Scholar Publishing Group
International Journal of Neural Network
https://doi.org/10.38007/NN.2026.050102
ISSN 2790-2692 Vol. 5, Issue 1: 11-21

Research on Big Data Security Governance and Privacy Protection Strategies for Cross-Domain Data Circulation Scenarios

Yongqiang Ma

School of Computer and Big Data, Jining Normal University, Ulanqab 012000, Inner Mongolia, China

Keywords: Big data; data security; privacy protection; cross-domain data flow; governance strategies

Abstract: In an era of expanding cross-domain data flow, algorithm-driven decision-making, and intelligent services, traditional data security solutions based on boundary protection and static authorization are no longer adequate for the governance needs of the big data era. This article, based on a review of international and industry research reports from recent years, analyzes the increasing cost of data breaches, stricter privacy regulations, the development of privacy-enhancing technologies, and the risks of secondary inference brought about by artificial intelligence. Research indicates that the main contradictions in data security and privacy protection currently lie in the conflict between releasing data value and the principle of minimum necessity, the conflict between the cost of privacy computing and business real-time requirements, and the imbalance between the black box nature of algorithms and accountability. Therefore, this paper proposes a comprehensive strategy system that combines classification and grading, zero-trust access, differential privacy, and federated learning, supported by a trusted execution environment, establishing complete on-chain auditing, and scoring governance capabilities. This research can provide a reference for government agencies, healthcare institutions, financial institutions, and platform enterprises to create data governance systems that balance security, compliance, and business value.

1 Introduction

In the era of big data, data activities have shifted from internal institutional processing to continuous flow across clouds, platforms, regions, and entities. Data has transformed from a static resource supporting business processes into a production factor in various fields such as precision marketing, risk identification, urban governance, medical decision support, and intelligent manufacturing. As data scale, connection density, and processing frequency increase, security boundaries become increasingly blurred, and privacy breaches and compliance violations can have cascading effects.

According to IBM's cost report released in 2025, the global average cost of a data breach is

around US$4.44 million. Although it has decreased, the risk exposure of ungoverned AI systems and distributed environments has increased [1]. IBM's research results in 2024 showed that the global average cost of a breach reached US$4.88 million, a significant increase from the previous year. This means that in the highly digitalized stage, the cost of security has not decreased naturally due to the popularization of technology [2]. IBM's report in 2023 pushed the average cost of a breach to a record high of US$4.45 million, and data breaches have changed from accidental security incidents to ongoing business risks [3].

From an organizational governance perspective, the Cisco 2025 Data Privacy Benchmark Study shows that 90% of respondents believe local storage is more secure, 91% believe global service providers are better at protecting data, and 96% believe that the benefits of privacy investment outweigh the costs [4]. This requires enterprises to reach a higher standard in terms of data localization, global collaboration, and compliance investment. The interaction between artificial intelligence and data privacy is becoming increasingly complex. According to research, AI technology will not only enhance the ability to collect and infer big data, but also reshape the individual's perception of privacy control.

In this context, static access control, perimeter firewalls, or post-event compliance filing alone are no longer sufficient to meet the requirements of a traceable, explainable, quantifiable, and operable security system in the big data environment. This paper chooses the cross-domain data circulation scenario as its starting point, discussing it in three parts: current situation analysis, problem identification, and strategy construction. Based on incorporating the latest research results in data governance, federated learning, differential privacy, and big data security, it establishes a comprehensive management system that can both leverage data value and protect user privacy.

2. Current Status Analysis of the Research Topic

2.1 The cost of data breaches continues to rise, and security issues are transforming into business issues.

International reports over the past three years have consistently shown that data security incidents are no longer emergency issues that can be handled by a single technology department; they have become systemic problems affecting brand reputation, customer trust, regulatory costs, and operating cash flow. Especially in hybrid cloud, data platform, and API interconnection scenarios, attackers can spread the virus from a single vulnerability to multiple business areas, thus increasing both the scope of the breach and the cost of handling it.

