GAN-Based Financial Data Generation and Prediction: Improving The Authenticity and Prediction Ability of Financial Statements

Feng Qi
School of Management, Guangdong University of Foreign Studies South China Business College Guangzhou 510545, China
E-mail: as_453520854@163.com

Keywords: GAN, financial data, financial statements, authenticity, prediction ability

Received: October 16, 2024

The research on the mining algorithm of financial data association relationship mainly explores a certain kind of association relationship in depth, but it is not suitable for the attributes and characteristics of financial data itself, and there are few comprehensive analysis and application for financial data association relationship mining. In order to overcome the above problems, this paper proposes a financial data generation and prediction model based on GAN. Based on WGAN network, this paper improves the authenticity of the generated virtual samples by increasing the cyclic consistency loss term and selecting intermediate samples for the generated samples to optimize the generated model. At the same time, in the system, this paper adopts intelligent data analysis research method, mines the correlation of different dimensions of financial statement data, and presents the mining results by using the correlation visualization method, so as to realize the risk assessment and trend prediction of enterprise financial status. In terms of the overall recognition accuracy of the model, the random forest model has the highest accuracy rate, reaching 74.46%. In terms of recall rate, the GBDT model is slightly higher than the random forest model, with a recall rate of 73.43%, but its accuracy rate, F1 value and AUC value are slightly lower than the random forest model. According to the comprehensive experimental analysis results, it can be seen that the model proposed in this paper has good performance in the authenticity analysis and prediction of financial data. Generally speaking, the model proposed in this paper provides a reliable tool for the authenticity audit of financial data, and can provide a reference for the formulation of subsequent schemes and policies through financial data prediction.

Povzetek: A deep transfer learning-based model (DTL-MD) enhances malicious code detection using ResNet50V2, GAN-generated variants, and online learning, achieving 95.8% accuracy and improving detection speed and robustness against evolving threats.

1 Introduction

After the formation of the capital market, the financial analysis system has been gradually improved, and the regulatory agencies have clearly defined and required the scope, frequency and caliber of financial data that enterprises need to disclose. At the same time, internal and external audits will severely punish the fraud of financial data, so that the financial statement data publicly disclosed by enterprises can truly reflect the operating status of enterprises, and the scope of financial analysis is correspondingly expanded to analyze the financial status, operating results and cash flow of enterprises. The traditional analysis method is to quantitatively or qualitatively evaluate the financial status of an enterprise in the field of financial accounting according to the key indicators formulated by the enterprise's solvency, operational ability and profitability, as well as the year-on-year situation of the indicators, but the ability to predict financial risk exposure and financial development trend is weak. Therefore, relevant experts began to try to use more mature artificial intelligence and data mining methods for financial analysis and prediction, but there are few studies on judging or predicting the operating conditions of enterprises by mining the association relationship between enterprise financial data [1].

The association relationship between enterprise financial data will produce various manifestations according to different data objects. Among them, the distance attribute of enterprise financial indicators in different dimensional spaces is the spatial association relationship of enterprise finance, and enterprises with closer distance in multi-dimensional spaces have higher financial similarity. The group dependence attribute between financial indicators in all enterprises is the static time association relationship of financial indicators. Based on the frequently occurring financial indicator groups, that is, the frequent item sets of financial indicators, the abnormal financial data of enterprises can be found [2]. The related attribute of the historical trend of financial indicators in different industries is the dynamic time correlation between industries. The change

of the financial situation of upstream industries will have an impact on downstream industries through a period of transmission, and then the financial indicators of upstream and downstream related industries will show positive or reverse correlation in the time trend. Through the analysis of trend correlation, the future financial situation of downstream industries can be predicted. According to different association relationships, predecessors have developed corresponding algorithms for data mining, but at present, various algorithms mainly explore a certain association relationship in depth, and there are few comprehensive analysis and applications for all association relationships [3].

In order to overcome the above problems, this paper proposes a financial data generation and prediction model based on GAN. Based on WGAN network, this paper improves the authenticity of the generated virtual samples by increasing the cyclic consistency loss term and selecting intermediate samples for the generated samples to optimize the generated model. At the same time, in the system, this paper adopts intelligent data analysis research method, mines the correlation of different dimensions of financial statement data, and presents the mining results by using the correlation visualization method, so as to realize the risk assessment and trend prediction of enterprise financial status.

2 Related works

In view of the data results mined by association relationships, some scholars further adopt visual analysis methods to release the value of data, make it easier to understand and make the essence of things more prominent. They also integrate visual interactive interfaces into the process of data mining, and combine expert experience in the execution steps of the algorithm, so as to improve the interpretability of the algorithm and make the data mining results more match the business reality. However, the existing visual analysis methods of association relationships cannot well adapt to the characteristics of financial data. At present, most of the various methods focus on the association relationships themselves, and the visual analysis of financial data needs to be further combined with the analysis objectives of indicator trends, indicator proportions, group changes, group order, etc. Moreover, the existing methods lack a comprehensive visual analysis solution for financial data association that matches the above objectives [4].

Scholars have used flow chart research, management evaluation scoring, stage analysis, etc., but qualitative research can't be applied to the changeable needs of financial risk early warning, and its accuracy can't meet the needs of enterprises. Therefore, scholars began to study the financial indicators of enterprises by quantitative methods, and developed univariate early warning model, multivariate early warning model, Logistic linear regression model, neural network analysis model and other methods [5].

Reference [6] put forward that a single financial index should be used as the judgment basis of financial risk early warning. By comparing and analyzing the results of a single financial index of enterprises with financial risks, it finally found that the two financial indexes, return on net assets and property rights ratio, had the best effect on financial early warning, and put forward a univariate early warning model. Reference [7] added indicators such as cash flow debt ratio and asset-liability ratio to the model research. Then, reference [8] puts forward new suggestions for improving financial indicators, which contributes to the early warning model of financial indicators. In addition, reference [9] put forward in the research that the effects of asset-liability ratio, return on total assets and working capital ratio are the most effective.

With the deepening of early warning research, many scholars have found that a single index can't fully reflect the financial risks of enterprises, and the accuracy of univariate financial early warning model still can't meet the needs of enterprises. Reference [10] proposed to apply multiple financial indicators to the research of financial risk early warning model, optimized five optimal comprehensive indicators among 22 financial indicators, calculated the weight coefficient of each indicator, established the Z-value model, and achieved great achievements. The Z-value model has made great achievements in the follow-up enterprise financial risk early warning analysis. Reference [11] put forward the concept of multivariate linearity, which proved that the multivariate linear model has higher accuracy than the multivariate early warning model and is more suitable for the existing enterprise financial early warning.

The logistic regression model is developed from the idea of multivariate linearity. Through the Logistic linear regression model, reference [12] performed linear analysis in combination with the current economic environment and model characteristics, and believed that financial risk early warning should accumulate experience with the increase of research samples and quantity, and the early warning results will become more accurate. After that, scholars put forward that the combination of factor analysis and Logistic regression model can more accurately reflect the possible financial risks in financial indicators, and reduce the excessive weight caused by the repetition of index factors, which also proves that it is more accurate and scientific.

