Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Financial fraud identification model construction based on topological data analysis

Yu Guan[1,✉], Zhijuan Zong[2]

[1] *Business School of Fuyang Normal University, Fuyang, Anhui, 236041, China*
[2] *School of Economics, Fuyang Normal University, Fuyang, Anhui, 236041, China*

ABSTRACT
With the increasing complexity of the financial market, corporate financial fraud events occur frequently, posing a serious challenge to investors and market regulators. Aiming at the limitations of traditional financial fraud recognition methods, this paper constructs a financial fraud recognition model MCN based on the topological data analysis method. The model consists of two parts: the Mapper algorithm and one-dimensional convolutional neural network (1DCNN), which combines the global topology extracted by the Mapper algorithm with the local features of the IDCNN to realize the effective identification of financial fraud samples. In order to evaluate the recognition performance of the model, this paper controls the topological feature extraction method unchanged and the classifier unchanged respectively, and compares the performance of the MCN model with other financial fraud recognition models. The results show that the Acc and F1-score of the MCN-based financial fraud recognition model in this paper are 98.69% and 97.64%, respectively, which are better than other models in both perspectives, proving the superiority of the financial fraud recognition model based on topological data analysis constructed in this paper, and thus providing powerful technical support for the regulation of the financial market and the risk management of enterprises.

Keywords: topological data analysis, Mapper algorithm, one-dimensional convolutional neural network, financial fraud identification

1. Introduction

Topological data analysis is a very powerful technique which helps to analyze complex network relationships in the real world and visualize the topology. Topological data analysis techniques involve topology modeling, topology exploration and topology visualization, which can be used to extract graphs and charts, as well as in a variety of application areas such as search engine optimization and social network analysis [1-2]. In addition, topological data analysis can provide insights about social

✉ Corresponding author.
E-mail address: guanyu20090828@sina.com (Y. Guan).
Received 25 February 2024; Revised 10 May 2024; Accepted 15 December 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-469
© 2025 The Author(s). Published by Combinatorial Press. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

networks in order to discover how computational network analysis and machine learning can be performed to solve more complex problems. Therefore, topological data analytics has an important application in various industries [3-4].

Financial fraud refers to the behavior of financial transactions or financial statements with deceptive intent to mislead investors, creditors or other stakeholders. Financial fraud is one of the serious challenges facing businesses and organizations, and it is crucial to screen and prevent financial fraud [5-6]. Through measures such as topological data analysis, internal control assessment and due diligence, enterprises can effectively identify potential financial fraud issues and take appropriate measures to deal with them. Only by maintaining transparency, integrity and prudence can enterprises avoid the risk of financial fraud and achieve sustainable development and business success [7-9].

Literature [10] discusses the different detection techniques used in financial fraud detection in many fraud areas such as credit card fraud detection, telecommunication fraud detection and so on. Indicated the possible existence of real transactions in fraudulent transactions, which cannot be accurately detected by simple detection techniques. Literature [11] created a fraud recognition model based on structured raw data from financial reports, which is able to detect fraud quickly. And by comparing the fraud detection effect of different single classification machine learning algorithms and multiple integrated learning algorithms, integrated learning algorithms perform better in fraud detection. Literature [12] constructed a model for customers based on their behavior, by using outlier identification techniques and VA to start the analysis in order to achieve the misreporting rate of financial fraud identification. And this method was used for financial institutions and the results of the study verified the effectiveness of the proposed method. Literature [13] indicates that with the development of technology and socio-economic growth, financial fraud has become common. Popular and effective anomaly detection techniques for detecting financial frauds are comprehensively described, especially in the area of semi-supervised and unsupervised learning. Literature [14] emphasizes the serious threat that financial fraud poses to businesses. Taking the research method of combining normative analysis and empirical research, listed companies with fraudulent behaviors are taken as an example, and a financial fraud identification model is established based on the CRIME theory, and strategies to deal with financial fraud of listed companies are proposed in response to the results of the study. Literature [15] through the use of random forest for financial fraud technology detection and feature selection, variable importance measures, and the use of parametric models and non-parametric models of a variety of statistical methods to build the detection model, the results concluded that the random forest model shows the highest detection accuracy, has an important practical significance. Literature [16] based on bibliometrics concluded that there are 34 data mining techniques used to identify a wide range of financial frauds and support vector machine is the most used detection technique and summarized the outstanding results of different researchers in the field of financial fraud. Literature [17] outlines the application of machine learning techniques in financial statement fraud detection, especially computer intelligence techniques, and collects a large number of detection samples and also shows many innovations in the research process. Literature [18] explored the application of a variety of data mining techniques in financial fraud and found that the accuracy of fraud detection is affected by the model construction method and sample selection in empirical studies. And with the development of accounting information system, taking the fraud identification in the data mining techniques and text mining combined application will also become the trend of development. Literature [19] proposed a fraud detection framework CoDetect, which is not only able to detect financial fraud through the network information and feature information, but also can simultaneously detect fraudulent activities and feature patterns associated with it, through experiments to prove that the framework is conducive to combating financial fraud, in particular, money laundering, and other fraud detection effectiveness. Literature [20] aims to explore the theoretical and conceptual foundations of financial fraud, and it distinguishes between moral and psychological, economic and legal aspects. It is emphasized that even though the concepts related to financial fraud are widely used,

