IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE SYSTEMS, MAN AND CYBERNETICS SOCIETY SECTION

Received 22 February 2023, accepted 12 April 2023, date of publication 24 April 2023, date of current version 2 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3269438

RESEARCH ARTICLE

A Big Data-Driven Financial Auditing Method Using Convolution Neural Network

HAO ZHAO[1] AND YU WANG[2]
[1]Chinese Academy of Fiscal Sciences, Beijing 100142, China
[2]Office of Academic Research, Changjiang Polytechnic, Wuhan 430074, China

Corresponding author: Yu Wang (Z0001915@zuel.edu.cn)

This work was supported by the 2020 China University Production-University-Research Innovation Fund through the New Generation Information Technology Innovation Project.

ABSTRACT In the big data era, traditional auditing methods are facing challenge such as limited audit scope, uneven distribution of audit power, and insufficient audit analysis. To pursue high efficiency, the utilization of big data analysis technique in financial auditing has been a novel tendency in this area. The deep learning has been popular in many areas due to its high freedom degree. Thus, this paper employs a typical deep learning model convolution neural network (CNN), and proposes a big data-driven financial auditing method using CNN. Specifically, the strong ability of feature abstraction of CNN is leveraged to extract multi-level features in materials, such as visual features, textual features, etc. Then, the multi-source features from auditing materials can be well fused for final discrimination. Some simulation experiments are conducted on real-world financial auditing scenes for assessment. And the results show that the designed the proposed financial auditing method possesses relatively high auditing accuracy.

INDEX TERMS Big data, convolution neural network, financial auditing, deep learning.

I. INTRODUCTION

With the development of big data technology and the rapid growth of business data, governments and enterprises consciously try to use big data technology to improve their governance and operation levels [1]. The government hopes to improve audit efficiency and audit coverage by building an audit big data platform to collect important electronic audit data [2]. It is embodied in the full collection of various structured and unstructured data such as public funds, state-owned assets, and state-owned resources [3]. In the modern economy, the core role of finance affects the healthy development of a country to a certain extent, and the significant negative effects of financial risks will endanger the country's financial and economic security [4]. As a result, it has been more and more important to develop effective financial auditing methods for relevant administration departments, in order to improve such supervision ability [5]. Especially in era of big data, the increasing data volume has brought more challenges to human expertise. In contemporary world, it remains a promising idea to utilize artificial intelligence algorithms to realize smart auditing affairs [6].

With the diversified development of the financial market and the increasing financing needs of small, medium, and micro enterprises, many Internet financial enterprises have emerged [7]. These enterprises mainly include consumer finance companies, online micro-loan companies, and P2P platforms [8], [9]. It has played an important role in meeting the capital needs of SMEs and promoting the development of SMEs [10]. Correspondingly, with the increase in the demand for funds, the Internet credit business has also shown an explosive growth phenomenon [11]. But in the process of the rapid expansion of the Internet finance industry, many problems inevitably arise [12]. Using cloud computing and financial auditing methods, the quality and comparability of auditable work papers can be improved. In the cloud auditing model, the financial transaction data of each financial unit is stored on the cloud platform established by the cloud service provider at the same time. Auditors The design, maintenance,

The associate editor coordinating the review of this manuscript and approving it for publication was Laura Celentano.

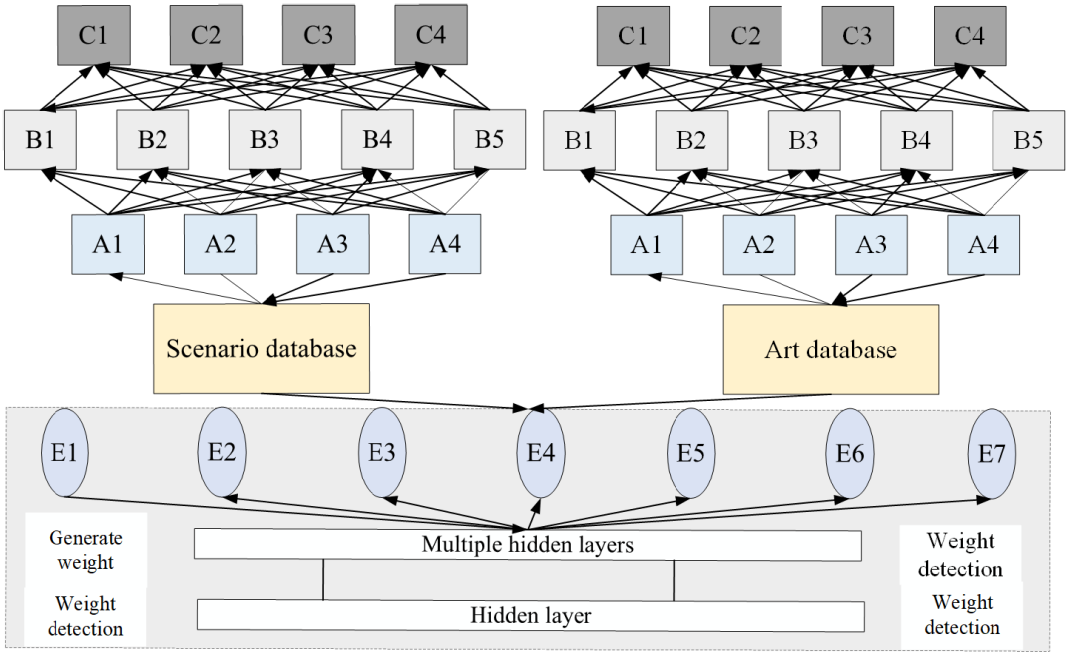

FIGURE 1. Algorithm framework of convolutional neural network-based data fusion technology.

and upgrade of the audit software and programs used are no longer undertaken by a certain developer. And technical issues can be completely handed over to professional cloud software developers, thus improving the compatibility of audit software. On the other hand, each financial unit uses the cloud platform as a user. Thus, they can obtain the advanced audit software manipulation experience, and the audit efficiency and quality can also be greatly improved. The problem left by the previous work lies in the low efficiency, as well as the large amount of data. We use a new method to geometrically reduce the amount of data applied to its efficiency to achieve the required results.

