ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Research on financial data fraud recognition system based on the joint application of Benford's law and autoencoder

Hui Xia[1,*], *Jinyu Shen*[1], *Qin Wang*[1], *Jinhong Jiang*[1] *and Tingting Zhou*[1]

[1] School of Accounting, ChongQing University of Technology, ChongQing, 400054, China

SUMMARY: *Although existing implemented regulatory responses have been effective in reducing the incidence of financial fraud, the regulatory measures are still not comprehensive enough. In this paper, Benford's law is used as the entry point of the research, and the collected conducted financial data are preprocessed to circumvent the influence of interfering information on the research results. With the theoretical support of relevant definitions, the financial selection algorithm based on conditional dynamic mutual information is designed, and the idea and process of the algorithm are described in detail. On this basis, the variational autoencoder and one-dimensional graph convolutional neural network are used to construct a financial data fraud recognition model, and the model is analyzed by example verification. After analyzing and calculating, it can be concluded that the precision, recall, and F-Score of SVM are 0.8923, 0.8517, and 0.8715, and the values of RF indicators are 0.7672, 0.7314, and 0.7489, while the values of this paper's method are 0.9728, 0.8896, and 0.9293, so the OCGVAE algorithm financial data fraud recognition process has the highest priority. The research in this paper is conducive to the good development of data quality in the financial market, effectively protect the rights and interests of investors and reduce the investment risk caused by information asymmetry.*

KEYWORDS: *Benford's law; variational autoencoder; graph convolutional neural network; financial fraud recognition model*

1 Introduction

In the last few years, the capital market has experienced rapid growth and companies in all sectors have continued to grow and expand their operations with great momentum [1]. However, the number of financial data fraud incidents has not decreased, but continues to grow. In most of the listed companies, financial data fraud, once revealed, can have a huge impact and is difficult to eliminate completely [2]. In addition, the consequences of financial data fraud are often very bad, which will not only bring huge losses to the interests of investors, but also cause the company to fall into a financial crisis, which will eventually be eliminated from the game [3, 4]. Therefore, once a company has financial data fraud, this will have a great negative impact on the normal operation of the securities market, bring immeasurable economic losses to major enterprises, and may even cause a serious impact on the country's economic growth [5, 6]. Since the last century, the problem of financial data fraud has gradually become a core issue of great concern at all levels of society.

Traditional anomaly detection methods are usually based on statistical methods, but these methods often rely on pre-defined thresholds or rules, which are difficult to be adapted to

*summertulip@126.com
https://doi.org/10.65102/is2026054
Published: 30 April, 2026

complex time series data [7]. In addition, traditional methods often fail to capture dynamic and complex correlations in time series, resulting in poor detection performance [8]. As a result, researchers have begun to explore machine learning and deep learning based methods for time series anomaly detection [9, 10]. These methods utilize the ability of machine learning algorithms and deep neural networks to automatically learn feature representations from data and detect anomalies by modeling normal data [11]. For example, an autoencoder can be used to learn a low-dimensional representation of a time series, which is an unsupervised learning neural network model that enables data compression and reconstruction through encoding and decoding processes [12, 13]. Its structure consists of an encoder and a decoder: the encoder is responsible for transforming the input data into a potential feature space, and the decoder reconstructs the input data based on these features [14]. The model can effectively downscale high-dimensional financial data to a low-dimensional space while preserving the main features of financial data and determining anomalies through reconstruction errors [15, 16].

In addition, Bamford's law was initially a mathematical conjecture of American scholar Simon Neweomb after a series of statistical analysis of different types of data [17]. That is, in different types of sample set data, the probability of 1 as the first digit is much larger than the probability of 2 as the first digit, at the same time, the probability of 2 as the first digit is equally larger than the probability of 3 as the first digit, and this probability will be with the first digit increasing the existence of gradually decreasing numerical state [18]. In recent years, the universal data description method represented by Benford's law has been developed in the fields of finance, medicine, sociology, etc., which holds that the probability of 1 to 9 as the first digit in the natural data shows a monotonically decreasing trend, and this statistical law becomes more and more obvious with the increase in the number of samples [19-21]. In 2001, Enron Corporation, the largest energy trader in the U.S., declared bankruptcy, and its financial bills did not comply with Benford's law, indicating that its management had committed financial data fraud [22]. The data auditing method based on Bentford's law can effectively detect data anomalies and provide new ideas for future auditing of various types of data [23].

Currently Bamford's law has been widely used in the financial field to study whether there is a possibility of financial data modification and falsification [24]. Clippe et al. (2012) [25] for the analysis of financial data research first used the Terrell transform for nonlinear remapping, remapping data and then by Bamford's law for further analysis, the Terrell transform simplifies and regularizes the characteristics of the financial data to improve the identification of abnormalities of the financial data. Sugiarto et al. (2016) [26] collected financial statements of companies in Indonesia and analyzed the first five effective digits of each indicator and concluded that if the observed frequency of the first five effective digits of the indicator is significantly not in accordance with the theoretical frequency of Benford's law, the company has a great potential for financial fraud. Jones et al. (2020) [27] examined violations of the Athletes' Disclosure Act database and the NCAA Today athletic department financial database, with Bamford's Law as the central tool for anomalous behavior detection, and found that the financial data from both databases essentially compounded the Bamford's Law expectations by investigating financial data over a five-year period. Qu et al. (2020) [28] assessed the effectiveness of Bamford's law in ranking abnormal behavior in financial reporting of nonprofit organizations and found that the entire sample was strongly consistent with Bamford's law, and that there existed 34% of the sample of individual organizations that did not conform to its consistency, suggesting that Bamford's law has a significant advantage in identifying abnormal indicators. Patel et al. (2022) [29] sampled the financial data of 220,583 Portuguese SMEs for the period 2010-2018, and firms whose financial indicators, such as net income, current liabilities, assets and sales, did not comply with Bamford's law had a higher chance of failure.

G. Harb et al. (2023) [30] applying Bamford's law to check for anomalies in accounting data before and after financial engineering, the study finds that capital adequacy, liquidity, and asset quality are affected by fraudulent manipulation of banks before and after financial engineering, which calls for the Lebanese government to improve the organization of its banks. Capalbo et al. (2023) [31] carried out a Bamford analysis of local elections and the quality of financial statements of municipally owned entities, in which financial data anomalies were observed around the election season, which sounded an alarm for the auditors to focus on financial data in a specific make and in a specific context. Cerqueti et al. (2024) [32] fused statistical analysis methods based on data laws and Bamford's law to systematically analyze the financial data, in order to cope with the risk of the financial market to make scientific decisions, Bamford's distributional test shows that there is a certain relationship between the average market financial returns and the level of risk. Mućko et al. (2025) [33] tested the consistency of Bamford's law on the financial statements of Polish listed companies in order to complete the modeling analysis of financial data manipulation in financial statements, and the study found that Bamford's law has a good applicability in financial statement financial modeling.

Although most of the studies reported prove that Bamford's law shows a good value in detecting financial anomalies, there may be situations where it is not fully applicable to certain corporate financial data. Davydov et al. (2016) [34] pointed out that bank financial data testing through Bamford's law yields higher quality financial statements, and consistency testing yields some information about financial data anomalies, which need to be combined with other detection algorithms to identify the probability of bank insolvency. Ahmadi et al. (2020) [35] found that Bamford's law has some limitations in identifying fraudulent companies, they collected financial data of fraudulent and non-fraudulent companies and applied Bamford's law to detect them, Bamford's distribution was effective in identifying the fraudulent companies but at the same time it identified some non-fraudulent companies as fraudulent companies.

