical data in financial statements, which is relatively less studied. S. L. Humpherys et al. [5] believed that enterprises with financial fraud would use a lot of modifiers in financial statements to hide the real financial situation, they used NLP technology to research Management Discussions And Analysis of financial statements and finally proved the hypothesis with a precise rate of 67%, Successfully used NLP technology in financial fraud detection research. However, their research primarily focuses on the impact of textual information on financial fraud and does not explore the role of financial ratios, resulting in low accuracy in identification. C. S. Throckmorton et al. [6] hypothesized that combinations of different categories of financial features could enhance overall fraud detection performance. They adopted a method that combines financial ratios, management discussions, and analysis with text to detect financial fraud, achieving an AUC value of 0.75, which is an improvement of approximately 0.07 compared to using only financial numerical features. Because, the authors used the GLRT model as a classifier, and traditional machine learning models have limited capabilities when dealing with complex language and speech information. Additionally, collecting speech samples in practical research is extremely challenging.

X. G. Wu and S. Y. Du [12] utilized 74 numerical data features along with managerial analysis and discussion text data as inputs for a deep learning model for fraud detection. They used word embedding models to extract text information and, in handling imbalanced data, employed Focal_loss as the model's loss function. In experiments, the AUC value increased by approximately 0.066 compared to the model that did not use text data. Regarding text feature extraction, this paper employs the Bert model, which has stronger contextual information extraction capabilities compared to simple word embedding models. In dealing with imbalanced data, this paper draws inspiration from their use of the Focal_Loss function but differs in that we handle text data and financial numerical data separately. We utilize the Bert model of deep learning incorporating Focal_Loss to digitize text information, while we employ the more mature and stable statistical method "Smote + RandomUnderSampler" to address data sample imbalance.

Currently, in authoritative publicly available articles on financial fraud identification using text features, the features mostly consist of sections related to management discussions and analysis in financial reports. There has been no research on using audit opinion features combined with financial statement data to identify financial fraud. But audit opinion is the impartial result issued by certified public accountants to the audited enterprise. Therefore, this research is of significant value as it utilizes a combination of financial statement audit opinions and financial statement data for financial fraud identification.

Table 1. Progress in the selection of financial fraud detection indicators

Indicators	Study	Data sources	Key findings
Fin	W. T. Mongwe &K. M. Malan (2020) [3]	South Africa	The ratio of debt to total operating income of fraudulent entities is very high, while non-fraudulent entities have high liquidity ratios. This means that financial numerical data are useful in the detection of fraud. Liquidity ratio and debt ratio will be important foundations for our financial fraud detection research.
Fin	E. Hytis et al. (2022) [4]	Athens	They used financial numerical indicators to calculate 22 financial ratios representing profitability ratio, capital structure ratio, current ratio, and operating ratio. These ratios provide references for our research.
Fin	X. Z. Yuan et al. (2022) [2]	China	The author utilized Gibbs random search to extract 8 feature factors that characterize financial fraud risk from a pool of 183 financial features, achieving an AUC value of 0.76. These features also serve as a basis for our research.
TXT	S. L. Humpherys et al. (2011) [5]	America	The author believes that fraudulent companies tend to use a significant number of qualifiers in their financial statements to conceal their true financial condition. Therefore, they employed NLP (Natural Language Processing) techniques to extract the management discussion and analysis section of the financial reports, achieving an accuracy rate of 67%. This substantiates the usefulness of text in financial fraud analysis.
Fin+TXT	C. S. Throckmorton et al. (2015) [6]	America	The author, for the first time, combined speech, text, and financial ratios in research for financial fraud detection. The experimental results demonstrated that if each category provides independent and complementary information about financial fraud, then cross-category feature combinations can enhance overall detection performance. This confirms the validity of our research.
Fin+TXT	X. G. Wu & S. Y. Du (2022) [12]	China	The author collected 74 numerical data features along with managerial analysis and discussion text data to construct a financial indicator system as input data for the deep learning model. It's worth noting that they utilized word embedding models to extract text features and used the Focal_Loss function to address data imbalance issues. In the experiments, compared to the model that did not utilize text data, the AUC value increased by approximately 0.066.

To sum up, this paper believes that financial digital information is the key to identifying corporate financial fraud, while text information is the key auxiliary index to improve the recognition rate. Therefore, the combination of financial statements audit opinions, and financial data is the focus of this paper.

2.2 Financial Fraud Detection Model

In essence, the detection of financial fraud is a classification process, which classifies enterprises to fraud and non-fraud. There are many machine learning classification models for reference available now. Several mainstream classification models will be introduced.

A. A. Akinyelu and A. O. Adewumi [13] adopted an improved support vector machine (SVM) model to detect electronic fraud. To solve the problem that SVM classification speed decreases with the increase of data set size, they introduced two filter-based instance selection techniques. Although the classification speed is significantly improved, classification accuracy is sacrificed. S. Noels et al. [14] proposed a new graph distance metric based on the earth mover's distance to calculate the similarity between two enterprises, and the experimental results reached the expected results in many cases. However, the authors only focused on the balance sheet portion of the financial statements to construct the graph distance measurements, and many financial ratios were not utilized. F. K. Alarfaj et al. [15] adopted the three-layer architecture based on convolutional neural network (CNN) to conduct fraud detection in the European credit card benchmark data set, and the accuracy of the model reached 99.72%. Although neural network models have good performance in classification tasks, their training is time-consuming, unsuitable for real-time trading environments, and requires a large number of training samples. M. N. Ashtiani and B. Raahemi [16] analyzed 47 financial fraud articles and conducted a large number of comparative experiments. One of their conclusions indicated that neural network models perform best among 17 models, and language models that extract text information (Word2Vec, Doc2Vec, BERT) are the future development trend. Therefore, the adoption of the Bert model is one of the main studies in this article. Considering the training time of the model, the neural network model is only used for text feature extraction. G. Ke et al. [10] proposed the LightGBM model to solve the inefficiency of the GBDT model which traverses all sample data every time it learns features. It not only inherits the astonishing performance of GBDT model classification, but also improves the training speed nearly 20 times, far exceeding the neural network model, and has a wide range of applications in the financial industry. Y. Zhang et al. [17] used the improved LightGBM model to detect Ponzi schemes in the blockchain field, and solved the problem of incomplete features and insufficient algorithms to detect Ponzi schemes. Experiments were conducted on the real data set of Ethereum, and the results demonstrated significant improvements in the F-value and AUC index compared with the most advanced methods. This study also used this model for the final classification task.

The model proposed in this paper is a combination of a neural network model and statistical learning model, which not only satisfies the effective extraction of text features but also satisfies the high-speed classification efficiency. In the experiment, we selected and compared the above classification models.