With the rapid development of artificial intelligence, with the powerful technical support of Internet big data, neural network began to be used in financial risk early warning. Reference [13] proposed to use the empirical risk minimization principle of neural network to early warning enterprises. At the same time, with the rapid development of computer technology, the prediction effect of neural network early warning model based on machine learning technology is getting better and better. Then, through the empirical analysis of past data, the

computer can quickly summarize the abnormal rules of financial risk companies' indicators and reflect them, and its accuracy is far better than that of previous models. Moreover, it can quickly adapt to a single category of samples, but it is not possible to accurately analyze the situation where the number of samples is small and the data is insufficient.

Financial indicators are the most widely used indicators in financial early warning models, and they are also indicators that objectively reflect the operating and financial status of an enterprise. Furthermore, it is easy to obtain, so as early as when the univariate early warning model was put forward, it received enough attention. In addition, the selection of financial indicators has also changed from a single indicator such as asset-liability ratio and equity ratio at the beginning to multiple indicators in parallel later, and then to the ability to classify specific financial indicators into multiple indicators later, so as to further improve the efficiency of the model [14].

With the improvement of financial risk early warning model system based on financial indicators, scholars have found that many external factors such as industry environment, national policies, competitive environment and other external factors will greatly affect the early warning results of the early warning model. Therefore, they put forward to add non-financial indicators to the financial early warning model, among which the addition of non-financial indicators such as company ownership structure, external economic indicators, market industry development status, etc. greatly improves the accuracy of the model. After that, stock fluctuation, inflation rate, equity concentration, etc. were all included in non-financial indicators, and the prediction effect of the model was further improved. It is not difficult to see that non-financial indicators play an important role in various enterprise financial early warning models, and their own early warning research value cannot be underestimated [15].

Regarding the purpose and function of financial diagnosis, by finding another way to focus on the strategic perspective, reference [16] pointed out that financial diagnosis must stand at the strategic height to play a role in the strategic development of the company. Reference [17] indicated that the purpose of financial diagnosis is to improve the company's ability to obtain operating profits, control financial risks and operational risks, and help the company better operate and manage. Reference [18] hold that financial diagnosis is a dynamic rather than a static process, strategy and financial diagnosis are intertwined, and financial diagnosis needs to focus on strategy, and strategy also needs to be matched with finance to achieve the desired effect. On the main content of financial diagnosis, reference [19] pointed out that financial diagnosis includes three major activities: operation, investment and financing. Only by starting from these three aspects and supplementing suitable methods can we evaluate and calculate the situation of enterprises. Reference [20] proposed that financial diagnosis should also include prospect diagnosis, establish an interval evaluation system, use the discounted cash flow method in financial management to evaluate prospects, and optimize the content of financial diagnosis. Reference [21] pointed out that massive data without aggregation will limit financial diagnosis, so useful information should be extracted through data mining technology, which can better ensure the accuracy of financial diagnosis results and apply it to enterprise decision-making.

In summary, the contents of the relevant work are shown in Table 1.

Table 1: Summary of related work.

Serial Number	Research Method	Shortcoming
1	Visualization of Association Relationships	Lack of adaptability
2	Single indicator research	Insufficient accuracy
3	Multi indicator research	Insufficient intelligence
4	Logistic regression model	Underfitting problem
5	Neural Network Model	Not suitable for small samples
6	Financial diagnosis	Insufficient intelligence and low accuracy

In summary, traditional financial data analysis methods have the problem of insufficient intelligence. Several common intelligent algorithms are prone to underfitting in financial data processing and require a large amount of data for training, which is not suitable for financial data analysis of small sample data. Therefore, this paper combines decision tree classification algorithm to process financial management data in universities, constructs a four-dimensional dynamic investment decision game system, improves the risk management

effect of financial data in universities, and enhances the accuracy of financial risk warning.

3 Zero-sample generation model based on cyclic invariance

Because there is serious domain offset and pivoting problems in the embedding model and the data imbalance problem in the zero-sample problem itself, based on WGAN network, this paper improves the authenticity of the generated virtual samples by increasing the cyclic consistency loss term and selecting intermediate samples for the generated samples to optimize the generated model.

3.1 Zero-sample generation model based on cyclic invariance

In this paper, n marked visible class samples are set, which have financial data characteristics $X \in ¡^{d\times n}$ and semantic description $A \in ¡^{m\times n}$ at the same time. Zero-sample learning aims to identify n_u invisible class samples $X_u \in ¡^{m\times n_u}$ that only have semantic attribute $A_u \in ¡^{d\times n_u}$ at the time of training. Y and Y_u are labels of X and X_u, respectively, and in a zero-sample environment, $Y \cap Y_u = \varnothing$ exists. It is assumed that the labels of visible class and invisible class are C and C_u respectively. In the traditional zero-sample learning, it is only necessary to correctly identify X_u in C_u, but under the generalized zero-sample condition, it is necessary to search and identify in $C \cup C_u$ space. Moreover, each semantic description a is a description of a category *y*. Formally, $\{X, A, Y\}$ and $\{A_u, Y_u\}$ are given to train the model, the goal of zero-sample learning is to learn the mapping function $f : X_u \to Y_u$, and the goal of generalized zero-sample learning is to learn the mapping function $f : \{X, X_u\} \to Y \cup Y_u$ [22].

The underlying generative model used is the WGAN model. The visible class sample $\{X, A, Y\}$, the attribute A_u of the invisible class, and the random noise $z \sim N(0,1)$ are given. The GAN generator G synthesizes false features through input class embedding a and noise z. At the same time, the GAN discriminator D takes the real financial data x and the features of the generated financial data $G(z,a)$ as inputs to distinguish whether the input features are true or false. The loss function of WGAN is as follows:

$$L_{WGAN} = E\left[D(x,a)\right] - E\left[D(\tilde{x},a)\right] - \lambda E\left[\left(\left\|\nabla_{\hat{x}} D(\hat{x},a)\right\|_2 - 1\right)^2\right] \tag{1}$$

Among them, $\tilde{x}$ is the generated false sample, $\hat{x}$ is the interpolation of x and $\tilde{x}$, and $\hat{x} = \alpha x + (1-\alpha)\tilde{x}, \alpha \in (0,1)$.