With the gradual network and virtualization development of the financial industry, big data finance has become an emerging financial model. In the era of big data, the financial industry has taken the lead in data center construction, software, and hardware system upgrades. With the continuous innovation of financial services, financial risks also arise gradually. As an important form of financial supervision, financial auditing needs to keep up with the pace of the time. The most urgent concern is to use big data technology to deal with new problems in the financial auditing process. The key points of the auditing can be quickly locked in the massive data, and the efficiency of the audit can be improved. Given this, problems can be found more efficiently, accurately, and timely. Government financial auditing is of great significance to national governance. Therefore, in this paper, a big data-driven financial auditing method using convolution neural network is proposed in this paper. Specifically, the strong ability of feature abstraction of CNN is leveraged to extract multi-level features in materials, such as visual features, textual features, etc. Then, the multi-source features from auditing materials can be well fused for final discrimination. The main contributions of our work are as follows:

- This paper introduces the development status of big data and the main big data technologies, combined with the problems faced by financial auditing in the era of big data, and analyzes the shortcomings of traditional auditing methods.
- The application of computer-aided auditing is proposed. Based on the financial auditing practice, the case analysis of the real data of the bank is carried out.
- The related technologies have reference value in the construction of an audit big data platform. This paper combines traditional auditing methods with new technologies and new ideas in the big data environment and makes a prospect for the future development of financial auditing in the era of big data.

II. RELATED WORK

Since the concept of big data was proposed, many developed countries and some developing countries have carried out active exploration and research. The theoretical aspects of big data in audit work, Chu introduced some concepts and methods of remote data auditing in the era of big data by discussing some achievements related to remote data auditing [13]. A new data structure is designed to effectively support operations such as dynamic data insertion, addition, deletion, and modification, to meet the frequent update calculation of big data, and to minimize the communication cost of data operation and auditors. In addition, through modeling analysis, it is proved that this remote data audit can provide better data security and future practical application prospects [14]. Perez elaborated on the definition of big

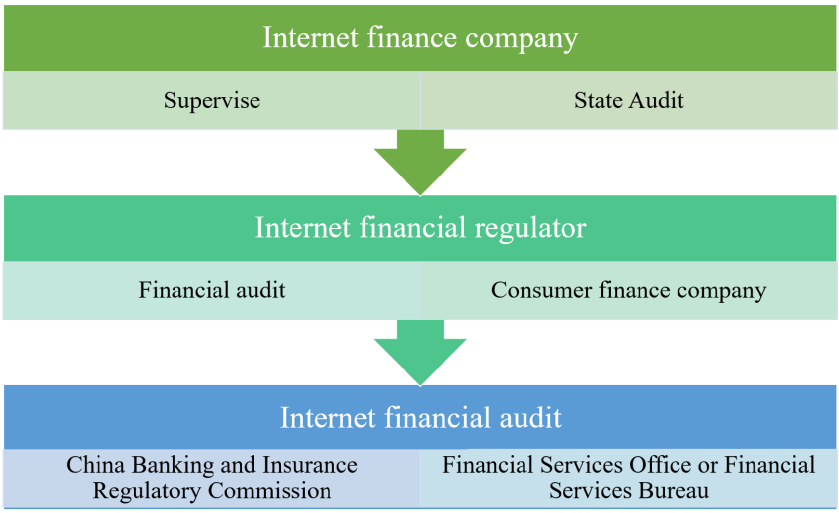

FIGURE 2. Internet financial audit supervision object diagram.

data, and introduced the impact of big data on the auditing industry in terms of improving accuracy, expanding audit scope, and improving analytical models, including improving the efficiency of financial statement auditing [15]. At the same time, it combines the characteristics and results of big data analysis with the needs of audit work, analyzes the problems encountered in the use of big data in audit work, and finally proposes that auditors should not rely too much on data relationships, because due to data sources [16]. Due to the influence of data volume and time changes, many analysis conclusions may be misled by the analysis results, so big data should be applied in combination with traditional audit methods.

Based on the development status of commercial banks, Wang analyzed that in the era of big data, the risk management of commercial banks has changed from internal control to external prevention, the data effect has changed from a loose association to a close association, and data security has changed from clear and controllable to fuzzy and difficult to control [17]. Risk characteristics, regard data and risk as the two major elements of bank operation, and propose methods to maintain the competitiveness of commercial banks by establishing a big data audit system that can predict bank risks, attaching importance to internal management process control audits, and strengthening bank risk-related audits [18]. Maintain the safe operation of the banking industry in the era of big data. Yang analyzed the problems faced by audit work under the background of big data, put forward work measures such as innovative audit organization methods, audit technical methods, and audit personnel training models in response to these difficulties, and summarized three development trends of data-based auditing in the era of big data [19]. Apply statistical methods to analyze audit data, apply big data processing technology to make audit analysis develop from query verification to mining, and use big data analysis thinking to promote audit analysis from causality analysis to correlation analysis [20].

At the same time, in the gradual development of computer-aided auditing, financial auditing also pays more attention to the application of computer software technology and popularizes the use of emerging technologies such as computers, the Internet, and automated services [21]. Therefore, how to apply new technologies and new methods in the big data environment to the practical work of financial audit, innovate in the aspects of big data collection, analysis, mining, etc., use big data technology to troubleshoot problem data, track capital flow, and determine audit focus At the same time, avoiding the risks brought by big data is an important research direction of computer-aided auditing in my country. In the advanced model used, we have optimized some parameters of the model to improve its operating efficiency. Given the business situation of my country's banking financial enterprises in the big data environment, this paper analyzes and explores the background and feasibility of the application of big data technology in financial auditing, and uses software to carry out practical case applications, to provide financial big data in the future. The audit provides ideas and methods for reference, to attract others.

III. CONVOLUTION NEURAL NETWORK-BASED DATA FUSION

According to the big data processing cycle, the big data technology system includes unstructured data collection technology, data cleaning, and screening technology, data distributed storage system, data-parallel computing analysis technology, data visualization technology, etc. In the context of big data, data sources are very wide, including mobile phones, computers, satellites, networks, media, social platforms, means of transportation, radio frequency signals, etc. However, the data of these channels usually have different formats, which is inefficient for the format conversion of large amounts of data and increases the difficulty of data collection [22]. According to statistics, in the existing big data storage system, the proportion of unstructured data and

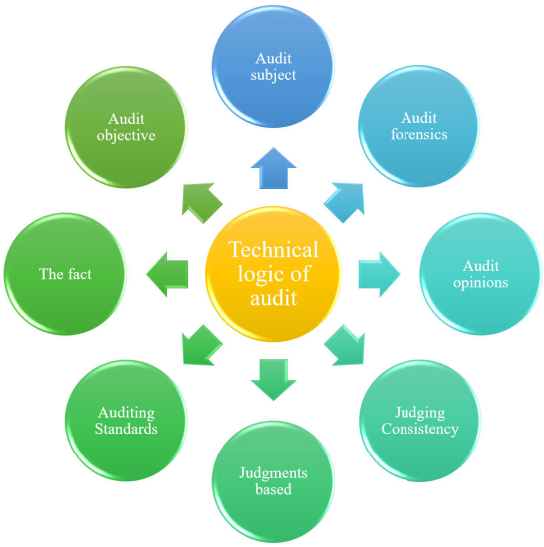

FIGURE 3. Technical logic of audit.

semi-structured data has accounted for about 80%. Therefore, traditional data collection tools are currently unable to meet the needs.