In order to quantify the exposure of various data assets, risk assessment can be reduced to a multi-indicator weighted approach.

$$R = \Sigma (w_i \times x_i), i = 1 \ldots n \quad (1)$$

In equation (1), R represents the comprehensive risk value of data assets, x_i represents the weights of indicators such as sensitivity level, scope of access, frequency of sharing, interface exposure, and historical events, and w_i represents the weights of each indicator. This equation indicates that big data governance should not only consider the amount of data, but also the data flow paths and usage scenarios.

If an organization can quantify risks and obtain comparable scores, it can prioritize security budgets, audit frequency, and critical security data domains.

Figure 1 shows that as the scale of data breaches grows and rises, the global average cost of data breaches also increases annually. When enterprises move towards cloudification, intelligentization, and cross-domain collaboration, the losses caused by the increased attack surface do not completely disappear. For big data platforms, reducing risks cannot rely solely on post-incident patching;

instead, defenses should be strengthened through data classification, access control, interface auditing, and AI governance.

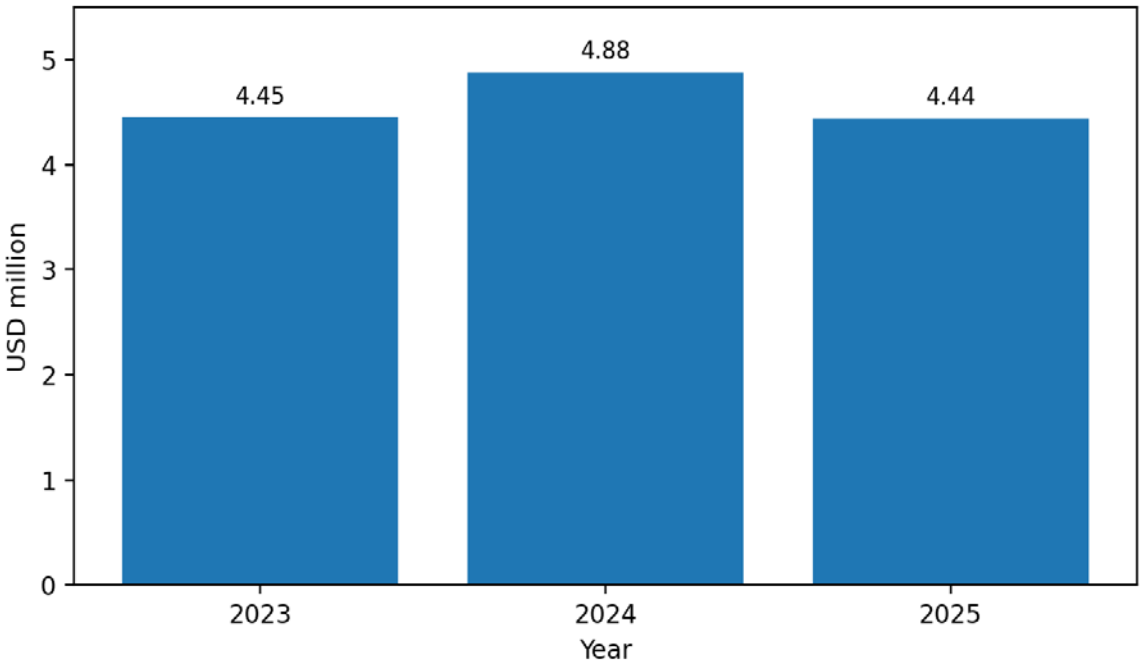

Figure 1. Global Average Data Breach Cost Statistics, 2023-2025

2.2 Data governance shifts from compliance-driven to value-driven

In recent years, data governance research has gradually shifted from "how to meet regulatory requirements" to "how to leverage the value of data while ensuring security." Bernardo et al., through a systematic review, argue that data governance has evolved from simple data quality management into a comprehensive system encompassing responsibility allocation, metadata management, process constraints, and audit loops. This means that the object of governance is no longer just the data itself, but also the way data is generated, the rules for sharing it, and the chain of responsibility.