Considering the diversity of generated samples, the model will generate multiple intermediate samples for each category here. Therefore, the samples will be divided into multiple clusters by clustering, and the central sample of each cluster will be calculated respectively. $\{x_1^c, x_2^c, \mathrm{L}, x_k^c\}$ is set as k clusters of class c, and the intermediate sample is $S^c = \{S_1^c, S_2^c, \mathrm{L}, S_k^c\}$.

$$S_k^c = \frac{1}{\left|x_k^c\right|} \sum_{x_i \in x_k^c} x_i \tag{2}$$

Similarly, for the generated virtual samples $\tilde{x}_i$, intermediate samples can also be defined:

$$\tilde{S}_k^c = \frac{1}{\left|x_k^c\right|} \sum_{\tilde{x}_i \in \tilde{x}_k^c} \tilde{x}_i \tag{3}$$

In order to encourage that each generated sample should be close to at least one intermediate sample S^c, this paper introduces a regularization term to deal with a single sample and has the following form, where n_1 is the number of generated samples and k is the number of intermediate samples per class.

$$L_{R1} = \frac{1}{n_1} \sum_{i=1}^{n_1} \min_{j\in[1,k]} \left\|\tilde{x}_i - S_j^c\right\|_2^2 \tag{4}$$

At the same time, it should also be ensured that the intermediate samples of each class should be close to their real samples, so that the samples of the whole cluster are close to the real samples. Then, a regularized cluster sample is introduced, and its form is as follows. Among them, C is the total number of categories.

$$L_{R2} = \frac{1}{C} \sum_{i=1}^{C} \min_{j\in[1,k]} \left\|\tilde{S}_j - S_j^c\right\|_2^2 \tag{5}$$

At this stage, through the above two regularizations, the correlation between the generated sample and the real sample is guaranteed from the financial data feature domain. However, in this process, the relevant semantic description is used to generate the corresponding virtual samples, so the cyclic consistency loss is introduced into the model, and the correlation between the generated virtual samples and the real samples is further measured from the aspect of semantic description, so as to further improve the authenticity of the generated samples. Among them, the cyclic consistency loss converts the generated virtual samples into semantic description information by adding a regressor R after the discriminator, calculates the loss between the virtual semantic information and the real semantic information, and thus feeds it back to the generator to optimize the generation process. Its calculation form is as follows:

$$L_{cyc} = E\left(\left\|a - R(G(a,z))\right\|_2^2\right) \tag{6}$$

After the generator is trained to generate enough financial data features for the invisible class, zero-sample learning can be transformed into a supervised learning task in the traditional sense.

In addition, this model combines two softmax classifiers into a cascade classifier to perform

classification tasks. In some other zero-sample classification models, a generated virtual financial data sample is used to train a softmax classifier to correctly classify the true invisible class samples at the time of testing. The loss of the classifier is as follows:

$$L_{cls} = -E\left(log\, P\left(y|x;\theta\right)\right) \tag{7}$$

Among them, $P\left(y|x;\theta\right)$ is the probability that the financial data sample x is correctly predicted as class y. The parameter θ is obtained by training with the following formula. The parameter T is related to the task type. When it is a traditional zero-sample classification task, T is an unknown class sample used for testing. However, when it is a generalized zero-sample classification task, T is the set of unknown class samples and known class samples used for testing.

$$P\left(y|x;\theta\right) = \frac{exp\left(\theta_y^T x\right)}{\sum_{i=1}^{N} exp\left(\theta_i^T x\right)} \tag{8}$$

Before this, a softmax classifier is added to this model, which is used to evaluate the confidence of the classifier. Since the output of the softmax layer is a probability vector, the uncertainty of the measurement result can be determined by the entropy of this result. Therefore, the sample with low classification entropy can be used as a reference to classify other unseen samples. The calculation method is as follows.

$$E\left(y\right) = -\sum_{c=1}^{C} y_c\, log\, y_c \tag{9}$$

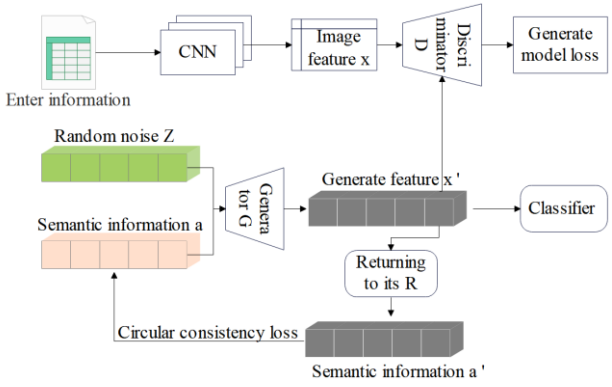

Figure 1: Model structure diagram.

The specific model structure is shown in Figure 1. The overall model loss function is as follows:

$$L = L_{WGAN} + \alpha L_{CLS} + \gamma L_{R1} + \gamma L_{R2} + \beta L_{CYC} \tag{10}$$

The evaluation index of zero-sample learning has different calculation methods under different settings.

In the case of traditional zero-sample learning, the same evaluation criterion, TOP-1 accuracy rate, is used to evaluate the accuracy of the model as the normal machine learning single-label financial data classification. However, because the number of samples in each category in the zero-sample dataset is not balanced, the average accuracy cannot be used for the overall dataset to evaluate the model. At present, in various zero-sample learning methods, the class average accuracy is usually used as the zero-sample evaluation standard. The average accuracy rate of each category is calculated first, and then the average accuracy rate of each category is calculated by finding the average value of the sum of all categories. The calculation formula is as follows

$$A_{cc} = \frac{1}{M}\sum_{i=1}^{M} Acc_i \tag{11}$$

Among them, M is the number of unseen classes and Acc_i is the classification accuracy on the i-th invisible class.

In the case of generalized zero-sample learning, the test set includes not only invisible class samples, but also some visible class samples. Therefore, this paper uses the harmonic mean accuracy proposed by Xian et al. as the evaluation index of generalized zero-sample learning, and the calculation formula is as follows

$$H = \frac{2 \times Acc_{y^s} \times Acc_{y^u}}{Acc_{y^s} + Acc_{y^u}} \tag{12}$$

Among them, Acc_{y^s} and Acc_{y^u} represent the class average accuracy of visible and invisible classes in the test set, respectively.

3.2 Attention-based FS-f-VAEGAN-D2 zero-sample learning method

In zero-sample learning, both VAE and GAN have certain defects as generators. The financial data generated by VAE model is rather fuzzy, and the expression effect of some more complex data is poor. For the GAN model, the input of the generator is random Gaussian noise, which will be difficult in the training process and the model will be difficult to converge. Therefore, this paper uses VAE-GAN model to build the system, including an encoder, a decoder/generator, and a discriminator. The VAE-GAN model diagram is shown in Figure 2.

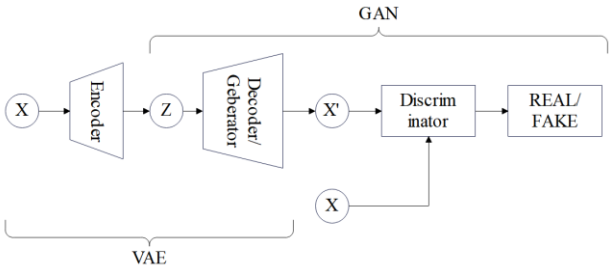

Figure 2: VAE-GAN model diagram.

