Scholar Publishing Group
International Journal of Multimedia Computing
https://doi.org/10.38007/IJMC.2026.070109
ISSN 2789-7168 Vol. 7, Issue 1:73-81

Research on the Design and Application of Homomorphic Encryption Privacy-Preserving k-means Clustering Algorithm for Cross-Institutional Collaborative Risk Control

Ziting Mai

Guangzhou College of Commerce, School of Information Technology & Engineering, Guangzhou, 511363, China

Keywords: Homomorphic encryption; cross-institutional risk control; privacy protection; k-means clustering; joint modeling

Abstract: Cross-institutional collaborative risk control aims to achieve risk pattern sharing among entities such as banks, consumer finance companies, payment providers, and guarantee companies. However, directly aggregating raw samples can lead to constraints related to compliance, competition, and data sovereignty. To address this contradiction, this paper proposes a homomorphic encryption-based privacy-preserving k-means clustering method for risk control scenarios. Homomorphic encryption is used to calculate cross-institutional ciphertext distances and update cluster centers. Batch encoding and ciphertext pipelined processing reduce rounds and communication costs, creating a closed loop for clustering, profiling, and early warning in the risk control application layer. The paper investigates four aspects: ciphertext similarity calculation, cross-institutional center aggregation, convergence judgment, and risk cluster interpretation, and presents a deployable distributed system architecture. A comparative analysis of experimental results from publicly available literature over the past three years shows that using homomorphic encryption-based clustering methods can significantly reduce the risk of plaintext leakage during collaboration between different institutions while ensuring controllable accuracy loss. Furthermore, it achieves relatively low communication time and latency requirements even with large datasets. The above research shows that applying the privacy-preserving k-means algorithm to pre-loan customer segmentation, mid-loan abnormal group identification, and post-loan collection strategy allocation can provide a safe and scalable technical approach for joint risk control.

1 Introduction

As financial activities become increasingly online, platform-based, and ecosystem-driven, risk information is dispersed among numerous entities, including banks, consumer finance companies, third-party payment institutions, internet platforms, and guarantee institutions. When a single

institution relies solely on local samples to create risk profiles, problems arise such as sample bias, outdated labeling, and insufficient scenario coverage. In areas like long-tail customers, micro and small enterprises, users migrating across platforms, and identifying organized fraud, traditional single-point risk control exhibits a significant "information silo" phenomenon. Simultaneously, financial regulators are imposing higher requirements on personal information protection, data export, the principle of minimum necessity, and algorithm auditing. The practice of directly concentrating raw data on a single platform is increasingly unable to meet both institutional and market constraints.

Therefore, in this context, privacy-preserving machine learning has become a technical approach for cross-institutional collaborative risk control. K-means clustering, due to its clear structure, easily interpretable results, and low deployment cost, is used for unsupervised analysis tasks such as customer segmentation, anomaly identification, suspicious transaction profiling, and post-loan strategy allocation. Compared to supervised default prediction, clustering methods do not require high-quality labels, making them more suitable for financial scenarios with complex data sources, high labeling costs, and constantly changing anomaly patterns. However, standard k-means repeatedly calculates the distance between samples and cluster centers during iteration, while continuously updating the center positions. This prevents it from directly accessing cross-institutional features, thus inherently posing a risk of privacy breaches.

Homomorphic encryption can perform addition, multiplication, or approximate real number operations within ciphertext, making data usable but invisible. Research over the past three years has shown that optimized methods for multi-key fully homomorphic, CKKS approximate homomorphic, symmetric homomorphic, and outsourced ciphertext clustering are significantly reducing the communication overhead and running speed of privacy-preserving clustering [1-4]. Meanwhile, privacy-preserving credit analysis and cross-institutional risk modeling also demonstrate that introducing cryptographic mechanisms into financial modeling can guarantee business availability with a relatively high degree of privacy [5-7]. Therefore, this paper mainly studies the secure and efficient execution of k-means in cross-institutional scenarios, transforming clustering results into implementable risk control strategies.