In the 1990s, due to the rapid rise of artificial intelligence technology, the drawbacks of statistical methods became increasingly obvious, and research scholars began to focus on applying deep learning models to the financial field [36]. In recent years, the application of autoencoder-based and its derivative models in the financial field has also become a research hotspot. Lv et al. (2019) [37] designed an autoencoder-based graph convolutional network in order to prevent online financial fraud, and trained the model by multi-task objective function to maximize the model's performance in identifying fraudulent behaviors, and the simulation results confirmed the reliability of the model to reach the level of the current state-of-the-art model. Demestichas et al. (2021) [38] propose a financial anomalous behavior detection engine with a set of autoencoders, which is based on a knowledge base of financial transaction investigations that correlates different levels of financial data to discover evidence of illegality in financial data. Mohanty et al. (2021) [39] jointly applied an autoencoder and a kernel-limit learning machine with stock prediction in the financial market, and simulation tests verified the effectiveness of this hybrid model, whose average absolute percentage error in predicting stocks was only 1.074%, which is better than many methods in the market. Muthukumaran et al. (2023) [40] developed financial crisis prediction based on optimal feature selection and optimal variational autocoder, in which the variational autocoder is able to automatically classify financial data in the test dataset into financial crisis data and non-financial crisis data, and the research has made good progress, providing a new and reliable method for financial crisis prediction. Almahadeen et al. (2024) [41] used a hybrid model of autoencoder and multilayer perceptron to detect threatening behaviors in financial network security, and the hybrid model can accurately detect threatening behaviors such as fraud, data leakage, and unauthorized access attempts, which effectively enhances the security of financial networks. Buchdadi et al. (2025) [42] explored the efficacy of autoencoders for anomaly detection in blockchain digital

transactions, autoencoders have a significant advantage in complex data processing, and their accuracy for anomaly identification reaches 0.87, which can effectively differentiate between normal and abnormal transaction behavior.

Based on the perspective of mathematical formulas, this paper provides a detailed overview of Benford's Law to further strengthen the association between this law and financial data fraud. Subsequently, financial data collection is carried out, and it is found that there is a lot of interfering information in the financial data, which needs to be data cleaned to ensure the usability of the research data. With the relevant definitions, a financial feature selection algorithm based on conditional dynamic mutual information is designed, and the research idea and computational process of the algorithm are also given. Combining Benford's law, deep learning algorithm, and autoencoder technology, a financial data fraud recognition model based on graph convolutional neural network and variational autoencoder is constructed. Finally, with the support of financial data in this paper, the research scheme of this paper is analyzed by example verification from multiple dimensions.

2 Research on Financial Data Fraud Recognition System

2.1 Overview of Benford's Law

2.1.1 Basic Theory

An intrinsic law of numerical statistics that states that the probability that the first digit of each sample will be each of the numbers 1 through 9 is stable within a certain range for all natural random variables, provided the sample space is large enough. The main founders of Benford's Law found that statistical analysis of various phenomena, such as birth rates, death rates, physical and chemical constants, and prime numbers, revealed that data obtained from the system of units of measure complied with the first law of numbers. Since financial data is information measured in terms of money, and to some extent also conforms to the above law, the author organizes and analyzes the first digit of financial data information to study its compliance with the law. Mathematicians once found that the frequency of the occurrence of the first digit in the data on the books conforms to Benford's law, and if the person who made the false accounts changed the real data, it would make the frequency of the occurrence of the first digit on the books change and deviate from the frequency in Benford's law. This has important implications for the study of the truthfulness, legality and integrity of financial data.

2.1.2 Mathematical definitions

Benford's law is an empirical law, also known as the first law of numbers because of its first digit extraction properties. It describes a natural phenomenon in which the probability distribution of the first digit 1 to 9 of a collection of data produced by a natural system is not uniform, but obeys a logarithmic distribution. Equation (1) shows the definition of the distribution of numbers in this law:

$$P_d = \log_{10}\left(1+\frac{1}{d}\right) \tag{1}$$

where d represents the first digit 1, 2, ..., 9, and P_d represents the probability of the occurrence of the corresponding d, the distribution of Benford's law digits is shown in Fig. 1, which gives a more intuitive distribution of the probability of each digit. The discovery of

Benford's law can be traced back to 1881, through the statistics of a number of naturally occurring data sets of numbers, found that the distribution of their first digits are subject to the law shown in Equation (1). These included river basins, stock market prices, census data, the heat capacity of chemicals, and even a bunch of numbers lifted from a newspaper, after which the discovery was eventually named. Although the statistic started out as purely experimental, it has now been established that it also applies to various mathematical series such as the Fibonacci sequence, while a number of natural science phenomena, such as the depth of earthquakes, have also been shown to conform to the law. Since the publication of Benford's law, many researchers have endeavored to explain the reason for this phenomenon, which is that the logarithmic distribution is the limiting distribution of random variables when they are constantly multiplied, divided, or taken to integer powers, and that once this limit is reached, the logarithmic distribution remains unchanged for all further multiplications and divisions, and for all integer powers.

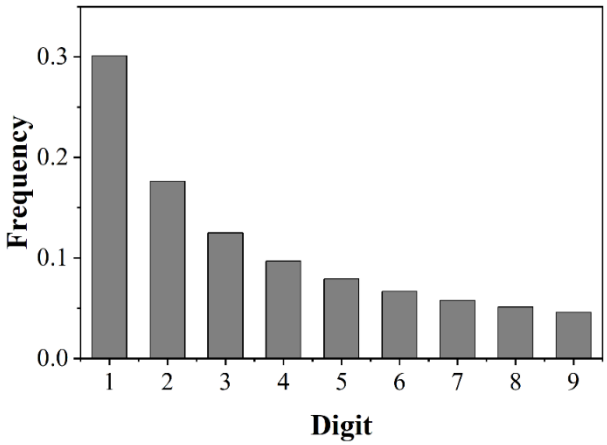

Figure 1: Benford's Law, numerical distribution

2.2 Financial data collection and cleansing

2.2.1 Financial data collection

The banking system database contains a large number of data tables, for example, the credit card dataset involves tables consisting of cardholder information tables, transaction record tables, etc. In order to form anomaly detection recognizable data, different tables need to be integrated. In addition we need to add some statistical fields, such as credit card each transaction record is a single, can not reflect the user's transaction patterns, so we need to add some historical transaction situation fields, such as the number of daily transactions than: the number of daily transactions / history of the maximum number of daily transactions, etc., where the transaction time can be calculated in units of months, days, hours. For example, on the Kaggle credit card fraud dataset, this paper introduces temporal indicators $f_m\left(m=d+1,\cdots d+l\right)$, such as the number of daily transactions/historical daily maximum number of transactions to reflect the impact of historical transactions on the current one, and also changes the Time field to be measured in hours when using it, and statistically finds that every morning Between 9:00am and 11:00pm is the high frequency time period of credit card spending.

2.2.2 Data cleansing

Data cleaning accomplishes the following tasks: dirty data removal, missing value processing, and correcting erroneous data. For example, the Kaggle loan default dataset has missing values

in the variables MonthlyIncome and NumberOfDependents, which are filled in this paper using the mean value. At the same time age in the minimum value of 0, and the bank is unlikely to give loans to customers under the age of 18, so age is less than 18 for the wrong data, this paper uses the method of judgment based on the value of other attributes, and if it is not possible to determine the direct deletion of the record.

2.3 Financial feature selection algorithm based on conditional dynamic mutual information

When dealing with the anomaly detection problem for large-scale datasets such as financial datasets, redundant and irrelevant features can degrade the classification performance, resulting in the problem of high false alarm rate and low detection rate of anomaly detection algorithms. This chapter proposes a Conditional Dynamic Mutual Information-based Feature Selection Method for Financial Data (CDMIFS) for the anomaly detection problem of financial datasets, which takes into account the contribution of features to the determination of anomaly classes in historical data when constructing the evaluation function, and uses Conditional Dynamic Mutual Information (CDMIFS) as a metric to measure the relevance of features to the anomaly classes on unrecognized samples, and removes irrelevant as well as redundant features.

