Available Online: 10 July 2025 DOI: 10.54254/2977-3903/2025.25182

Defense strategies against data poisoning attacks in AI financial risk control models

Yina Liu

Chengyi College, Jimei University, Xiamen, China

764520494@qq.com

Abstract. Against the backdrop of rapid fintech development, Artificial Intelligence (abbreviated as AI) financial risk control models have been widely applied in financial risk assessment and management due to their efficiency and accuracy. However, data poisoning attacks, as a malicious means targeting model training data, severely threaten the reliability and security of these models. From a professional and technical perspective, this paper deeply analyzes the principles of AI financial risk control models and data poisoning attacks, systematically sorts out the existing problems in the current response process, including incomplete data source verification mechanisms, insufficient abnormal data identification capabilities, to be improved model robustness, imperfect dynamic defense systems, and lagging attack traceability technologies. Aiming at these issues, specific response strategies are proposed, such as constructing a multi-dimensional data verification system, strengthening abnormal data detection algorithms, optimizing model architecture design, establishing dynamic monitoring and response mechanisms, and enhancing attack traceability technology, with the aim of providing theoretical and practical references for ensuring the safe and stable operation of AI financial risk control models.

Keywords: AI financial risk control model, data poisoning attack, response strategy, data verification, model robustness

1. Introduction

With the continuous deepening of digital transformation in the financial industry, the application of artificial intelligence technology in the field of financial risk control has become increasingly widespread [1]. AI financial risk control models can quickly and accurately assess customers' credit risks, fraud risks, etc., by learning and analyzing large amounts of historical data, providing strong support for decision-making such as credit approval and risk control in financial institutions. However, with technological development, attack methods against AI financial risk control models have become more diverse, and data poisoning attack is one of the most threatening attack methods. Data poisoning attacks mislead the model training process by injecting malicious data into the training data, causing the model to make wrong decisions in practical applications, thus bringing huge losses to financial institutions [2]. Therefore, in-depth research on response strategies for data poisoning attacks on AI financial risk control models is of great theoretical and practical significance.

2. Principles of AI financial risk control models and data poisoning attacks

2.1. Overview of financial risk control models

Financial risk control models are important tools for financial institutions to assess and manage risks, with their core being the establishment of mathematical models through the analysis and processing of large amounts of data to predict and evaluate risks. Before the application of AI technology, financial risk control models were mainly based on traditional statistical methods such as logistic regression and decision trees. These models have advantages such as simple structure and strong interpretability, but have certain limitations in handling complex data and non-linear relationships. With the development of AI technology, technologies such as deep learning and machine learning have been introduced into the field of financial risk control, forming AI financial risk control models. AI financial risk control models can automatically learn features and patterns from large amounts of data, have stronger non-linear fitting capabilities and generalization abilities, and can more accurately assess risks [3]. Common

AI financial risk control models include neural network models, random forest models, support vector machine models, etc. These models play important roles in credit evaluation, fraud detection, market risk prediction, and other aspects.

2.2. Principle analysis of data poisoning attacks

Data poisoning attack refers to an attacker injecting malicious data into the model training data to change the distribution and characteristics of the data, thereby misleading the model training process and causing the model to make wrong decisions in the testing and application stages. The principle of data poisoning attacks is mainly based on the model's learning mechanism, that is, the model establishes the mapping relationship between input and output through learning from training data. If there is malicious data in the training data, the model will learn the features of these malicious data into the model, thereby affecting the model's prediction results. Data poisoning attacks can be divided into two types: polluting training data and polluting test data [4]. In the field of financial risk control, data poisoning attacks mainly target training data because the model training process is a long-term process, and attackers can gradually change the model's parameters and decision boundaries by injecting malicious data for a long time. Attackers usually design specific malicious data according to the characteristics of the target model and attack objectives, such as forging false customer information, transaction records, etc., so that these malicious data have certain representativeness and influence in the training data, thereby achieving the purpose of misleading the model.