In the context of big data, compliance is not an additional cost; it is a fundamental prerequisite for ensuring business continuity and user trust. Especially when sharing data across domains, the more frequent the data flow, the greater the compliance conflicts and privacy disputes will be if there is a lack of definition regarding data ownership, purpose of use, retention period, and boundaries of reuse.

2.3 Privacy-enhancing technologies are moving from single-point use to collaborative deployment.

Privacy protection technology has gradually evolved from early anonymization and access control to the current era of multiple technologies such as differential privacy, federated learning, multi-party secure computation, and trusted execution environment. Mohammadi et al. believe that although federated learning greatly reduces the need for the aggregation of original datasets, there are still complex trade-offs between communication costs, model accuracy, convergence speed, and privacy strength [7]. Hu et al. believe that federated learning is not inherently secure, and problems such as model inversion, gradient leakage, and malicious client poisoning still exist [8].

Therefore, current research trends no longer emphasize the "omnipotence" of a single technology, but rather place greater emphasis on scenario adaptability and the synergy between different technologies. For risk control platforms that require real-time performance, minimum available anonymization and hierarchical access control can be used; while for cross-institutional modeling tasks, adding secure aggregation and differential privacy within a federated learning framework is more appropriate.

Table 1DataLifecycle Risks ofcross-domain Scenarios

Stage	Typical data	Main risk	Control focus
Collection	Logs, IDs, GPS	Over-collection	Purpose binding
Storage	Profiles, records	Mass leakage	Encryption at rest
Sharing	APIs, files, models	Unauthorized reuse	Least privilege
Analysis	Features, labels	Re-identification	PET integration
Archival	Backups, snapshots	Long-tail exposure	Retention control

Based on the five stages of data collection, storage, sharing, analysis, and archiving, the main risks in cross-domain scenarios are identified. It can be seen that these risks do not arise at a single isolated node but exist throughout the entire data lifecycle. In the sharing and analysis stages, unauthorized reuse and re-identification pose the highest risks; therefore, purpose binding, least privilege, and privacy enhancement technologies should be incorporated into the process design phase.

3. Raise questions

3.1 The exposure of data throughout its entire lifecycle continues to expand.

Big data platforms are characterized by a wide variety of sources, rapid flow, frequent replication, and complex interfaces. During the processes of data collection, caching, cleaning, labeling, sharing, modeling, and backup, a large number of copies, indexes, and derived features are generated. If any link fails to control, it will cause a chain of leaks. This problem is more pronounced in medical and highly sensitive data. Relevant system reviews show that as the value of data use increases, the exposure also expands [9].

Therefore, even if the direct identifier is deleted, the combination of quasi-identifiers can still lead to re-identification. To illustrate the degree of anonymity of a single record within an equivalence class, it can be written in the following form:

$$k(q) = |E(q)|, \text{ and } k(q) \geq k_0 \tag{2}$$

In equation (2), q is the quasi-identifier combination, E(q) is its equivalence class, and k0 is the minimum anonymity threshold value predetermined by the system. When the equivalence class size is too small, attackers can easily use external data to associate and identify it.

However, for high-dimensional sparse data and multi-source fusion data, relying solely on k-anonymity is no longer sufficient to resist attribute inference and background knowledge attacks, and the shortcomings of traditional anonymization schemes in the big data environment are becoming increasingly prominent.

3.2 There is a structural conflict between sharing efficiency and the principle of minimum necessity.

Cross-domain data flow aims to achieve the goals of "being able to share, knowing how to share, and being traceable." However, business systems generally prioritize efficiency and availability, leading to problems such as overly broad shared fields, excessively granted permissions, and excessively long retention times. Although many organizations have established approval processes, they have not made detailed restrictions on the boundaries of field-level, interface-level, and task-level usage. Therefore, the principle of minimum necessity remains only at the level of policy documents.