3. Data Acquisition and Processing

Research data will be introduced in this section: firstly, the source of the data, secondly the establishment of the database, and finally the data enhancement techniques to solve the problem of imbalance between the proportion of fraud and non-fraud samples.

3.1 Acquisition of Financial Fraud Data

To obtain the experimental data of the model, this study obtained corporate financial statement information and financial fraud detection information. The data of enterprise financial

statements mainly came from the ChinaStock Market & Accounting Research Database (CSMAR), which starts from the needs of academic research, drawing on the professional standards of the global authoritative database, and combined with China's national conditions to develop the financial field database. This study extracted basic data from the topics of "Financial Statements," "Annual, Interim, and Quarterly Release Date" and "Audit Opinion on Financial Reports" in the CSMAR database, covering from 2014 to 2020 and including non-financial enterprises. As for the data of the fraud information tables, they were mainly from penalty information from the relevant official website in China and identified enterprises that falsified financial data.

3.1.1 "FINANCIAL STATEMENTS"

The data of "financial statements" are mainly derived from the company's financial statements, especially the income statement, balance sheet, and cash flow statement. In terms of functionality, they can be divided into four categories: solvency, operating capacity, profitability, and growth capacity. **Table 2** shows the categories of some of the main indicators in "financial statements".

Table 2. Financial analysis indicators

Solvency	Asset-liability ratio
	Cash ratio
Operational Capacity	Turnover of total assets
	Days of accounts receivable turnover
Profitability	Return on equity
	Operating cost ratio
Growth ability	Revenue growth rate
	Net profit growth rate

3.1.2 "DATE OF PUBLICATION OF ANNUAL, MID-TERM AND QUARTERLY REPORTS"

"Date of Publication of Annual, Mid-term and Quarterly Reports" gives the date of first disclosure of financial reports. According to the requirements of the measures for The Administration of Information Disclosure of Listed Companies, listed companies are required to publish their financial statements within a specified time. Therefore, this article assumes that high-quality enterprises will disclose their financial information on schedule, while problematic enterprises will disclose their financial statements later.

3.1.3 "AUDIT OPINIONS ON FINANCIAL REPORTS"

"Audit Opinions on Financial Reports" provides the type of audit and the auditor's opinion of an accounting firm. It is a relatively objective evaluation of the financial situation of a company. **Table 3** presents the financial audit opinion of a fraudulent company. From the comments, it can be seen that the accounting firm has a pessimistic sentiment towards the company's operations. By using NLP technology, it is possible to digitize this textual information.

Table 3. Audit opinions of an enterprise involved in financial fraud

Stkcd	AccountingDate	Audittyp	Adtremark
A	YYYY-MM-DD	Reservation of opinion	Matters resulting in reservations: Material deficiencies in your internal control over the complete identification of related party relationships in your financial reporting process. Omit intermediate information. At the same time, we cannot determine whether related party transactions that may not be identified will have an impact on the accounting treatment in the financial statements.

3.1.4 Acquisition of Corporate Punishment Information

For the acquisition of financial fraud enterprise information, the study collected all penalty information from the SMAR database "Violation Information Summary Table". Due to the small amount of such information and a time delay, we also used crawler technology to collect information on penalties from the China Securities Regulatory Commission (CSRC), the Shanghai Stock Exchange (SSE), and the Shenzhen Stock Exchange (SZSE); then identified the information involving "fictitious profit", "falsified records", "major omission", "false disclosure", "fraudulent listing", "unauthorized change of the use of funds" and "fictitious profit" and other keywords as counterfeit enterprises; and used regular expressions to match the enterprise code, enterprise name, and fraud date; and finally built the “fraud detection information base”.

3.2 The Establishment of Financial Fraud Database

In the paper, the collected "financial statements" and "annual, interim, and quarterly release dates" were used to establish a "digital database of financial statements", while "Audit Opinion of Financial Reports" was used to establish "Audit Opinion Information Database of financial statements". In the establishment of the two databases, a "fraud detection Information database" was used to label each enterprise sample, and the results are shown in **Table 4** and **Table 5**.

Table 4. Sample digital database of financial statements

Stkcd	StkName	AccountingDate	Operating_cost	FLAG
000001	Ping An Bank	2019/1/1	84411000000	0
000002	Vanke	2019/1/1	234000000000	0

Table 5. Sample form of Audit Opinion Information Base for financial statements

Stkcd	AccountingDate	Audittyp	Adtremark	FLAG
A	YYYY-MM-DD	Reservation of Opinion	Matters resulting in reservations: Material deficiencies in your internal control over the complete identification of related party relationships in your financial reporting process. Omit intermediate information. At the same time, we cannot determine whether related party transactions that may not be identified will have an impact on the	1

			accounting treatment in the financial statements.	
B	YYYY-MM-DD	Standard unqualified opinion	N/A	0

Finally, we obtained 385 indicators for financial data and 1 item for audit opinion information, with a total of 4,153 listed enterprises in China, 18 industries and 19,003 samples. The model feature learning is mainly based on numerical indicators. Through longitudinal analysis of financial numerical indicators, it is found that there are two problems to be solved: (1) there are different degrees of missing values in 385 indicators and 19,003 samples; (2) The sample distribution of non-counterfeiting enterprises and counterfeiting enterprises was extremely imbalanced, reaching 85:1.

To solve the problem of missing values, the study firstly deleted the indicators with more than 30% missing values in the financial data and secondly deleted the sample data of non-fraudulent enterprises with more than 20% missing values. **Table 6** shows the data distribution before and after deletion. For the remaining missing values, the K-nearest neighbors (KNN) algorithm [18] was used for interpolation. Finally, a complete sample data was obtained, and we normalized it to control the numerical granularity between 0 and 1. After processing for missing values, the total number of samples was 11116, and including 10988 no-fake samples and 128 fake samples.

Table 6. Comparison before and after missing values treatment

Items	Before	After
Number of indicators	385	86
Number of data samples	19,003	11,116

3.3 Treatment of Imbalanced Financial Fraud Data

The study used data enhancement technology to solve the problem of imbalanced sample distribution, that is, the gap in the proportion of sample number between counterfeiting enterprises and non-counterfeiting enterprises was too large. In this paper, the proportion of positive and negative samples collected was 85:1. It should be clear that the function of the training set was to let the model learn the segmentation hyperplane of positive and negative samples, but the data imbalance interfered with the learning process of the model, so the data enhancement technology was used in the training set. The function of the test set was to test the generalization ability of the model after training by using the original distribution of historical samples, so the test set had to maintain the original distribution.