it is necessary to clarify the phenomenon. Literature [21] indulges in a literature review proposing the conceptual framework FraudFind, which enables the identification of fraud with the support of triangulation theory and advocates for efforts in the field of cybersecurity to reduce financial fraud. Literature [22] creates a text classification model applicable to the capital market based on the social media platform UGC, through aspects such as topic features, number of breaking news features in order to detect fraudulent companies. The analysis results emphasize that the constructed model is able to screen fraudulent behaviors of listed companies in a timely manner, which is of great significance in preventing market risks. Literature [23] stated that with the development of technology and the popularization of non-cash payments, fraud in the financial sector has gradually increased, and various financial tools have also led to the spread of fraudulent schemes, resulting in the growth of a variety of transnational financial crimes, so it is very necessary to look for favorable tools to prevent financial fraud. Literature [24] lies in the application of fraud pentagonal model to detect financial statement fraud with external pressure, financial objectives, etc. as independent variables and the ability to detect financial statement fraud as dependent variable, and 31 fraudulent and non-fraudulent companies were studied.

In this paper, a financial fraud recognition model based on topological data analysis is constructed by combining the Mapper algorithm and one-dimensional convolutional neural network, and its performance is examined. First, a paired dataset containing 720 sample enterprises is constructed by defining financial fraud samples and non-financial frauds and determining the selection criteria of sample data. Then, the variables are controlled with two components of the model, respectively. On the one hand, the IDCNN classifier is controlled unchanged and the extraction method of sample topological features is changed. On the other hand, the Mapper algorithm is controlled unchanged and a different classification algorithm is used. Thus, based on the accuracy and F1 score, the reliability and effectiveness of the MCN model for financial fraud recognition is demonstrated.

2. Application of topological data analysis in financial fraud identification

2.1. Definition and dangers of financial fraud

Financial fraud, refers to the accounting activities, the relevant parties in order to pursue personal interests, and take violations, fraud and other means to mislead and conceal financial information, so as to achieve the purpose of tax avoidance and other purposes, and to advance through the falsification of vouchers, false accounting, concealment of income, set up a series of off-the-books accounts, private remittances of public funds and other preparations and arrangements to intentionally create and provide false accounting information to deceive the user of the financial statements behavior. Financial fraud may involve false bookkeeping, concealment of important information, fictitious transactions and other means, resulting in the distortion of the enterprise's financial position, damage to credibility, falling stock prices, investors' interests and other consequences.

2.2. Limitations of traditional financial fraud identification methods

The traditional methods of corporate financial fraud identification are divided into four main categories: financial analysis, cash flow analysis, accounts receivable and inventory analysis.

1) Financial analysis method. Financial fraud identification methods based on financial analysis is mainly an analysis of the past period of the enterprise's operating performance and efficiency, profitability, solvency and development potential. Financial statement analysis is based on a comprehensive analysis of the existing assets of the enterprise, based on the financial position of the enterprise or monetary funds, physical assets or labor and other resources as the object, the enterprise in the past period and a certain period of time in the future related to assets, liabilities, owners' equity and other reasons for the change of economic interests to carry out a comprehensive analysis and

judgment. Financial analysis methods generally have a different focus, there are a variety of ratio indicators to calculate and analyze the method, but also through the financial statements of the proportion of each project to analyze the financial structure of the enterprise and the trend of change and so on.

2) Cash flow analysis. In reality, most enterprises will use cash flow analysis to analyze the quality of corporate profits. In short, it is to compare and analyze the cash generated by the enterprise in various activities and the net cash flow with the profit and investment income respectively, so as to identify and judge whether the enterprise's profit quality and net profit reach the predetermined target.

3) Accounts Receivable Analysis. Accounts receivable analysis is an analysis of current assets, i.e., the proportion of accounts receivable to main business revenue to analyze the ability of the enterprise's accounts receivable to bring in revenue. After comparing the accounts receivable turnover ratio by period, it is used to analyze the payback situation, and accordingly analyze whether the enterprise's accounts receivable management strategy is reasonable or not. It can also analyze the credit situation of accounts receivable customers, whether there is a balance between the size and capacity of each, and so on.

4) Inventory analysis. Inventory analysis can analyze the market share, production and sales volume, and the strength of control over upstream and downstream enterprises of the enterprise under study through the specific amount of raw materials, products in process, finished goods, etc. in the enterprise's inventory items. It is mainly an analysis of the inventory turnover rate. The faster the inventory turnover rate is, the lower the occupancy level of the inventory is reflected, the more liquid it is, and at the same time the faster it is converted to other forms of money such as cash. By comparing it to the optimal level of inventory for a given set of operating conditions, the activities of the business and its profits are then analyzed.

These traditional methods of financial fraud identification rely heavily on financial statement analysis, audit opinions, and the internal control mechanisms of the enterprise, which have a number of limitations. First, financial statements, while containing a wealth of financial information, are susceptible to manipulation and whitewashing. Second, audit opinions are limited by the auditor's professional judgment and audit procedures, making it difficult to detect all financial fraud. Finally, the internal control mechanism of an enterprise may be ineffective due to management override. As a result, traditional methods are inadequate in identifying complex frauds.