In a platform-based business environment, once data is accessed through APIs, SDKs, or model services, its use cases often deviate from their original purpose, falling into a governance trap of "one-time compliance authorization, continuous reuse." This is one of the major reasons for the large number of complaints about data misuse and privacy violations.

3.3 The contradiction between the cost of privacy enhancement technologies and the real-time nature of business operations is prominent.

Privacy-enhancing technologies can significantly improve protection levels, but they also incur computational, communication, deployment, and governance costs. In particular, technologies like federated learning, secure aggregation, and homomorphic encryption can easily lead to longer training times, increased system resource consumption, and more difficult engineering implementation in environments with numerous terminals and high concurrency. Without a layered deployment strategy, big data platforms oscillate between high protection and high performance.

$$P(B|X) = 1 / (1 + e^{-(\beta_0 + \Sigma\beta_i x_i)}) \quad (3)$$

Equation (3) treats data breaches as a conditional probability problem, where X represents variables such as the number of high-risk interfaces, the proportion of unmasked fields, the number of abnormal accesses, and the frequency of external sharing. This equation can be used for risk warning and governance priority ranking.

If the leakage probability obtained by the model is consistently greater than the threshold, then the scope of sharing should be reduced first, the authentication strength should be increased, and then further auditing of the relevant chain should be carried out.

3.4 The application of artificial intelligence has exacerbated the challenges of privacy inference and liability identification.

After the widespread use of large models, recommendation algorithms, and intelligent agents, the risk of data privacy has changed from whether the original records are leaked to whether individual characteristics can be inferred. Martin and Zimmerman believe that because AI changes the way data is collected, processed, and output, the logic of privacy decisions is also changing [5]. This means that even if the original data is not directly presented, the model's output can still infer a person's privacy situation based on profiles, attributions, and related correlations.

Another scenario is that complex model chains can weaken the determination of responsibility. If training data, fine-tuning data, knowledge bases, and external plugins all participate in decision-making, organizations cannot determine whether a particular privacy violation was caused by data input, rule settings, or model behavior.

3.5 Fragmented organizational governance; technology and systems have not yet formed a closed loop.

Many organizations have deployed security products and established privacy policies, approval systems, and auditing procedures. However, there remains a lack of clear division of responsibilities among departments: business departments prioritize efficiency, legal departments focus on compliance boundaries, technical departments prioritize system stability, and data management departments prioritize definitions and quality. The lack of unified metrics and a clear chain of responsibility leads to a disconnect between systems, processes, technology, and auditing, hindering the construction of a sustainable governance system.

4. Problem Solving/Strategies

4.1 Construct a dynamic risk governance system based on classification and grading

In the era of big data, how to solve security and privacy issues? First, we must establish a governance system that emphasizes classification and grading, scenario identification, and dynamic evaluation. Classification and grading cannot be a one-time event; they must be continuously adjusted according to changes in business operations, sharing frequency, external exposure levels, and compliance requirements. Highly sensitive data domains should implement a linkage mechanism of field-level labeling, usage-level approval, and task-level authorization to ensure that risks are visible, priorities are clearly defined, and the scope of dissemination is controlled.

Differential privacy is an important tool in dynamic governance. Its basic constraints can be written as:

$$\Pr[M(D)\in S] \leq e^{\varepsilon} \cdot \Pr[M(D')\in S] + \delta \tag{4}$$

In the formula, M is random, D and D′ are symmetric datasets differing by one record, and ε and δ are privacy budget parameters. This means that adding or deleting any single record will not change the distribution of the query results, thereby reducing the probability of an individual being identified or inferred.

In practical use, the privacy budget can be divided into statistical release tasks, internal risk control and operational tasks according to data sensitivity and business real-time requirements, so as to ensure privacy while meeting legal requirements.