One advantage of the GAN model is that its discriminator network can be measured by the similarity of financial data to distinguish them from "non-financial data". Specifically, since the reconstruction error of the

elements in the VAE model is not sufficient for the expression of financial data invariance, the VAE-GAN model replaces the VAE reconstruction (expected log-likelihood) error term with the reconstruction error expressed in the GAN discriminator. The end result is a method that combines the features of GAN as a high-quality generative model and VAE as a method to generate data encoders into latent space z.

The attention module proposed in this paper is shown in Figure 3. Firstly, financial data features are fed into a $1*1$ convolutional layer with three different weight values to obtain three attention features. After transposing one of the attention features, it is multiplied by the other attention feature softmax gets an attention map, and the calculation formula of the attention map is as follows

$$\beta = \frac{exp\left(\left(W1x_i\right)^T \times \left(W2x_i\right)\right)}{\sum_{i=1}^{N} exp\left(\left(W1x_i\right)^T \times \left(W2x_i\right)\right)} \tag{13}$$

Finally, the attention map is multiplied by the last attention feature, and then input into a $1*1$ convolutional layer again, and finally the financial data feature x' with attention is obtained. Its calculation formula is as follows

$$x' = W\left(\sum_{i=1}^{N} \beta\left(W3x_i\right)\right) \tag{14}$$

Firstly, the FS-f-VAEGAN-D2 model is briefly introduced. The model structure diagram is shown in Figure 4. In the case of inductive model, based on the VAEGAN model, VAE and GANs are combined to use a shared decoder and generator to enhance the feature generator. The definition of zero-sample learning in this paper is consistent with the previous description.

For the first VAE-GAN model, the advantages of the VAE model and the GAN model can be utilized to learn complementary information to generate features. When the target data follows a complex multi-modal distribution, VAE loss and GAN loss are able to capture different modalities of the data. It mainly trains the whole model generator and discriminator through visible class samples. Among them, the loss function of the VAE model is as follows:

$$L_{VAE}^{S} = KL\left(q\left(z|x,a\right)\|p\left(z,a\right)\right) - E_{q(z|x)}\left[logp\left(x|z,a\right)\right] \tag{15}$$

The loss function of the GAN model is as follows, where $\%$ is the generated sample of the visible class, and $\hat{x}$ is the interpolation of x and $\%$.

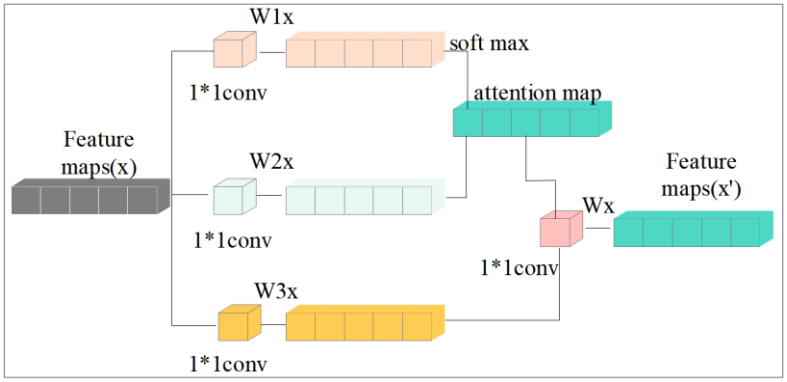

Figure 3: Attention module diagram.

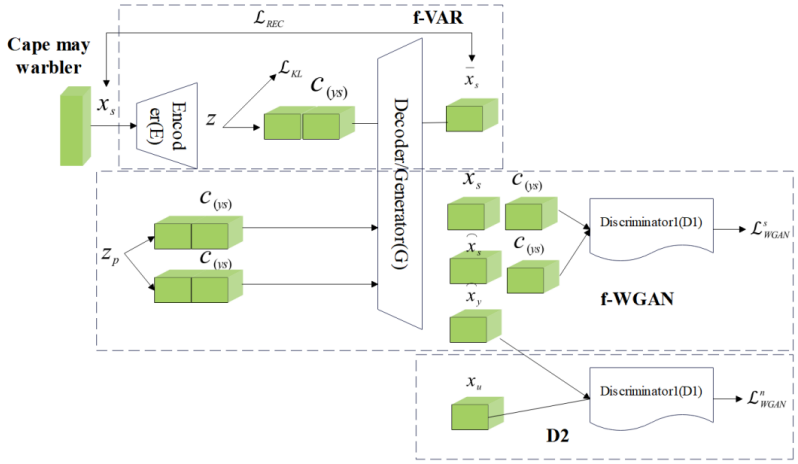

Figure 4: FS-f-VAEGAN-D2 model diagram.

$$L_{WGAN}^{S} = E\left[D(x,a)\right] - E\left[D(\%,a)\right] - \lambda E\left[\left(\left\|\nabla_{\hat{x}} D(\hat{x},a)\right\|_2 - 1\right)^2\right] \tag{16}$$

Under the inductive method, the objective function of the overall f-VAEGAN model is as follows. At this time, the model only contains one discriminator D1.

$$L_{VAEGAN}^{S} = L_{VAN}^{S} + \gamma L_{WGAN}^{S} \tag{17}$$

When the unlabeled samples of the invisible class are available at this time in the case of the direct push model, the model adds an additional discriminator D2 to distinguish the real features of the unseen class from the generated virtual features. By inputting the true unlabeled features of the unseen class into the discriminator, the manifold structure of the unseen class can be obtained, thus generating more true unseen class features. The loss of this discriminator is as follows. Among them, % is the generated sample of the visible class, and $\hat{x}$ is the interpolation of x and %.

$$L_{WGAN}^{n} = E\left[D(x,a)\right] - E\left[D(\%,a)\right] - \lambda E\left[\left(\left\|\nabla_{\hat{x}} D(\hat{x},a)\right\|_2 - 1\right)^2\right] \tag{18}$$

Under the direct deduction method, the objective function of the whole FS-f-VAEGAN-D2 is as follows:

$$\min_{G,E} \max_{D_1,D_2} L_{WGAN}^{n} + L_{WGAN}^{n} \tag{19}$$

In the modified model, the attention module mentioned herein is added to the FS-f-VAEGAN-D2 model (Figure 5). Before adding the generated visible and invisible financial data features and the real financial data features to the discriminator network, the corresponding financial data features with attention are generated by the attention module. Then, the features of financial data after selective attention are input into the discriminator to improve the discrimination ability of the discriminator. Furthermore, by inputting unlabeled samples into the attention module to generate selectively noticed financial data features, the ability to properly classify financial data features can be improved in the discriminator. This is very important in the direct inference model. If the unlabeled samples are classified into the wrong category, it will affect the subsequent generation process and the accuracy of the model will be greatly reduced. Next, this paper will analyze the performance of the model after adding attention mechanism through experiments.