2. Current Status Analysis of the Research Topic

Existing research can be broadly categorized into three types. The first type consists of privacy-preserving clustering schemes built using multi-party secure computation or secret sharing techniques. These methods offer high security, but often require multiple interactions, leading to a rapid increase in communication volume as the number of samples, feature dimensions, and iterations increase. This can easily result in latency amplification on wide area networks spanning different regions and institutions. The second type is outsourced clustering schemes based on fully homomorphic or near-homomorphic encryption. These schemes encrypt and upload data to the computing nodes in a single operation, completing the main computations within a few interaction rounds. Representative works include multi-party k-means schemes based on multi-key FHE, outsourced ciphertext k-means for distributed data, and fully outsourced clustering schemes using CKKS to enhance batch processing capabilities [1-4]. The third type of research focuses on privacy-preserving modeling related to financial risk control, such as credit scoring, federal credit analysis, and fraud detection. These schemes aim to ensure compliance while maintaining model feasibility.

In terms of algorithms, the privacy protection of k-means clustering focuses on three aspects. First, distance calculation involves squaring, multiplying and adding, comparing, and selecting the minimum value; none of these calculations can be directly handled by most homomorphic schemes. Second, cluster center updating involves counting and averaging all samples in a cluster, then taking

the minimum value to represent the cluster center. Third, clustering scenarios applied across different institutions, where vertical or mixed feature distributions generate significant costs for key management, ciphertext alignment, relinearization, and key rotation. To overcome these problems, new research uses batch encoding, vectorized packaging, and approximate computation to reduce the cost of each iteration, while employing an outsourced system architecture to reduce participant online time, thus making privacy-preserving clustering theoretically feasible and practically implementable.

While existing research has laid the foundation for this study, two shortcomings remain. First, most studies focus on the cryptographic protocols themselves, neglecting features such as heterogeneity, label delay, real-time policy linkage, and model interpretation that arise in financial risk control. Second, while existing schemes can achieve ciphertext clustering, they do not address the closed-loop application of clustering results to risk control decisions. Although business departments obtain many ciphertext cluster centers, they cannot rely solely on these centers to manage risk; instead, they should integrate risk clusters with credit granting, limit setting, monitoring, collection, and manual review processes. Therefore, it is necessary to propose a cross-institutional privacy-preserving k-means scheme that considers algorithms, systems, and business applications.

Table 1 Comparison of the capabilities of relevant programs in the past three years

Scheme	Core crypto	Cross-institution	Accuracy gap	Rounds	Bottleneck
MK-FHE KMeans	MK-FHE	Yes	Low	Medium	Ciphertext comparison
PPOKC	PHE+SS	Yes	Low	High	Interaction cost
COPPk-means	CKKS FHE	Yes	Very low	Low	Bootstrap overhead
SHE-KMeans	SHE	Single owner / cloud	Low	Low	Model update precision
Proposed	HE+batching	Yes	Controllable	Low	Key management

Table 1 summarizes and compares multi-key fully homomorphic clustering, distributed outsourced clustering, fully outsourced CKKS clustering, and the proposed scheme. Overall, in the past three years, the focus has shifted from "whether ciphertext k-means can be completed" to "how to reduce interaction rounds, improve batch processing capabilities, and control accuracy loss." Among these, the CKKS-based and batch encoding approach is more suitable for cross-institutional financial scenarios because risk control features are mostly standardized real-valued vectors, and approximate homomorphism is appropriate for such operations. However, key management and comparison operations remain bottlenecks in system operation.