4.2 Establish a zero-trust access control and continuous authentication mechanism

To address the issues of excessive authorization, prolonged usage time, and distorted identity information in cross-domain data sharing, the traditional credit model based on network boundaries should be replaced with zero-trust access control based on identity, device, task, and environment. Access authorization should not be limited to entering the system; instead, the criterion should be whether access is necessary within the current task, terminal, time period, and field scope.

Therefore, an access control policy based on attributes can be created on the data gateway, taking into account factors such as user responsibilities, data sensitivity, purpose of the call, terminal health, and location, and linking them with log auditing, anomaly monitoring, and session cancellation measures.

Figure 2 leads to the following conclusions regarding the current international consensus on privacy governance: first, privacy regulations and investment are the foundation of trust; second, there is a growing demand for data localization, while global service capabilities remain recognized; and third, businesses generally believe that privacy governance yields positive benefits. From this,

we can conclude that data security and privacy protection are not obstacles to data flow, but rather the foundation and conditions for improving the quality of data flow.

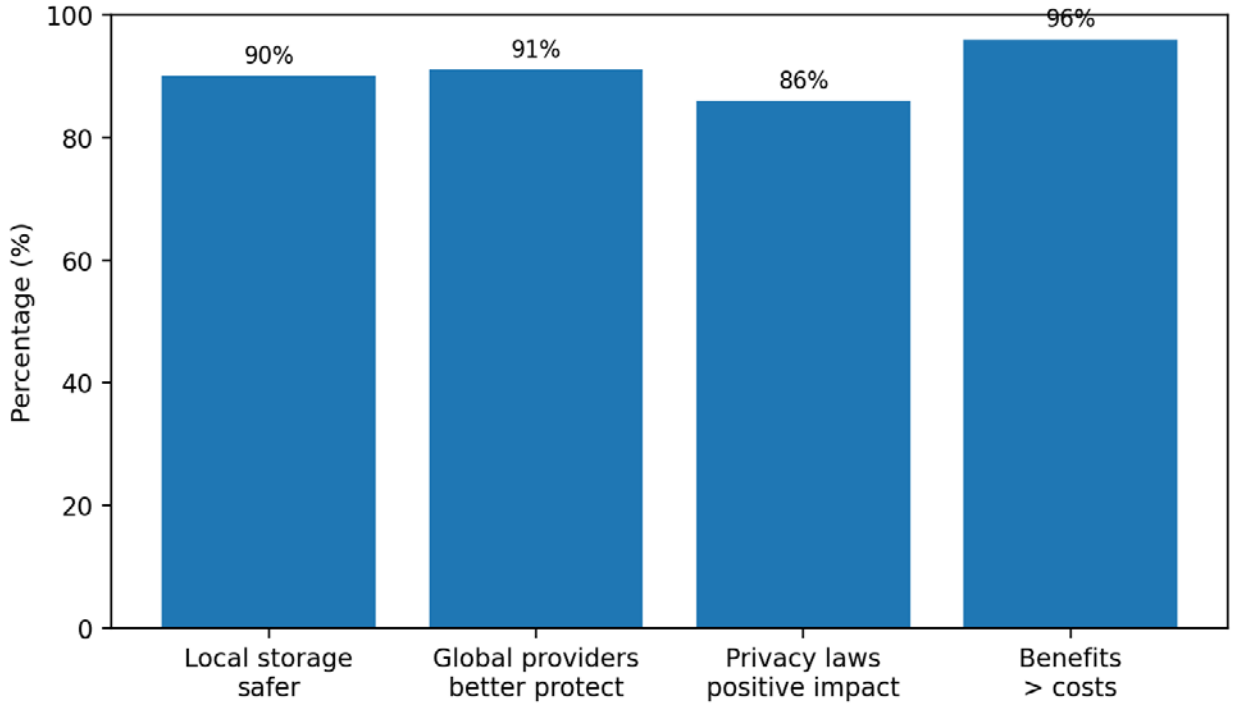

Figure 2. Statistical chart of survey results for key indicators of privacy governance