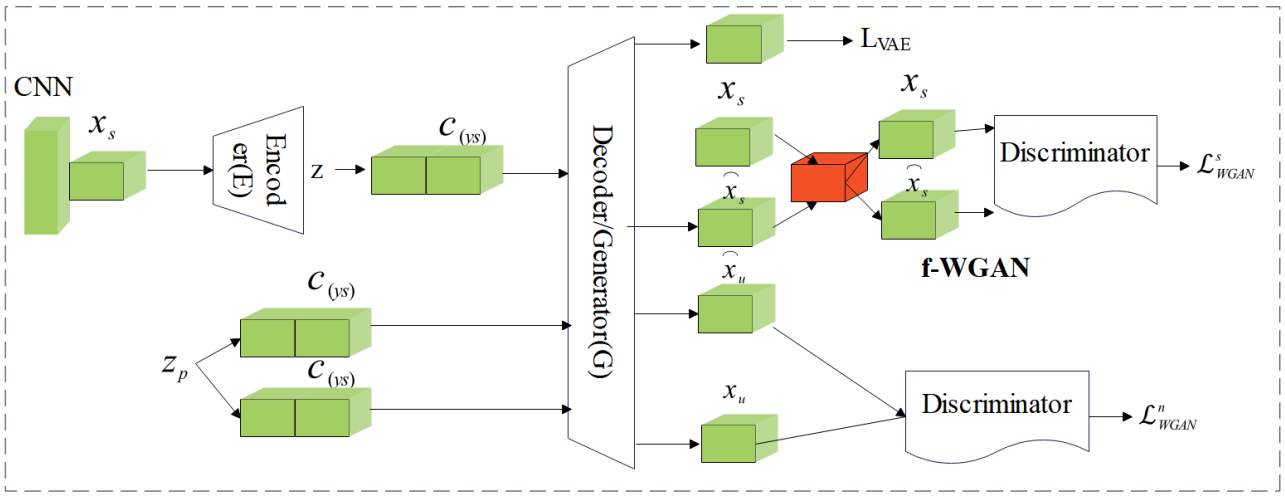

Figure 5: Schematic diagram of FS-f-VAEGAN-D2 model after adding attention mechanism.

4 System construction and test

4.1 System model

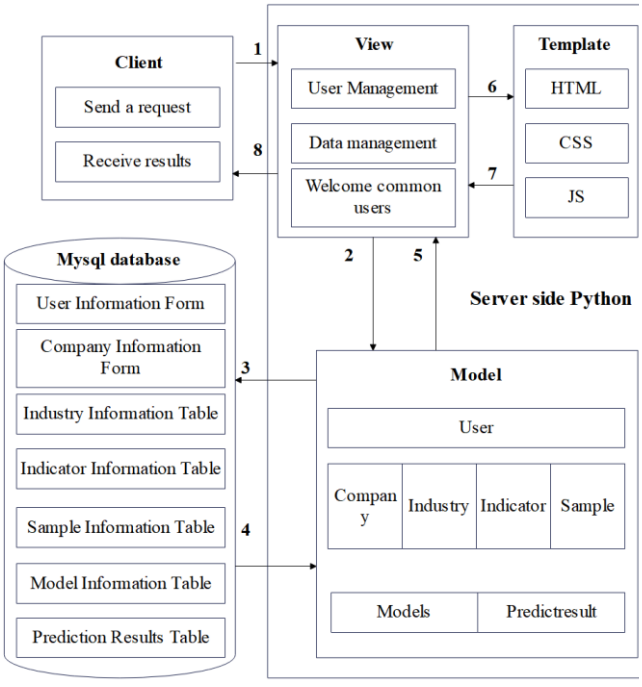

Figure 6: System architecture diagram

This system uses Python language for programming and development in PyCharm. The database adopts the overall architecture of MySQL system and is divided into three parts: client, server and database. The client and server interact through URLconf, and the server operates the database by calling the Model corresponding to the data table one-to-one. The system architecture design is shown in Figure 6. The client page display part is implemented using the Bootstrap30 framework, and the server is implemented using the Django framework.

Django is a Web framework based on MVT design pattern, in which M stands for Model, which is used to encapsulate the Model for accessing the database, V stands for View, which is responsible for receiving requests forwarded by URLconf, processing business logic, accessing the Model and returning processing results, T stands for Template, which is responsible for data display and encapsulates HTML, CSS and other files. In the server based on MVT mode, View receives the request sent by the client, calls the corresponding business logic processing method to respond, and operates the classes in the Model if it needs to access the database. Then, Model defines a class corresponding to the data table one-to-one, and operates the database by instantiating the objects of the class. Finally, Template receives the parameters passed by View, embeds them into the front-end page, and completes the data display

work.

According to the analysis of system functional requirements, this paper divides the system functions into four modules: login module, user management module, basic data management module and model management module. The basic data management module includes four sub-modules: company management, index management, industry management and sample management. The specific design is shown in Figure 7.

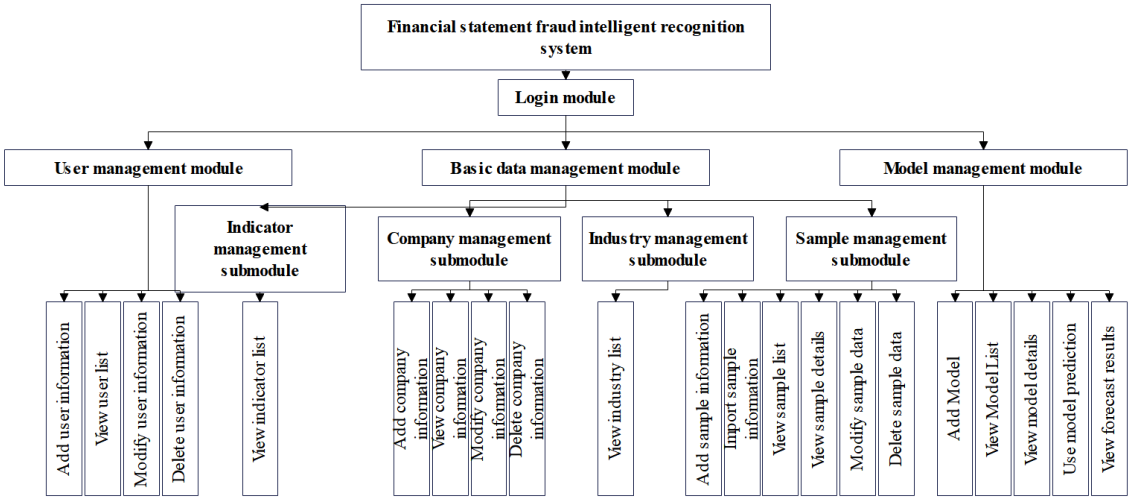

Figure 7: System functional module design.