3. Raise questions

Based on the above analysis, this paper identifies four main problems with cross-institutional privacy-preserving k-means clustering in financial risk control applications. First, cross-institutional samples cannot be directly aligned, thus precluding the use of traditional center initialization, distance calculation, and cluster allocation processes. Second, encrypted comparison and center updates are costly, and excessive interactions reduce the system's near real-time risk control capabilities. Third, clustering results need to be mapped to business strategies; if risk clusters are uninterpretable, it will be difficult to support credit approval, transaction monitoring, and collection decisions. Fourth, the system should be scalable and compliant, meeting the needs of multiple

parties while minimizing plaintext exposure. To address these issues, this paper proposes a cross-institutional privacy-preserving k-means framework for risk control.

$$J = \Sigma(i=1\rightarrow k)\ \Sigma(x\in Ci)\ ||x - \mu i||^2 \tag{1}$$

Equation (1) is the standard k-means objective function, which is to find the objective function that minimizes the sum of squared distances from samples within each cluster to the cluster center. In risk scenarios, the joint feature vector of customer, account, or transaction samples can be used as x, and the center of the corresponding cluster can be denoted as μi.

$$Jw = \Sigma(m=1\rightarrow M)\ \omega m\ \Sigma(i=1\rightarrow k)\ \Sigma(x\in Ci(m))\ ||x - \mu i||^2 \tag{2}$$

Equation (2) gives the cross-institutional weighted objective as the institutional weighted objective of M institutions, where M represents the number of participating institutions and ωm represents the credibility weight or sample quality weight of the m-th institution. The weights are used to reduce the impact of differences in sample size and feature quality among institutions on the results.

$$Enc(a)\ \oplus\ Enc(b) = Enc(a + b),\ Enc(a)\hat{}c = Enc(ca) \tag{3}$$

Equation (3) is a summary of the basic properties of additive homomorphism. For systems using Paillier or related additive homomorphism components, weighted summation and counting aggregation can be performed without decryption; while with CKKS, multiplication and addition operations on approximate real numbers can be implemented.

$$dij = \Sigma(l=1\rightarrow p)(xil - \mu jl)^2 = \Sigma xil^2 - 2\Sigma xil\mu jl + \Sigma\mu jl^2 \tag{4}$$

Equation (4) decomposes the squared distance between the sample and the center into three parts, performing vector multiplication and addition and batch summation respectively in the ciphertext domain. This decomposition reduces redundant computation and is the basis for reducing the overhead of homomorphic distance in this paper.

$$Rj = \alpha\cdot PDj + \beta\cdot OVDj + \gamma\cdot FRj + \delta\cdot EXj \tag{5}$$

The cluster-level risk score (Equation 5), which is the sum of default rate (PDj), delinquency intensity (OVDj), fraud correlation (FRj), and multiple application exposure (EXj) multiplied by their respective business weights (α, β, γ, δ), is used as the business risk score for the entire system. This score transforms the clustering results into interpretable and implementable risk control strategies.

4. Problem Solving/Strategies

4.1 System Overall Architecture

This paper proposes a system consisting of a data owner, a key service, ciphertext processing nodes, and risk control strategy nodes. Each institution completes sample standardization, feature alignment, and pseudo-identifier mapping locally, then encrypts and uploads the feature vectors using a homomorphic encryption scheme. The key management service is responsible for distributing and auditing public and private keys, rotation keys, and relinearization keys. The ciphertext computation nodes perform center initialization, ciphertext distance calculation, ciphertext cluster allocation, and center update. The strategy nodes only receive cluster-level statistics after threshold control and interpretation enhancement, without accessing the original plaintext samples. This structure adheres to the principle that "computation can be centralized, but data cannot," effectively meeting the deployment requirements of multi-institutional joint risk

control.