2.3. Advantages and application prospects of topological data analysis

Topological data analysis is an emerging data analysis method, which reveals the laws and patterns hidden behind the data by studying the topology of the data. Compared with traditional methods, topological data analysis has the following advantages:

1) The ability to handle high latitude data and capture nonlinear relationships in the data.
2) Able to identify outliers and noise in the data and improve the robustness of the model.
3) Ability to provide intuitive visualization results for users to understand and interpret the model.

In the field of financial fraud identification, topological data analysis is expected to become a new and effective tool.

3. Financial fraud identification model based on topological data analysis

In this paper, we use the topological data analysis method for the construction of financial fraud recognition model, extracting the global topology by Mapper algorithm and combining the local features of one-dimensional convolutional neural network (IDCNN) to get the financial fraud recognition model based on MCN (Mapper+IDCNN).

3.1. Theory of topological data analysis

The Topological Data Analysis (TDA) approach combines computational topology and data science, using geometric and topological theories as tools to enable qualitative analysis of data [25].

3.1.1. Topological space. In the research methodology of analyzing and processing data, the essence of each field is defined by the mathematical object of study, just as the object of study in linear algebra is the vector space, the mathematical object of interest in the analysis of topological data is the topological space, and these mathematical objects are simple sets with specific rules about set formation, transformation, and manipulation. Topological spaces can be defined from the point of view of sets:

Suppose there is a set of ordered pairs (X,τ), where X is a set and τ is a cluster of subsets of X, which satisfies the following properties:

1) The empty sets $\varnothing$ and X belong to τ.
2) The union of any number of elements in τ still belongs to τ.
3) The intersection of finitely many elements of τ still belongs to τ. Then the elements of τ become open sets and the cluster τ becomes a topology on X.

For this study, the topological space, i.e., the set of data point clouds obtained after phase-space reconstruction of time-series signals of gait features using delayed time embedding methods.

3.1.2. Metric space. A pure topological space is an extremely abstract concept, and at this point a topological space with an explicit notion of distance, i.e., a metric space, is needed. It is worth noting that all metric spaces are topological spaces, but not all topological spaces are metric spaces, i.e., topological spaces do not always have an exact concept of distance. The topological data analysis methods used in this paper are all based on metric spaces for research. In this paper, metric spaces are defined by means of sets and distance functions:

Suppose there is an ordered pair (X,d), where X is a set, and define function $d: X\times X\to$ □ to be the metric of X, i.e., function d maps every ordered pair of elements in X to an element of the set of real numbers □, for any element $x,y,z\in X$:

1) $d(x,y)\geq 0$.
2) $d(x,y)=0$ if and only if $x=y$.
3) $d(x,y)=d(y,x)$
4) $d(x,z)\leq d(x,y)+d(y,z)$.

3.2. Basic steps in topological data analysis

The basic steps of TDA are as follows:

Step 1: A filter function is used to calculate a filter value for each data point.

Step 2: The data points are divided into different filter value intervals according to their filter values, from small to large. Neighboring filter intervals are set to have a certain overlap region, and the points in the overlap region belong to two intervals at the same time.

Step 3: Cluster analysis is done on the data in each interval.

Step 4: The subclasses obtained from the clustering of the intervals are put together, and each subclass is represented by a circle of different sizes. If there are the same raw data points between the 2 classes (this is the reason why the intervals need to overlap each other), then an edge is added between them.

Step 5: Apply a layer of mechanical layout to the graph of circles and edges to balance it and get the final "data graph".

The schematic representation of the basic workflow of TDA is shown in Fig. 1. Where X is representing the high dimensional data and X_r is representing the nested simplex complexes, the process converts the high dimensional data into simplex complexes in which the homology of each

complex is considered to give the persistence $H_i(X_{r_0}) \to H_i(X_{r_1}) \to H_i(X_{r_2}) \to \ldots$ The persistence module is converted into a barcode or topology analysis graph of the data by visualization.

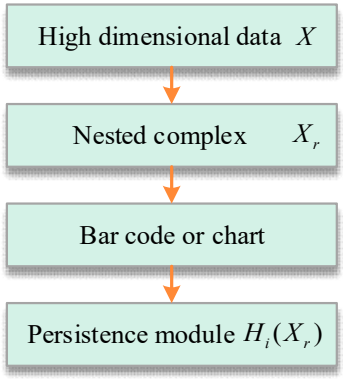

Fig. 1. Basic workflow of TDA

3.3. Mapper's algorithm

High-dimensional data cannot be visualized directly, although there are many methods to extract low-dimensional structures from a dataset, such as principal component analysis, local linear embedding and multidimensional scaling. However these methods do not perform well when dealing with high dimensional data as many different topological features can be found in the same dataset, therefore TDA is crucial for the study of high dimensional spatial visualization, in this paper we use a general method called Mapper, which not only demonstrates the clusters, relationships between variables but also obtains a higher level of understanding of the structure of high dimensional data.

3.3.1. Principles of the Mapper Algorithm. The Mapper algorithm summarizes the topology of a dataset onto a graph through a mapping $f: X \to G$ is a way to build graphs from data, it reveals topological features of high-dimensional data spaces and can estimate important connections in the underlying space of the data Data mining and exploration in visual topological maps are applicable to any simple or complex data and are very flexible [26].Mapper can be Visualize or cluster data, an unsupervised method for generating visual representations of data, often revealing new insights and discoveries about the data that are not available by other methods, and can be used by data analysts to explore the structural and topological properties of complex high-dimensional data sets.