The validation of the circular consistency term and the contribution of intermediate sample clustering to the improvement of authenticity in financial warning models is mainly achieved through a series of rigorous steps and methods.

For the validation of loop consistency items, the following steps can be followed:

(1) Clear cycle consistency indicator: In financial warning models, cycle consistency usually refers to the consistency between the model's predicted results and actual financial data. To verify this, it is necessary to first determine specific metrics for measuring consistency, such as accuracy, recall, F1 score, etc.

(2) Historical data validation: The cycle consistency of the model is verified using known historical financial data. The model's prediction results are compared with the actual historical data, and consistency indicators are calculated to evaluate the performance of the model on the real data.

(3) Sensitivity analysis: A sensitivity analysis is performed to observe the response of the model prediction results to changes in the input data. This helps to understand the stability and consistency of the model in different situations.

This paper analyzes parameters such as accuracy and recall, validates the model using an actual dataset, and discusses and analyzes it in conjunction with actual data

The data used in this paper are all from the penalty announcements published by China Securities Regulatory Commission, Stock Exchange and Securities Regulatory Bureau for fraudulent behaviors of listed companies. The fraud announcements of listed companies publicly criticized or punished from 2012 to 2023 published by CSMAR contain 16 types of violations, which are diverse and cross-cutting. The higher the complexity of the violation types, the greater the impact on the data and the empirical effect. Therefore, this paper focuses on selecting fraudulent companies with false records, fictitious profits and false assets as fraud samples. Only by comparing fraudulent companies with non-fraudulent companies can we observe the obviousness of fraudulent behaviors, and the non-fraudulent companies matching with each fraudulent company are selected as non-fraudulent samples. Then, the data of Choice Financial Terminal and Juchao Information Network are selected as the stability verification data set, and they are named Choice and cninf respectively.

In the analysis of the authenticity of financial statements, this paper balances the data set, then brings it into the random forest model, GBDT model and XGBoost model to construct the identification model, and uses the FS-f-VAEGAN-D2 (financial statements-f-VAEGAN-D2) in this paper to construct the fusion model, and tests the stability of each model to explore the optimal financial fraud identification model on this data set.

In the analysis of the predictive ability of the model, this paper uses the model proposed in this paper to conduct experiments on the data sets of SSE 50 and CSI 300, and compares them with the LSTM model, CNN-LSTM model, LSTM-Attention model, VMD-LSTM model, TCN model, BiLSTM model and TCCFR model.

4.2 Results

In this paper, the accuracy rate, recall rate, accuracy rate, F1 value and AUC value of each model on the CSMAR

test set are counted, and the statistical results are shown in Table 2.

Table 2: Comparative evaluation of models.

Model Category	Accuracy	Recall	Precision	F1-Score	AUC
RF	0.7446	0.7218	0.2893	0.4130	0.7996
GBDT	0.7422	0.7270	0.2879	0.4124	0.7953
XGBoost	0.7056	0.7041	0.2529	0.3722	0.7730
FS- f-VAEGAN-D2	0.7748	0.8034	0.2993	0.4361	0.8581

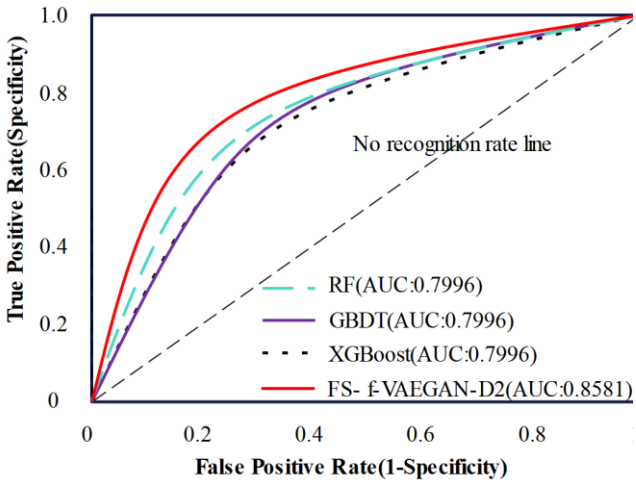

Figure 8: Comparison chart of ROC curves.

Financial forecasting models usually do not directly apply AUC curves for evaluation. The AUC curve (Area Under the Curve) is mainly used to evaluate the performance of binary classification models, especially when dealing with imbalanced datasets, which can provide stable and reasonable evaluation results. In binary classification problems, the AUC curve measures the model's ability to distinguish between positive and negative classes by showing the relationship between true case rate (TPR) and false positive case rate (FPR) at different thresholds. The ROC curve comparison chart is shown in Figure 8.

This article selects three publicly available anomaly detection table column datasets as experimental datasets, with dimensions ranging from less than 10 dimensions to hundreds of dimensions, and quantities ranging from hundreds to hundreds of thousands. The following is a specific introduction to the dataset and related experimental settings.

(1) CSMAR China Stock Market & Accounting Research Database is a research-oriented and precise database developed by Shenzhen Xishima Data Technology Co., Ltd. based on academic research needs and practical situations. There are a total of 385213 samples in CSMAR, most of which are abnormal samples, accounting for 78.36% of the total. CSMAR is a classic network intrusion detection dataset that includes four types of attacks in addition to normal data: denial of service, monitoring activities, remote unauthorized access, and local unauthorized access. Positive samples consist of data samples from four types of attacks, which are the superposition of multiple Gaussian distributions. Therefore, positive samples with complex distributions can also be used. This article will use single hot encoding to transform discrete features into continuous data that can be processed. The continuous features will be normalized and reduced to [1,1] -. The final CSMAR dataset consists of 119 features, 20000 training data, and 5000 testing data.

(2) Choice database is a financial database that provides comprehensive professional data services. It covers various financial products such as Shanghai and Shenzhen listed companies, funds, New Third Board, macro, industry, wealth management, bonds, futures, options, US stocks, Hong Kong stocks, etc., providing a variety of financial data including basic information, announcements, financial data, etc. This paper randomly samples and fills all data containing NAN and Inf to ensure that each data dimension is the same. Perform single hot encoding on discrete data columns, convert timestamp columns to relative time and normalize them, and remove useless information columns that are all single values. The final data dimension is 73 dimensions, with 40960 training data and a total of 20000 positive and negative test samples. Among them, each of the seven attack methods contains 1000 cases (if the total is less than 1000, they will all be classified as the test set).

(3) The cninf database provides corresponding programming interfaces or data processing tools for users to access and process data more conveniently This article will perform single hot encoding on discrete data columns, convert timestamp columns to relative time and normalize them, and remove useless information columns that are all single values. The final data dimension is 70 dimensions, with 35214 training data and 19525 positive and negative samples in the test data. Among them, each of the seven attack methods contains 1000 cases (if the total is less than 1000, they are all included in the test set).

