The Model and Method of Information Security Data Leakage Detection and Prevention

1st Hao Feng
State Grid Information&Communication Branch of Hubei Electric Power
Wuhan, China
fenghao223@outlook.com

2nd Hanlin Jiao
State Grid Information&Communication Branch of Hubei Electric Power
Wuhan, China
jiaohanlin1@outlook.com

3rd Chenyan Zhang
State Grid Information&Communication Branch of Hubei Electric Power.
Wuhan, China
zhangchenyan2@outlook.com

4th Zhikun Li
State Grid Information&Communication Branch of Hubei Electric Power
Wuhan, China
lizhikun2@outlook.com

5th Shuang Qiu
State Grid Information&Communication Branch of Hubei Electric Power
Wuhan, China
ray19892022@163.com

2024 International Conference on Data Science and Network Security (ICDSNS) | 979-8-3503-7311-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDSNS62112.2024.10691002

***Abstract*—In the rapidly developing era of informatization, data leakage has become a major threat faced by institutions. Traditional data leakage detection techniques have shortcomings in terms of real-time and adaptability. This article focuses on the shortcomings of existing algorithms and studies a new model for information security data leakage detection and prevention. This article adopts a network security monitoring network based on deep learning technology, constructs a network security monitoring network, realizes continuous monitoring and analysis of data flow, and effectively monitors and processes it, providing reliable security guarantees for network security. This article uses the method of interactive verification to test the generalization performance of the constructed model, and evaluates the stability and reliability of the model. The cross validation results show that for accuracy, the average value is 0.9, and the average recall rate is 0.87. The model for detecting and preventing information security data leaks is beneficial for timely detection and prevention of data leaks.**

Keywords—information security, data leakage detection, cross validation, model accuracy

I. Introduction

In the era of the internet, information security is a very important field. With the widespread use of data in various industries, the problem of information leakage is becoming increasingly serious. Data leakage not only affects user privacy, but also brings huge commercial benefits and reputation losses to enterprises. Therefore, studying an efficient data leakage detection and protection model and method is of great significance.

Preventive measures mainly include data encryption, access control and audit. Data encryption refers to converting confidential data into unreadable ciphertext to avoid unauthorized access. On this basis, this paper puts forward a new method, that is, accessing the system. At the same time, the mechanism can effectively prevent data leakage by detecting users' access behavior in real time and responding accordingly. To establish an effective data leakage detection and prevention mode, we must adopt various methods and establish a perfect protection system. For example, you can build a comprehensive system including data classification, encrypted storage, access control, real-time monitoring and log audit. Such a system can completely protect all data, thus ensuring the security of important data. Generally speaking, based on traditional feature matching and machine learning, the detection and prevention of information security data leakage are realized. In this process, we should pay attention to the implementation of preventive measures in order to build a complete protection system and better ensure the security of data.

This article mainly studies the models and methods for effectively detecting and preventing data leaks in information security. Firstly, this article analyzes the enormous harm caused by information leakage to individuals and businesses, and points out its urgency. Then, this article analyzes some issues in existing research and explores the detection and prevention of information security data leakage. On this basis, the method used in this article and the expected results achieved are finally verified.

II. Related Work

Several existing methods have been proven to be effective in addressing the current information security issues. For example, data leakage detection methods based on deep learning have achieved good results in real applications. Ni Huikang explored a plan for building personal information security capabilities[1]. He Yifan explored privacy leakage risk assessment methods for reversible neural networks [2]. Wu Yubao designed a ship network information security management system[3]. Yu Wenliang analyzed the detection and security of operator user information [4]. Liu Bowen analyzed the data information security guarantee technology in network communication [5]. Kayode A B studied a distributed model for preventing information leakage based on mobile agents [6]. Zhou X explored the application of simulated encryption boxes in network multimedia data security [7]. Getman A P studied the interrelationship between information security and social culture [8]. Alshurideh M studied the impact of information security on the electronic supply chain of the logistics and distribution industry in the United Arab Emirates [9]. Culot G explored information security management standards from a literature review and theoretical research perspective [10]. However, their methods all have their own shortcomings, such as weak processing ability for small sample data. To this end, we will explore how to use models for information security data leakage detection and prevention to address the aforementioned issues.