2.3. Impact of data poisoning attacks on financial risk control models

The impact of data poisoning attacks on AI financial risk control models is mainly reflected in the accuracy, reliability, and security of the models (see Figure 1). First, data poisoning attacks will cause the distribution of training data to shift, making the model learn wrong features and patterns during the training process, thereby reducing the accuracy of the model. For example, in a credit evaluation model, if an attacker injects a large number of false high-credit customer data, the model will overestimate the credit level of customers, leading financial institutions to issue too many high-risk loans. Second, data poisoning attacks will affect the reliability of the model, making the model perform unstable when facing new data, and prone to misjudgments and omissions. Finally, data poisoning attacks also pose a threat to the security of the model. Attackers can obtain sensitive information of the model through data poisoning attacks or use the model's vulnerabilities for further attacks.

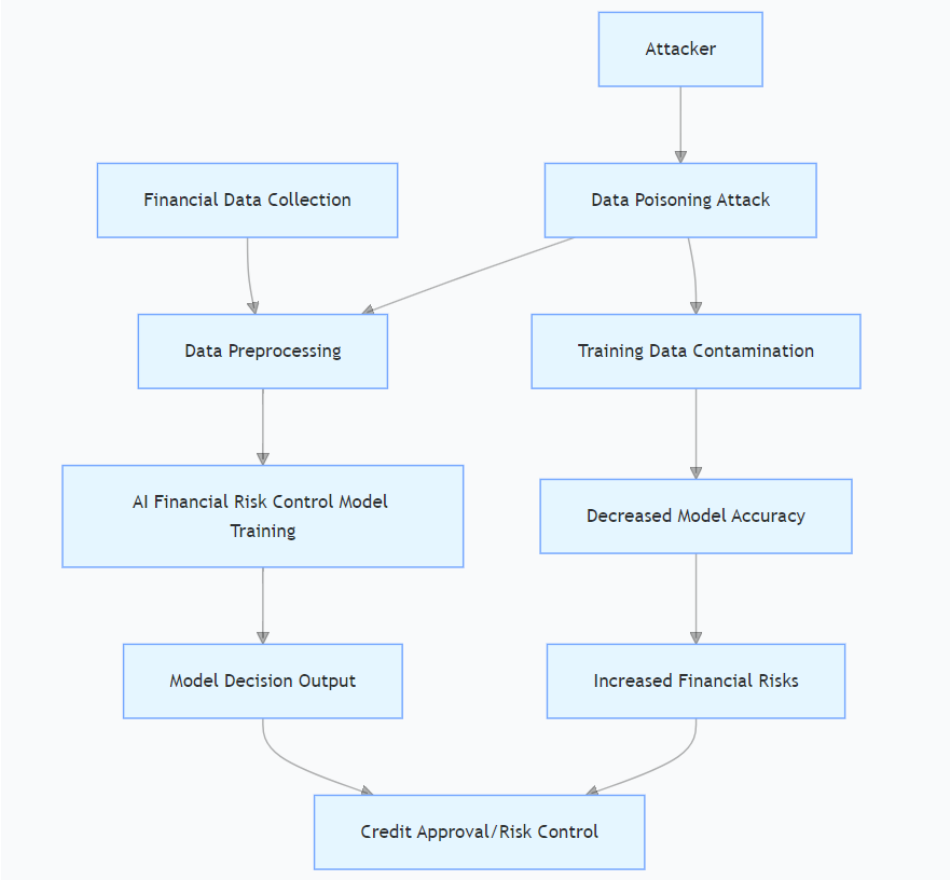

Figure 1. Flowchart of the relationship and impact between AI financial risk control models and data poisoning attacks

2.4. Common means of data poisoning attacks

Common means of data poisoning attacks include data injection, data modification, and data deletion, as shown in Table 1. Data injection refers to the attacker adding new malicious data to the training data, which can be forged false data or tampered real data. Data modification refers to the attacker tampering with the existing data in the training data, such as modifying customers' credit records, transaction amounts, etc., making the data lose authenticity and reliability. Data deletion refers to the attacker deleting key data in the training data, such as deleting high-risk customer records, so that the model cannot learn these key information, thereby affecting the model's risk assessment ability. In addition, attackers can also use data generation technologies such as Generative Adversarial Networks (GAN) to generate malicious data similar to real data, thereby improving the effect and concealment of data poisoning attacks [5].