Two techniques for data enhancement are available: undersampling and oversampling. Undersampling is the random sampling of a large class of samples so that the data are similar to the number of a small class. Oversampling is to use an algorithm to analyze and simulate the distribution of a small class of samples, and automatically generate new samples to add to the small class. A single undersampling will destroy the original distribution of data, while a single oversampling may lead to over-fitting of the model.

Financial statement digital database and audit opinion information database need to be handled separately. This paper adopted the combination of oversampling and random undersampling in the statement digital database [19]. The SMOTE algorithm [20, 21, 22] was used in oversampling. SMOTE oversampling technique calculated the European distance between each fraudulent enterprise and all other fraudulent enterprises and selected the nearest few in accordance with formula (2) to generate new fraudulent enterprise samples.

$$x_{new}^{i} = x + rand(0,1) * (x_i - x) \quad (2)$$

In this paper, the pipeline mechanism was used to combine SMOTE oversampling and random undersampling. On the premise of ensuring the original distribution law of data, the difference between the proportion of positive and negative samples was minimized. For the selection of proportion, this study used the method of manual parameter adjustment. If a certain sampling ratio maximizes AUC, then that ratio is the optimal sampling ratio value, and some results of parameter adjustment are shown in **Table 7**.

Table 7. Results of data augmentation parameters

SMOTE	Random undersampling	AUC	Choose or not
30%	30%	78.62%	×
30%	40%	79.27%	✓
30%	50%	78.69%	×
20%	40%	77.73%	×
40%	40%	77.68%	×

In the specific procedure, we first divided the samples into a training set and a test set. Data augmentation was applied only to the training set. We adjusted the sampling values for SMOTE and Random undersampling and observed changes in AUC. When the AUC value was maximized, that set of proportions was determined as the optimal sampling ratio.

To illustrate how data augmentation was used to address the issue of data sample imbalance, consider the optimal sampling ratio data from **Table 7** (the second row) as an example. First, SMOTE oversampling was performed on the fraudulent samples. For each fraudulent sample "x" the algorithm calculated the nearest neighboring sample "x_i" based on Euclidean distance. Using "x", "x_i" and Formula 2, Generate a new set of fraudulent samples with a quantity of "30% of the non-fraudulent samples", and add the newly added fraudulent samples to the original fraudulent samples in the training set, and form the final fraudulent samples in the training set. Then, Random undersampling was conducted on the non-fraudulent samples, in the non-fraudulent samples, randomly selecting a number of samples equal to "the final fraudulent sample / 40%" to replace the original non-fraudulent samples in the training set. And generate a new training set.

After data augmentation, the ratio of non-fraudulent to fraudulent company samples changed from 85:1 to 2.5:1. The data augmentation technique used for the audit opinion information database (see **Table 5**) employed Focal_Loss, which will be discussed in Section 4.2 and Section 5.1.2.

4. The construction and training of a financial fraud detection model integrating audit opinions.

In this section, the overall framework and workflow of the model is introduced, then the audit opinion feature extraction method is analyzed, and lastly, the detection model of fraudulent financial enterprises is studied.

4.1 Workflow of Financial Fraud Detection Model

Show in **Fig. 1**, this section demonstrates the workflow of the proposed financial fraud detection model in this paper:

1. First,deal with missing values in financial statement data, and then apply data augmentation techniques to deal with imbalanced sample data (see Sections 3.2 and 3.3).

2. Building a Bert model incorporating the Focal_loss function. While training this Bert model, adjusting the parameters of the Focal_loss function to address the issue of imbalanced audit opinion samples and using the trained Bert model to extract features from audit opinion text.

3. Incorporating the AUC evaluation function into the LightGBM algorithm to construct a fused financial fraud detection model that combines audit opinions. This model is jointly trained using financial indicator data and digitized audit opinions to achieve fraud detection.

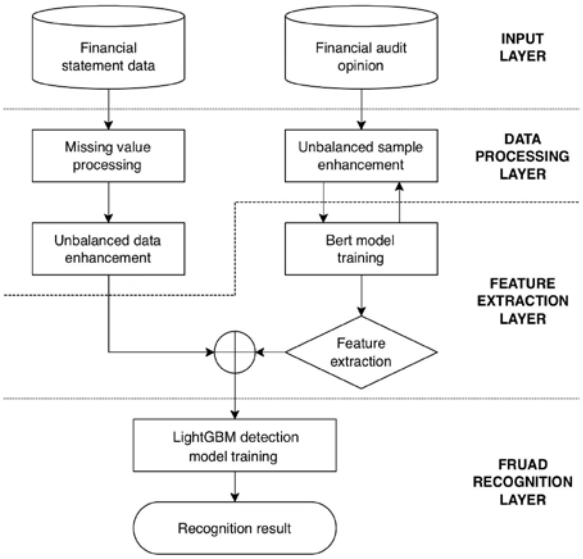

Fig. 1. Workflow of financial fraud detection model

4.2 Construction and Training of BERT Model for Extracting Audit Opinion Text Features

To extract features from audit opinion text and address the issue of imbalanced audit opinion samples, we incorporate the Focal_loss function into the Bert algorithm to build a new Bert model. By adjusting the parameters of the Focal_loss function, α and γ, to resolve the problem of sample imbalance.

4.2.1 Analysis of the BERT Model and Focal_Loss Function

The Bert semantic encoding model [7] is a language model that can generate dynamically richer word vectors by leveraging contextual information, leading to a significant improvement in the accuracy of natural language processing tasks, and the model architecture is shown in **Fig. 2**.

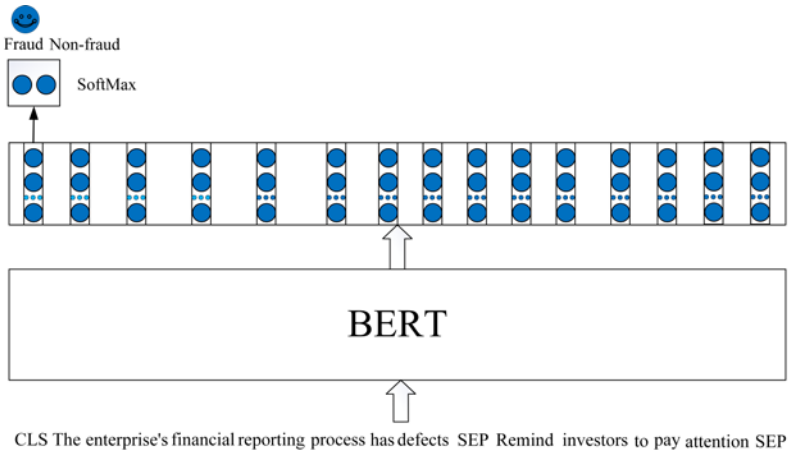

Fig. 2. Bert Model

The input to the BERT model is a sequence of words. It takes the sum of three-word vectors as its input, namely, the token_embeddings vector, the segment_embeddings vector, and the position_embeddings vector. These three vectors form the input layer of the BERT model.