After big data collection, simple data preprocessing is required, including cleaning technology and screening technology. These two big data technologies aim to completely clean up a large amount of corrupted, redundant, and useless data in the network, optimize multi-source data and multi-modal data, integrate various types of data collected, convert high-quality data into usable information, and Extract valid information for subsequent analysis. Therefore, data cleaning and screening technology can control the quality of data from different sources and provide basic technical support for data analysis. Unstructured data is stored in distributed file systems, so in the era of big data, distributed storage systems are very important. Most traditional data storage systems adopt a centralized approach, and all data are stored on a separate server [23]. The reliability and security of the storage server are the bottlenecks for improving system performance, and cannot meet the needs of big data storage applications. Input of the proposed model includes all the initial materials that need to be audited, such as instruments, accounting records, etc. The initial contents will be transformed into digital features that can be calculated inside neural network models. Output of the proposed model is the discriminative auditing result. Similar to general machine learning models, the proposal is also a typical peer-to-peer structure from features to results. The system structure adopted by the distributed data storage system is scalable, with multiple servers sharing the storage load pressure. The method of using the location server to locate and store information can not only improve the security, reliability, availability, and access efficiency of the entire system, but also can also take advantage of the extensible features.

Data visualization technology uses forms, images, colors, animations, and other forms to visually explain data information, which can clearly and effectively convey data information. With the development of cloud computing and big data, data visualization technology is no longer satisfied with using traditional data visualization tools to extract data from the database, and summarize and simply demonstrate it. New data visualization products must meet the big data needs of the Internet explosion, quickly collect, filter, analyze, summarize, and present the information required by decision-makers, and update them in real-time according to new data. In the era of big data, in the face of massive data information, the use of data visualization technology can display the results of data analysis more intuitively and conveniently, and further help researchers analyze and my big data. Entity extraction is a subtask of named entity recognition in information extraction techniques. Extract meaningful noun phrases from the acquired data, effectively identify word boundaries, and output valuable structured knowledge. The accuracy of the steps directly affects the quality of the constructed map. Named entity recognition has many technical means [24]. In this paper, the main structural data sources for entity extraction are supported, and the extraction is carried out utilizing regularization, which realizes the data support for the expected functions, as shown in Figure 1.

The data center system is the core configuration of the audit big data platform. This system realizes the acquisition, storage, management, and simple query application of all information resources of the platform, provides data support for the application analysis system layer and provides analysis and modeling for large data sets. The data center system mainly includes functions such as information resource planning and management, data collection and processing, data authority management, and data analysis applications. The introduction of big data technology in the audit data center is to satisfy the full amount of submitted data and increase the industry coverage of data-based auditing [25].

Finally, an organizational system of structured data in data resources is formed to provide big data computing support capabilities for data analysis and processing. To solve the storage problem of massive data and achieve full coverage of audit data, the introduction of distributed data warehouse technology is an important solution. An efficient SQL-on-Hadoop is an effective means. By comparing several distributed computing engine technologies, it is believed that HAWQ has significant advantages in many aspects, and is more suitable for auditing structured data analysis and modeling. The composition, structure, and training methods of these models are different, but they are all artificial neural networks composed of artificial neurons. Except for ELM, most of the other ANN models have a deep network structure, which can extract the deep features of the data. Compared with the LR and LDA models, a typical feature of the ANN model is that it is a nonlinear model.

The input layer is used to receive input sample features, so the number of neurons in the input layer is determined according to the feature dimension of the input samples. After

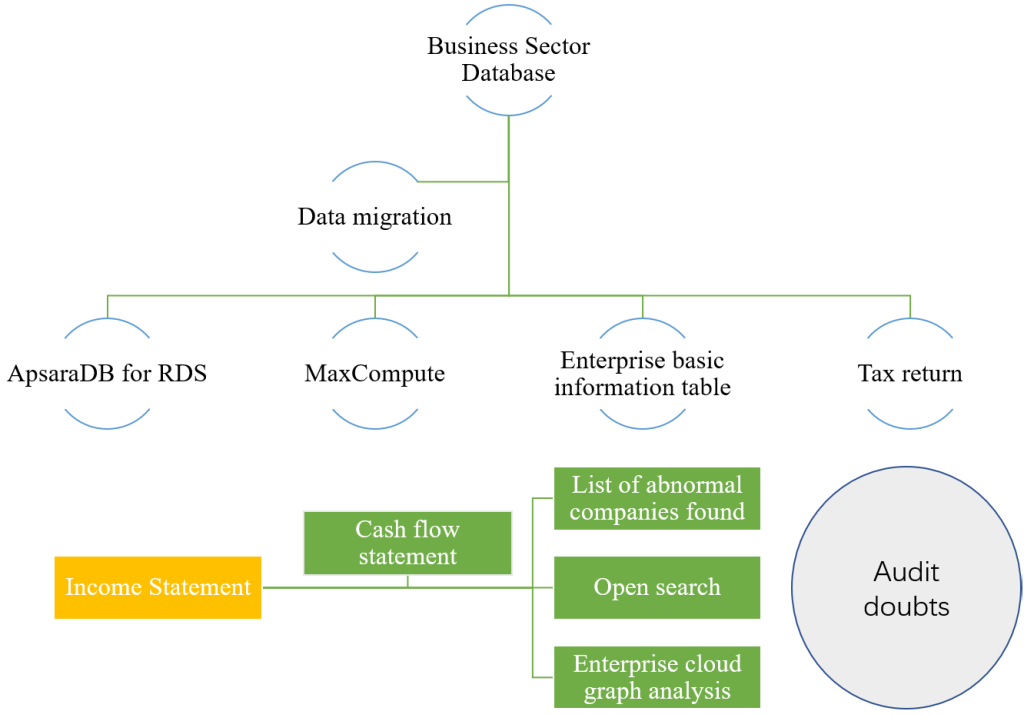

FIGURE 4. Flow chart of audit method.

FIGURE 5. Algorithm performance results.

the input layer processes the sample features, the processing result is passed to the hidden layer. The hidden layer passes the result to the final output according to the information passed by the input layer and the connection weights and bias parameters between the neuron layers. The output layer gets the final processing result. In the field of personal credit scoring, the output layer is generally set to 2 neurons, because personal credit scoring is a binary classification problem. After the above analysis, in the ANN model, the connection weights and biases between the artificial neurons have a significant impact on the output of the model. When building an ANN model, the number of neurons in each layer can be determined, but the connection weights and bias sizes need to be adjusted by iterative training of the model.