2.3.1 Relevant definitions

For the financial dataset O, assuming C is the anomaly category, F is the subset of candidate features, and S is the subset of selected features, the conditional dynamic mutual information is defined for $\forall f_i \in F$ as:

$$\begin{aligned} CDMI &= \sum_{j=1}^{t} \alpha_j MI_j(f_i;C) \\ &\{MI(f_i;C/S) + MI(S, f_i;C) - MI(f_i;S)\} \end{aligned} \tag{2}$$

where $MI_j(f_i;C)$ denotes the mutual information between the feature f_i and the anomaly category in the previous j periods, which indicates the contribution of the feature to the classification of the anomaly category in the historical data, the larger the value of the mutual information, the more favorable f_i is to the classification of the anomaly in the historical data. $\alpha_j = 2(t-j+1)/(t(t+1))$ denotes the weight of the previous j periods, and j denotes the number of historical data available. $\alpha_j MI_j(f_i;C)$ denotes the effect of the contribution of the attribute feature to the anomaly discrimination in the historical data on the current data, e.g., in the financial statements of the previous 2 years, it was found that the crisis of the firm was partly due to the fact that there was a significant decline in the cash current debt ratio and the profitability of the main business was almost unchanged, whereas in the previous 1 year, it was found that the cash current debt ratio rebounded and the The profitability of main business decreased significantly, so when forecasting future statements, it is important to synthesize the analysis by measuring the impact of the characteristics of the previous years on the category, and selecting the characteristics that have the greatest relevance to the category. $MI(f_i;C/S)$ represents the mutual information between the candidate features and the anomaly category C under the current selected feature set; $MI(S \cup \{f_i\};C)$ represents the mutual information between the subset of features and the anomaly category C when the candidate features are

added to the selected feature set, which indicates the correlation between the selected feature set and the anomaly category C in the process of dynamic change, and the bigger the values of the two indicate that the selected features are related to the anomaly category C. The larger these two values indicate that the selected features are related to the anomaly category and the redundancy of the feature subset is minimized. The $MI(f_i;S)$ represents the mutual information between the candidate features and the selected features.

2.3.2 Algorithm description

In this section, the feature selection algorithm for financial data based on conditional dynamic mutual information is analyzed and described in detail, and the detailed research ideas of the algorithm are given. The details are shown as follows:

The idea of the algorithm can be described as follows: assume the optimal feature subset S, the candidate feature set F, the identified sample set O_u, the unidentified sample set O_l, and the anomaly category C. Firstly, the mutual information between each feature in the candidate feature set and the anomaly category is derived, and the feature with the largest mutual information value is selected to be added to the feature subset S and removed from F. When the candidate feature set F satisfies $|F|\neq 0$ or the unrecognized sample O_u satisfies $|O_u|\neq 0$, compare $MI(f_i;S)$ and $MI(f_i;C/f_s)\ (\forall f_s \in S)$ in the candidate feature set in turn, if $MI(f_i;S) > MI(f_i;C/f_s)$, then the feature is directly removed from F. Otherwise, select f_i corresponding to the maximum value DCMI, add it to S and remove it from F, and remove the samples identified by the feature value f_i from the unrecognized samples; loop the above operation until the set of F is empty or the set of unrecognized samples is empty.

The search strategy of CDMIFS is sequential forward search, so the subset of self-selected features obtained by CDMIFS is the optimal approximation subset. Assuming that the training dataset contains n features, although the number of unrecognized samples in the CDMIFS algorithm may not necessarily decrease at the end of each loop, it selects one feature at a time, and the number of features in the set of candidate features is on a decreasing trend, and therefore the algorithm will eventually terminate. The first step of the algorithm needs to compute the mutual information of all the candidate features with the category, which has a time complexity of $O(n)$, and the time complexity in the loop statement is $O\left(n^2\right)$, the time complexity of this algorithm is $O\left(n^3\right)$.

2.4 Financial Data Fraud Recognition Model

Due to the sensitivity of financial data and the specificity of data samples, the traditional financial fraud research methods are no longer appropriate for the current situation. In this regard, on the basis of Benford's law, deep learning algorithm, and autoencoder technology, a financial data fraud recognition model based on graph convolutional neural network with variational autoencoder is designed. The design process is as follows:

2.4.1 Principles of the autoencoder algorithm

Auto-encoder consists of encoder and decoder, where the input layer and the hidden layer form the encoder and the hidden layer and the output layer form the decoder. The process of auto-encoder is as follows: the encoder encodes the high dimensional input data into low dimensional

hidden variables by means of an activation function. The decoder reduces the hidden variables to their initial state, and the best state of the auto-encoder is that the output of the decoder can maximize the reduction of the original input. In this paper, we use an improved self-encoder: the Variable Auto-Encoder (VAE) to reduce the overfitting phenomenon of the neural network, and to achieve the effect of predicting the sample category by training on only one type of data.

Figure 2 shows the graph model of VAE, the data that can be observed in this paper is x, while x is generated by the hidden variable z, from $z \to x$ is the generative model $p_\theta(X \mid Z)$, which from the point of view of the autocoder, is the decoder, while from $x \to z$ is the recognition model $q_\phi(Z \mid X)$, which is analogous to the encoder of the autocoder.

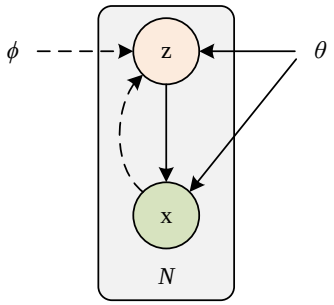

Figure 2: VAE graph model

In this paper, it is assumed that all the data are independently and identically distributed, and the better to make the generative model is to do parameter estimation of the generative model $p_\theta(X \mid Z)$ by using the log maximum likelihood method to maximize the log-likelihood function of the following equation (3).

$$\log p_\theta\left(x^{(1)}, x^{(2)}, \ldots, x^{(N)}\right) = \sum_{i=1}^{N} \log p_\theta\left(x^{(i)}\right) \tag{3}$$

In practical applications, the data can be easily obtained, but the distribution of the data is not known. So in VAE, this paper uses the encoder $q_\phi\left(Z \mid X^{(i)}\right)$ to approximate the true posterior probability $p_\theta\left(Z \mid X^{(i)}\right)$, and in this paper, we use the KL dispersion to measure the similarity between the two distributions, that is Equation (4) is derived:

$$\begin{aligned} KL\left(q_\phi\left(Z \mid X^{(i)}\right) \| p_\theta\left(Z \mid X^{(i)}\right)\right) &= E_{q_\phi(Z|X^{(i)})} \log \frac{q_\phi(Z \mid X^{(i)})}{p_\theta(Z \mid X^{(i)})} \\ &= E_{q_\phi(Z|X^{(i)})} \log \frac{q_\phi(Z \mid X^{(i)}) p_\theta(X^{(i)})}{p_\theta(Z \mid X^{(i)}) p_\theta(X^{(i)})} \\ &= E_{q_\phi(Z|X^{(i)})} \log \frac{q_\phi(Z \mid X^{(i)})}{p_\theta(Z, X^{(i)})} + E_{q_\phi(Z|X^{(i)})} \log p_\theta(X^{(i)}) \\ &= E_{q_\phi(Z|X^{(i)})} \log \frac{q_\phi(Z \mid X^{(i)})}{p_\theta(Z, X^{(i)})} + \log p_\theta\left(X^{(i)}\right) \end{aligned} \tag{4}$$

So get:

$$\log p_\theta\left(X^{(i)}\right) = KL\left(q_\phi\left(Z \mid X^{(i)}\right) \middle\| p_\theta\left(Z \mid X^{(i)}\right)\right) + \mathcal{L}\left(\theta, \phi, X^{(i)}\right) \tag{5}$$