First, the sentence is segmented, and the " [CLS]" tag is added at the beginning of the sentence, the " [SEP]" tag is added at the end, and between sentences" [SEP]" tag is added, forming the vector of " [CLS] The enterprise's financial reporting process has defects [SEP] Remind investors to pay attention [SEP]". This vector is processed through token_embeddings, segment_embeddings, and position_embeddings of the Bert model, resulting in three tensors. The three tensors are added correspondingly by element and input to the Bert model. After processing by the intermediate layer of the Bert model, the semantic numerical value of the entire sentence is output at the hidden output layer. Then, the semantic numerical value is input to the SoftMax normalization exponential function (equation 3), This function maps semantic numerical values to emotional numerical values between 0 and 1. This process is the digitization of audit opinion text information, namely, the extraction of audit opinion text features.

$$SoftMax\left(z_{[\mathrm{CLS}]}\right) = \frac{e^{z_{[\mathrm{CLS}]}}}{\sum_{c=1}^{2} e^{z_c}} \tag{3}$$

In the formula, z represents the node output value. For text data, two schemes are prepared in this study to deal with the problem of data imbalance. First, random sampling from fake text data was used as the corresponding audit opinion of fake samples in the digital database, and then the Focal Loss function [23] proposed by K. He team was used as the learning function of the Bert model. The loss function was modified on the basis of the standard cross-entropy loss function. Focal_Loss function (Formula 4) can reduce the weight of samples with more categories so that the model can focus more on samples with fewer categories during training.

$$L_{fl} = \begin{cases} -\alpha(1-y')^{\gamma}\log y', y = 1 \\ -(1-\alpha)y'^{\gamma}\log(1-y'), y = 0 \end{cases} \tag{4}$$

In this context: y represents the real label, $y \in \{0,1\}$, y' represents the predicted label, the difference between the predicted value y' and the real value y is referred to as loss, α is a category weight factor used to adjust the proportion of fake samples (minority) to non-fake samples (majority) participating in the model training. Increasing its value enhances the weight of fake samples during the model training process, $\alpha \in [0,1]$.

Both fake and non-fake sample sets contain samples that are hard to distinguish and easy to distinguish. The parameter γ is a difficulty weight factor, γ is used to adjust the ratio of hard-to-distinguish samples to easy-to-distinguish samples. By increasing its value, the weight

of easy-to-classify samples is reduced while the weight of hard-to-classify samples is increased, enabling the model to focus more on learning from challenging samples, $\gamma \in [0,5]$.

Analysis of Bert Model Parameters: "Batch_size" represents the size of enterprise samples input into the model during each training round, used to expedite the model training process. "Epochs" denotes the number of rounds in which the model repeats learning from enterprise samples. "Self-loss" is Bert model's feature learning evaluation function. In this study, this parameter's value is set as "Focal_loss", thereby integrating the Focal_loss function into the Bert model. The values of the two parameters, α and γ, dynamically update during the model learning process.

4.2.2 Training method of Bert model and Extraction of Audit Opinion Features

How is the Bert model incorporating Focal_Loss trained? This study adopts a strategy of training with minimum loss. Initially, a set of parameter values for the Bert model is predefined (see 5.1.2). Audit opinion texts are input into the Bert model in batches. The forward transmission calculates the loss between the predicted value y' and the true value y, and the model uses backpropagation to automatically update the values of hidden parameters (including α and γ) to continuously reduce the loss between the predicted value y' and the actual value y, gradually bringing the predicted value y' closer to the actual value y, When the loss reaches its minimum value, the Bert model training is complete, and the optimal parameter values of the model are saved.

After the model is trained, all audit opinions in the audit opinion information base are input into the model, and all output values are collected as data characteristics corresponding to audit opinions, and save these numeric values in correspondence with the audit opinion text in **Table 5**.

4.3 Construction and Training of the Financial Fraud Detection Model Integrating Audit Opinions

To obtain a more effective financial fraud detection model, we incorporate the AUC evaluation function (Equation 8) into the LightGBM algorithm to build a financial fraud detection model. And combine digitized audit opinions with processed financial indicator data to train the LightGBM model.

4.3.1 Analysis of the LightGBM Algorithm

Gradient Boosting Decision Tree (GBDT) is a long-standing model in machine learning. Its main idea is to combine multiple base learners (i. e. Weak decision tree) [24] to achieve the final strong classification results. This model has the advantage of a good training effect and is not easy to overfit. GBDT has been widely used in industry applications and has also achieved good performance in the financial quantitative competition of Kaggle, an international data mining competition, which is suitable for multi-dimensional financial data. However, GBDT needs to go through the whole training data multiple times in each iteration, which cannot be divided into batches like neural network and other algorithms. Especially in the face of industrial-level massive data, the GBDT algorithm still cannot meet the demand.

LightGBM implements the framework of the GBDT algorithm [10]. It adopts the improved histogram algorithm to discretize the continuous floating point feature values into several values first, and then selects the optimal segmentation points among the values, and constructs the classification tree. This algorithm outperforms the GBDT algorithm in terms of training speed and memory consumption, and the pseudocode of this algorithm is as follows **Fig. 3**.

Algorithm 1: Histogram-Based Algorithm

Input: I:training data, d: max depth
Input: m:feature dimension
$nodeSet \leftarrow \{0\}$ # Tree nodes in current level
$rowSet \leftarrow \{\{0,1,2,\ldots.\}\}$ # Data indices in tree nodes
$for\ i = 1\ to\ d\ do$
  $for\ node\ in\ nodeSet\ do$
    $usedRows \leftarrow rowSet[node]$
    $for\ k = 1\ to\ m\ do$
      $H \leftarrow new\ Histogram()$
      $for\ j\ in\ uesdRows\ do$
        $bin \leftarrow I.f[k][j].bin$
        # Sum of gradients of samples in each bin
        $H[bin].y \leftarrow H[bin].y + I.y[j]$
        # Number of samples in each bin
        $H[bin].n \leftarrow H[bin].n + 1$
      Find the best split on histogram H.
  Update $rowSet$ and $nodeSet$ according to the best split points.