However, these models and methods also have some limitations. First of all, there are various forms of data leakage, including internal leakage, external attack, etc., each form has its own unique characteristics and rules. Therefore, it is difficult for a single model and method to cover all leakage

2024 International Conference on Data Science and Network Security (ICDSNS) | 979-8-3503-7311-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDSNS62112.2024.10691002

scenarios. Secondly, with the continuous development of technology, hacker attack methods are also constantly updated and upgraded, which requires us to constantly update and improve the technology and methods of data leak detection and prevention. In addition, data privacy protection is also a major concern. In the process of detecting and preventing data leaks, we need to ensure that sensitive data is not misused or leaked, which requires strict controls at the technical and management levels. Data security is related to social stability and economic development. Data leakage will cause serious consequences such as personal privacy and trade secret leakage, and have a significant impact on social and economic development. Studying methods for detecting and preventing data leaks can effectively reduce the adverse social and economic consequences of such accidents. The detection and prevention of data leaks is a multidisciplinary research topic, and modeling research in this field can promote innovation and development of related technologies.

III. METHOD

A. Design and Implementation of Deep Learning Models

This paper presents a new and effective solution to the problem of data leakage in information security. Nowadays, with the rapid development of information technology, data leakage events occur frequently, causing significant economic losses and reputation risks to enterprises and individuals. Therefore, establishing a set of effective data leakage detection and protection model, and conducting scientific analysis of it, is the key to ensure enterprise information security.

First, we must define the data breach detection and protection model. In this process, the establishment of the model includes a lot of data collection, processing, analysis and prediction processes. At the data collection stage, we need to determine whether the data is sensitive and what information is at risk of disclosure. This requires us to better understand the business process, data flow and data security policy. In the process of data processing, the data should be cleaned, integrated and labeled to prepare for future analysis and prediction. In terms of data analysis, machine learning, data mining and other methods are used to detect possible leakage risks through pattern recognition and association analysis. Finally, in the forecasting stage, using the previous information and the current situation, the oil spill accident that may occur in the future is forecasted and prevented.

Secondly, for the detection and protection of data leakage, we can study from different perspectives. On the one hand, information security is enhanced through technical measures such as encryption, access control and intrusion detection. In addition, it is necessary to strengthen the management of employees, through training, education and other means to enhance the security awareness of employees, improve the level of business, and reduce the risk of data leakage caused by humans. In all aspects of data collection, storage, transmission, use, etc., establish a sound security management system and procedures to ensure the legal and safe data.

Some problems should also be noted in the implementation. First, it is necessary to deeply understand the service needs and security requirements of enterprises to ensure that the detection and protection models and methods of information leakage are targeted and efficient. The second is to pay attention to the privacy of the data and prevent the leakage of sensitive information in the detection and protection links. Third, it is necessary to strengthen exchanges and cooperation with relevant units and individuals to form a joint force to deal with the risk of information leakage.

1) Step 1. Data preprocessing: We will perform preprocessing work such as data cleaning and normalization to provide support for subsequent deep learning models.

2) Step 2. Feature selection and extraction: Based on this, deep neural networks, recurrent neural networks, and other methods are used to extract keywords from text and visual features from images.

3) Step 3. Model construction: This article intends to study data leakage detection methods based on deep learning technology, including: anomaly detection models based on autoencoders; An image leakage detection model based on convolutional neural networks.

4) Step 4. Training strategy: Selecting appropriate loss functions, optimization algorithms, etc., design learning rate adjustment, data augmentation and other methods to improve the learning rate and detection performance of the model.

5) Step 5. Model Integration: Based on this, researching and integrating multiple deep learning models to improve the robustness and accuracy of the algorithm.

6) Step 6. Real time monitoring system deployment: Based on this, a deep learning based information security monitoring network is adopted to construct a network security monitoring network, providing reliable security guarantees for network security [11-12].

In information security, information entropy $H(X)$ can be used to evaluate the degree of data chaos or the strength of encryption:

$$H(X) = -\sum_{i=1}^{n} p\,(x_i) log_2 p(x_i) \qquad (1)$$

Among them, X is a random variable, $\mathrm{p}(x_i)$ is the probability that X takes a value of x_i, and n is a constant.

Conditional entropy $H(X|Y)$ is the uncertainty of random variable Xgiven another random variable Y.

$$H(X|Y) = -\sum_{y\in Y} p\,(y) \sum_{x\in X} p\,(x|y) log p(x|y) \qquad (2)$$

$p(x|y)$ is the conditional probability that X takes a value of x under the condition of Y taking a value of y, and $\mathrm{p}(\mathrm{y})$ is the probability that the random variable Y takes a value of y.