3.3.2. Mapper Algorithm Implementation Steps. In practical applications mainly distributed Mapper algorithm is used and to ensure that the output of distributed Mapper is same as the sequential Mapper, some processing on the overlay is required to obtain the final Mapper output. In Distributed Mapper, consider N chains of overlays $A_1, A_2, ..., A_N$ spaced one $[a,b]$ and their overlays $u_1, u_2, ...u_N$. After completing the preprocessing of the overlays and obtaining the set $\langle A_i, U_i \rangle_{i=1}^N$, each pair (A_i, U_i) is then mapped to a specific processor P_i which P_i performs some computations to produce subgraphs G_i and finally merges the subgraphs into a single graph G.

1) Sequential Mapper

Input: dataset x with metric concepts between data points.

Scalar function: $f: X \to R^n$

A finite $f(X)$ covering $u = \langle U_1, U_2, ...U_k \}$

Output: a graph representing $N_1(f^*(u))$.

Step 1: For each set $X_i := f^{-1}(U_i)$, its cluster $X_{i,j} \subset X_i$ is computed by using clustering algorithm.

Step 2: Each cluster is considered as a vertex in the Mapper graph. Moreover, an edge is inserted between two nodes X_{ij} and X_{kl} whenever $X_{ij} \cap X_{kl} \neq \phi$.

2) Coverage Preprocessing

Input: point cloud data X.

Scalar function: $f: X \to [a,b]$.

Set of N processors (P):

Output: a set of pairs of $\{A_i, U_i\}_{i=1}^{N}$ where $\{A_i\}_{i=1}^{N}$ is a $[a,b]N$-chain cover and U_i is a A_i-cover.

Step 1: Construct a N-chain covering of $[a,b]$, $[a,b]$ covered by N open intervals $A_1, A_2, ..., A_N$ when $|i-j|=1$ and empty set $A_i, j := A_i \cap A_j \neq \phi$.

Step 2: For each open set A_i construct an open covering $U_i, \langle U_i \rangle_{i=1}^{N}$ covering satisfying the following conditions:

1) $A_{i,i+1}$ is the open set covering U_i and U_{i+1}, i.e., $U_i \cap U_{i+1} = \{A_{i,i+1}\}$

2) If $U_i \in u_i$ and $U_i + 1 \in u_{i+1}$ make $U_i \cap U_i + 1 \neq \phi$; then for each $i = 1,2,...,N-1$; there is $U_i \cap U_{i+1} = A_{i,i+1}$.

Algorithm 1: Distributed Mapper

Input: point cloud data X. scalar function: $f: X \to [a,b]$. N processor (P) set. Pairs of sets $\{A_i, U_i\}_{i=1}^{N}$ obtained from the coverage preprocessing algorithm.

Output: distributed Mapper graph.

Step 1: for $(i \leftarrow 1 \; to \; i = N)$ do.

Step 2: $P_i \leftarrow (A_i, u_i)$ //Map each A_i and its overlay u_i to processor P_i.

Step 3: Determine the set of points $x_i \in X$ through f it is mapped to A_i and run the Sequential Mapper construction on the overlay $(f \mid x_i)^*(u_i)(i = 1,2,...,N)$ at the same time to get N graphs $G_1, G_2, \ldots, G_N$, and $N = 1$, then return graph G_1.

Step 4: Let $C_{j_1}^i, C_{j_2}^i, ..., C_{j_i}^i$ be the cluster obtained from $f^{-1}(A_{i,j+1})$. By choosing to cover u_i and u_{i+1}, these clusters are represented by vertices $v_{ji}^{\;i}, v_{j_2}^j, ...v_{j_i}^j$ in G_i and G_{i+1} (each v_k^i corresponding to cluster C_k^i).

Step 5: Merge graph $G_1, G_2, ...G_N$ in the following way by constructing $A_{i,i+1}, u_i$ and u_{i+1} such that each $f^*(u_i)$ and $f^*(u_{i+1})$ share a cluster $C_{ji}^{\;i}$ in $f^*(A_{i,i+1})$ and hence $C_{j_i}^i$ is represented by a vector in graphs G_i and G_{i+1}, the merge is done by considering disjoint union graphs $G_1 \cup G_2 \cup ... \cup G_N$ and then taking the quotient of this graph to determine the corresponding vertices in G_i and $G_i + 1 (1 \le i \le N-1)$.

3.4. One-dimensional convolutional neural networks

One-dimensional convolutional neural network (1DCNN) is a multilayer neural network model that combines feature extraction and classification and recognition layers [27]. 1DCNN has the characteristics of sparse connection and weight sharing, which can significantly reduce the network parameters and avoid algorithm overfitting, and its convolution operation can obtain the typical features of the input data, and ultimately realize the feature expression based on the invariant translation, rotation and scaling of the input data, which has excellent robust characteristics and generalization ability.