Among them:

$$\begin{aligned} \mathcal{L}\left(\theta, \phi, X^{(i)}\right) &= -E_{q_\phi(Z|X^{(i)})} \log \frac{q_\phi(Z \mid X^{(i)})}{p_\theta(Z, X^{(i)})} \\ &= E_{q_\phi(Z|X^{(i)})} \log p_\theta\left(Z, X^{(i)}\right) - E_{q_\phi(Z|X^{(i)})} \log q_\phi\left(Z \mid X^{(i)}\right) \\ &= -KL\left(q_\phi\left(Z \mid X^{(i)}\right) \| p_\theta(Z)\right) + E_{q_\phi(Z|X^{(i)})} \log p_\phi\left(X^{(i)} \mid Z\right) \end{aligned} \tag{6}$$

Since the KL scatter is non-negative, when the two distributions are the same, the KL scatter is zero. Thus there is $\log p_\theta(X^{(i)}) \geq \mathcal{L}(\theta, \phi, X^{(i)})$, and $\mathcal{L}(\theta, \phi, X^{(i)})$ is known as the lower bound of the variance of the log-likelihood function. Following this, the optimization problem for $\log p_\theta(x^{(i)})$ is transformed into the optimization of the variational lower bound $\mathcal{L}(\theta, \phi, X^{(i)})$, and the corresponding optimization function is transformed into $\mathcal{L}(\theta, \phi, X) = \sum_{i=1}^{N} \mathcal{L}(\theta, \phi, X^{(i)})$. In order to solve the optimization function more conveniently, the reparameterization technique is introduced. Let the recognition model $q_\phi(Z \mid X)$ be the differentiable function $g_\phi(\varepsilon, X)$, where ε is the noise and $\varepsilon \sim p(\varepsilon)$. Thus, $\mathcal{L}(\theta, \phi, X^{(i)})$ can be used to estimate expectations using Monte Carlo methods:

$$\begin{aligned} \mathcal{L}\left(\theta, \phi, X^{(i)}\right) = &-KL\left(q_\phi(Z \mid X^{(i)}) \| p_\theta(Z)\right) \\ &+ \frac{1}{L} \sum\nolimits_{l=1}^{L} \log p_\theta\left(X^{(i)} \mid Z^{(i,l)}\right) \end{aligned} \tag{7}$$

where $Z^{(i,l)} = g_\phi(\varepsilon^{(i,l)}, X^{(i)})$ and $\varepsilon^{(i,l)} \sim p(\varepsilon)$.

In the actual calculation of Eq. (7), the following parameters are taken in this paper:

$$\begin{aligned} p(\varepsilon) &= \mathcal{N}(\varepsilon; 0, I) q_\phi(Z \mid X^{(i)}) \\ &= \mathcal{N}(Z; \mu^{(i)}, \sigma^{2(i)} I) p_\theta(Z) \\ &= \mathcal{N}(Z; 0, I) g_\phi(\varepsilon^{(i,l)}, X^{(i)}) \\ &= \mu^{(i)} + \sigma^{(i)} \odot \varepsilon^{(l)} \end{aligned} \tag{8}$$

According to equation (7), the first term on the right hand side of the equal sign of equation (6) can be calculated as:

$$
\begin{aligned}
-KL\left(q_\phi(Z\mid X)\middle\|p_\theta(Z)\right) &= \int q_\phi\left(Z\mid X\right)\log p_\theta(Z) \\
&\quad -q_\phi\left(Z\mid X\right)\log q_\phi\left(Z\mid X\right)dZ \\
&= \int \mathcal{N}(Z;\mu,\sigma^2)\log \mathcal{N}(Z;0,I)dZ \\
&\quad -\int \mathcal{N}(Z;\mu,\sigma^2)\log \mathcal{N}(Z;\mu,\sigma^2)dZ \\
&= -\frac{J}{2}\log(2\pi)-\frac{1}{2}\sum\nolimits_{j=1}^{J}\left(\mu_j^2+\sigma_j^2\right) \\
&\quad -\left(-\frac{J}{2}\log(2\pi)-\frac{1}{2}\sum\nolimits_{j=1}^{J}\left(1+\log\sigma_j^2\right)\right) \\
&= \frac{1}{2}\sum\nolimits_{j=1}^{J}\left(1+\log\left(\sigma_j^2\right)-\mu_j^2-\sigma_j^2\right)
\end{aligned}
\tag{9}
$$

where J is the dimension of the hidden variable. (The second term on the right side of the equals sign of Eq. (6) is chosen according to whether the actual data is binary or real-valued, and in this paper, we define our own generative model, and the second term is computed using cross-entropy summation.

2.4.2 Graph Convolutional Neural Networks

The main idea of Graph Convolutional Neural Networks (GCN) is to perform a convolution operation on the features of a node and its neighboring nodes to generate a new representation of the node's features.The GCN model usually consists of multiple graph convolution layers, each of which learns a more abstract and complex representation of the node's features. This convolutional operation is implemented through the adjacency matrix of the graph. The definition is shown in equation (10):

$$
H^{(l+1)}=\sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{10}
$$

where l denotes the number of layers of the GCN, $H^{(l)}$ is the feature matrix of the l th layer, with one feature vector for each node, $W^{(l)}$ is the weight matrix of the l th layer, which denotes the convolution kernel of the l th layer, $\tilde{A}=A+I$ is the sum of the adjacency matrix A and the unitary matrix I, $\tilde{D}$ is the diagonal matrix, σ is the activation function, commonly used are ReLU and Sigmoid. The final feature matrix $H^{(l)}$ output can be obtained through continuous iterative computation.

The GCN model can be used for node classification, link prediction, social network analysis and other tasks. In recent years, GCN and its extended models have been widely used in various real-world scenarios, including social media, product recommendation, urban planning, financial fraud recognition, and other fields.

2.4.3 Mathematical modeling

One-dimensional graph convolutional neural network and variational autoencoder based fraud recognition model for financial data (OCGVAE), OCGVAE consists of three main modules, each of which has an important role and is interrelated. The first module of the framework is the Graph Convolutional Neural Network (GCN), which is the key hub connecting the data features to the P-D relational network. It contains an input layer and two hidden layers, the

second hidden layer has two parallel structures which share the first layer parameters. The second module is the variational autoencoder (VAE), whose inputs are the outputs of the GCN, one output is the mean vector and the other is the standard deviation vector. They constitute the distribution of the hidden variables. The last module is about the output of the framework, where labels of nodes and links between nodes can be predicted in the architecture proposed in this paper.

Using the algorithm GCN to construct an undirected graph $G=(V,E,W)$, define a graph convolution operation "*G" based on the spectral domain convolution, for any patient node $x\in\mathbb{R}^n$ and a convolution kernel function Θ, this paper defines the convolution operation as follows:

$$\Theta *_G x=\Theta(L)x=\Theta(U\Lambda U^T)x=U\Theta(\Lambda)U^T x \tag{11}$$

In equation (4-1) above, the Laplace matrix $L=I_n-D^{-\frac{1}{2}}AD^{-\frac{1}{2}}=U\Lambda U^T\in\mathbb{R}^{n\times n}$, I_n is the unit matrix , $D\in\mathbb{R}^{n\times n}$ is the degree matrix, and $D_{ii}=\sum_j^n A_{ij}$ is the formula for the degree matrix, A is the adjacency matrix (the weighting information between the financial data), and $\Lambda\in\mathbb{R}^{n\times n}$ is the diagonal matrix consisting of the eigenvalues of the Laplace matrix; $U\in\mathbb{R}^{n\times n}$ is the eigenvectors of the Laplace matrix; and the filter $\Theta(\Lambda)$ is the diagonal matrix about the Laplace matrix. Through the convolution operation on the financial data node, the information of this node and the nodes connected to it can be intercommunicated, so that the distribution of nodes of the same type can be more closely. So GCN has a layer-by-layer propagation rule as follows in equation (12):

$$H^{(l+1)}=\sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{12}$$

where $H^{(0)}=X$, which is the financial data node information, and $H^{(l)}$ denotes the output of the lth layer of said graph convolutional neural network. $W^{(l)}$ is the weight matrix of the lth layer of the network of said graph convolutional neural network, and $\sigma(\cdot)$ is the sigmoid activation function.