4.3 Achieving Collaborative Protection through Federated Learning, Trusted Execution Environment, and Security Aggregation

For complex scenarios such as cross-institutional joint modeling, regional collaborative analysis, and industry-wide joint risk control, simply centralizing data collection is insufficient. A combined structure of federated learning, secure aggregation, and a trusted execution environment can be adopted: raw data is stored locally, parameter updates are uploaded via secure aggregation, and critical computations are performed in a secure execution environment. This reduces the risk of direct data leakage and alleviates the compliance pressure that centralized storage places on data collection.

However, collaborative protection cannot only consider the strength of privacy protection; it must also consider its impact on model accuracy, communication latency, and operational costs. Therefore, the governance objective can be expressed as:

$$U = \alpha A + \beta C + \gamma I - \lambda T - \mu M \tag{5}$$

In equation (5), U represents the overall governance effect, A represents availability, C represents compliance, I represents the benefits of integrity and credibility, T represents the delay cost, and M represents the operation and maintenance cost. By adjusting the parameters, more suitable protection methods can be used in various application scenarios such as government affairs, healthcare, and finance.

This theory suggests that an excellent privacy protection solution is not simply about increasing the strength of protection, but about finding the optimal solution that can be verified and explained among the four factors of security, compliance, performance, and cost.

Table 2 Comparison of privacy-enhancing technologies

Technology	Privacy strength	Performance cost	Best-fit scenario	Maturity
Masking	Medium	Low	Reporting	High
Differential privacy	High	Medium	Statistics release	High
Federated learning	High	Medium-High	Joint modeling	Medium
Secure aggregation	High	Medium	Cross-site training	Medium
Trusted execution	High	Medium-High	Sensitive computing	Medium

Table 2 presents the protection strength, performance cost, and applicable scenarios of several common privacy enhancement techniques. It can be observed that there is no single optimal solution for all business needs. Anonymization and masking are suitable for low-cost deployments, differential privacy is suitable for statistical disclosures, federated learning is suitable for cross-institutional joint modeling, and trusted execution environments are suitable for hardware-level isolation of sensitive computing links.

4.4 Supporting accountability tracking and compliance orchestration with end-to-end auditing

In data security governance, the real challenge is not preventing a single attack, but defining boundaries beforehand, identifying deviations during an attack, and assigning responsibility afterward. Therefore, it's crucial to establish a comprehensive auditing system encompassing data request, field approval, API calls, model training, and result output. Audit logs should not only record who accessed which data, but also the purpose of using that data, the algorithms used, and the resulting constraints.

For high-risk tasks, an automated compliance orchestration approach can be adopted. The retention period, purpose binding, desensitization rules, cross-border conditions, and deletion policies can be written into the process engine in the system. The system requirements will automatically affect the technical execution layer, reducing the gap between manual approval and actual use.

4.5 Promoting the shift of governance maturity from project-based governance to operation-based governance

In the long run, data security and privacy protection cannot rely on one-off rectifications, but should be developed into a sustainable operational capability. Organizations should create a comprehensive governance model from the perspectives of systems, technology, processes, personnel, auditing, and value assessment, and advance governance in stages. The initial stage addresses the issue of existence, the intermediate stage addresses the issue of accuracy, and the advanced stage addresses the issue of stability.

Table 3 presents a governance maturity roadmap. Its key point is that organizations should not view security governance as an isolated project acceptance process, but rather as a continuous operation. Only by establishing a closed-loop system for data security and privacy, encompassing classification and grading, process approval, technical control, audit trails, and value evaluation, can data flow effectively support new business innovations.