3.4.1. Convolutional layers. Convolutional layers utilize multiple convolution kernels to perform local region convolution operations with the input signals of the layer to produce a sequence of

features for each local region to achieve feature extraction of the signal. In a convolutional layer, each layer output corresponds to the convolution result of multiple inputs, which is mathematically modeled as follows:

$$X_j^n = f\left(\sum_{i\in M_i} X_i^{n-1} * W_{ij}^n + B_j^n\right) \tag{1}$$

where: i and j are the position indexes of the input and output data respectively. X_i^{n-1} is the ith input feature of layer $n-1$. * is the convolution operation. W is the weight matrix. B is the bias matrix. X_j^n is the jth output feature of layer n. M_i is the ith convolution region. $f(\cdot)$ is the activation function.

3.4.2. Activation functions. In 1DCNN, a nonlinear activation function is usually applied to the results of convolutional operations or neuron outputs to enhance the representational capability of 1DCNN models. The common activation functions are ReLU, Tanh, and Sigmoid functions. The three activation function algorithms are shown in Fig. 2, the Tanh function and Sigmoid function have a slow change trend when the input value is large, and the derivative value is close to 0. The ReLU function has a constant derivative value of 1 in the positive interval, which can effectively avoid the gradient dispersion problem. Therefore, this paper adopts the ReLU activation function to improve the nonlinear expression ability of the model and accelerate the convergence speed of the model, and its mathematical expression is as follows:

$$f(x) = \max\{0, x\} \tag{2}$$

where x is the output feature or neuron output after the convolution operation.

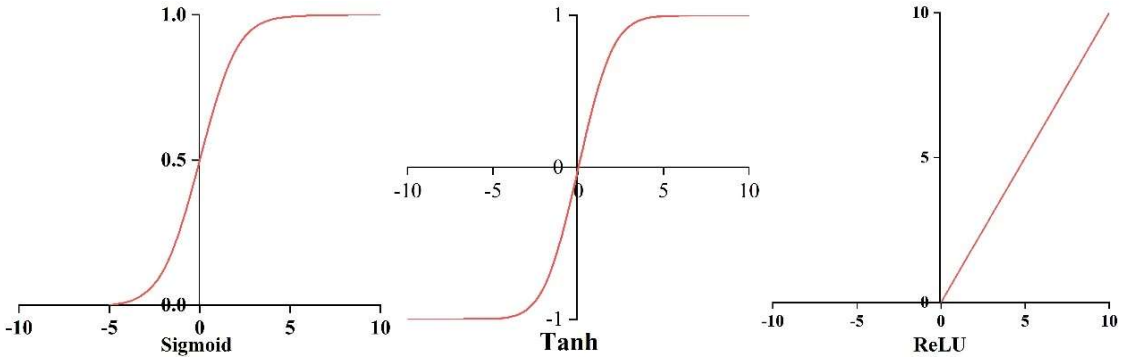

Fig. 2. Three activation functions

3.4.3. Pooling layer. The pooling layer uses the pooling function to realize the downsampling of the target layer data features, which can reduce the number of training parameters, enhance the significant characteristics of the sample data, and constrain the overfitting behavior to a certain extent. In this paper, the maximum pooling operator is used, and its main idea is to capture the maximum value in the overall features of a certain spatial scale to replace the original overall spatial output characteristics, and its function expression is:

$$H_i^{n+1}(j) = \max_{(j-1)d+1\le m\le jd}\left\{F_i^n(m)\right\} \tag{3}$$

where: F_i^n and H_i^{n+1} are channel i input features of layer n and channel i output features of layer $n+1$, respectively. j is the pooling region increment. m is the data location variable of the pooling region. d is the pooling region width.

3.4.4. Classification layer. The classification layer mainly consists of a fully connected layer and an activation function. The ultimate goal of the classification layer is to solve binary or multiple classification problems. Sigmoid activation function is usually used for binary classification problems, and its mathematical model is as follows:

$$S(x)=\frac{1}{1+e^{-x}} \tag{4}$$

3.4.5. Objective function. The objective function, i.e., the loss function, is used to indicate the degree of agreement between the predicted values and the actual values.1 The DCNN model performs several transformations and operations on the set of input values to obtain a definite set of output values. The smaller the value of the objective function, the better the degree of consistency between the set of output values and the set of actual values, and the better the robustness of the model. Financial fraud identification is a binary classification problem, so the binary cross entropy loss function is used, and its mathematical expression is:

$$J=\sum_{i=1}^{k} y_i \log \hat{y}_i + (1-y_i)\log\left(1-\hat{y}_i\right) \tag{5}$$

where: k is the number of sample categories. y_i is the sample label value. The label value of the true sample is {0,1}, $\hat{y}_i$ is the probability that the sample label is predicted to be 1, and $1-\hat{y}_i$ is the probability that the sample label is predicted to be 0.

4. Analysis of the application of the financial fraud identification model

4.1. Selection of model samples

4.1.1. Definition of the financial fraud sample. The definition criteria of financial fraud will directly affect the empirical analysis of the model, and different judgment criteria will screen out different research samples. Therefore, this paper determines whether the listed company is publicly punished by the regulator as the judgment standard of financial fraud.

The financial fraud sample of this study is selected from the list of penalties published by the regulatory bodies, that is, the A-share listed companies whose annual financial reports have been penalized by the Securities and Futures Commission (SFC) and other regulatory authorities for reasons such as fictitious assets, inflated profits, false statements (misleading statements), material omissions, etc., and if the listed company is involved in a multi-year fraud, then the year of the first occurrence of financial fraud is selected as the year of the research data. The years of the research sample are set from 2012 to 2022, and all sample data are obtained from the Cathay Pacific China Listed Company Violation Handling Research Database.