In this paper, the input financial data node information X, and their relationship matrix A, through the $GCN(\cdot)$ operation can effectively combine the node information and the network topology information, can greatly reduce the dimension of the topology graph and capture the information of the domain nodes.

Consider the features of each node, whose feature dimension is denoted as D, whose label is denoted as y, and for the hidden coding dimension in VAE is denoted as F. The node feature matrix $X\in N\times D$, the node's adjacency matrix $A\in N\times N$, and the hidden node code $Z\in N\times F$ are obtained. Based on the theory of variational self-coding above, the following model is proposed in this paper:

Theoretical Model:

$$\begin{aligned} q(Z\mid X,A,y)&=\prod_{i=1}^N q(z_i\mid X,A,y)\\ q(z_i\mid X,A,y)&=\mathcal{N}\left(z_i\mid \mu_i,diag(\sigma_i^2)\right)\end{aligned} \tag{13}$$

where $\mu = GCN_{\mu}(X, A; W_{\mu})$, $\log\sigma = GCN_{\sigma}(X, A; W_{\sigma})$, $GCN(X, A; W)$ is a graph convolution operation, μ and σ share the first layer weight parameter W_0; and y is the node label. Generate the model:

$$\begin{aligned} p(y, A \mid Z) &= p(y \mid Z)\, p(A \mid Z) \\ &= p(y \mid Z) \prod_{i=1}^{N} \prod_{j=1}^{N} p\left(A_{ij} \mid z_i, z_j\right) \end{aligned} \tag{14}$$

where $p(A_{ij} = 1 \mid z_i, z_j) = \sigma(z_i^T z_j)$, $p(y_i = 1 \mid z_i) = \sigma(W_i z_i + b)$, and $\sigma(\cdot)$ is an activation function, which is used as sigmoid function in this paper. W_i and b are logistic regressors and bias terms, respectively. Optimization objective:

$$\begin{aligned} \mathcal{L} &= E_{q(Z|X,A,y)} \log p(y, A \mid Z) - KL\left[q(Z \mid X, A, y) \| \, p(Z)\right] \\ &= \underbrace{E_{q(Z|X,A,y)} \log p(y \mid Z) + E_{q(Z|X,A,y)} \log p(A \mid Z)}_{reconstruction} \\ &\quad - \underbrace{KL\left[q(Z \mid X, A, y) \| \, p(Z)\right]}_{regularization} \end{aligned} \tag{15}$$

The objective optimization function is divided into three formulas, the first two formulas compute the reconstruction loss, that is, the loss of the generative model, which is divided into the node labeling loss model and the loss model of the edges between nodes. The third eqn. is a regularization term, an eqn. that de-approximates the hidden coding distribution according to the distribution specified in this paper, and its computation is given by Eqn. (16):

$$\begin{aligned} & E_{q(Z|X,A,Y)} \log p(A \mid Z) = \int q\left(Z \mid X, A, Y\right) \log p\left(A \mid Z\right) dZ \\ & = \int \prod_{i=1}^{N} q(z_i \mid X, A, y) \log p(A \mid Z) dZ \\ & = \int \prod_{i=1}^{N} \mathcal{N}\left(z_i \mid \mu_i, diag(\sigma_i^2)\right) \log p(A \mid Z) dZ \\ & = \int \prod_{i=1}^{N} \mathcal{N}\left(z_i \mid \mu_i, diag(\sigma_i^2)\right) \prod_{i=1}^{N} \prod_{j=1}^{N} \log\sigma\left(z_i^T z_j\right) dz \\ & = \int \prod_{i=1}^{N} \mathcal{N}\left(\eta_i \mid 0,1\right) f(\eta; W) d\eta_i = \frac{1}{S} \sum_{s=1}^{S} f\left(\eta_i^{(s)}; W\right) \end{aligned} \tag{16}$$

where $\eta_i = (z_i - \mu_i(W)) / \sigma_i(W) \sim \mathcal{N}(\eta \mid 0,1)$, due to the similarity in form of the first and the second equations, in this paper, we only give $E_{q(Z|X,A,y)} \log p\left(A \mid Z\right)$ the derivation process, where reparameterization techniques and Monte Carlo sampling estimation are used.

For loss calculation in machine learning, cross entropy is often used in this paper. The first equation:

$$\begin{aligned} & E_{q(Z|X,A,y)} \log p(y \mid Z) \\ & = -\frac{1}{m} \sum_{i=1}^{m} \left[y_i \log\left(\sigma(\theta z_i + b)\right) + (1 - y_i) \log\left(1 - \sigma(\theta z_i + b)\right) \right] \end{aligned} \tag{17}$$

where m is the number of samples known to be financial data fraud. The second equation:

$$E_{q(Z|X,A,y)} \log p(A|Z) = norm\_weight$$
$$*\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[-A_{ij} * \log\left(\sigma(z_i^T z_j)\right) * pos\_weight\right. \tag{18}$$
$$\left.-(1-A_{ij}) * \log\left(1-\sigma(z_i^T z_j)\right)\right]$$

Among them:

$$norm\_weight = \frac{N*N}{\left(N*N-\sum_{i=1}^{N}\sum_{j=1}^{N}A_{ij}\right)^2} \tag{19}$$

$$pos\_weight = \frac{N*N-\sum_{i=1}^{N}\sum_{j=1}^{N}A_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{N}A_{ij}} \tag{20}$$

N is the number of nodes in the entire network (the total number of financial data).

3 Example analysis

3.1 Financial feature selection algorithm validation

3.1.1 Data sets

Based on the financial data collection and cleaning in Section 2.2, the dataset for this study was obtained. It can be seen that it is divided into twelve characteristics, namely, profitability, risk, liquidity, security, maturity, leverage, creditworthiness, volatility, mobility, diversity, correlation, and regulatory nature. The statistical analysis of the dataset is shown in Table 1. After financial data collection and cleaning, a total of 6,578 financial characteristic data were obtained, with a sample size of 13,973. Among the twelve financial characteristics, the numbers of profitability, risk, liquidity, safety, maturity, leverage, creditworthiness, volatility, mobility, diversity, correlation and regulation are 389, 734, 361, 900, 539, 959, 604, 489, 184, 187, 624 and 608 respectively. The corresponding sample sizes are 1816, 1581, 1418, 460, 1584, 1536, 2027, 864, 158, 2086, 95, and 348. This provides a more intuitive display of the detailed information on the number of features and samples of the dataset, offering data support for the following research and analysis, and facilitating the progress of the subsequent research work.

Table 1: Statistical analysis of the dataset

Project	Number of features	Sample size.
Profitability	389	1816
Risk	734	1581
Liquidity	361	1418
Safety	900	460
Term	539	1584
Leverage	959	1536
Creditworthiness	604	2027
Volatility	489	864
Mobility	184	158
Diversity	187	2086
Relevance	624	95
Regulatory	608	348
Total	6578	13973

3.1.2 Algorithm performance evaluation metrics

Usually, we need to evaluate the classification performance of the feature subset finally selected by the feature evaluation criteria in order to judge the advantages and disadvantages of the feature selection algorithm. In this experiment, we will use three evaluation criteria, namely, accuracy, t-test test and F1 value, to measure the algorithm from different perspectives.