Table 3 Governance maturity roadmap

Level	Core target	Key capability	Main metric
L1 Basic	Control exposure	Inventory, labeling	Coverage rate
L2 Managed	Standardize use	Workflow, review	Policy hit rate
L3 Integrated	Enable sharing	PET, zero trust	Risk reduction
L4 Optimized	Operate by value	Automation, metrics	ROI, trust

5. Conclusion

In the era of big data, data security and privacy protection have transformed from secondary technical issues into fundamental problems of the digital governance system. Given the rapid development of cross-domain data flow, algorithm-driven approaches, and intelligent applications, traditional governance methods relying primarily on static boundaries, coarse-grained permissions, and post-event filing are no longer sufficient to meet the demands of high-quality data utilization.

Based on a review of English literature and international reports from the past three years, this paper identifies five main contradictions in current big data security governance: increased exposure throughout the entire lifecycle, difficulties in implementing the minimum necessity principle, performance costs associated with privacy enhancement technologies, increased risks from secondary inference due to artificial intelligence, and fragmented organizational governance. To address these issues, this paper proposes a comprehensive governance strategy that starts with classification and grading and dynamic risk assessment, uses zero-trust access and end-to-end auditing as its framework, employs differential privacy, federated learning, and trusted execution environments as technical tools, and utilizes maturity management and value assessment.

In general, data governance cannot simply be a choice between "open utilization" and "strict protection." It requires a combination of institutional, technological, and procedural approaches, ensuring the security of data flow based on traceability, explainability, and quantifiability. Only in this way can the value of big data for social governance, industrial collaboration, and public service innovation be continuously realized while protecting individual rights and organizational compliance.

Funding

This work was supported by PhD Innovation Research Fund Project of Jining Normal University and Higher Education Research Project of the Inner Mongolia Autonomous Region Higher Education Association (Number: jsbsjj2335, jsbsjj2336, jsbsjj2413,NMGJXH-2025XB160); Intelligent Recognition and Image Processing Research Center (Number: jskypt2436).

References

[1] *Wu, W. (2025, June). Construction and optimization of intelligent gateway software management platform based on jenkins cluster management under cloud edge integration architecture in industrial internet of things. In International Conference on 6G Communications Networking and Signal Processing (pp. 633-645). Singapore: Springer Nature Singapore.*

[2] *Hua, X. (2024, November). Design and Implementation of a Game QoE Monitoring and Evaluation System Driven by Network Traffic Analysis. In International Conference on Cognitive based Information Processing and Applications (pp. 149-161). Singapore: Springer Nature Singapore.*

[3] *Huang, J. (2025, September). Performance Evaluation Index System and Engineering Best Practice of Production-Level Time Series Machine Learning System. In 2025 International Conference on Intelligent Communication Networks and Computational Techniques (ICICNCT) (pp. 01-07). IEEE.*

[4] *Sun, Q. (2026). Research on a Robotic Natural Language Intelligent Decision-Making Framework Based on Large Language Models, Thinking Chain Reasoning, and Multi-Agent Collaboration.*

[5] *Wang, Y. (2026). Research on the Application of Artificial Intelligence in Supply Chain Risk Early Warning.*

[6] *Wu, W. (2025, June). Construction and optimization of intelligent gateway software management platform based on jenkins cluster management under cloud edge integration architecture in industrial internet of things. In International Conference on 6G Communications Networking and Signal Processing (pp. 633-645). Singapore: Springer Nature Singapore.*

[7] *Hua, X. (2024, November). Design and Implementation of a Game QoE Monitoring and Evaluation System Driven by Network Traffic Analysis. In International Conference on Cognitive based Information Processing and Applications (pp. 149-161). Singapore: Springer Nature Singapore.*

[8] *Huang, J. (2025, September). Performance Evaluation Index System and Engineering Best Practice of Production-Level Time Series Machine Learning System. In 2025 International Conference on Intelligent Communication Networks and Computational Techniques (ICICNCT) (pp. 01-07). IEEE.*