4.1.2. Definition of the non-financial fraud sample. In this paper, we propose to select listed companies as a paired sample of fraudulent companies from those that have not been publicly penalized by regulators for financial fraud. Since it is impossible to exclude the situation that the financial fraud of some listed companies has not been punished by the regulators for the time being, in order to improve the credibility of the sample and the prediction accuracy of the model, this paper makes the following requirements for the definition of the sample of non-fraudulent enterprises:

1) Enterprises have not been penalized by regulators for financial fraud during the period from 2012 to 2022.

2) The audit opinion of the annual report of the enterprise in that year is “standard unqualified opinion”.

In addition, in order to make the fraud identification model more stable, the following principles are followed when pairing fraud samples with non-fraud samples:

1) Fraud samples and non-fraud samples are selected in the ratio of 1:1.

2) Non-financial fraud companies are prioritized to select companies that are in the same industry and same year as the financial fraud companies.

4.1.3. Sample selection results. According to the above guidelines for defining research samples, this paper selects financial fraud samples from listed companies that have been publicly penalized by regulators for financial fraud during the period from 2012 to 2022, and collects a total of 360 financial fraud companies. And according to the above matching principle, 360 non-financial fraud companies in the same industry in the same year were selected as matching samples according to the ratio of 1:1, and finally a total of 720 research samples were used to establish the financial fraud identification model.

The industry distribution of the 360 listed companies selected in this paper that have been publicly penalized by regulators due to financial fraud during the period from 2012 to 2022 is shown in Figure 3, of which less than 10 are uniformly categorized within other industries.

As can be seen in Figure 3, the largest share is in the manufacturing industry, with 176 fraudulent companies, accounting for 48.89%. This is followed by the wholesale and retail industry with 51 or 14.17%. The real estate industry has 37 fraudulent enterprises, accounting for 10.28%. The transportation, energy, and mining industries accounted for a relatively small number of fraudulent enterprises, 15, 22, and 14 fraudulent enterprises, or 4.17%, 6.11%, and 3.89%, respectively. The largest number of fraudulent enterprises in the manufacturing sector is due to the large base and wide coverage of enterprises in the manufacturing sector, which accounts for a relatively large proportion.

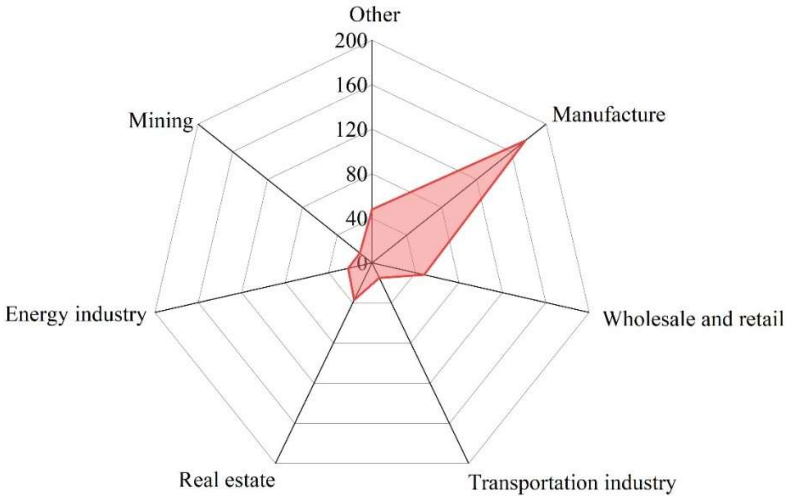

Fig. 3. Fraud sample industry distribution

4.2. Comparative analysis of financial fraud identification

4.2.1. Comparison of topological feature extraction methods. In this study, the topological features extracted by the Mapper algorithm are used as inputs to the 1D-CNN model to obtain the financial fraud recognition model MCN, and then its effectiveness and superiority in financial fraud recognition are verified through experiments. In the experiment, 720 samples are divided into 80% training set and 20% test set, the ratio of financial fraud samples and non-financial fraud samples in both training set and test set is kept 1:1, the iteration rounds are 1000 times, and 5 trials are conducted to take the average value. Experiments are conducted using Tensorflow framework, i7-10700F CPU, NVIDIARTX3060 GPU and 32GB RAM. Pattern recognition results were evaluated using overall accuracy and F1 score. The Mapper algorithm with Principal Component Analysis (PCA), DBSCAN Cluster Analysis, Persistent Homotopy (PH), and Self-Organizing Mapping (SOM) are all used to extract topological features from the samples and the extraction results are fed into 1DCNN to compare the

recognition of financial fraud. The comparison of financial fraud recognition results of different topological feature extraction methods is shown in Table 1.

As can be seen from Table 1, the topological feature extraction method based on Mapper's algorithm is 98.69% and 97.64% in Acc and F1 score indexes, respectively, which are greater than other methods, indicating that this paper's method can achieve advantages in pattern recognition, which can effectively reduce the error and optimize the performance of pattern recognition.