Fig. 3. Histogram-Based Algorithm

According to Algorithm 1, first, it is necessary to clarify the training data I, the maximum tree depth d, and the feature dimension m. Then, in each tree layer, the algorithm will traverse the nodeSet to find the best feature splitting point until reaching the maximum tree depth. During this process, the algorithm will iterate through the corresponding training data set rowSet for each node based on the feature dimension, and calculate the bin values for all data using the histogram algorithm, as well as the updated data quantity and gradient sum. These values will be used to find the optimal splitting point in the subsequent steps. Finally, after obtaining the feature splitting point for the current node, the algorithm will update the current tree node and training data, and proceed to calculate the splitting point for the next node.

In terms of reducing training data, the LightGBM algorithm adopts leaf-wise growth strategy. Different from the level-wise growth strategy of traditional GBDT, the leaf-wise growth strategy with its depth limit growth strategy, which enables it to find the leaf with the largest splitting gain from all the current leaves each time, then splits, and repeats the same process, thus avoiding the invalid overhead caused by "egalitarianism", show in **Fig. 4**.

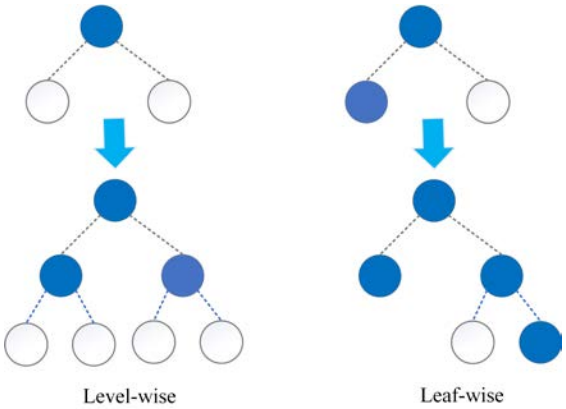

Fig. 4. Growth strategy optimization of LightGBM algorithm

4.3.2 Main Parameter Analysis of the LightGBM Detection Model

In this study, we incorporate the AUC evaluation function into LightGBM using the parameter "Metric" to build the financial fraud detection model. Here's the analysis and application of model parameters:

(1) "Boosting_type" represents the implementation algorithm of the LightGBM model. In this research, we choose its value as GBDT (Gradient Boosting Decision Tree).

(2) "Objective" represents the learning function that calculates the difference between real values and predicted values. Since it's a binary classification problem, its value is chosen as "binary".

(3) "Metric" represents the evaluation function that assesses the results of each training round of the detection model. Its value is selected as "AUC" for integrating the AUC evaluation function into LightGBM.

(4) "Learning_rate" represents the learning rate of the model, $Learning\_rate \in (0.1,0.3)$.

(5) "Num_leaves" represents the maximum number of leaves generated by the leaf-wise algorithm, $Num\_leaves < 2^{Max\_depth} - 1$, to prevent overfitting.

(6) "Max_depth" represents the maximum depth of leaves generated by the leaf-wise algorithm. $Max\_depth \in (1,8)$, with a default value of -1 indicating no restriction.

(7) "Feature_fraction" represents the proportion of randomly selected parameters in each iteration. For example, when this value is 0.8, it means 80% of the parameters are randomly chosen to generate the decision tree. $Feature\_fraction \in (0,1)$.

(8) "Bagging_fraction" represents the number of samples selected for training in each iteration without duplicate sampling. For instance, if its value is 0.8, it means 80% of the samples are chosen for training before each tree is trained. $Baggiing\_fraction \in (0,1)$.

(9) "GBDT_num_boost_round" represents the maximum number of rounds for model training.

(10) "GBDT_early_stopping_rounds" represents the maximum number of consecutive rounds during model training in which the AUC value decreases continuously.

(11) The evaluation function "AUC" (Equation 9) represents the evaluation metric for the LightGBM detection model. A higher "AUC" value indicates a better model detection performance.

4.3.3 Training method of the LightGBM Model Integrating Audit Opinions

How to combine processed financial numerical data and digitized textual data? First, we define the "Accounting Date" and "Stkcd" columns in **Table 4** as the composite primary key "key1", and the "Accounting Date" and "Stkcd" columns in **Table 5** as the composite primary key "key2". And use "key1" and "key2" to establish a link between **Table 4** and **Table 5**, resulting in a new dataset. This dataset is then used to train the LightGBM detection model.

How is the model trained? The training of this model employs an early stopping strategy without setting a fixed number of training rounds. In the experiments, the maximum training rounds are set to 1000, with an early stopping criterion of 5. This means that during the 1000 training iterations, the training stops if the AUC evaluation metric decreases continuously for 5 consecutive rounds. The values of parameters corresponding to the best training performance among the last 5 rounds of decreases are used to define the optimal tree structure for the model. This results in a well-trained financial fraud detection model that incorporates audit opinions.

5. Experiment and Results Analysis

The Python 3.7 platform was used for model implementation, the LightGBM package was used for Lightgbm model implementation, the keras_bert package and tensorflow-gpu 2.5.0 framework were used for Bert model implementation, and the Sklearn package was used for data processing and control experiment model.

5.1 Model Setting

The preprocessing process of training data: In the experiment, Total number of samples were 19003 samples and including 385 indicators. After data preprocessing, the total number of samples were 11116 (including 86 indicators), and including 10988 no-fake samples and 128 fake samples. Then, the training and testing sets were divided in a ratio of 0.7:0.3, with 7692 non fraudulent samples and 90 fraudulent samples in the training set. The sample size of the test set is 3296 non fraudulent samples and 38 fraudulent samples.

5.1.1 Index of Evaluation

Since the label distribution of data samples is imbalanced, and the test set maintains the distribution of original data, we used accuracy, recall, precision, F1 value, AUC value, and training duration as the evaluation indicators of the model, and paid more attention to the recognition performance of fraud samples.

$$Accuracy = \frac{TP+TN}{M+N} \tag{5}$$

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

$$F1 = \frac{2*precision*recall}{precision+recall} \tag{8}$$

$$AUC = \frac{\sum_{i \in positiveClass} rank_i - \frac{M(1+M)}{2}}{M*N} \tag{9}$$

TP represents true fraud samples, TN represents true non-fraud samples, FN represents false fraud samples, FP represents fake fraud samples, N represents fake samples and M represents non-fake samples. Accuracy is used to measure the proportion of samples that are correctly predicted, while Recall and Precision are used to measure the proportion of samples that are correctly predicted to be fraud and the proportion of samples that are classified as fraud and actually to be so. The F1 value is a combined measure of Recall and Precision. AUC value is the area under the ROC curve, often due to the evaluation of sample imbalance classifier performance, the greater the AUC value of the classifier, the better the recognition performance.