It can be seen from the sampling process of the algorithm that the ADASYN algorithm considers the distribution of minority samples [26]. The algorithm determines the samples that each sample needs to generate according to the location of each minority sample and the number of samples that need to be generated. However, when nearest neighbors are almost all samples of the majority class, the algorithm will synthesize more samples near, which may increase the difficulty of classification.

$$t_i = \Delta_i / h, i \in [1, m] \tag{1}$$

$$t_i' = t_i / \sum_{i=1}^{m} t_i^2 \tag{2}$$

It can be seen from the above sampling process that the Borderline-SMOTE sampling algorithm mainly solves the problem of which minority class samples should be expanded. However, augmenting the minority class samples near the boundary is likely to change the original classification boundary of the dataset, which may increase the difficulty of the classifier.

$$\phi(x, M, N, \gamma) = \sum_{i=1}^{M} e^{\left(\frac{M_i + x}{\gamma}\right)^2} + \sum_{j=1}^{N} e^{\left(\frac{N_i - x}{\gamma}\right)^2} \tag{3}$$

It can be seen from the above sampling process that each time a new minority class sample is generated by the RBO algorithm, the potential energy relationship between the sample and the entire dataset will be calculated to decide whether to discard it. Since the RBO algorithm uses an iterative method to generate samples, when the scale of the data set is large, the algorithm's demand for computing resources will also increase, so more computing resources are required.

$$P_x = a_0 - a_1 x_1 - a_2 x_2 - a_3 x_3 -, \ldots, -a_n x_n \tag{4}$$

When a large file is uploaded, many fragmented data will be generated, and the fragmented data information needs to be persisted through the database. One is to check the integrity of the shards when merging files; the other is to check the file upload progress through the shard information when resuming the upload is required. However, when there are many simultaneous upload tasks, there will be many sharded data read and write database tasks. The data comes from Hummingbird Data, which is an open-source financial database that aggregates 10,000+ time series from mainstream financial markets and provides high-quality free data. As the access frequency increases, the database often falls into a performance bottleneck, which eventually becomes a concurrency bottleneck for file uploading.

For this problem, there are generally three ways to solve it. One is to reduce the number of concurrent clients and reduce data collection tasks at the same time, and the other is to adjust the size of the fragmented file. Third, a high-performance cache database is used to complete the persistence of fragmentation information. After testing and verification, the shard size is set to 10 MB, and Redis is used instead of the relational database to record shard information, as shown in Figure 2. Although the classic principal-agent theory is initially applicable to discussing cash holding behavior, it is also applicable to studying the behavior of enterprises using monetary funds to allocate financial assets. The investment decisions of the management of the enterprise will be affected by the uncertainty of economic policies. Since the shareholders of the company transfer the management rights to the management and do not participate in the management of the enterprise, the management will have greater discretion over investment behaviors such as financial asset allocation.

In response to changes in the external environment, the management may not fully follow the goal of maximizing shareholder value and allocating high-risk, high-yield financial assets by changing the motivation of financial asset allocation to ensure the maximization of the management's interests. To understand the scientific and rationality of corporate management's decision-making, shareholders have created a series of systems such as an information disclosure system, internal control system, and company investor relationship management, to alleviate the principal-agent problem between the two [27]. At the same time, as an external supervision and governance factor, auditing can effectively transfer information between the company's shareholders and management, and further, alleviate the principal-agent problem by involving a third party in supervision. Online small and micro loan companies and P2P platforms have reached several thousand. The audit investigation object cannot cover all Internet financial companies. Except for licensed consumer finance companies, the other two types of companies use sample surveys.

IV. DESIGN OF FINANCIAL AUDITING MODEL

Auditing is based on a mobile Internet-based auditing vocational education platform, which provides audit practitioners with a platform for online discussion and learning of auditing technologies and methods in the new era, the audit technology and methods can keep pace with the times and keep up with the pace of informatization and claudication. Auditing builds an online and offline interactive audit communication platform, applies the latest Internet and cloud technology platforms, forms audit training big data, and finally achieves the goal of comprehensively serving the professionalization of auditing. Electronic data auditing is a new type of auditing method proposed in the era of cloud computing. It takes the fast processing and computing methods of cloud computing as the premise and takes electronic data as the auditing carrier. Its technical attribute is verification, not mining. It analyzes electronic data according to audit objectives and analyzes data to serve audit objectives, rather than excavation-type data analysis conducted away from audit objectives [28]. The main audit process is still divided into three stages: audit preparation, audit implementation, and audit report.

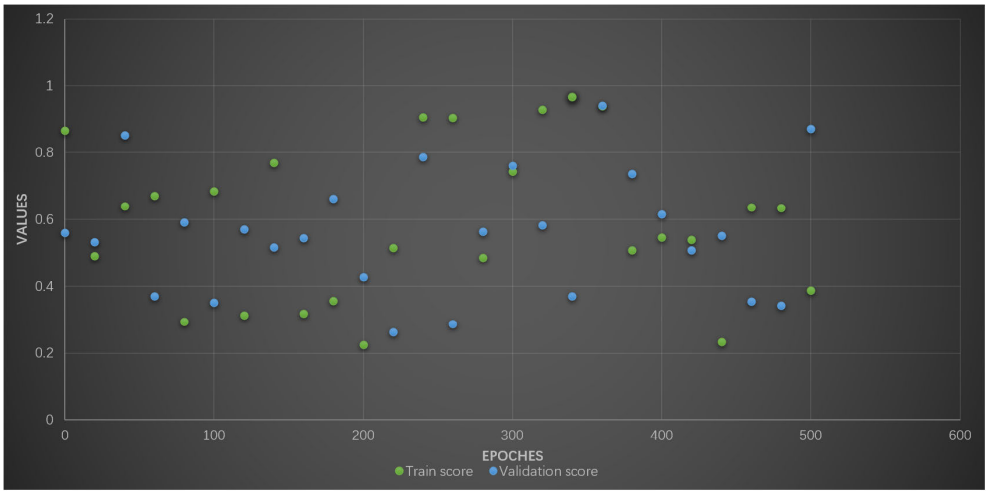

FIGURE 6. ELM model validation curve.