Table 1. Comparison of identification results of different methods

Method	Evaluation index	
	Acc	F1-score
SOM	94.57%	93.46%
DBSCAN	95.98%	94.76%
PCA	96.12%	95.84%
PH	97.32%	96.45%
Mapper	98.69%	97.64%

4.2.2. Comparison of different models for financial fraud identification. In order to verify the advantages of Mapper extracted topological features in financial fraud recognition, this paper compares the 1DCNN with Mapper topological features as input with the traditional financial fraud recognition methods of PRPD mapping and statistical features STA, and also the point cloud image without feature extraction and the combined features combining the topological features with the statistical features are respectively inputted into different classifiers to be Comparison. The classifiers used are 2DCNN, RF, GBDT, GAN and SVM, PC represents the local discharge 3D point cloud image without feature extraction, and STA+TDA represents the combined features of statistical features and topological features. The financial fraud recognition results of different models are shown in Table 2.

As can be seen from Table 2, the MCN financial fraud recognition model in this paper performs better than other mainstream models, and its accuracy can reach 98.69%, F1 score reaches 97.64%, and the differentiation is relatively good. Compared to the model with statistical features as input, the model in this paper can improve the accuracy by up to 14.37%. Compared to the model with PRPD mapping as input, the accuracy of the model in this paper is improved by 7.18%. Although the STA+TDA+1DCNN and PC+2DCNN models perform relatively well, there is still an accuracy gap of 0.82% and 2.56% compared to this paper's method, and the F1 score gap is 1.66% and 5.22%, respectively.

Table 2. Comparison of identification results of different models

Model	Evaluation index	
	ACC/%	F1-score/%
Mapper+1DCNN(MCN)	98.69%	97.64%
STA+1DCNN	92.96%	91.87%
STA+TDA+1DCNN	97.87%	95.98%
PC+2DCNN	96.13%	92.42%
PRPD+2DCNN	91.51%	85.37%
STA+RF	84.32%	75.66%
STA+GBDT	86.43%	79.35%
STA+GAN+SVM	87.08%	78.44%
STA+GAN+RF	92.75%	85.16%

At the same time, in order to fully prove the superiority of the model in this paper, the financial fraud recognition performance of the IDCNN classifier used in this paper is comparatively analyzed. Under

the condition that the same Mapper algorithm is used to process the data, the financial fraud recognition performance of different classifiers is shown in Figure 4.

As can be seen in Fig. 4, 1DCNN shows a clear advantage in processing the dataset with financial fraud topological features, with an overall accuracy of 98.69%. In contrast, the F1 scores of the other classifiers are relatively low, with the gap ranging from 2.80% to 6.64%. Therefore, 1DCNN can further enhance the superiority of Mapper's algorithm in differentiating financial fraud samples from non-financial fraud samples.

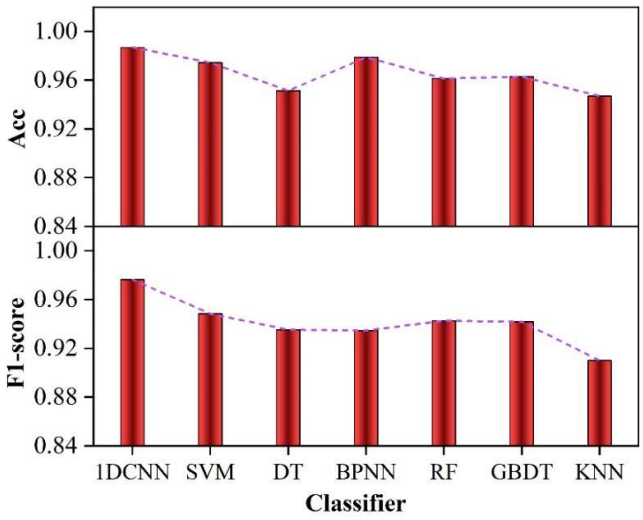

Fig. 4. The identification performance of different classifier

5. Conclusion

In this paper, based on the Mapper algorithm of topological data analysis and one-dimensional convolutional neural network, the construction of financial fraud recognition model is realized and its recognition performance is examined. Its main steps and conclusions are shown below:

First, the judgment criteria of financial fraud samples and non-financial fraud samples are defined, and 360 financial fraud enterprise samples are selected from the penalty list published by the regulatory agencies, and 360 enterprises in the same industry and the same year as the financial fraud companies are selected as the paired non-financial fraud samples.

Secondly, this paper evaluates the financial fraud identification effect of the model. First, the IDCNN model is used as a classifier for financial fraud recognition with control invariance, and different topological feature extraction methods are selected to compare the recognition performance of the model obtained by combining it with 1DCNN. In terms of recognition accuracy and F1 score, the topological feature extraction method based on Mapper algorithm is 98.69% and 97.64% respectively, which is better than Principal Component Analysis (PCA), DBSCAN cluster analysis, Persistent Homotopy (PH), and Self-Organizing Mapping (SOM) topological feature extraction methods. Then the Mapper algorithm is controlled unchanged and different classifiers are replaced to compare their recognition effects. The results show that 1DCNN exhibits a significant advantage in dealing with the dataset of financial fraud topological features, with an overall accuracy of up to 98.69%, while the F1 score is significantly better than the other classifiers, with a gap ranging from 2.80% to 6.64%. It indicates that 1DCNN can further enhance the superiority of Mapper's algorithm in financial fraud recognition.