3.1.3 Comparison of average accuracy

The comparison algorithms chosen for the experiments in this subsection are the four classic traditional feature selection algorithms, CIFE, CMIM, JMI and mRMR, and the dynamic feature selection algorithm, DCSF, and the comparison of the average accuracies is shown in Table 2. Each feature selection algorithm has its own advantages and disadvantages, but a closer look reveals that the ordering of the average classification accuracy of all financial data feature selection algorithms has a similar trend: CDMIFS>DCSF>JMI>mRMR>CMIM>CIFE, and the overall average accuracy values are 0.8524, 0.7549, 0.6778, 0.6633, 0.6375, 05500, compared with other algorithms, the accuracy of this paper's algorithm is more prominent in financial data feature selection. For example, in the actual financial detection process, the algorithm can be utilized to accurately extract financial anomaly data, which in turn helps users to avoid financial fraud and risk.

Table 2: Comparison of average accuracy rates

Project	CIFE	CMIM	JMI	mRMR	DCSF	CDMIFS
Profitability	0.5755	0.6750	0.7416	0.6556	0.7949	0.8398
Risk	0.5450	0.6178	0.6120	0.6730	0.7634	0.8633
Liquidity	0.5531	0.6174	0.6066	0.7066	0.7288	0.8503
Safety	0.5788	0.6179	0.6605	0.7402	0.7318	0.8822
Term	0.5201	0.6639	0.7195	0.7228	0.7049	0.8310
Leverage	0.5948	0.6133	0.6939	0.6363	0.7701	0.8462
Creditworthiness	0.5241	0.6852	0.7386	0.7224	0.7644	0.8098
Volatility	0.5080	0.6724	0.7116	0.6078	0.7332	0.8765
Mobility	0.5371	0.6050	0.7003	0.6084	0.7728	0.8563
Diversity	0.5238	0.6123	0.7378	0.6252	0.7912	0.8422
Relevance	0.5790	0.6589	0.6057	0.6484	0.7360	0.8997
Regulatory	0.5610	0.6107	0.6055	0.6124	0.7678	0.8317
Total	0.5500	0.6375	0.6778	0.6633	0.7549	0.8524

3.1.4 t-test test

In order to facilitate a more intuitive comparison of the degree of superiority and inferiority of the performance of the six algorithms involved in this chapter, this paper conducted a significant difference test, i.e., a statistical t-test test, on the average accuracy of the algorithms, and the t-test results are shown in Table 3. The performance of the data in the table shows that the p-value between the algorithm CDMIFS and any comparison algorithm is less than 0.05 in the vast majority of cases, and its overall p-value is 0.0157, 0.0208, 0.0237, 0.0207, 0.0205, respectively, i.e., the performance of the CDMIFS algorithm performs better than the other five algorithms. For example, when financial firms use the method to detect problems with the financial statements of companies that have been disclosed as fraudulent, i.e., when the data clearly contradict Benford's Law, there is a high probability that the company is suspected of financial fraud.

Table 3: t-test results

Project	CIFE	CMIM	JMI	mRMR	DCSF	CDMIFS
Profitability	0.02	0.008	0.014	0.007	0.018	-
Risk	0.032	0.02	0.023	0.018	0.024	-
Liquidity	0.026	0.004	0.026	0.033	0.013	-
Safety	0.011	0.029	0.037	0.038	0.016	-
Term	0.019	0.01	0.027	0.015	0.026	-
Leverage	0.004	0.024	0.017	0.024	0.037	-
Creditworthiness	0.002	0.04	0.021	0.002	0.012	-
Volatility	0.003	0.024	0.014	0.001	0.016	-
Mobility	0.019	0.021	0.027	0.036	0.018	-
Diversity	0.039	0.008	0.031	0.033	0.031	-
Relevance	0.006	0.038	0.037	0.039	0.024	-
Regulatory	0.007	0.023	0.01	0.002	0.011	-
Total	0.0157	0.0208	0.0237	0.0207	0.0205	-

3.1.5 Comparison of F1 values

Box-and-whisker plots can be used to analyze the shape of the distribution of comparative samples as well as the concentration trend of the sample distribution. A box-and-whisker plot consists of minimum, lower quartile, median, upper quartile, and maximum, where the upper quartile and lower quartile work together to draw the box, the median is located in the middle of the box, and the maximum and minimum are connected to the upper quartile and the lower quartile with a dotted line, respectively. Fig. 3 shows the box-and-whisker plots of the F1 values of the different algorithms, in which the horizontal coordinates represent the six algorithms for feature selection, and the vertical coordinates show the F1 values obtained by the different algorithms on the financial data set. The vertical coordinates are the F1 values obtained by different algorithms on the financial dataset. After observation, it is observed that CDMIFS>DCSF>mRMR>CMIM>JMI>CIFE on twelve features of financial dataset with F1 values of 0.855, 0.744, 0.694, 0.628, 0.624, 0.504. Therefore, it can be concluded that for the F1 value evaluation index CDMIFS algorithm achieves the best performance in terms of financial data feature selection. In financial enterprises, the current stage of financial data is often massive high-dimensional data, some attributes are not needed for anomaly detection, these attributes do not have any effect on anomaly detection, but will increase the difficulty of anomaly detection, this time, the CDMIFS algorithm can be used to select attributes conducive to the detection of anomalies from the high-dimensional data set, and examine whether the data

conforms to the law of Bentham Ford, thus providing favorable data support for the diagnosis of risks and frauds of the financial enterprises. Risk and fraud diagnosis to provide favorable data support.

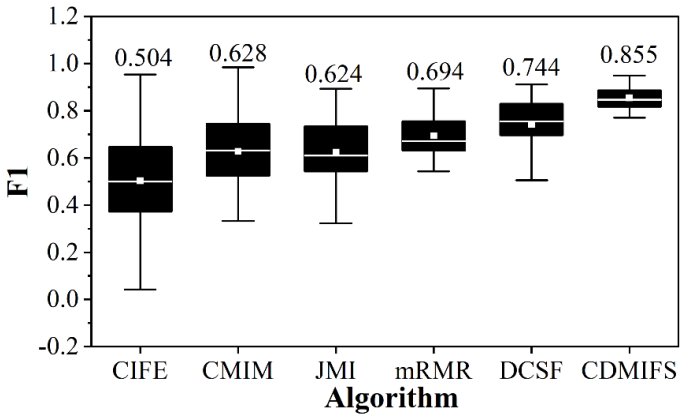

Figure 3: The F1 values of different algorithms

3.2 Financial Data Fraud Recognition Model Validation

3.2.1 Autoencoder Training Analysis

The variational self-encoder training in this paper was performed using the Keras library, an open source neural network library written in Python. The hidden layer is the core part of the variational auto-encoder, which is responsible for converting the input data into a compact representation and reducing it back to the original data in the decoder. In this paper, the hidden layer of the variational autocoder has three layers, and the neurons in the hidden layer learn how to extract and represent the useful information of the input data through the training process, so the number of neurons in each hidden layer is particularly important. In the pre-experiment, it is found that the number of neurons between 16 and 24 can maintain a better convergence state, so the optimization test of the number of neurons is based on this, and the training loss value of the number of neurons in each layer of the hidden layer varies with the number of iterations as shown in Fig. 4. In the experiments of different neuron number combinations, the number of neurons in the three hidden layers are 20, 16, 20, respectively, the training loss curve converges the fastest and the loss value is the smallest, and in the subsequent iterations to maintain a smooth state. Therefore, in this paper, the number of neurons in the self-encoder hidden layer is set to 20, 16, and 20, respectively.