Mutual information $\mathrm{I}(\mathrm{X},\mathrm{Y})$ is a measure of the correlation between two random variables:

$$I(X,Y) = \sum_{y\in Y} \sum_{x\in X} p\,(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right) \qquad (3)$$

$\mathrm{p}(\mathrm{x},\mathrm{y})$is the joint probability of random variables X and Y taking values of x and y simultaneously, and $\mathrm{p}(\mathrm{x})$ is the probability of random variable X taking values of x.

However, although our country has made some progress in this area, many problems and challenges remain. With the popularity of big data, cloud computing and other technologies, the scale of data is getting larger and more complex, and the detection and prevention of data leaks is becoming more and more difficult. At the same time, due to the continuous update and change of the network attack mode, the technology and method of information leakage detection

and protection in the network environment need to be updated and improved.

On this basis, a new solution is proposed. The first is to increase the research and research on data leaks, and explore more effective and accurate data leak detection and protection methods. The second is to increase the training and introduction of relevant personnel, and cultivate a group of highly skilled and experienced information security teams. Third, it is necessary to strengthen international cooperation and exchanges, absorb the successful experience of countries or regions in data security, and join hands to deal with data security issues on a global scale.

Therefore, this paper presents a new method of information security analysis based on network. In order to deal with the increasingly serious data security issues, we must continue to explore, innovate, and increase technical research and personnel training. In order to build a safe, reliable and efficient information society, China should strengthen exchanges and cooperation with other countries and regions in the world.

B. *Application of Deep Learning Models in Information Security Data Leakage Detection*

Data feature extraction: Based on deep learning networks, extracting features of different formats from data such as text, images, audio, etc. Deep learning can learn high-level features from data and effectively distinguish between normal and abnormal data. The research content of abnormal behavior detection includes: anomaly detection of multi-source heterogeneous data such as network traffic and user behavior based on deep learning. By monitoring abnormal patterns and behaviors in the data flow, potential data leakage issues can be effectively detected. Model training and optimization: This article intends to use methods such as deep neural networks, recurrent neural networks, and self-coding to study deep learning models for data leakage detection. On this basis, a new neural network-based adaptive learning method was adopted.

Real time monitoring system construction: Based on this, building a deep learning based real-time monitoring system to achieve continuous monitoring and analysis of data streams, and effectively monitor and process them. This system can respond quickly to potential data leakage issues, thereby enhancing the security performance of data. Model performance evaluation and indicator calculation: By evaluating and evaluating the accuracy, recall, F1 value, and other performance of the constructed model, we can better understand the application effect of the model in leak detection and further optimize and improve the model.

IV. Results and Discussion

A. *Recall Rate Evaluation*

A statistical model was established with known leakage data, which refers to the probability of being correctly identified in all actual leakage accidents. The collected data for recall rate evaluation is shown in Table 1.

TABLE I. Recall Rate Evaluation Collected Data

Serial number	Data type	Feature 1	Feature 2	Feature 3	Prediction result	Actual result
1	Leak data	0.2	0.1	0.9	Leak	Leak
2	Leak data	0.1	0.3	0.8	Leak	Leak
3	Normal data	0.7	0.6	0.4	Normal	Normal
4	Leak data	0.3	0.2	0.7	Leak	Leak
5	Normal data	0.6	0.5	0.2	Normal	Normal
6	Leak data	0.4	0.5	0.6	Normal	Leak
7	Normal data	0.5	0.8	0.3	Normal	Normal
8	Leak data	0.2	0.4	0.5	Leak	Leak

Numbers 1, 2, 4, 6, and 8 are samples with leaked data, and the predicted results of number 6 do not match the actual results. Therefore, the recall rate of the model is 0.8.

B. *Cross Validation*

Using interactive testing methods to test the generalization performance of the constructed model, and evaluating the stability and reliability of the model.

Cross validation is a widely used model evaluation method, whose main purpose is to partition the sample set, train and test the sample set, and evaluate its generalization performance. This method can effectively reduce overfitting, underfitting and other issues, and enhance the stability and reliability of the model.

Experimental steps:

1) Step 1. Dataset partitioning: Dividing a sample set into k similar sized subsets.

2) Step 2. Training and validation: Taking a test set from each subset, and use the remaining k-1 subsets as the training set to model the training set and evaluate its performance.

3) Step 3. Performance evaluation: Recording each confirmation (such as accuracy, recall, F1 value, etc.).

4) Step 4. Summary of Results: Through statistical analysis of k confirmed results, calculating their mean and standard deviation, and evaluate their stability and reliability.