The algorithm uses local candidate center encrypted voting aggregation to generate the initial center, preventing any single institution from monopolizing the power of initialization. Then, the computing nodes calculate the distance from each sample point to each center point in the ciphertext according to equation (4), and select the nearest center using an approximate comparison protocol. Unlike the traditional plaintext implementation, this paper emphasizes the vectorization and packaging of samples in the same batch, mapping multiple distance components to the same ciphertext slot, thereby reducing the number of rotations and relinearizations. In the center update stage, the ciphertext and ciphertext count are calculated for each cluster using two methods: "intra-cluster sum + count". The results are then decrypted within the security boundary before updating the center. If the CKKS approximate homomorphic encryption algorithm is used, the numerical error can be controlled by scaling and resampling, keeping it within the acceptable range for the business.

In pre-loan scenarios, privacy-preserving k-means clustering can be used for cross-institutional customer segmentation, categorizing customers into low-risk and stable groups, medium-risk and volatile groups, high-risk and multi-risk groups, and suspicious and abnormal groups. Unlike traditional credit scoring cards that directly provide a single probability of default, clustering is more suitable for characterizing new types of risk patterns that have not yet been fully labeled. During the loan process, the system identifies abnormal customer groups based on characteristics such as fund flows, device behavior, geographical changes, and transaction frequency, providing real-time alerts for sudden cluster migrations. Differentiated collection methods are designed based on the characteristics of risk clusters: gentle reminders are used for customers with high exposure but still active assets, while increased manual review and transaction interception thresholds are applied to customers with high exposure and fraudulent connections.

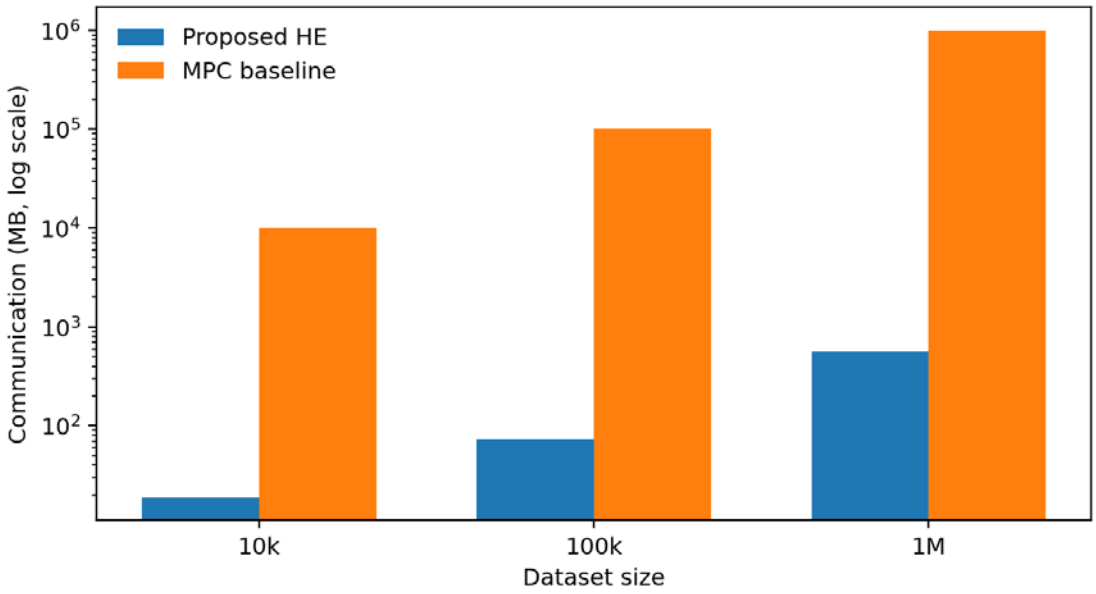

Figure 1. Comparison of communication overhead between the homomorphic scheme and the MPC baseline under different data volumes

Figure 1 is a redrawing of communication data from Mazzone et al.'s 2025 published paper. As the figure shows, as the sample size increases from 10,000 to 1 million, the communication volume of the homomorphic encryption-based scheme is significantly less than the traditional MPC baseline, and the difference widens with increasing sample size. For cross-institutional risk control, this

means that participants do not need to frequently exchange large amounts of intermediate results in each iteration, making it more suitable for joint modeling in wide area network environments. It also helps to retain audit traces and control network costs under regulatory requirements.