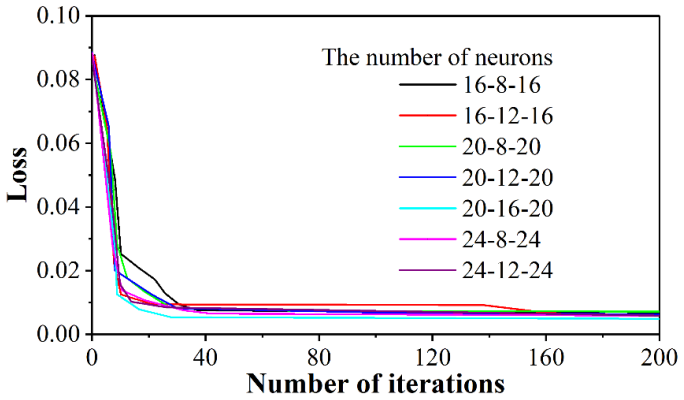

Figure 4: Training loss value variation curve

The variational autoencoder determines whether the instance is abnormal by calculating the reconstruction error of the instance, if the reconstruction error is greater than the threshold, it is abnormal, otherwise it is normal, so it is necessary to use the training data to find the autoencoder threshold. The distribution of the training error loss values of the variational self-encoder for the financial data training set is shown in Figure 5. The horizontal coordinates represent the distribution intervals of the variational autocoder error loss values, and the vertical coordinates represent the number of instances of error loss values that fall within different intervals. The error loss values of the variational self-encoder training data are distributed in the range of 0~0.24, and the experiments by selecting different error loss values found that the threshold value of 0.20 has the best effect, so this paper uses this as the threshold value of the self-encoder for basic detection.

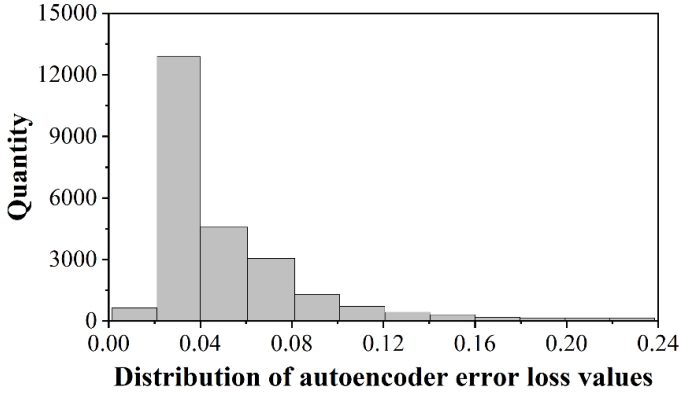

Figure 5: The distribution of error loss values in autoencoder training data

3.2.2 Graph Convolutional Neural Network Training

The graph convolutional neural network training is performed using the Scikit-learn library, which is an open software machine learning library for Python. After the financial data training set is trained by graph convolutional neural network, the distribution of abnormal financial data is shown in Figure 6. The horizontal coordinates represent the distribution intervals of abnormal financial data, and the vertical coordinates represent the number of instances of abnormal financial data falling in different intervals. Most of the abnormal financial data are in the range of more than 0.8, and very few are less than 0.12. By selecting different abnormal score values for experiments, it is found that the threshold value of 0.8 can achieve a better comprehensive effect, so this paper selects 0.8 as the threshold for graph convolutional neural network to be used in the identification of financial data fraud.

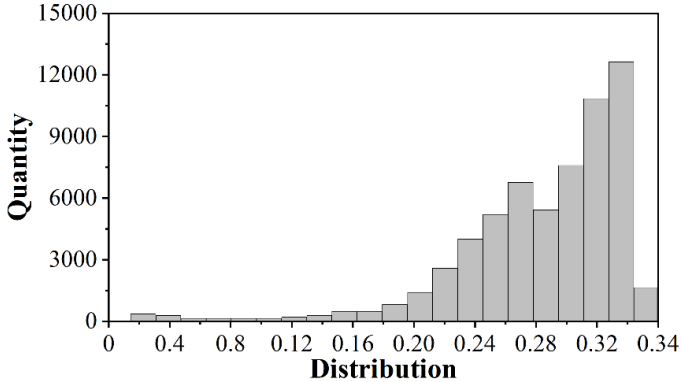

Figure 6: Abnormal distribution of financial data

3.2.3 Comparative analysis of recognition results

Under the same sample conditions, we use three different machine learning algorithms to construct a financial data fraud recognition model to predict the same dataset, and the prediction results show differences. Although the machine learning algorithms have strong learning ability, the models need to be optimized continuously due to the limitation of sample data. The effectiveness of this paper's model is verified by comparing the model's precision, recall and comprehensive index F-Score, and the comparative analysis of financial data fraud recognition results is shown in Table 4. According to the data in the table, it can be seen that the recognition effect of the three machine learning algorithms is very close. The highest precision of model recognition is OCGVAE 0.9728, which indicates that there is 0.9728 possibility of recognizing financial fraud data in the result of recognizing as a fraudulent user. The highest recall metric is OCGVAE 88.96%, indicating that most of the financial fraud data is successfully identified. Comparing the comprehensive evaluation metric F-Score, OCGVAE is 0.9293, SVM is 0.8715, RF is 0.7489, and OCGVAE still has the highest F-Score. It indicates that under the sample conditions given in this paper, the financial data fraud recognition model based on OCGVAE is optimal, both in terms of precision, recall index, and comprehensive index F-Score. For example, in the actual financial transaction activities, financial enterprises use the model in this paper to screen out user data with fraud possibility and mark them out, which can improve the anti-fraud efficiency of financial enterprises and establish a new security barrier for Internet finance.

Table 4: Comparative analysis of Financial Data fraud Identification results

Index	OCGVAE	SVM	RF
Accuracy	0.9728	0.8923	0.7672
Recall rate	0.8896	0.8517	0.7314
F-Score	0.9293	0.8715	0.7489

4 Conclusion

The number of financial data frauds revealed in the financial market is not to be underestimated, which urges regulators to increase supervision and improve the regulatory system. In this paper, under the theoretical guidance of Benford's law, we carry out financial data collection and cleaning to ensure the availability of research data. Combined with the relevant definitions of financial data features, the financial feature selection algorithm based on conditional dynamic mutual information is designed, and a financial data fraud identification model based on graph convolutional neural network and variational autoencoder is further constructed. Finally, the research content of this paper is exemplified and analyzed by combining financial data sets and evaluation indexes. Its research results are:

(1) The analysis of the financial feature selection algorithm shows that on the twelve features of the financial dataset, the algorithm's F1 value is ranked as CDMIFS>DCSF> mRMR>CMIM>JMI>CIFE, with the values of 0.855, 0.744, 0.694, 0.628, 0.624, and 0.504, which confirms that the CDMIFS algorithm has been used in the financial data feature selection The application value in terms of In the usual financial transactions, the feature data that conforms to Benford's law can be selected from the financial data by CDMIFS algorithm, thus realizing the intelligent diagnosis of the risk and fraud of financial enterprises.

(2) The precision, recall, and F-Score of the OCGVAE algorithm are superior to SVM and RF, with values of 0.9728, 0.8896, and 0.9293, respectively, i.e., it shows that the OCGVAE algorithm financial data fraud identification process is superior. Financial enterprises use the

model in this paper to accurately identify user data with fraud possibility and mark it out, which has theoretical guidance value for improving the anti-fraud efficiency of financial enterprises.

Funding

This work was supported by the Social Science Project of Chongqing Educational Research Institute (2022CJG07) and the Graduate Education Reform Research Project of Chongqing (YJG23117).

References

[1] Ngo, T., & Le, T. (2019). Capital market development and bank efficiency: a cross-country analysis. International Journal of Managerial Finance, 15(4), 478-491.

[2] Albashrawi, M. (2016). Detecting financial fraud using data mining techniques: A decade review from 2004 to 2015. Journal of Data Science, 14(3), 553-569.

[3] Roszkowska, P. (2021). Fintech in financial reporting and audit for fraud prevention and safeguarding equity investments. Journal of Accounting & Organizational Change, 17(2), 164-196.

[4] Grandstaff, J. L., & Solsma, L. L. (2021). Financial statement fraud: A review from the era surrounding the financial crisis. Journal of Forensic and Investigative Accounting, 13(3), 421-437.

[5] Du, M. (2021). Corporate governance: five-factor theory-based financial fraud identification. Journal of Chinese Governance, 6(1), 1-19.