The cross validation results are shown in Figure 1. For accuracy, the average value is 0.9, and the average recall rate is 0.87.

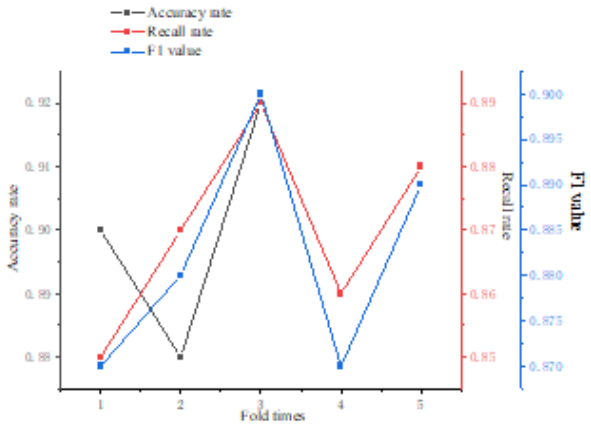

Fig. 1. Cross validation results

C. *Time Efficiency Evaluation*

Real time monitoring of the built model and evaluating its temporal validity.

1) Average prediction time/sample: It reflects the average time required for a model to predict each sampling point, and is an important indicator for evaluating the time effectiveness of the model in real-time monitoring systems. The evaluation results of time efficiency are shown in Figure 2.

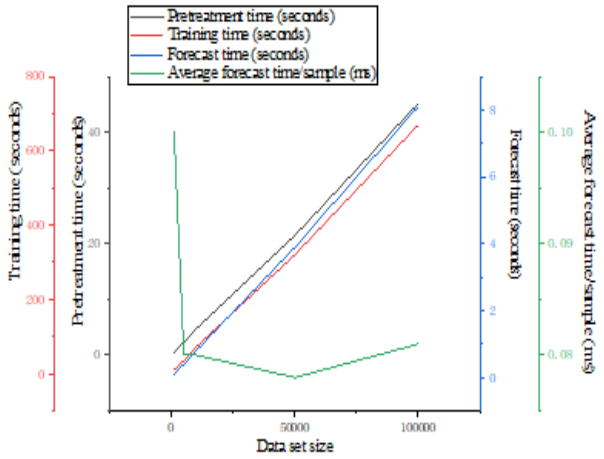

Fig. 2. Time efficiency evaluation results

As the size of the dataset increases, the preprocessing time, training time, and prediction time all increase accordingly. This is because the more data to be processed, the more computational resources are required. The dataset size is 1000, with a preprocessing time of 0.5 seconds, a training time of 10.2 seconds, and a prediction time of 0.1 seconds.

D. *Evaluation of False Alarm Rate*

The false alarm rate when modeling normal data, which is the rate at which normal data is mistaken for abnormal data, is used to evaluate the robustness and credibility of the model.

Numbers 1, 2, 4, 6, and 8 use normal data, while numbers 3, 5, and 7 use abnormal data. The normalized data statistics are shown in Figure 3. Feature 1 in sequence number 1 is 0.7, feature 2 is 0.6, and feature 3 is 0.4.

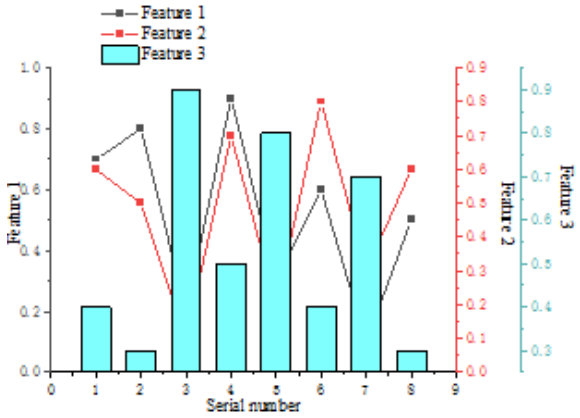

Fig. 3. Normalized data statistics

The predicted and actual results under different serial numbers are shown in Table 2. The predicted result for number 2 is abnormal, but the actual result is normal. The false alarm rate is 0.125.