The work of the audit planning phase is the implementation of an on-site audit plan for the audit matter. As is shown in Figure 3, the work in the audit report stage includes two aspects: making audit conclusions with reasonable assurance and analyzing and determining the final audit opinion type. The big data anti-money laundering audit system is suitable for auditing institutions to conduct compliance audits on the business data of large state-owned and commercial banks. With big data correlation analysis as the core technology, the system achieves efficient and accurate minute-level processing capability, breaking through the current limit of hourly-level processing capability in foreign countries [29]. Analyze and evaluate the efficiency effects of other special services, such as the new credit evaluation system, related customer discovery, and product recommendation. Industry experts said that the system represents a new generation of anti-money laundering audit technology in the future. Compared with the hourly ''rule chain technology'' widely used abroad, it is revolutionary progress and has a high application value.

The audit implementation system is an information system in which auditors use computers as the basic hardware to carry out audit projects, and is subdivided into two parts: an on-site audit implementation system and an online audit implementation system according to different implementation methods. The on-site audit implementation system is a fully functional and easy-to-use information system developed to meet the requirements of government auditors to conduct on-site audits on audit objects. The online audit implementation system is the basic information system software used by audit institutions to implement online audits. In the standardized management of the audit unit after the audit, the method of dynamic and remote audit is adopted to achieve the benefits of pre, in, and post-audit audits. It can also conduct trend analysis and forecast evaluation on the past audit data combined with the current data, and propose corresponding measures. The audit opinion is shown in Figure 4.

Since few companies in the financial industry use the ''private cloud'' of the cloud service platform to store their financial data, auditors seldom can directly obtain the audit information of the audited entity from the cloud service platform when conducting financial audits. The three-party supervision departments first obtain the necessary audit information and then import the data into the cloud database through data migration. However, at present, most government auditors use general-purpose collection software, and their open interfaces have not been redeveloped according to actual work [30]. The built-in functions and models may not satisfy all types of data projects, so the software has a secondary development interface. However, in actual work, auditors are a bit conservative when using the software, and are accustomed to using the same set of templates to process all data. Audit work results in low efficiency and effectiveness of audit work. And in the data collection and migration stage, the auditors lack necessary data cleaning and verification steps, which makes the collection and migration data incomplete. It is common for personnel to only import the collected data into the audit software without logical judgment, which often leads to the failure of the project to progress.

V. ALGORITHM PERFORMANCE RESULTS

Since the NCR sampling algorithm does not set the number of redundant samples to be removed, the redundant samples to be removed are only related to the selected samples and their neighboring samples, so the cleaning effect is relatively limited. After the German dataset is down-sampled by NCR, there are still 476 samples in the majority class. In the Default data set after RU and RBU down-sampling processing, the number of samples in the two categories is the same, both of which are 6636 cases. Like German, after NCR processing, there are still 13,214 samples in the majority class of the Default dataset. The experimental results show that the ability of NCR to clean up redundant samples is not as good as the other two algorithms.

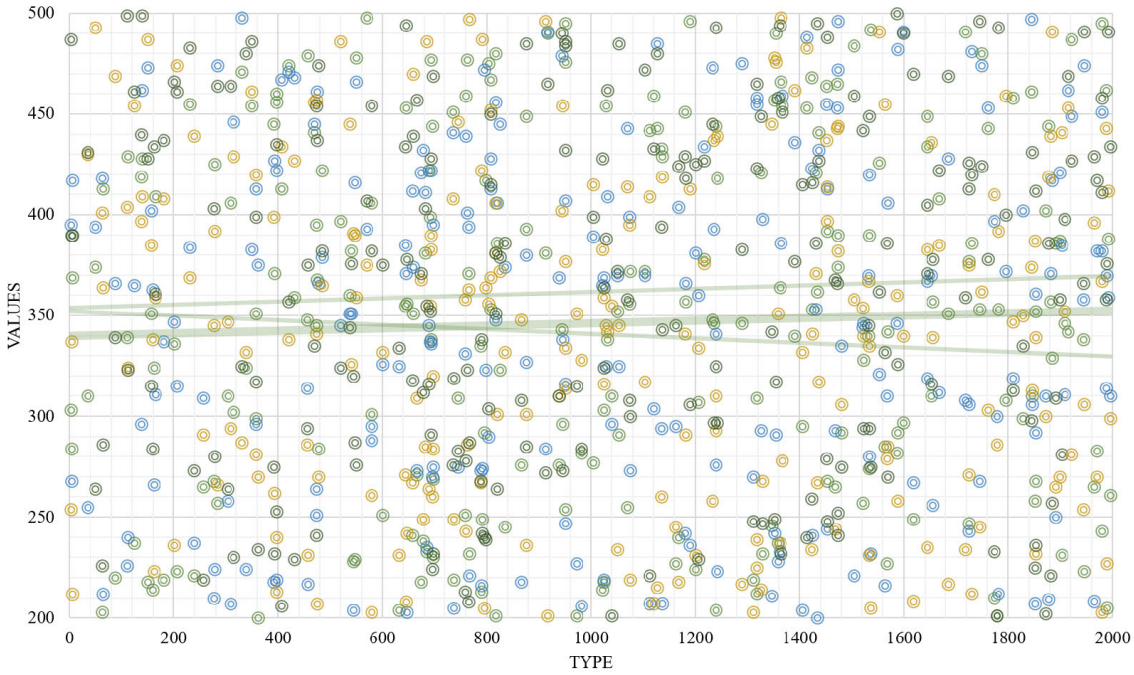

FIGURE 7. Scatter plot of loans under all loan accounts.

After being processed by SMOTETomek, there are 626 samples of both types in the German dataset. Not only that, but the two types of samples in the Default dataset are also the same, with 22706 cases. Using Kmeans SMOTE-RBU to process the German data set, the two types of samples obtained are both valued as 500. The default data set after processing has the same number of two types of samples, which is 15000. Compared with the original data set, the number of samples has not changed, which avoids the impact of changes in the overall number of samples on the classification results. After sampling and normalizing the dataset, a classifier is used to measure the performance of each sampling algorithm, as shown in Figure 5.

KmeansSMOTE-RBU sampling algorithm designed in this paper has the highest Accuracy index and outperforms other sampling algorithms on all five classifiers. On the five classifiers, the highest classification accuracy achieved was 0.8211. After sampling processing, the gap between the four evaluation indicators has narrowed significantly. Compared with None, the results of Kmeans SMOTE-RBU have obvious advantages. The Accuracy index has increased from 0.7060 to 0.8051, and the value has increased by 0.0991, which is 14% higher than that of None. The value of the F-measure has increased by 0.1, and the improvement rate exceeded 15%. The value of G-mean increased by 0.24, which shows a relatively obvious improvement. The AUC index value is also increased by nearly 0.2. This section mainly analyzes the above two experimental results. The performance evaluation indicators used in this experiment are still four indicators: Accuracy, F-measure, G-mean, and AUC. First, to set the appropriate number of ELM hidden layer nodes, this paper draws the verification curve of the ELM model, and the change of the curve is shown in Figure 6.