Table 1. Common types of data poisoning attacks

Attack Type	Operation Mode	Impact on Financial Scenarios
Data Injection	Forging false data (such as high-credit customer records) and injecting them into the training set	The model misjudges low-risk customers, increasing the risk of credit defaults
Data Modification	Tampering with key features such as transaction amounts and device fingerprints	The fraud detection model misses malicious transactions, reducing risk identification accuracy
Data Deletion	Deleting high-risk customer or abnormal transaction records	The model underestimates the risk probability, leading to an expansion of risk exposure
Generative Adversarial Poisoning	Using GAN to generate malicious samples similar to real data	The model has difficulty distinguishing between normal and fraudulent transactions, and the decision boundary shifts

3. Problems in responding to data poisoning attacks on AI financial risk control models

3.1. Incomplete data source verification mechanisms

In the training process of AI financial risk control models, the diversity and complexity of data sources increase the risk of data poisoning attacks. At present, many financial institutions lack effective data source verification mechanisms in the process of data collection and integration. For example, for externally purchased data, the qualifications of data suppliers and data quality are not strictly reviewed and evaluated, resulting in some problematic data entering the training dataset [6]. At the same time, for internal data, a perfect data generation and storage record mechanism has not been established, and it is impossible to accurately trace the source and generation process of data. This allows attackers to mix malicious data into the training data by forging data sources or tampering with the data generation process, and financial institutions find it difficult to detect and identify. For example, when a financial institution used customer credit data provided by a third party, due to the lack of strict verification of the data source, the attacker forged a large number of false high-credit customer data and injected them into the training data, resulting in serious deviations in the model's credit evaluation and bringing huge economic losses to the institution.

3.2. Insufficient abnormal data identification capabilities

Abnormal data is an important manifestation of data poisoning attacks, and accurate identification of abnormal data is a key link in responding to data poisoning attacks. However, many current AI financial risk control models have deficiencies in abnormal data identification (see Table 2). On the one hand, existing abnormal detection algorithms are mainly based on traditional statistical methods and machine learning algorithms, and the detection effect on complex and new data poisoning attack methods, such as malicious data generated by using generative adversarial networks, is not good. These malicious data have features and distributions similar to real data, and traditional detection algorithms are difficult to distinguish. On the other hand, in the model training process, financial institutions often pay too much attention to the overall performance indicators of the model, such as accuracy and recall rate, but ignore the special analysis and processing of abnormal data. A perfect abnormal data detection process and standards have not been established, resulting in some malicious data hidden in a large amount of normal data cannot be found in time. For example, in the fraud detection model of a financial institution, the attacker used a generative adversarial network to generate a batch of malicious transaction data very similar to normal transaction data and injected them into the training data [7]. Due to the insufficient abnormal data identification capability of the model, these malicious data were not detected, resulting in a significant decrease in the recognition rate of similar fraud transactions in the actual application of the model.

Table 2. Different detection technologies show different characteristic differences in responding to poisoning attacks

Detection Method	Typical Algorithms	Advantages	Disadvantages
Statistical Methods	Z-score, Box Plot	Simple and fast, suitable for linear data	Difficult to handle complex patterns, susceptible to distribution interference
Traditional Machine Learning	Isolation Forest, One-Class SVM	Automatically learn local features	Relies on manual features, limited generalization ability
Deep Learning	Autoencoder, GAN	Efficiently detect complex anomalies	High training cost, need to prevent overfitting