TABLE II. PREDICTION AND ACTUAL RESULTS UNDER DIFFERENT SERIAL NUMBERS

Serial number	Prediction result	Actual result
1	Normal	Normal
2	Abnormal	Normal
3	Abnormal	Abnormal
4	Normal	Normal
5	Abnormal	Abnormal
6	Normal	Normal
7	Abnormal	Abnormal
8	Normal	Normal

V. CONCLUSION

In the context of informatization, the problem of data leakage is becoming increasingly prominent. Therefore, this article adopts a model for information security data leakage detection and prevention. This article intends to use methods such as deep neural networks, recurrent neural networks, and self-coding to study deep learning models for data leakage detection. On this basis, a real-time monitoring system based on deep learning is constructed to achieve continuous monitoring and analysis of data streams, and to effectively monitor and process them. This system can respond quickly to potential data leakage issues, thereby enhancing the security performance of data. Faced with the diversity and complexity of data, future modeling methods will focus more on the processing and fusion of multi-source data, including text, images, videos, and other multi-source data, in order to improve the comprehensiveness and effectiveness of detection. In the information age, data has become an important asset of enterprises and organizations, but frequent data leakage incidents have seriously affected the economy and reputation. With the development of information technology, traditional protective measures have been difficult to cope with the increasingly complex data leakage problem, and more advanced detection and prevention models and methods are urgently needed. By introducing detection methods based on machine learning and big data analysis, the accuracy and timeliness of data leakage detection can be significantly improved. These methods can automatically identify abnormal behaviors, monitor data flow in real time, and find potential leakage risks in time. In addition, the comprehensive application of data encryption, access control and audit mechanism has effectively enhanced the data protection capability. Although the new technology has made progress in detection and prevention, its computational complexity is high, and the efficiency still needs to be improved when dealing with large-scale data. In addition, relying on a large number of historical data for machine learning model training may lead to a lag in response to new leakage means. In the future, detection algorithms should be further optimized to improve processing efficiency, and new technologies such as blockchain should be combined to enhance data transparency and traceability. In addition, it is necessary to strengthen the construction of a multi-level comprehensive protection system to improve the overall data security level, so as to better cope with the evolving threat of data leakage.

AUTHOR CONTRIBUTIONS

The corresponding author is Hao Feng

REFERENCE

[1] Ni Huikang, He Fei. Research on Building Personal Information Security Capability [J]. Microcomputers and Applications, 2020, 039 (002): 19-22,33.

[2] He Yifan, Zhang Jie, Zhang Weiming, et al. Privacy leakage risk assessment of reversible neural networks [J]. Journal of Network and Information Security, 2023, 9 (4): 29-39.

[3] Wu Yubao. Ship Network Information Security Management System Based on SOA Architecture [J]. Ship Science and Technology, 2020, v.42 (22): 173-175.

[4] Singh, D., Roy, D. and Mohan, C.K., 2016. DiP-SVM: distribution preserving kernel support vector machine for big data. IEEE Transactions on Big Data, 3(1), pp.79-90.

[5] Liu Bowen. Analysis of Data Information Security Technology in Network Communication [J]. Computer Application Abstract, 2023, 39 (1): 99-101.

[6] Kayode A B , Dayo A O , Uthman A A .A Review on Distribution Model for Mobile Agent-Based Information Leakage Prevention[J]. Communication and Networking (English), 2021, 013(002):P.68-78.

[7] Zhou X, Li B , Qi Y ,et al.Mimic Encryption Box for Network Multimedia Data Security[J].Security and Communication Networks, 2020, 2020(2):1-24.

[8] Getman A P, Danilyan O G, Dzeban A P, et al. Information security in modern society: Sociocultural aspects[J]. Amazonia Investiga, 2020, 9(25): 6-14.

[9] Srinivas, Dava, I. Bhuvaneshwarri, G. P. Ramesh, Shankar Nayak Bhukya, and I. Poonguzhali. "An improved cuckoo search algorithm with deep learning approach for classifying arrhythmia based on ECG signal." Internet Technology Letters: e477.

[10] Culot G, Nassimbeni G, Podrecca M, et al. The ISO/IEC 27001 information security management standard: literature review and theory-based research agenda[J]. The TQM Journal, 2021, 33(7): 76-105.

[11] Koohang A, Anderson J, Nord J H, et al. Building an awareness-centered information security policy compliance model[J]. Industrial Management & Data Systems, 2020, 120(1): 231-247.

[12] Cardarilli, G.C., Di Nunzio, L., Fazzolari, R., Giardino, D., Matta, M., Patetta, M., Re, M. and Spanò, S., 2019. Approximated computing for low power neural networks. Telkomnika (Telecommunication Computing Electronics and Control), 17(3), pp.1236-1241.