The above figure contains 4 small figures, each small figure represents the change in a data set. The horizontal axis in the figure is the number of neurons in the ELM hidden layer. The vertical axis represents the classification accuracy. As can be seen from Figure 6, on the Australian, Japanese, and German datasets, as the number of nodes increases, the training accuracy of the ELM model continues to increase, but the validation accuracy does not. On these three datasets, when the number of hidden layer nodes increases to around 50, the validation accuracy no longer increases with the number of nodes. On the Default data set, when the number of nodes increases to more than 50, the increase in the verification accuracy is not obvious. Based on the above four data sets, this paper sets the number of hidden layer nodes of the ELM network to 50.

After being combined with the 1DCNN network, the classification performance of other basic models has also been improved to varying degrees. Among these basic models, the performance improvement of the NB model is the most obvious. Its Accuracy index increased from 0.8478 to 0.9058, which receives an increase of 6.84%. The F-measure increased from 0.8476 to 0.9064, which receives an increase of 6.93%. The G-mean (0.9113) and the AUC (0.9120) index are increased even more, reaching 8.3%. In addition to the NB model, the four indicators of the SVM model are increased by 1.64% (Accuracy), 1.63% (F-measure), 1.33% (G-mean), and 1.08% (AUC). The increases of the indicators are 2.63% (Accuracy), 2.48% (F-measure), 1.53% (G-mean) and 1.61% (AUC). The increase of the XGBoost model is

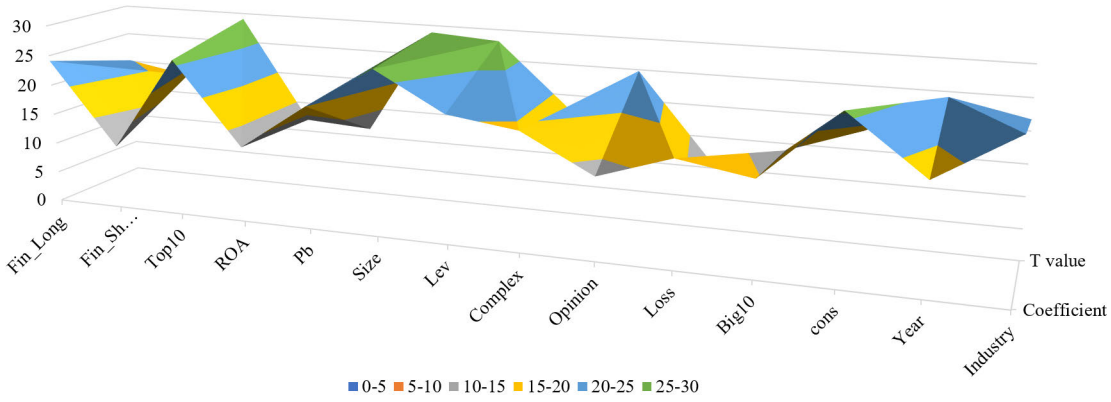

FIGURE 8. Financial asset allocation, economic policy uncertainty, and audit fees with different maturities.

4.20% (Accuracy), 4.10% (F-measure), 3.17% (G-mean) and 2.94% (AUC), which are higher than the SVM and Adaboos models. In addition, the other basic models have an increase of 1% within.

After combining with the 1DCNN network, the classification performance of LR, KNN, SVM, NB, DT, Adaboost and XGBoost have been improved to varying degrees. The four indicators of the LR model have increased by about 1%, and the KNN model has achieved a performance improvement of more than 3%. The Accuracy and F-measure indicators of the SVM model have increased by 1.7% and 1.5% respectively. The Accuracy index of the NB model has not increased significantly, but the G-mean and AUC indexes have increased by 5.97% and 6.59% respectively. Besides, the F-measure value has also increased by 1.8%, the four indicators of the DT model are increased about 1.83%, and the Adaboost model is increased about 2.7%. The four indicators of XGBoost increased by 0.85% (Accuracy), 1.2% (F-measure), 2.07% (G-mean) and 1.95% (AUC), which is not obvious.

VI. ANALYSIS OF THE APPLICATION RESULTS OF THE FINANCIAL AUDIT MODEL

The case data in this paper comes from the credit loan database of XX Bank. Considering the confidentiality of bank data, this paper only selects part of the data for analysis, including three data tables: loan sub-account, loan sub-account, and loan issuance and recovery registration book. Among them, a total of 4,961 pieces of data are collected in the loan sub-account, a total of 26,050 pieces of data are collected in the loan sub-account sub-account, and a total of 41,366 pieces of data are collected in the loan issuance and recovery register. The accuracy rate of the unsuccessful results of the model is very low, and it can be seen from the result graph whether it is successful or not. When analyzing through traditional Excel tools, in the process of processing the loan issuance and recovery registration book with a large amount of data, it will encounter problems such as slow running speed, disordered information filtering, and unclear graphic display.

When applying the R language to analyze these data, first make a macro-observation of the data, and then further process the items that need to be focused on according to the situation, and try to visualize the data results to facilitate the discovery and exploration of subsequent problems. Finally, according to some functional relationships, suspicious problems are checked, and suggestions for follow-up work are put forward in combination with traditional auditing methods. The specific code and analysis process are as follows. To analyze the loan distribution, first, extract the two variables of the loan account V2 and the loan amount V15, and use the ggplot2 package to draw a scatter diagram of the loan situation under all loan accounts, where the abscissa is the loan account number, and the ordinate is the loan amount.

Due to a large amount of selected data, the overlapping coverage between the scatter points, and the different loan conditions under different loan accounts, the scatter points in the figure have low concentration, no regularity, and poor visualization effect. However, according to the statistical results of the description in the previous step, all loan data are divided into 46 loan types. According to different loan types, all loan information is further visualized, as shown in Figure 7. Based on the previous perturbation point map, load the IDPmisc package to color the concentration of the data. From the legend on the right side of the figure, we can see that the darker the color block in the figure, the higher the repetition of the loan information. Among the loan types starting with 20, the data with the loan amount of 5 million yuan is the most concentrated, followed by 0 yuan, which needs to be checked based on the actual business situation.