4.2 System Performance Analysis and Engineering Optimization

From a performance perspective, the overall cost of privacy-preserving k-means can be divided into four parts: encryption/decryption, ciphertext computation, network communication, and key management. Unlike solutions that pursue stronger cryptographic security alone, financial scenarios require a compromise solution that is both "secure enough and fast enough." Therefore, this paper proposes three engineering optimizations: first, using batch encoding to map the distances of multiple samples or clusters to the same ciphertext in parallel to improve single-operation throughput; second, using a central update cache, intra-cluster incremental statistics, and batch processing pipelined methods to reduce redundant computations; and third, the policy node only outputs cluster-level statistics and thresholded risk labels to prevent information leakage caused by fine-grained ciphertext decryption.

Simultaneously, key management and deployment flexibility must be considered. For joint risk control platforms with multiple participating institutions, an institution-level hierarchical key strategy can be adopted. Business institutions have local key shares, and the platform only handles orchestration and log storage. For high-concurrency end-of-day batch processing tasks, multi-queue task orchestration and elastic node scaling can be used to avoid concentrating ciphertext operations within a certain time window. In this case, ciphertext can be clustered and incorporated into the existing unified scheduling system of the risk control platform while ensuring that privacy boundaries are not violated.

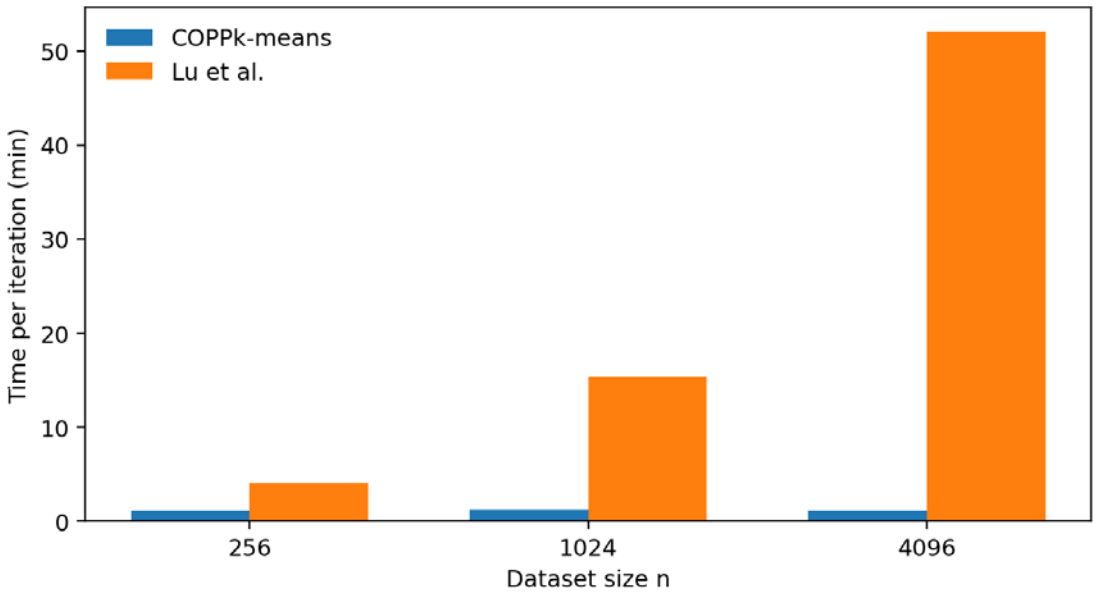

Figure 2 Comparison of single-round iteration time under different data sizes

Figure 2 is a reconstruction based on the single-round iteration time results reported by Yang et al. in their 2024 paper, comparing three sets of data with n=256, 1024, and 4096 when k=8. The results show that the single-round time of COPPk-means remains essentially constant with increasing sample size, but the time of the comparison scheme increases faster. This indicates that vectorized packaging and low-round outsourcing structures can effectively suppress the interference

caused by the increase in scale to the single-round latency. For risk control systems, this smoother latency curve is beneficial for end-of-day batch processing, hourly profile updates, and event-triggered review.