In order to verify the stability of the model proposed in this paper, the original test data set is increased from one to three, and the stability of the model proposed in this paper is measured. The stability test results shown in Table 3 are obtained.

Table 3: Stability test results.

Test set	Model Category	Accuracy	Recall	Precision	F1-Score	AUC	P
CSMAR	RF	0.7446	0.7218	0.2893	0.4130	0.7996	<0.05
	GBDT	0.7422	0.7270	0.2879	0.4124	0.7953	<0.05
	XGBoost	0.7056	0.7041	0.2529	0.3722	0.7730	<0.05
	FS- f-VAEGAN-D2	0.7748	0.8034	0.2993	0.4361	0.8581	<0.05
Choice	RF	0.7419	0.7214	0.2874	0.4158	0.7968	<0.05
	GBDT	0.7390	0.7332	0.2851	0.4095	0.8007	<0.05
	XGBoost	0.7044	0.7111	0.2510	0.3725	0.7739	<0.05
	FS- f-VAEGAN-D2	0.7720	0.8090	0.2996	0.4322	0.8549	<0.05
cninf	RF	0.7423	0.7262	0.2874	0.4106	0.7939	<0.05
	GBDT	0.7458	0.7204	0.2893	0.4125	0.7876	<0.05
	XGBoost	0.6997	0.7043	0.2524	0.3741	0.7663	<0.05
	FS- f-VAEGAN-D2	0.7681	0.8026	0.2967	0.4341	0.8558	

The prediction error values of each model are shown in Table 4.

4.3 Analysis and discussion

Among the compared single integrated models, the evaluation indexes of XGBoost model are lower than those of random forest model and GBDT model, and the ability of identifying fraudulent companies and non-fraudulent companies is weak. In terms of the overall recognition accuracy of the model, the random forest model has the highest accuracy rate, reaching 74.46%, followed by the GBDT recognition rate at 74.22%, and the XGBoost model has the lowest accuracy rate, only 70.56%. In terms of recall rate, the GBDT model is slightly higher than the random forest model, with a recall rate of 72.70%%, but its accuracy rate, F1 value and AUC value are slightly lower than the random forest model. Therefore, in general, the overall performance of the random forest model is better than that of the GBDT model. To sum up, the random forest model has the best recognition performance, followed by the GBDT model, and finally the XGBoost model.

Table 4: Comparison of prediction effects of each model.

Dataset	Models	RMSE	MAE	MAPE/%	InferenceTime/ms
SSE 50	LSTM	35.442	25.622	0.907	3.477
	CNN-LSTM	40.418	29.563	1.048	2.154
	LSTM-Attention	36.163	26.748	0.947	2.323
	VMD-LSTM	39.109	30.404	1.072	11.816
	TCN	58.182	49.319	1.72	6.768
	BiLSTM	36.102	25.912	0.916	5.62
	TCCFR	34.804	24.711	0.875	4.272
	Logistic regression	38.251	29.351	1.021	9.321
	Neural networks	36.231	31.214	1.035	13.214
	FS-f-VAEGAN-D2	34.489	24.148	0.855	10.118
CSI 300	LSTM	47.387	37.968	0.982	5.904
	CNN-LSTM	51.008	40.354	1.044	3.483
	LSTM-Attention	40.727	31.972	0.831	3.843

	VMD-LSTM	49.918	41.728	1.089	18.564
	TCN	50.194	41.334	1.067	9.384
	BiLSTM	48.633	40.238	1.047	9.391
	TCCFR	40.635	31.339	0.814	7.295
	Logistic regression	39.214	30.474	1.024	9.761
	Neural networks	37.557	31.280	1.058	13.354
	FS-f-VAEGAN-D2	37.77	28.869	0.749	8.019

According to the results, in the fraud samples of the test set, the recognition rate of FS-f-VAEGAN-D2 model is 77.48%. It can be seen that FS-f-VAEGAN-D2 model has obviously improved the recognition effect of fraudulent companies, which is three percentage points higher than that of GBDT model, and has a good recognition ability for companies with fraudulent behaviors. Moreover, the area of ROC curve and coordinate axis also reached 0.817, which is higher than the area of the three single models established above, and has good overall performance.

Judging from the stability test results, the AUCs obtained by the RF model on the test sets CSMAR, Choice, and cninf are 0.7996, 0.7968, and 0.7939, respectively, and the AUCs obtained by the GBDT model on the test sets CSMAR, Choice, and cninf are 0.7953, 0.8007, 0.7876, the AUCs obtained by the XGBoost model on the test sets CSMAR, Choice, and cninf are 0.7730, 0.7739, and 0.7663, respectively, and the AUCs obtained by the FS-f-VAEGAN-D2 model on the test sets CSMAR, Choice, and cninf are 0.8581, 0.8549, and 0.8558, respectively, and the test results have little fluctuation and are basically stable. On the whole, the test method proposed in this paper has high stability.

The data in Table 4 shows that compared with existing neural network models and logistic regression models, our model has certain advantages in RMSE, MAE, MAPE, and Inference Time, which verifies that our model has certain advantages in financial early warning compared to existing models. It can be seen from Table 4 that the RMSE, MAE and MAPE values of the FS-f-VAEGAN-D2 model on the two data sets are the smallest. It shows that the prediction error of FS-f-VAEGAN-D2 model is small, and the prediction accuracy is higher than that of other models. However, the FS-f-VAEGAN-D2 model does not have advantages in terms of algorithm computational overhead. The InferenceTime of this model is 10.118 milliseconds on the SSE 50 dataset and 8.019 milliseconds on the CSI 300 dataset. The CNN-LSTM model has the shortest average time required to process a single test sample. The InferenceTime of the CNN-LSTM model is 2.154 ms on the SSE 50 dataset and 3.483 ms on the CSI 300 dataset. Although the FS-f-VAEGAN-D2 model takes more time to process a single sample on average than the CNN-LSTM model, InferenceTime is calculated in milliseconds. Compared with the millisecond time gap, the FS-f-VAEGAN-D2 model has a more significant improvement in prediction effect.

Based on the above analysis results, the FS-f-VAEGAN-D2 model proposed in this paper has good performance in the authenticity analysis of financial data and the prediction of financial data. Therefore, this paper increases data characteristics through financial indicators to improve the prediction ability. On the premise of considering historical data, prediction research is carried out on various financial indicators integrated into the company's financial statement information. Overall, the model proposed in this paper provides a reliable tool for financial data authenticity audit, and can use financial data forecasting to provide a reference for the formulation of subsequent plan policies.

5 Conclusion

Different from traditional financial analysis methods, the intelligent data analysis research method proposed in this paper mines the correlation of different dimensions of financial statement data, and presents the mining results by using the correlation visualization method to realize the risk assessment and trend prediction of enterprise financial status. Based on the basic WGAN model, the quality of generated samples is improved in the process of generating samples. Moreover, a cascade classifier is set in the classification stage to improve the classification accuracy. Through the experiments on each data set, it can be seen that the two regularizers and classifiers proposed by the model have the ability to improve the accuracy of zero-sample learning classification. According to the comprehensive experimental analysis results, it can be seen that the model proposed in this paper has good performance in the authenticity analysis and prediction of financial data. Generally speaking, the model proposed in this paper provides a reliable tool for the authenticity audit of financial data, and can provide a reference for the formulation of subsequent schemes and policies through financial data prediction.