The first is the positioning of small loan companies themselves. The relevant administrative measures announced by the China Banking and Insurance Regulatory Commission did not define micro-loan companies as financial institutions. Small loan companies are private lending units. After the small loan company is registered with the local industry and commerce department, it will be registered with the

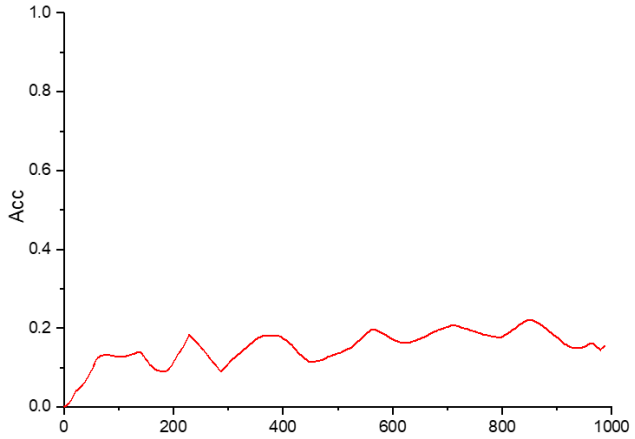

FIGURE 9. An example of an unsuccessful training result for the model.

local financial office or financial bureau. And small loan companies have not accepted the direct supervision of the China Banking and Insurance Regulatory Commission and the People's Bank of China. Therefore, small loan companies cannot enjoy the relevant preferential policies of financial institutions. And there are no precise and detailed regulations on the preferential tax policies given to small loan companies by the financial system. Similarly, when it comes to legal issues such as administrative litigation, small loan companies, and financial institutions have very different ways of handling them. Such a living environment is not conducive to the healthy development of small loan companies, as shown in Figure 8.

Secondly, the rise of the Internet financial market has intensified the fierce competition among different players in the industry. The market for small loan companies is mainly divided into two directions. One is the personal loan market and the other is the corporate loan market. But licensed consumer finance companies and P2P platforms, are aggressively vying for market share. Even traditional financial institutions such as banks are eyeing the Internet financial credit business. As shown in Figure 9, it is the result of an unsuccessful model. When the model is unsuccessful, we usually adjust the parameters in the model, such as the number of training steps and the size of each training session. As a result, both the quantity and quality of customer groups of small loan companies have been seriously affected. In addition, many small loan companies have weak risk control capabilities, limited pre-loan approval and post-loan supervision, and high overdue rates.

Although convolution operations are involved in the proposal, calculation is mainly in the format of numerical analysis. Unlike computer vision tasks, there is no too complex tensor transformation. Inside the hardware with one ''Tesla V100'' GPU, the running time for the training process is averagely within the range of 15-20 seconds. For testing, the running time for one piece of sample is certainly less than 0.1 second. Compared with typical machine learning-based methods, the time complexity is not increased. While the proposal can have better recognition performance, which can be regarded as a more practical way inside the investigated problem scenario. Under this circumstance, if an enterprise allocates short-term financial assets with strong liquidity and high liquidity, it can alleviate the financial distress caused by financing constraints to a certain extent. On the contrary, the allocation of long-term financial assets with the characteristics of poor liquidity, large capital occupation, and long time, will further increase the risk of corporate cash flow breakage and increase the inherent risk of audit risk. To reduce inspection risk, auditors need to implement more Adequate audit procedures, which in turn require higher audit premiums.

VII. CONCLUSION

In the era of big data, computer-aided auditing has received more and more attention from audit departments and auditors, especially financial auditing. The application of computer statistical analysis software has gradually deepened, and the development and application of the R language have also received increased attention. And from this, the application of computer-aided auditing is proposed. Based on the financial auditing practice, the case analysis of the real data of the bank is carried out. From the perspective of a financial auditor, in the software environment of the R language, the bank loan data has been comprehensively made. In the process of case application, the meaning of R language code is fully and comprehensively explained, the common functions and models in the R language are found in combination with audit practice, and the content of data visualization is mainly realized. At the same time, the data analysis of each step also combines traditional audit methods to think about the possible problems behind the case data and put forward suggestions for the follow-up audit work. Finally, through the code writing and visual presentation of the R language, the application of financial big data auditing is realized, and the working ideas of financial big data auditing are tentatively planned to provide certain help to auditors.

REFERENCES

[1] Z. Guo, K. Yu, A. K. Bashir, D. Zhang, Y. D. Al-Otaibi, and M. Guizani, ''Deep information fusion-driven POI scheduling for mobile social networks,'' *IEEE Netw.*, vol. 36, no. 4, pp. 210–216, Jul. 2022.

[2] Q. Li, L. Liu, Z. Guo, P. Vijayakumar, F. Taghizadeh-Hesary, and K. Yu, ''Smart assessment and forecasting framework for healthy development index in urban cities,'' *Cities*, vol. 131, Dec. 2022, Art. no. 103971.

[3] L. Yang, Y. Li, S. X. Yang, Y. Lu, T. Guo, and K. Yu, ''Generative adversarial learning for intelligent trust management in 6G wireless networks,'' *IEEE Netw.*, vol. 36, no. 4, pp. 134–140, Jul. 2022.

[4] Z. Zhou, Y. Su, J. Li, K. Yu, Q. M. Jonathan Wu, Z. Fu, and Y. Shi, ''Secret-to-image reversible transformation for generative steganography,'' *IEEE Trans. Dependable Secure Comput.*, early access, Oct. 27, 2022, doi: 10.1109/TDSC.2022.3217661.

[5] Q. Zhang, Z. Guo, Y. Zhu, P. Vijayakumar, A. Castiglione, and B. B. Gupta, ''A deep learning-based fast fake news detection model for cyber-physical social services,'' *Pattern Recognit. Lett.*, vol. 168, pp. 31–38, 2023.

[6] Z. Guo, K. Yu, N. Kumar, W. Wei, S. Mumtaz, and M. Guizani, ''Deep-distributed-learning-based POI recommendation under mobile-edge networks,'' *IEEE Internet Things J.*, vol. 10, no. 1, pp. 303–317, Jan. 2023.