4.3 Risk Control Application Examples and Strategy Output

Table 2. Examples of Business Interpretation of Clustering Results in Risk Control Scenarios

Cluster	Default rate	Overdue days	Inquiry freq.	Suggested action
C1	Low	Short	Low	Fast approval
C2	Medium	Short	High	Limit control
C3	High	Long	Medium	Manual review
C4	High	Long	High	Alert and block

Table 2 shows the basic mapping method from clustering results to business strategies. Unlike supervised models that directly provide pass or reject, clustering results represent the risk group structure, making them more suitable for pre-stratification and decision support modules. In actual deployment, cluster-level risk scores Rj, along with existing scorecards, rule engines, and graph features, can be fed into a unified decision layer. For C1 clusters, automatic approval efficiency can be maintained; for C2 clusters, quota limits and transaction frequency monitoring measures should be implemented; and for high-risk C3 and C4 clusters, joint manual review, device verification, and list management should be implemented. On the one hand, the advantages of unsupervised clustering in discovering new patterns can be fully utilized; on the other hand, business actions cannot be directly determined by a single clustering result.

In terms of compliance, the value of this proposed solution lies not only in "encrypting, decrypting, and recompiling data," but also in redefining the boundaries of cooperation between various institutions. Participants can input their statistical information into the system, contributing statistical information while ensuring that control over the original data is not transferred. Through key separation and the use of audit logs, institutions can prove that they did not access sensitive fields without authorization during model training. Strategy improvements are based on group characteristics rather than individual plaintext data. Through cluster-level interpretation output, risk control departments can refine strategies based on group characteristics. This mechanism has significant practical value for cross-bank anti-fraud, joint credit granting, supply chain finance, and collaborative risk governance among platform merchants.

5. Conclusion

This paper studies a cross-institutional privacy-preserving k-means clustering algorithm based on homomorphic encryption and its application in risk control. Addressing the challenges of data sharing, model coordination, and easily interpretable results in cross-institutional joint risk control, the paper presents a comprehensive technical solution based on homomorphic encryption, with batch processing and low-round outsourced computation as the main technical approaches, and cluster-level risk interpretation as the final solution.

Research shows that the k-means algorithm, after encrypted distance decomposition, center update reconstruction, and system-level pipelined optimization, can complete cross-institutional clustering without exposing the original samples. According to publicly available research results from the past three years, schemes using CKKS, MK-FHE, and SHE can complete medium- to large-scale joint analysis within acceptable communication latency and range. In the financial sector, privacy-preserving k-means is more suitable for customer segmentation, anomaly detection, and

strategy-assisted decision-making, and can improve the ability to discover new risk patterns.

Further research can be conducted in three areas: first, combining clustering with graph learning and federated representation learning to improve the ability to identify complex association fraud; second, using differential privacy and security auditing mechanisms to improve the provable privacy of the output stage; and third, conducting more refined system stress testing and A/B verification based on real risk control processes, so that privacy-preserving clustering truly becomes a routine basic capability in the joint risk control platform.