[9] *Sun, Q. (2026). Research on a Robotic Natural Language Intelligent Decision-Making Framework Based on Large Language Models, Thinking Chain Reasoning, and Multi-Agent Collaboration.*

[10] *Wang, Y. (2026). Research on the Application of Artificial Intelligence in Supply Chain Risk Early Warning.*

[11] *Liu, H. (2026). Research on Dynamic Price Prediction of E-commerce Based on Time Series Modeling.*

[12] *Yu, X. (2026). Strategy Models and Practical Research of Growth Marketing under the Background of Digital Transformation.*

[13] *Hou, Y. (2026). Research on Server Performance Stability Assurance Mechanisms during Cross-Generation Computing Platform Upgrades.*

[14] *Han, X. (2026). Research on Process Decision-Making Behavior under Incomplete Information Conditions in Automobile Manufacturing Systems.*

[15] *Han, X. (2026). Research on Automotive Manufacturing Process Optimization Methods for Multi-Supplier Collaboration.*

[16] *Yu, X. (2026). Exploration of Multi-Channel Conversion Path Optimization Methods Based on A/B Testing.*

[17] *Zheng, H. (2026). Research on Edge Computing Network Task Scheduling and Resource Management Optimization Based on Artificial Intelligence Technology.*

[18] *Zhang, Z. (2026). Research on the Design of Scalable Enterprise-Level AI Systems Data Platform Architectures from an SDE Perspective.*

[19] Xiao Ma. Engineering Study of Disaster Recovery and Fault Self-Healing Mechanisms for Distributed Systems under Cross-Regional Deployment Conditions. International Journal of Engineering Technology and Construction (2026), Vol. 7, Issue 1: 1-7.

[20] Zheng, H. (2026). Research on Edge Computing Deep Neural Network Task Unloading Based on Resource Collaboration Framework and Multi Strategy Optimization.

[21] Zinuo Wang. Value Reassessment Logic of Resource-Based Enterprises in the Context of Energy Transition. International Journal of Social Sciences and Economic Management (2026), Vol. 7, Issue 1: 141-149.

[22] Weiyao Ma. Automated Operation Approach for Scalable Cloud Data Platform. International Journal of Big Data Intelligent Technology (2026), Vol. 7, Issue 1: 131-139.

[23] Zelin Wang. Data Analysis and Risk in Supply Chain Management. International Journal of Social Sciences and Economic Management (2026), Vol. 7, Issue 1: 132-140.

[24] Truong, T. H. (2025). Research on the Application of Digital Healthcare Platforms in Chronic Disease Management. Advances in Computer and Communication, 6(5).

[25] Ye, J. (2025). Optimization of Neural Motor Control Model Based on EMG Signals. International Journal of Engineering Advances, 2(4), 1-8.

[26] Ye, J. (2025). Design of a Non-Invasive Brain Computer Interface System for Handwritten Text Based on L2 Regularization and Attention Supervision Paradigm, and Optimization of EEG Signal Decoding. International Journal of Big Data Intelligent Technology, 2025, 6 (1), 126, 134.

[27] Yin, J. (2026). Research on a CLO Secondary Market Spread Volatility Prediction Model Based on RoBERTa Sentiment Factors. Advances in Computer and Communication, 7(1).

[28] Liu, H. (2025). Research on the Application of Sentiment Analysis in Customer Segmentation and Precision Marketing. Advances in Computer and Communication, 6(4).

[29] Liu, H. (2025). Research on the Evaluation of User Safety Intervention Measures Based on Causal Inference. Engineering Advances, 5(4).

[30] Wang, B. (2025). Research on Load Balancing Technology in Distributed System Architecture. International Journal of Multimedia Computing (2025), 6(1), 152-159.

[31] Wang, B. (2025). Application of Efficient Load Test Strategies in Infrastructure. Journal of Computer, Signal, and System Research, 2(4), 69-75.