5.1.2 Parameter Configuration and Training Process of the Detection Model

(1) The process of handling imbalanced samples:

After data preprocessing in Section 5.1, the training dataset consists of 7,692 unfaked samples and 90 faked samples. Using the method in Section 3.3, we oversampled the fake samples using the SMOT method and generated a new set of "7,692*30%=2,307" fake samples using the 90 fake samples, their neighboring samples, and Equation 2. Based on 90 original fake samples, 2,307 new fake samples were added, resulting in a total of 2,397 fake samples. Then, using the RandomUnderSampler method, we randomly selected "2,397/40% = 5,992" non-fake samples from the non-fake samples to replace the non-fake samples in the training set. Finally, the training dataset has 5,992 non-fake samples and 2,397 fake samples

(see **Table 8**), the total number of samples were 8,389.

Table 8. Resampling strategy

Steps	Sampling Strategy	Value	Comparison of the number of non-fraudulent and fraudulent samples
0	Null	Null	7,692: 90
1	SMOTE	30%	7,692: 2,397
2	RandomUnderSampler	40%	5,992:2,397

(2) The training process for the financial fraud detection model integrating audit opinions is as follows:

For the BERT model, first, from the audit opinion samples of 128 fraudulent enterprises, randomly select and expand the number of fraudulent audit opinion samples to 2397. Then, from the audit opinion samples of 7692 non fraudulent enterprises, randomly select 5992 non fraudulent audit opinion samples and the 2397 fraudulent audit opinion samples to train the Bert model. So the training samples are 8,389 audit opinion text. The preset values for the Bert model parameters "Batch_size, Epochs, α, γ " are "4, 30, 0.75, 2.0" (see **Table 9**). These parameter values are used for a total of 30 rounds of training, with 4 randomly selected samples input into the model each round. Therefore, there are a total of 2,097 inputs per round. Following the training method described in Section 4.2.2, when the loss reaches its minimum value, the BERT model training ends, and the optimal parameter values of the BERT model are saved.

For the LightGBM model, after integrating financial numerical data and digitized audit opinion text data, the number of training samples is 8,389, each with 87 features. The preset values for LightGBM model parameters "GBDT_num_boost_round", "GBDT_early_stopping_rounds" and "Learning_rateare" are "1000", "5" and"0.1" respectively. Following the training method described in Section 4.2.2, the training is terminated if the "AUC" evaluation metric continuously decreases for 5 rounds before reaching the maximum training round limit of 1000. The parameter values from the best training performance among the last 5 rounds are used as the model decision tree's optimal values of parameter "Max_depth, Num_leaves, Feature_fraction, Bagging_fraction" (see **Table 9**).

Table 9. Parameters of the Model

Model	Parameters	Value	Description
Bert	Batch_size	4	Training batch size for each round
Bert	Epochs	30	Training rounds
Bert	Selfloss	Focal_loss	Function of loss
Bert	Alpha	0.75	Positive and negative sample balance moderator
Bert	Gamma	2	Difficult and easy sample imbalance moderator
LightGBM	Boosting_type	Gbdt	Model improvement algorithm
LightGBM	Objective	Binary	Objective function
LightGBM	Metric	Auc	Evaluation function
LightGBM	Learning_rate	0.1	Rate of learning
LightGBM	Num_leaves	20	Number of leaf nodes
LightGBM	Max_depth	5	Maximum tree depth
LightGBM	Feature_fraction	0.9	Selection ratio of tree features
LightGBM	Bagging_fraction	0.8	Data proportions used for each

			iteration
LightGBM	GBDT_num_boost_round	1000	Maximum training rounds
LightGBM	GBDT_early_stopping_rounds	5	Number of consecutive decreasing AUC values (Early Stopping Rounds.)

5.2 Validity Test of Audit Opinion Characteristic

In this section, the changes of each index before and after the test were integrated into the opinion of the audit text to prove the effectiveness of the research. First, we visualized the ROC curve changes of the model before and after the feature fusion of deliberation opinions, and the results are shown in **Fig. 5**.

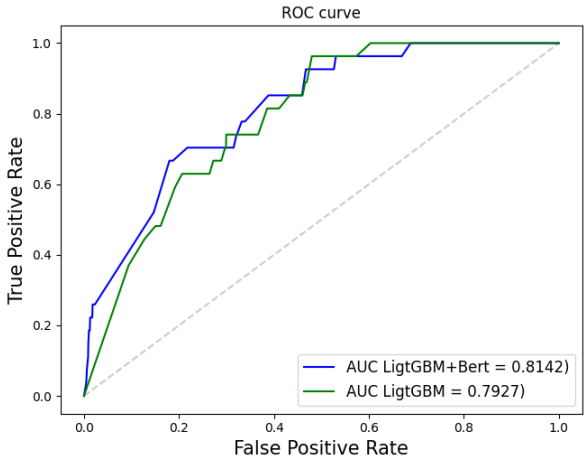

Fig. 5. Changes in the ROC curve

The horizontal axis represents the error probability of the model in identifying fraudulent companies, while the vertical axis represents the correct probability of the model in identifying fraudulent companies. It can be seen from **Fig. 5** that after Bert was added to extract the features of deliberation opinions, the ROC curve of the model covers a larger area, reaching 81.42%, with a significant effect. Below, we tested the evaluation indicators of the model before and after the integration of audit opinion characteristics, and the results are shown in **Table 10**.

Table 10. Audit opinion effect test

Model	Accuracy	Precision	Recall	F1	AUC
LightGBM	72.71%	50.98%	66.67%	57.78%	79.27%
LightGBM+ Bert(ours)	78.15%	49.21%	70.37%	57.92%	81.42%

It can be seen from **Table 10** that the model with audit opinion features has better performance in accuracy, F1, and AUC. With significant performance improvement, it is evident that the ideas proposed in the study are effective. This is because audit opinion is the most direct and fair expression of emotion on the financial situation. For example, an audit opinion of an enterprise: "Matters leading to reservations: there are major deficiencies in the internal control of your enterprise's financial reporting process regarding the complete detection of related party relationships." This information is of great significance in improving

the detection effect of the model.

5.3 Comparison Test of Mainstream Models

In this section, four mainstream classification models for comparative experiments were used to test the performance of the LightGBM model in financial fraud detection with the same data set, namely the logistic regression model [25], SVM [13], Random Forest, KNeighbors, and Three-layer CNN [15]. The results are shown in **Table 11**.