References

[1] *Q. Xu, "Implementation of Intelligent Chatbot Model for Social Media Based on the Combination of Retrieval and Generation," 2025 2nd International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS), Hassan, India, 2025, pp. 1-7, doi: 10.1109/IACIS65746.2025.11210989.*

[2] *M. Zhang, "Research on Joint Optimization Algorithm for Image Enhancement and Denoising Based on the Combination of Deep Learning and Variational Models," 2025 International Conference on Intelligent Communication Networks and Computational Techniques (ICICNCT), Bidar, India, 2025, pp. 1-5, doi: 10.1109/ICICNCT66124.2025.11232800.*

[3] *Wu Y. Software Engineering Practice of Microservice Architecture in Full Stack Development: From Architecture Design to Performance Optimization[J]. 2025.*

[4] *Sun J. Quantile Regression Study on the Impact of Investor Sentiment on Financial Credit from the Perspective of Behavioral Finance[J]. 2025.*

[5] *Wang Y. Application of Data Completion and Full Lifecycle Cost Optimization Integrating Artificial Intelligence in Supply Chain[J]. 2025.*

[6] *Chen M. Research on Automated Risk Detection Methods in Machine Learning Integrating Privacy Computing[J]. 2025.*

[7] *Wu Y. Optimization of Generative AI Intelligent Interaction System Based on Adversarial Attack Defense and Content Controllable Generation[J]. 2025.*

[8] *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.*

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

[10] *Liu, H. (2026). Research on the Application of Causal Reasoning Method in Content Compliance Experimental Evaluation.*

[11] *Ding, J. (2025). Research On CODP Localization Decision Model Of Automotive Supply Chain Based On Delayed Manufacturing Strategy. arXiv preprint arXiv:2511.05899.*

[12] *Yu, X. (2025). Digital Transformation Empowers Growth Marketing with Marketing Data Analysis Integration and Real-Time Display Strategy.*

[13] *Yin, J. (2026). Research on Financial Time Series Prediction and Multiscale Correlation Based on the Fusion of Network Big Data and Deep Learning.*

[14] *Hou, Y. (2026). Research on BIOS and BMC Compatibility Optimization Methods for Cross-Generation Servers in Production Environments.*

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

[16] *Chang, Chen-Wei. "Compiling Declarative Privacy Policies into Runtime Enforcement for Cloud and Web Infrastructure." (2025).*

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

[18] Lu, Z. (2025). Design and Practice of AI Intelligent Mentor System for DevOps Education. European Journal of Education Science, 1(3), 25-31.

[19] Wu Y. Software Engineering Practice of Microservice Architecture in Full Stack Development: From Architecture Design to Performance Optimization[J]. 2025.

[20] 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.

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

[22] Liu, X., & Yang, D. (2025, March). LLM Data Strategy: Improving Data Availability and Efficiency. In Doctoral Symposium on Computational Intelligence (pp. 425-437). Singapore: Springer Nature Singapore.

[23] Zhang, C., Han, J., Zou, Y., Dong, K., Li, Y., Ding, J., & Han, X. (2024, April). Detecting the anomalies in LiDAR pointcloud. In WCX SAE World Congress Experience. SAE Technical Paper.

[24] 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.

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

[26] Zhang, Q. (2025, October). Application of Reinforcement Learning in Dynamic Advertising Content Generation. In 2025 2nd International Conference on Software, Systems and Information Technology (SSITCON) (pp. 1-5). IEEE.

[27] Wu, Y. (2025, October). Multi-Level Belief Rule Base Modeling Architecture and Intelligent Optimization Technology for Decision Support Systems. In 2025 2nd International Conference on Software, Systems and Information Technology (SSITCON) (pp. 1-8). IEEE.

[28] Huang, J. (2025, August). Research on Multi-Model Fusion Machine Learning Demand Intelligent Forecasting System in Cloud Computing Environment. In 2025 2nd International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS) (pp. 1-7). IEEE.

[29] Hou, Y. (2026). Research on Heterogeneous Server Upgrade Strategies and Resource Utilization Efficiency Oriented Toward Green Computing Objectives. Advances in Computer and Communication, 7(1).

[30] Yanchun Wang. (2025) Research on Enhancing ERP System Efficiency Through AI in Cross-border Supply Chain Environments. Advances in Computer and Communication, 6(5), 268-273.