However, the model proposed in this paper has too many regularization terms, which will affect the overall training process of the model and lead to model convergence failure in some extreme cases. Therefore, in the follow-up research, it is necessary to focus on finding a better regularization term to replace the proposed regularization term and reduce the complexity of the

model.

The current model in this article has certain limitations in convergence under extreme conditions, and actionable work will be taken in the future to address these issues. Therefore, further validation strategies need to be proposed in different financial environments to enhance the robustness of the model.

References

[1] Wasserbacher, H., & Spindler, M. (2022). Machine learning for financial forecasting, planning and analysis: recent developments and pitfalls. Digital Finance, 4(1), 63-88. https://doi.org/10.1007/s42521-021-00046-2

[2] Tang, Y., Song, Z., Zhu, Y., Yuan, H., Hou, M., Ji, J., ... & Li, J. (2022). A survey on machine learning models for financial time series forecasting. Neurocomputing, 512, 363-380. https://doi.org/10.1016/j.neucom.2022.09.003

[3] Masini, R. P., Medeiros, M. C., & Mendes, E. F. (2023). Machine learning advances for time series forecasting. Journal of Economic Surveys, 37(1), 76-111. https://doi.org/10.1111/joes.12429

[4] Wang, J., Hong, S., Dong, Y., Li, Z., & Hu, J. (2024). Predicting stock market trends using lstm networks: overcoming RNN limitations for improved financial forecasting. Journal of Computer Science and Software Applications, 4(3), 1-7. https://doi.org/10.5281/zenodo.12200708

[5] Foroni, C., Marcellino, M., & Stevanovic, D. (2022). Forecasting the Covid-19 recession and recovery: Lessons from the financial crisis. International Journal of Forecasting, 38(2), 596-612. https://doi.org/10.1016/j.ijforecast.2020.12.005

[6] Barbaglia, L., Consoli, S., & Manzan, S. (2023). Forecasting with economic news. Journal of Business & Economic Statistics, 41(3), 708-719. https://doi.org/10.1080/07350015.2022.2060988

[7] Bhat, A., Kulkarni, N., Husain, S., Yadavalli, A., Kaur, J. N., Shukla, A., ... & Seshadri, V. (2024). Speaking in terms of money: financial knowledge acquisition via speech data generation. ACM Journal on Computing and Sustainable Societies, 2(3), 1-35. https://doi.org/10.1145/3663775

[8] Ren, S. (2022). Optimization of enterprise financial management and decision-making systems based on big data. Journal of Mathematics, 2022(1), 1708506. https://doi.org/10.1155/2022/1708506

[9] Qi, Q. (2022). Analysis and forecast on the price change of shanghai stock index. Journal of Economics, Business and Management, 10(1), 72-78. https://doi.org/10.18178/joebm.2022.10.1.676

[10] Petrozziello, A., Troiano, L., Serra, A., Jordanov, I., Storti, G., Tagliaferri, R., & La Rocca, M. (2022). Deep learning for volatility forecasting in asset management. Soft Computing, 26(17), 8553-8574. https://doi.org/10.1007/s00500-022-07161-1

[11] Li, Y., & Pan, Y. (2022). A novel ensemble deep learning model for stock prediction based on stock prices and news. International Journal of Data Science and Analytics, 13(2), 139-149. https://doi.org/10.1007/s41060-021-00279-9

[12] Souto, H. G., & Moradi, A. (2023). Forecasting realized volatility through financial turbulence and neural networks. Economics and Business Review, 9(2), 133-159. https://doi.org/10.18559/ebr.2023.2.737

[13] Zhan, X., Ling, Z., Xu, Z., Guo, L., & Zhuang, S. (2024). Driving efficiency and risk management in finance through AI and RPA. Unique Endeavor in Business & Social Sciences, 3(1), 189-197. https://doi.org/10.69987/JACS.2024.40501

[14] Wei, L., Deng, Y., Huang, J., Han, C., & Jing, Z. (2022). Identification and analysis of financial technology risk factors based on textual risk disclosures. Journal of Theoretical and Applied Electronic Commerce Research, 17(2), 590-612. https://doi.org/10.3390/jtaer17020031

[15] Lei, Y., Qiaoming, H., & Tong, Z. (2023). Research on supply chain financial risk prevention based on machine learning. Computational Intelligence and Neuroscience, 2023(1), 6531154. https://doi.org/10.1155/2023/6531154

[16] Levytska, S., Pershko, L., Akimova, L., Akimov, O., Havrilenko, K., & Kucherovskii, O. (2022). A risk-oriented approach in the system of internal auditing of the subjects of financial monitoring. International Journal of Applied Economics, Finance and Accounting, 14(2), 194-206. https://doi.org/10.33094/ijaefa.v14i2.715

[17] Wang, H., & Budsaratragoon, P. (2023). Exploration of an" Internet+" grounded approach for establishing a model for evaluating financial management risks in enterprises. International Journal for Applied Information Management, 3(3), 109-117. https://doi.org/10.47738/ijaim.v3i3.58

[18] Rodríguez, C. E. L., De la Hoz Solano, V. M., & Rozo, C. A. B. (2022). Financial risks in the operation of special service transportation in the hotel sector in Bogota, Colombia. Revista Investigación, Desarrollo Educación, Servicio, Trabajo, 2(1), 63-88. https://doi.org/10.31876/idest.v2i1.32

[19] Vuletić, M., Prenzel, F., & Cucuringu, M. (2024). Fin-gan: Forecasting and classifying financial time series via generative adversarial networks. Quantitative Finance, 24(2), 175-199. https://doi.org/10.1080/14697688.2023.2299466

[20] Wu, W., Han, M., Hu, Y., Ma, J., & Zhang, X. (2024). Application of SOM-GAN based deep learning technology in the security protection of rural bank depositors' funds information. Highlights in Science, Engineering and Technology, 94, 639-646. https://doi.org/10.54097/jvmf7f78

[21] Bai, X., Zhuang, S., Xie, H., & Guo, L. (2024).

Leveraging generative artificial intelligence for financial market trading data management and prediction. Journal of Artificial Intelligence and Information, 1, 32-41. https://doi.org/10.20944/preprints202407.0084.v1

[22] Zhang, Y., Jiang, Z., Peng, C., Zhu, X., & Wang, G. (2024). Management analysis method of multivariate time series anomaly detection in financial risk assessment. Journal of Organizational and End User Computing, 36(1), 1-19. https://doi.org/10.4018/JOEUC.342094