[6] Qianru, Q. (2016). Trust and Financial Regulations: A Survey and Its Implication on Shanghai FTA. In New Strategic Research on China (Shanghai) Pilot Free Trade Zone (pp. 215-232).

[7] Yuan, Y., Feng, Y., & Lu, X. (2016). Statistical hypothesis detector for abnormal event detection in crowded scenes. IEEE transactions on cybernetics, 47(11), 3597-3608.

[8] Koren, O., Koren, M., & Peretz, O. (2023). A procedure for anomaly detection and analysis. Engineering Applications of Artificial Intelligence, 117, 105503.

[9] Hanif, M. A., Wadood, A., Ahmad, R. W., Shah, S. A., & Khan, R. (2025). Real-Time Anomaly Detection in IoT Sensor Data Using Statistical and Machine Learning Methods. ACADEMIA International Journal for Social Sciences, 4(3), 5203-5227.

[10] Yang, X., Qi, X., & Zhou, X. (2023). Deep learning technologies for time series anomaly detection in healthcare: A review. Ieee Access, 11, 117788-117799.

[11] Yahya, M. A., Moya, A. R., & Ventura, S. (2025). Deep learning for multivariate time series anomaly detection: an evaluation of reconstruction-based methods. Artificial Intelligence Review, 58(12), 400.

[12] Sun, C., Jia, Y., Song, H., & Wu, Y. (2020). Adversarial 3d convolutional auto-encoder

for abnormal event detection in videos. IEEE Transactions on Multimedia, 23, 3292-3305.

[13] Jin, L., Liu, Z., & Tang, Y. (2022, January). Auto encoder based= abnormal fluctuations monitoring method. In 2022 international conference on big data, information and computer network (BDICN) (pp. 768-777). IEEE.

[14] Wang, N., Chang, H., & Zhang, D. (2021). Theory-guided auto-encoder for surrogate construction and inverse modeling. Computer Methods in Applied Mechanics and Engineering, 385, 114037.

[15] Lu, S., Zhang, W., Zhao, H., Liu, H., Wang, N., & Li, H. (2024). Anomaly detection for medical images using heterogeneous auto-encoder. IEEE Transactions on Image Processing, 33, 2770-2782.

[16] Oh, D. Y., & Yun, I. D. (2018). Residual error based anomaly detection using auto-encoder in SMD machine sound. Sensors, 18(5), 1308.

[17] Pimbley, J. M. (2014). Benford's law and the risk of financial fraud. Risk Professional, 5, 1-7.

[18] Miller, S. J. (2015). A quick introduction to Benford's Law. Benford's Law: Theory and Applications, 3-22.

[19] Alali, F. A., & Romero, S. (2013). Benford's Law: Analyzing a decade of financial data. Journal of Emerging Technologies in Accounting, 10(1), 1-39.

[20] García-Sosa, A. T. (2024). Benford's Law and distributions for better drug design. Expert Opinion on Drug Discovery, 19(2), 131-137.

[21] Mir, T. A. (2014). The Benford law behavior of the religious activity data. Physica A: statistical mechanics and its Applications, 408, 1-9.

[22] Khan, M. A., Khan, U. N., Jamali, A. K., & Jamshed, J. (2022). The Factors Contributing to a Corporation's Demise: An Analysis of Enron. Journal of Management Practices, Humanities and Social Sciences, 6(2), 15-21.

[23] Nickell, E. B., Schwebke, J., & Goldwater, P. (2023). An introductory audit data analytics case study: Using Microsoft Power BI and Benford's Law to detect accounting irregularities. Journal of Accounting Education, 64, 100855.

[24] Riccioni, J., & Cerqueti, R. (2018). Regular paths in financial markets: Investigating the Benford's law. Chaos, Solitons & Fractals, 107, 186-194.

[25] Clippe, P., & Ausloos, M. (2012). Benford's law and Theil transform of financial data. Physica A: Statistical Mechanics and its Applications, 391(24), 6556-6567.

[26] Sugiarto, T., Budiman, A. I., & Rosini, I. (2016, December). The First Digits Analysis Until the Fifth Benford Law in Financial Statement. In International Conference on Ethics in Governance (ICONEG 2016) (pp. 1-4). Atlantis Press.

[27] Jones, W. A. (2020). A Benford Analysis of National Collegiate Athletic Association

Division I Finance Data. Journal of sports economics, 21(3), 234-255.

[28] Qu, H., Steinberg, R., & Burger, R. (2020). Abiding by the law? Using Benford's law to examine the accuracy of nonprofit financial reports. Nonprofit and Voluntary Sector Quarterly, 49(3), 548-570.

[29] Patel, P. C., Tsionas, M. G., & Guedes, M. J. (2022). Benford's law, small business financial reporting, and survival. Managerial and Decision Economics, 43(8), 3301-3315.

[30] G. Harb, E., Nasrallah, N., El Khoury, R., & Hussainey, K. (2023). Applying Benford's law to detect accounting data manipulation in the pre-and post-financial engineering periods. Journal of applied accounting research, 24(4), 745-768.

[31] Capalbo, F., Galati, L., Lupi, C., & Smarra, M. (2023). Local elections and the quality of financial statements in municipally owned entities: A Benford analysis. Chaos, Solitons & Fractals, 173, 113752.

[32] Cerqueti, R., Maggi, M., & Riccioni, J. (2024). Statistical methods for decision support systems in finance: how Benford's law predicts financial risk. Annals of Operations Research, 342(3), 1445-1469.

[33] Mućko, P. (2025). Exploring the Applicability of Benford's Law in Decision-Making Models for Assessing the Quality of Financial Reporting: A Case Study of Polish Public Companies. Procedia Computer Science, 270, 4274-4283.

[34] Davydov, D., & Swidler, S. (2016). Reading Russian tea leaves: Assessing the quality of bank financial statements with the Benford distribution. Review of Pacific Basin Financial Markets and Policies, 19(04), 1650021.

[35] Ahmadi, S. J., Faghani Makrani, K., & Fazeli, N. (2020). Providing a model for forecasting fraudulent financial statements and comparing financial statements and ratios with Benford Law. Journal of Management Accounting and Auditing Knowledge, 9(35), 221-237.

[36] Miao, Z. (2024). Financial fraud detection and prevention: Automated approach based on deep learning. Journal of Organizational and End User Computing (JOEUC), 36(1), 1-27.

[37] Lv, L., Cheng, J., Peng, N., Fan, M., Zhao, D., & Zhang, J. (2019, May). Auto-encoder based graph convolutional networks for online financial anti-fraud. In 2019 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr) (pp. 1-6). IEEE.

[38] Demestichas, K., Peppes, N., Alexakis, T., & Adamopoulou, E. (2021). An advanced abnormal behavior detection engine embedding autoencoders for the investigation of financial transactions. Information, 12(1), 34.

[39] Mohanty, D. K., Parida, A. K., & Khuntia, S. S. (2021). Financial market prediction under deep learning framework using auto encoder and kernel extreme learning machine. Applied Soft Computing, 99, 106898.

[40] Muthukumaran, K., Hariharanath, K., & Haridasan, V. (2023). Feature Selection with

Optimal Variational Auto Encoder for Financial Crisis Prediction. Computer Systems Science & Engineering, 45(1).

[41] Almahadeen, L., Mahadin, G. A., Santosh, K., Aarif, M., Deb, P., Syamala, M., & Bala, B. K. (2024). Enhancing Threat Detection in Financial Cyber Security Through Auto Encoder-MLP Hybrid Models. International Journal of Advanced Computer Science & Applications, 15(4).

[42] Buchdadi, A. D., & Al-Rawahna, A. S. M. (2025). Anomaly Detection in Open Metaverse Blockchain Transactions Using Isolation Forest and Autoencoder Neural Networks. International Journal Research on Metaverse, 2(1), 24-51.