Meanwhile, the MCN model is compared with other combination models. Compared with the model with statistical features as input, the model in this paper can improve the accuracy up to 14.37%. Compared to the model with PRPD mapping as input, the accuracy of this paper's model is improved by 7.18%. And comparing the STA+TDA+1DCNN and PC+2DCNN models, the accuracy of this paper's

method is improved by 0.82% and 2.56%, and the F1 scores are improved by 1.66% and 5.22%, respectively.

The comparison experiments of the three aspects fully proved the effectiveness of the MCN financial fraud identification model constructed in this paper, and provided a technical path for the effective identification of financial fraud.

Funding

The Key Research Projects of Humanities and Social Sciences in Universities of Anhui Province (2023AH052850, 2022AH052811).

References

[1] Wasserman, L. (2018). Topological data analysis. Annual Review of Statistics and Its Application, 5(1), 501-532.

[2] Skaf, Y., & Laubenbacher, R. (2022). Topological data analysis in biomedicine: A review. Journal of Biomedical Informatics, 130, 104082.

[3] Munch, E. (2017). A user's guide to topological data analysis. Journal of Learning Analytics, 4(2), 47-61.

[4] Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes. J. Mach. Learn. Res., 16(1), 77-102.

[5] Reurink, A. (2018). Financial fraud: A literature review. Journal of Economic Surveys, 32(5), 1292-1325.

[6] Karpoff, J. M. (2021). The future of financial fraud. Journal of Corporate Finance, 66, 101694.

[7] West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: a comprehensive review. Computers & security, 57, 47-66.

[8] Lv, B., Cheng, P., Zhang, C., Ye, H., Meng, X., & Wang, X. (2021, October). Research on modeling of e-banking fraud account identification based on federated learning. In 2021 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech) (pp. 611-618). IEEE.

[9] Hashim, H. A., Salleh, Z., Shuhaimi, I., & Ismail, N. A. N. (2020). The risk of financial fraud: a management perspective. Journal of Financial Crime, 27(4), 1143-1159.

[10] Richhariya, P., & Singh, P. K. (2012). A survey on financial fraud detection methodologies. International journal of computer applications, 45(22), 15-22.

[11] Chen, Y., & Wu, Z. (2022). Financial fraud detection of listed companies in china: A machine learning approach. Sustainability, 15(1), 105.

[12] Torres, R. A. L., & Ladeira, M. (2020, December). A proposal for online analysis and identification of fraudulent financial transactions. In 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 240-245). IEEE.

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

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

[15] Liu, C., Chan, Y., Alam Kazmi, S. H., & Fu, H. (2015). Financial fraud detection model: Based on random forest. International journal of economics and finance, 7(7).

[16] Al-Hashedi, K. G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. Computer Science Review, 40, 100402.

[17] Chimonaki, C., Papadakis, S., Vergos, K., & Shahgholian, A. (2019). Identification of financial statement fraud in Greece by using computational intelligence techniques. In Enterprise Applications, Markets and Services in the Finance Industry: 9th International Workshop, FinanceCom 2018, Manchester, UK, June 22, 2018, Revised Papers 9 (pp. 39-51). Springer International Publishing.

[18] Yao, S. (2021, September). Application of data mining technology in financial fraud identification. In 2021 4th International Conference on Information Systems and Computer Aided Education (pp. 2919-2922).

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

[20] KIZYMA, T., & KHAMYHA, Y. (2019). Financial fraud: theoretical conceptualization and economic basis. World of finance, (2 (59)), 109-123.

[21] Sánchez, M., Torres, J., Zambrano, P., & Flores, P. (2018, January). FraudFind: Financial fraud detection by analyzing human behavior. In 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 281-286). IEEE.

[22] Liu, H., Tao, L., & Liu, M. (2020, November). Research on financial fraud identification of listed companies based on text data mining. In 2020 International Conference on Image, Video Processing and Artificial Intelligence (Vol. 11584, pp. 457-462). SPIE.

[23] Gryazeva, E., Mayorova, O., Malchikova, N., Nemkova, M., & Paravina, M. (2021). International financial fraud: economic and psychological aspects, classification and ways of minimization. Economic Annals-XXI/Ekonomìčnij Časopis-XXI, 189.

[24] Andriani, K. F., Budiartha, K., Sari, M. M. R., & Widanaputra, A. A. G. P. (2022). Fraud pentagon elements in detecting fraudulent financial statement. Linguistics and Culture Review, 6(S1), 686-710.

[25] Serena Grazia De Benedictis, Grazia Gargano & Gaetano Settembre. (2024). Enhanced MRI brain tumor detection and classification via topological data analysis and low-rank tensor decomposition. Journal of Computational Mathematics and Data Science100103-100103.

[26] Ma Ling Yan, Feng Tao, He Chengzhang, Li Mujing, Ren Kang & Tu Junwu. (2023). A progression analysis of motor features in Parkinson's disease based on the mapper algorithm. Frontiers in Aging Neuroscience1047017-1047017.

[27] Marapatla Ajay Dilip Kumar & E Ilavarasan. (2024). An effective attack detection framework using multi-scale depth-wise separable 1DCNN via fused grasshopper-based lemur optimizer in IoT routing system. Intelligent Decision Technologies(3),1741-1762.