Table 11. Comparison of the effectiveness of mainstream classification models

Model	Accuracy	Precision	Recall	F1	AUC	Training Duration
Logistics Regression	66.82%	40.93%	62.43%	49.45%	69.81%	null
SVM	72.31%	45.74%	69.21%	50.08%	78.22%	null
Random Forest	68.89%	43.28%	66.53%	52.44%	71.31%	null
KNeighbors	75.89%	48.52%	70.18%	57.37%	79.83%	null
Three-layer CNN	77.92%	48.95%	72.19%	58.34%	81.87%	3.50 Min
LightGBM(ours)	78.15%	49.21%	70.37%	57.92%	81.42%	0.83 Min

From **Table 11**, it can be seen that, for the three evaluation metrics: Accuracy, F1 score, and AUC value, the experimental results of the logistic regression algorithm, SVM, and random forest algorithm are all significantly worse than the experimental results of our proposed model. Specifically, their AUC values are lower by 11.61%, 3.2%, and 10.11%, respectively. Although the KNeighbors algorithm lags behind by only 1.59% in terms of AUC value, the overall Accuracy of the samples is 2.26% lower.

These classic traditional classifiers mentioned above are clearly outperformed by LightGBM in terms of feature learning capabilities when dealing with high-dimensional, large-scale financial data. While the Three-layer CNN convolutional neural network model and our proposed model do not show significant differences in terms of Accuracy, F1 score, and AUC value, they have noticeable differences in training duration during the model training process. We also tested the training duration of both models, and it's clear that the Three-layer CNN convolutional neural network model takes 2.67 minutes longer than our proposed model.

Furthermore, from **Table 11**, it can be observed that all models have relatively low F1 values. The reason for this is that the models aim to improve AUC values and recall rates at the cost of reducing Precision values.

5.4 Analysis of Characteristic Indicators of Fraud Detection

In this section, the weights of the 11 features with higher weights are ranked and their falsification financial logic is analyzed.

Fig. 6. Ranking of importance weights

It is clearly indicated in **Fig. 6** that the text characteristics of audit opinions rank first, with a weight of about 4.44% and significant importance, and are followed by basic earnings per share and diluted earnings per share. Stock earnings can best reflect the actual operating conditions of a listed enterprise. If there is fraud in financial statements, the data reflected in its financial statements will be inconsistent with stock earnings. Additionally, off-balance-sheet expenses and undistributed profits are also one of the main methods used for financial fraud by disguising, their actual profits by manipulating off-balance-sheet income and undistributed profits. For the deadline, the actual disclosure date, and the release date, a well-run enterprise will release its financial statements on time, while a financial fraud enterprise will often delay. Receivables and paid-in indicators also account for a relatively high proportion, because when the cash flow of the enterprise is problematic, it can recognize the income in advance through receivables and accounts payable. When the increase in accounts receivable is found to be significantly higher than the increase in revenue, there is a high possibility that future earnings are transferred to the current period, which should attract the investors' attention. Through the above analysis, it can be proved that the top 13 fraud characteristics are in line with the business logic of enterprise fraud.

5.5 Detecting and Analysis of Fraud Enterprises in Different Genre of Industries

In this section, the vertical market of industry is used as the dimension to further statistically analyze the models of detecting fraud enterprises, and the results are shown in **Fig. 7**.

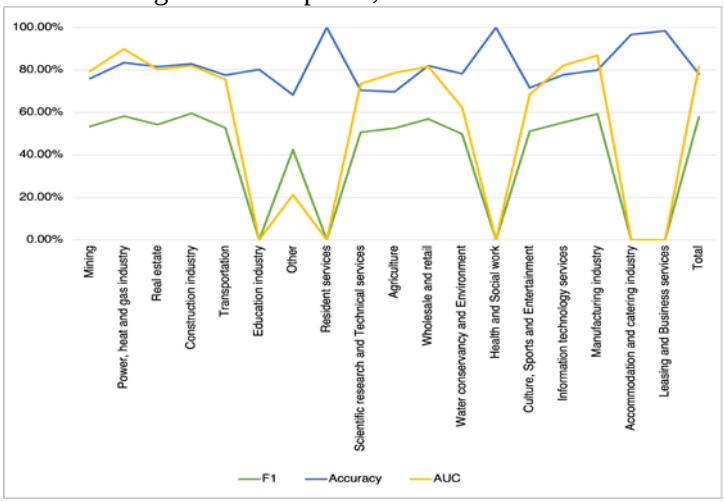

Fig. 7. Distribution of fraud detection performance in different genres of industries

Fig. 7 shows that the AUC and F1 values of the Education industry, Resident service industry, Health and Social work, Accommodation and catering industry, and Leasing and Business service industry are all very low, while the Accuracy value is high. This is due to the limited number of samples and fraud samples in such industries. Thus, the model has a weak ability to learn the characteristics of these industries. However, the F1, Accuracy, and AUC values of Power, heat and gas industry, Real estate industry, Transportation industry, Wholesale and retail, Information technology service industry and Manufacturing industry are basically the same as the total, because the sample base of these industries is large, prone to financial fraud cases, and the fraud characteristics have common characteristics. Thus the model has a strong ability to learn the characteristics of these samples.

6. Conclusion and Future Work

The model first utilizes the Bert model, a Natural Language Processing technique, to transform audit opinion information into numerical features with emotional sentiment information. These digitized audit opinion features are then combined with financial numerical features as training data for the LightGBM detection model. Additionally, in the task of financial fraud identification, imbalanced sample distribution is one of the main problems to be addressed. This article uses the "combination of oversampling Smote algorithm and Random undersampling + Focal_Loss function" technique to solve this problem. After processing, the ratio of positive to negative samples changes from 85:1 to 3:1.

In the experiment, this study collected a total of 19,003 original financial statement samples from Chinese companies, comprising 385 data indicators and 1 audit opinion text indicator. After handling missing values, a complete dataset was obtained, consisting of 11,116 samples and 86 data indicators. Compared to the LightGBM model without using audit features, the AUC value improved by approximately 2%, demonstrating the importance of audit opinions. This study also sequentially tested the performance of five mainstream fraud detection models: Logistics Regression, SVM, Random Forest, KNeighbors, and CNN. Only CNN performed comparably to the proposed model in terms of F1 score and AUC value. However, considering the duration of model training, CNN took about four times longer than LightGBM, highlighting the superiority of the LightGBM model. In future research, we will attempt to improve fraud detection rates by incorporating text information such as managerial personality traits, shareholder comments, and corporate gossip news.

Our study makes the following three contributions. 1) Different from previous research that primarily focuses on the utilization of financial numerical data or Management Discussion and Analysis in financial statements, this paper proposes an enhanced method for fraud detection using the audit opinion in financial statements as a feature and provides a text processing solution. 2) This paper reviews the current research status and conducts experimental comparisons with selected relevant models. The analysis highlights the advantages of the proposed model based on F1 score, accuracy, AUC score, and training duration. 3) This paper provides reference information for scholars studying the handling of imbalanced data in researching financial fraud.