3.3. Model robustness needs to be improved

Model robustness refers to the ability of the model to maintain stable performance in the face of noise, interference, and attacks. At present, many AI financial risk control models do not fully consider the impact of data poisoning attacks in the design and training process, resulting in poor model robustness. On the one hand, there are defects in the model architecture design, such as excessive reliance on certain key features or shallow features, and lack of deep learning and abstraction capabilities for data features. When malicious data is injected into the training data, these key features or shallow features are easily disturbed, thereby affecting the decision-making results of the model. On the other hand, the model training process lacks effective regularization and adversarial training mechanisms. Regularization methods can prevent model overfitting and improve model generalization ability, but many financial institutions do not apply regularization technology reasonably in model training. Adversarial training is to let the model learn the ability to resist attacks by adding adversarial samples to the training data, but the application of adversarial training in the field of financial risk control is not yet widespread and in-depth. For example, the credit evaluation model of a financial institution adopted a simple neural network architecture and did not use regularization and adversarial training technologies in the training process. When the attacker injected a small amount of malicious data into the training data, the credit evaluation results of the model changed significantly, showing that the model's robustness was seriously insufficient.

3.4. Imperfect dynamic defense system

Data poisoning attack is a dynamic process, and attackers will continuously adjust attack means according to the model's defense strategy, so it is necessary to establish a dynamic defense system to deal with it. However, many current financial institutions lack the concept and method of dynamic defense when responding to data poisoning attacks. On the one hand, the model monitoring and updating mechanism is not flexible enough to timely detect abnormal changes in model performance and make adjustments. For example, a real-time model performance monitoring system has not been established to compare and analyze the model's prediction results and actual risk situations in the production environment, making it difficult to detect long-term poisoning attacks by attackers. On the other hand, there is a lack of real-time analysis and research on attack patterns and means, and it is impossible to update defense strategies in a timely manner according to the latest attack trends. When new attack means appear, existing defense measures often fail and cannot effectively resist attacks. For example, after the risk control model of a financial institution was deployed, a dynamic monitoring and updating mechanism was not established, and the attacker gradually changed the model's decision boundary by injecting malicious data slowly for a long time. The financial institution did not find the problem until the model made a large number of wrong decisions, but serious losses had already been caused at this time.

3.5. Lagging attack traceability technology

Attack traceability refers to the ability to accurately track the source and path of an attack after a data poisoning attack occurs, providing a basis for subsequent investigation and processing. However, in the field of AI financial risk control, attack traceability technology is relatively lagging. On the one hand, there is a lack of effective data tracking and recording mechanisms, and it is impossible to completely record and monitor the entire life cycle of training data. For example, there are no detailed log records in the links of data collection, cleaning, labeling, storage, etc., and when malicious data is found, it is difficult to trace which link the data was injected into. On the other hand, the technology for analyzing and identifying attack behaviors is not mature enough to extract effective attack features and clues from a large amount of data and model operation logs. This makes it difficult for financial institutions to timely and accurately determine the source and means of the attack when facing data poisoning attacks, and it is difficult to take effective countermeasures. For example, after a financial institution found abnormalities in its risk control model, due to the lagging attack traceability technology, it was unable to determine whether it was an internal staff or an external attacker, nor could it determine by what means the attack was carried out, resulting in slow progress in the investigation and handling work.

3.6. Imperfect data security management system

In addressing data poisoning attacks on AI financial risk control models, the incomplete data security management system has emerged as a critical vulnerability. This manifests in loose access control, where permissions are not set based on the principle of least privilege, leading to cross-departmental unauthorized access and inadequate isolation between development-test and production environments—for instance, a financial institution suffered data injection by a former employee whose account privileges were not revoked timely. Inadequate data encryption and backup mechanisms leave sensitive data unprotected by strong encryption and lack automated off-site disaster recovery, hindering quick restoration after tampering. Weak employee security awareness, due to insufficient training on data poisoning, makes staff vulnerable to attacks like phishing emails, as seen in a case where a backdoor implanted via such an email increased a fraud model's false negative rate by 40%. Security loopholes in third-party collaborations, such as undefined data security responsibilities and lack of audit trails for external personnel, allow suppliers to illicitly tamper with data, causing "compliance poisoning." This aligns with the strategy of strengthening data security management, urging a closed-loop system integrating management processes, personnel training, and technical safeguards.