[7] Z. Guo, D. Meng, C. Chakraborty, X.-R. Fan, A. Bhardwaj, and K. Yu, "Autonomous behavioral decision for vehicular agents based on cyber-physical social intelligence," *IEEE Trans. Computat. Social Syst.*, early access, Oct. 27, 2022, doi: 10.1109/TCSS.2022.3212864.
[8] E. M. Hassib, A. I. El-Desouky, L. M. Labib, and E.-S.-M. El-Kenawy, "WOA + BRNN: An imbalanced big data classification framework using whale optimization and deep neural network," *Soft Comput.*, vol. 24, no. 8, pp. 5573–5592, Apr. 2020.
[9] Y. Li, H. Ma, L. Wang, S. Mao, and G. Wang, "Optimized content caching and user association for edge computing in densely deployed heterogeneous networks," *IEEE Trans. Mobile Comput.*, vol. 21, no. 6, pp. 2130–2142, Jun. 2022.
[10] L. Zhao, Z. Bi, A. Hawbani, K. Yu, Y. Zhang, and M. Guizani, "ELITE: An intelligent digital twin-based hierarchical routing scheme for softwarized vehicular networks," *IEEE Trans. Mobile Comput.*, early access, May 31, 2022, doi: 10.1109/TMC.2022.3179254.
[11] Z. Guo, K. Yu, K. Konstantin, S. Mumtaz, W. Wei, P. Shi, and J. J. P. C. Rodrigues, "Deep collaborative intelligence-driven traffic forecasting in green internet of vehicles," *IEEE Trans. Green Commun. Netw.*, early access, Jul. 26, 2022, doi: 10.1109/TGCN.2022.3193849.
[12] L. Zhao, Z. Yin, K. Yu, X. Tang, L. Xu, Z. Guo, and P. Nehra, "A fuzzy logic-based intelligent multiattribute routing scheme for two-layered SDVNs," *IEEE Trans. Netw. Service Manage.*, vol. 19, no. 4, pp. 4189–4200, Dec. 2022.
[13] M. K. Chu and K. O. Yong, "Big data analytics for business intelligence in accounting and audit," *Open J. Social Sci.*, vol. 9, no. 9, pp. 42–52, 2021.
[14] D. Appelbaum, D. S. Showalter, T. Sun, and M. A. Vasarhelyi, "A framework for auditor data literacy: A normative position," *Accounting Horizons*, vol. 35, no. 2, pp. 5–25, Jun. 2021.
[15] L. F.-R. Pérez and A. R. Blasco, "A data science approach to cost estimation decision making-big data and machine learning: Un enfoque de ciencia de datos para la toma de decisiones en la estimación de costes-big data y aprendizaje automático," *Revista de Contabilidad-Spanish Accounting Rev.*, vol. 25, no. 1, pp. 45–57, 2022.
[16] T. Sun, "Applying deep learning to audit procedures: An illustrative framework," *Accounting Horizons*, vol. 33, no. 3, pp. 89–109, Sep. 2019.
[17] J. Wang, Y. Yang, T. Wang, R. S. Sherratt, and J. Zhang, "Big data service architecture: A survey," *J. Internet Technol.*, vol. 21, no. 2, pp. 393–405, 2020.
[18] W. Qian and Y. Ge, "The implementation of leisure tourism enterprise management system based on deep learning," *Int. J. Syst. Assurance Eng. Manage.*, vol. 12, no. 4, pp. 801–812, Aug. 2021.
[19] F. Yang and M. Wang, "A review of systematic evaluation and improvement in the big data environment," *Frontiers Eng. Manage.*, vol. 7, no. 1, pp. 27–46, Mar. 2020.
[20] Y. Li, J. Yi, H. Chen, and D. Peng, "Theory and application of artificial intelligence in financial industry," *Data Sci. Finance Econ.*, vol. 1, no. 2, pp. 96–116, 2021.
[21] T. Sun and L. J. Sales, "Predicting public procurement irregularity: An application of neural networks," *J. Emerg. Technol. Accounting*, vol. 15, no. 1, pp. 141–154, Jul. 2018.
[22] M. Yildirim, A. Çinar, and E. Cengil, "Investigation of cloud computing based big data on machine learning algorithms," *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, vol. 10, no. 2, pp. 670–682, May 2021.
[23] M. Connolly-Barker, E. Gregova, V. V. Dengov, and I. Podhorska, "Internet of Things sensing networks, deep learning-enabled smart process planning, and big data-driven innovation in cyber-physical system-based manufacturing," *Econ., Manage., Financial Markets*, vol. 15, no. 2, pp. 23–30, 2020.
[24] J. Li, Z. Ye, and C. Zhang, "Study on the interaction between big data and artificial intelligence," *Syst. Res. Behav. Sci.*, vol. 39, no. 3, pp. 641–648, May 2022.
[25] I. Vagliano et al., "Open innovation in the big data era with the MOVING platform," *IEEE MultimediaMag.*, vol. 25, no. 3, pp. 8–21, Jul. 2018.
[26] A. Praveena and B. Bharathi, "An approach to remove duplication records in healthcare dataset based on mimic deep neural network (MDNN) and chaotic whale optimization (CWO)," *Concurrent Eng.*, vol. 29, no. 1, pp. 58–67, Mar. 2021.
[27] L. M. Cristea, "Emerging IT technologies for accounting and auditing practice," *Audit Financiar*, vol. 18, no. 160, pp. 731–751, Oct. 2020.
[28] Q. Yi, M. Xu, S. Yi, and S. Xiong, "Identifying untrusted interactive behaviour in enterprise resource planning systems based on a big data pattern recognition method using behavioural analytics," *Behav. Inf. Technol.*, vol. 41, no. 5, pp. 1019–1034, Apr. 2022.
[29] K. Valaskova, P. Ward, and L. Svabova, "Deep learning-assisted smart process planning, cognitive automation, and industrial big data analytics in sustainable cyber-physical production systems," *J. Self-Governance Manage. Econ.*, vol. 9, no. 2, pp. 9–20, 2021.
[30] Z. H. Munim, M. Dushenko, V. J. Jimenez, M. H. Shakil, and M. Imset, "Big data and artificial intelligence in the maritime industry: A bibliometric review and future research directions," *Maritime Policy Manage.*, vol. 47, no. 5, pp. 577–597, Jul. 2020.

HAO ZHAO received the B.S. degree in industry and business administration from the Dongbei University of Finance and Economics, Dalian, Liaoning, in 2013, and the M.S. degree in management from the Chinese Academy of Fiscal Sciences, Beijing, in 2022. From 2014 to 2022, he was researching intelligent audit by the help of an Associate Research Fellow with the Chinese Academy of Fiscal Sciences. His research interest includes the development of audit processing through internet techniques, such as big data, cloud computing, and block chain.

YU WANG was born in Jiangsu, China, in 1982. He received the Associate degree from Taizhou Teachers College, Jiangsu, in 2005, and the master's degree from the Zhongnan University of Economics and Law, in 2013, and the Ph.D. degree from the College of Business Administration, Zhongnan University of Economics and Law, in 2017. From 2005 to 2011, he was with Taizhou Teachers College. He is currently with the Office of Academic Research, Changjiang Polytechnic. He has published five papers, two of which has been indexed by SCI. His research interests include big data in financial business and industrial organizational theory.