References

[1] A. Bloomenthal, "Detecting Financial Statement Fraud," *Investopedia*, New York, USA, May. 2021. [Online] Available: https://www.investopedia.com/articles/financial-theory/11/detecting-financial-fraud.asp

[2] X. Z. Yuan, Y. P. Zhou, C. X. Yan, H. Y. Liu, G. Q. Qian, F. Wang, and L. J. Wei, "The establishment and application of financial fraud risk characteristics screening framework," *Chinese Journal of Management Science*, vol. 30, no. 3, pp. 43-54, 2022. Article (CrossRef Link)

[3] W. T. Mongwe and K. M. Malan, "The Efficacy of Financial Ratios for Fraud Detection Using Self Organising Maps," in *Proc. of 2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, Canberra, ACT, Australia, pp. 1100-1106, 2020. Article (CrossRef Link)

[4] E. Hytis, V. Nastos, C. Gogos, and A. Dimitsas, "Automated identification of fraudulent financial statements by analyzing data traces," in *Proc. of the 7th IEEE South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference*, Ioannina, Greece, pp. 1-7, 2022. Article (CrossRef Link)

[5] S. L. Humpherys, K. C. Moffitt, M. B. Burns, J. K. Burgoon, and W. F. Felix, "Identification of fraudulent financial statements using linguistic credibility analysis," *Decision Support Systems*, vol. 50, no. 3, pp. 585-594, 2011. Article (CrossRef Link)

[6] C. S. Throckmorton, W. J. Mayew, M. Venkatachalam, and L. M. Collins, "Financial fraud detection using vocal, linguistic and financial cues," *Decision Support Systems*, vol. 74, pp. 78-87, 2015. Article (CrossRef Link)

[7] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, "Revisiting pre-trained models for Chinese natural language processing," in *Proc. of the 2020 8th Conference on Empirical Methods in Natural Language Processing*, pp. 657-668, 2020. Article (CrossRef Link)

[8] Z. Liu, C. Liu, W. Lin, and J. Zhao, "Multi-task learning Pre-training financial language model for financial text mining," *Journal of Computer Research and Development*, vol. 58, no. 08, pp. 1761-1772, 2021. Article (CrossRef Link)

[9] W. Ali, W. Zuo, R. Ali, G. Rahman, X. Zuo, and I. Ullah, "Towards Improving Causality Mining using BERT with Multi-level Feature Networks," *KSII Transactions on Internet and Information Systems*, vol. 16, no. 10, pp. 3230-3255, 2022. Article (CrossRef Link)

[10] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Proc. of the 2017 Advances in neural information processing systems*, pp. 3146-3154, 2017. Article (CrossRef Link)

[11] G. Menardi and N. Torelli, "Training and assessing classification rules with imbalanced data," *Data mining and knowledge discovery*, vol. 28, pp. 92–122, 2014. Article (CrossRef Link).

[12] X. G. Wu and S. Y Du, "An Analysis on Financial Statement Fraud Detection for Chinese Listed Companies Using Deep Learning," *IEEE Access*, vol. 10, pp. 22516-22532, 2022. Article (CrossRef Link).

[13] A. A. Akinyelu and A. O. Adewumi, "On the Performance of Cuckoo Search and Bat Algorithms Based Instance Selection Techniques for SVM Speed Optimization with Application to e-Fraud Detection," *KSII Transactions on Internet and Information Systems*, vol. 12, no. 3, pp. 1348-1375, 2018. Article (CrossRef Link)

[14] S. Noels, B. Vandermarliere, K. Bastiaensen, and T. D. Bie, "An Earth Mover's Distance Based Graph Distance Metric For Financial Statements," in *Proc. of the IEEE Symposium on Computational Intelligence for Financial Engineering and Economics*, Helsinki, Finland, pp. 1-8, 2022. Article (CrossRef Link)

[15] F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan, and M. Ahmed, "Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms," *IEEE Access*, vol. 10, pp. 39700-39715, 2022. Article (CrossRef Link)

[16] M. N. Ashtiani and B. Raahemi, “Intelligent Fraud Detection in Financial Statements Using Machine Learning and Data Mining: A Systematic Literature Review,” *IEEE Access*, vol. 10, pp. 72504-72525, 2021. Article (CrossRef Link)
[17] Y. Zhang, W. Yu, Z. Li, S. Raza, and H. Cao, “Detecting Ethereum Ponzi Schemes Based on Improved LightGBM Algorithm,” *IEEE Transactions on Computational Social Systems*, vol. 9, no. 2, pp. 624-637, 2022. Article (CrossRef Link)
[18] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, “Missing value estimation methods for DNA microarrays,” *BIOINFORMATICS*, vol. 17, no. 6, pp. 520-525, 2001. Article (CrossRef Link)
[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321-357, 2002. Article (CrossRef Link)
[20] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, “A study of the behavior of several methods for balancing machine learning training data,” *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20-29, 2004. Article (CrossRef Link)
[21] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, “Boosting methods for multi-class imbalanced data classification: an experimental review,” *Journal of Big Data*, vol. 7, no. 70, pp. 1-47, 2020. Article (CrossRef Link)
[22] J. Park, S. Kwon, and S. P. Jeong, “A study on improving turnover intention forecasting by solving imbalanced data problems: focusing on SMOTE and generative adversarial networks,” *Journal of Big Data*, vol. 10, no. 36, pp. 1-16, 2023. Article (CrossRef Link)
[23] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal Loss for Dense Object Detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318-327, 2020. Article (CrossRef Link)
[24] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *The Annals of Statistics*, vol. 29, no. 5, pp. 1189-1232, 2001. Article (CrossRef Link)
[25] W. Z. Hong, X. X. Wang, and H. Q. Feng, "Research on fraud Identification in Financial Reporting of listed Companies based on Logistic Regression Model," *Chinese Journal of Management Science*, vol. 22 no. S1, pp. 351-356, 2014. Article (CrossRef Link)

Leiruo Zhou is currently working toward a bachelor's degree in accounting at the School of International Business, Yunnan University of Finance and Economics. Her main research interests include the application of machine learning and data mining in the financial field.

Yunlong Duan received his Ph.D. degree in management science and engineering from Kunming University of Science and Technology and his M.S. degree in Business Administration from Yunnan University of Finance and Economics. He is currently working as a professor, doctoral supervisor, and director of the Science and Technology Department of Yunnan University of Finance and Economics. His main research interests include knowledge management and innovation.

Wei Wei received the M.S. and Ph.D. degrees from Xi'an Jiaotong University. He is currently working as an associate professor with the School of Computer Science and Engineering, at Xi'an University of Technology. His current research interests include the area of wireless networks, wireless sensor network applications, image processing, mobile computing, distributed computing, pervasive computing, Internet of Things, and sensor data clouds.