4. Response strategies for data poisoning attacks on AI financial risk control models

4.1. Construct a multi-dimensional data verification system

In order to ensure the authenticity and reliability of training data, financial institutions should construct a multi-dimensional data verification system (see Figure 2). First, in the data collection stage, establish a strict data supplier audit mechanism, comprehensively evaluate the qualifications of data suppliers, data sources, data quality, etc., and choose to cooperate with suppliers with good reputation and reliable data quality [8]. At the same time, for internally generated data, establish detailed data generation records and audit processes to ensure that the data generation process complies with standards. Second, in the data integration stage, use data fingerprint technology, data hash algorithms, etc., to uniquely identify and verify data to ensure that data is not tampered with during transmission and storage. Third-party data verification institutions can also be introduced to independently verify important data to improve data credibility. For example, when collecting customer credit data, not only the qualifications of data suppliers should be reviewed, but also the collected data should be cross-verified with the central bank's credit investigation system and credit data of other financial institutions to ensure the accuracy and integrity of the data. By constructing a multi-dimensional data verification system, the possibility of malicious data entering the training dataset can be reduced from the source.

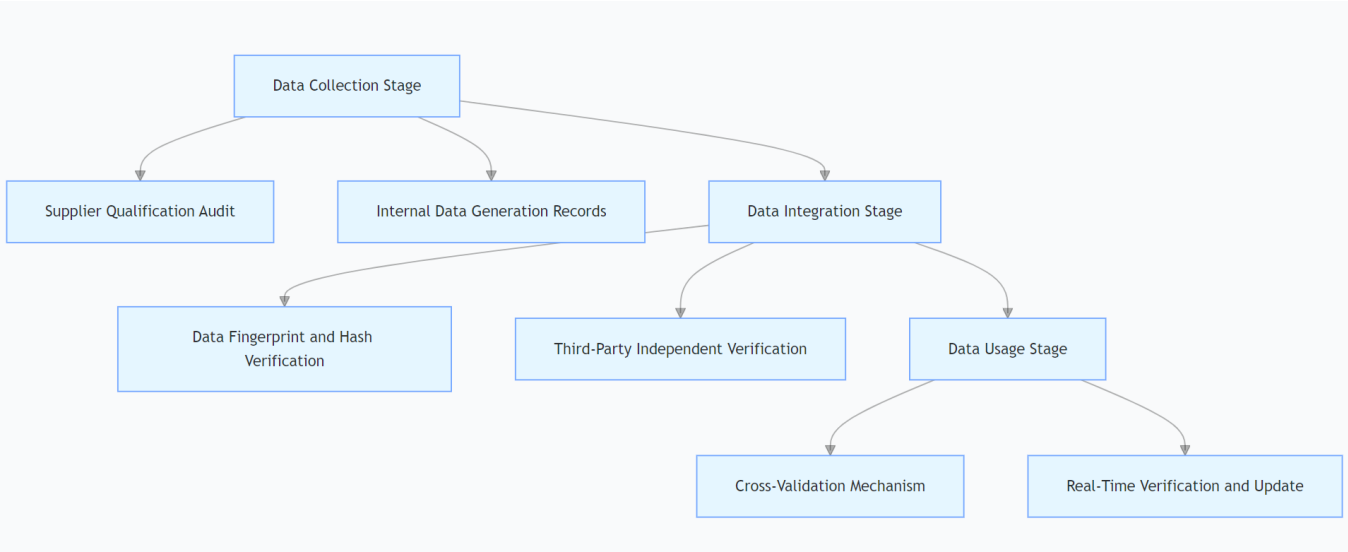

Figure 2. Framework diagram of multi-dimensional data verification system

4.2. Strengthen abnormal data detection algorithms

Aiming at the current problem of insufficient abnormal data identification capabilities, financial institutions should strengthen abnormal data detection algorithms. On the one hand, introduce advanced machine learning and deep learning algorithms, such as isolation forest, autoencoder, generative adversarial network, etc., to improve the detection ability of complex malicious data. These algorithms can automatically learn the features and distribution of data and have a better recognition effect on abnormal

data. On the other hand, combined with the characteristics of financial business scenarios, establish a special abnormal data detection rule and index system. For example, in fraud detection, a series of abnormal transaction detection rules can be formulated according to features such as transaction time, transaction amount, and transaction location, and combined with machine learning algorithms to improve the accuracy and efficiency of detection. In addition, the abnormal data detection model should be regularly updated and optimized, and the model parameters and detection strategies should be continuously adjusted according to new attack cases and data features. For example, a financial institution introduced an autoencoder algorithm to detect abnormal transaction data, and combined with business rules, effectively improved the detection ability of malicious transaction data generated by using generative adversarial networks, reducing the missed detection rate of abnormal data by more than 30% [9].

4.3. Optimize model architecture design

In order to improve the robustness of the model, financial institutions should optimize the architecture design of the model. First, adopt a multi-layer and multi-feature model architecture to avoid excessive reliance on single key features or shallow features. For example, in the neural network model, the depth and width of the network can be increased, and technologies such as attention mechanism and residual connection can be introduced to improve the model's deep learning and abstraction capabilities of data features, so that the model can better resist the interference of malicious data on key features. Second, in the model training process, reasonably apply regularization technologies, such as L1/L2 regularization, dropout, etc., to prevent model overfitting and improve model generalization ability. Adversarial training methods can also be used to add a certain proportion of adversarial samples to the training data, allowing the model to learn the ability to identify and resist attacks. For example, a financial institution used adversarial training technology in the training process of the credit evaluation model, and added adversarial samples generated by simulating data poisoning attacks to the training data, so that the robustness of the model was significantly improved. In the face of malicious data attacks, the decline range of the model's prediction accuracy was reduced from the original 20% to below 5%.

4.4. Establish a dynamic monitoring and response mechanism

In order to deal with the dynamics of data poisoning attacks, financial institutions should establish a dynamic monitoring and response mechanism. First, establish a real-time model performance monitoring system to conduct real-time monitoring and analysis of the model's prediction results, actual risk situations, data distribution changes, etc., in the production environment. When abnormal fluctuations in model performance are found, such as a sudden drop in accuracy and a significant increase in misjudgment rate, the early warning mechanism is triggered in a timely manner. Second, establish a dynamic update mechanism for the model, and timely adjust and update the model according to the real-time monitored data and attack situations. For example, when it is found that the distribution of training data has shifted, online learning methods can be used to gradually update the model parameters to make the model adapt to the new data distribution. An attack response team can also be established to be specifically responsible for the analysis and processing of data poisoning attacks, and formulate detailed emergency plans to ensure that effective countermeasures can be taken quickly when an attack occurs. For example, a financial institution established a real-time model performance monitoring system. When it was detected that the fraud recognition rate of the model dropped by more than 5% for three consecutive days, an early warning was immediately triggered. The attack response team quickly analyzed the data and model, found that the attacker had injected new malicious data, and timely updated and optimized the model, avoiding further expansion of losses.

4.5. Improve the level of attack traceability technology

In order to improve the ability of attack traceability, financial institutions should improve the level of attack traceability technology. First, establish a perfect data tracking and recording mechanism, and make detailed log records of the entire life cycle of data, including data collection, cleaning, labeling, storage, use, etc., including information such as data sources, operation time, operators, and operation contents. Through data traceability technologies such as blockchain technology, ensure that the operation records of data are tamper-proof and can accurately trace each change link of data [10]. Second, strengthen the research on attack behavior analysis and identification technologies, and use big data analysis, machine learning and other technologies to extract effective attack features and clues from massive information such as model operation logs and data operation records. For example, determine the source and means of the attack by analyzing the characteristics, injection time, and impact range of malicious data. In addition, cooperation with external security agencies and research institutions can be carried out to share attack cases and technical experience, and jointly improve the level of attack traceability technology. For example, a financial institution used blockchain technology to record the life cycle of training data. When malicious data was found, through the tamper-proof characteristics of blockchain, it was accurately traced that the data was injected by a third-party supplier in the data collection link, providing strong evidence for subsequent investigation and processing.

4.6. Strengthen data security management

Data security management is an important guarantee for responding to data poisoning attacks. Financial institutions should strengthen data security management and establish a perfect data security system and process. First, strengthen the management of data access rights, set different data access rights for personnel in different positions, and prevent unauthorized personnel from accessing and modifying training data. Second, use data encryption technology to encrypt stored and transmitted data to ensure the security and confidentiality of data. A data backup and recovery mechanism can also be established to regularly back up training data and models to prevent data loss and damage. In addition, strengthen the data security awareness training of employees, improve employees' understanding and prevention awareness of data poisoning attacks, and avoid data security incidents caused by employees' improper operations. For example, a financial institution strictly controlled the access to training data by strengthening the management of data access rights, and only authorized personnel could operate the data. At the same time, data encryption technology was used to protect the data, effectively reducing the risk of data poisoning attacks.

5. Conclusion

In the current vigorous development of fintech, AI financial risk control models play an increasingly important role in financial risk assessment and management. However, data poisoning attacks, as a highly targeted and harmful attack means, pose severe challenges to the security and reliability of AI financial risk control models. From a professional and technical perspective, this paper deeply analyzes the principles of AI financial risk control models and data poisoning attacks, points out the existing problems in the current response process, including incomplete data source verification mechanisms, insufficient abnormal data identification capabilities, model robustness to be improved, imperfect dynamic defense systems, and lagging attack traceability technologies. Aiming at these problems, specific response strategies are proposed, including constructing a multi-dimensional data verification system, strengthening abnormal data detection algorithms, optimizing model architecture design, establishing dynamic monitoring and response mechanisms, improving the level of attack traceability technology, and strengthening data security management. These strategies aim to comprehensively improve the ability of AI financial risk control models to resist data poisoning attacks from multiple levels such as data sources, the model itself, monitoring and response, and security management. In practical applications, financial institutions should comprehensively use these strategies in combination with their own business characteristics and technical levels, continuously improve the system for responding to data poisoning attacks, ensure the safe and stable operation of AI financial risk control models, and provide strong support for the healthy development of the financial industry.

References

[1] Mu, Y. C., Chen, A. W., Chen, G. R., Xu, J. M., Yan, X. M., & Duan, L. (2025). Security defense strategies for federated learning based on data poisoning attacks. *Systems Engineering and Electronics*, 1–14.

[2] Jiang, W. Q., Lu, X. H., Tang, S. L., Yuan, K. G., Cai, W. X., & Mei, S. (2025). Solutions to security risks of data poisoning in artificial intelligence. *Telecom Engineering Technology and Standardization, 38*(5), 69–71, 87. https://doi.org/10.13992/j.cnki.tetas.2025.05.005

[3] Yuan, J. H. (2025). Application of financial risk control models based on big data technology in financial investment. *Investment Beijing*, (5), 58–59.

[4] He, X. Y., & Zhang, Z. (2024). Research on the competency model of fintech risk control engineers. *Fintech Time, 32*(6), 21–25.

[5] Liu, L. (2023, August 28). Pre-trained large models will enhance the risk control capabilities of financial institutions. *Financial Times*, 9.

[6] Yin, W. (2022). *Research on risk monitoring and governance of internet financial platforms.* Southeast University Press.

[7] Li, K. H. (2022). *Research on financial risk control models based on machine learning* [Master's thesis, University of Electronic Science and Technology of China]. https://doi.org/10.27005/d.cnki.gdzku.2022.003102

[8] Cai, Z. X. (2021). *Intelligent risk control and anti-fraud*. Machinery Industry Press.

[9] Zhang, J., Zhang, Z., Fang, K. N., Shi, X. J., & Zheng, C. L. (2020). Research on consumer financial risk control based on sparse structure continuous ratio model. *Statistical Research, 37*(11), 57–67. https://doi.org/10.19343/j.cnki.11-1302/c.2020.11.005

[10] Chen, N., Guo, S. H., & Wang, J. Z. (2020). Research on internet financial risk control models based on middle office architecture. *China High-Tech*, (20), 118–119. https://doi.org/10.13535/j.cnki.10-1507/n.2020.20.53