Proceedings of CONF-FMCE 2025 Symposium: Semantic Communication for Media Compression and Transmission
DOI: 10.54254/2755-2721/155/2025.GL23426

Analysis of Defence Mechanisms for Security Federated Learning

Xinwei Mi

Faculty of Computer Science, Xi'an Shiyou University, Xi'an, China
202207070601@stumail.xsyu.edu.cn

Abstract: This review systematically examines security mechanisms in Federated Learning (FL), addressing the critical privacy-utility trade-offs in sensitive sectors such as healthcare, finance, and the Internet of Things (IoT). FL enables collaborative model training without raw data sharing, but its susceptibility to attacks like data poisoning and gradient leakage threatens its adoption in privacy-sensitive domains. The study evaluates four primary defence strategies: Secure Aggregation (SA), Differential Privacy (DP), Byzantine-robust aggregation, and Hybrid Frameworks, assessing them from theoretical and practical perspectives. SA utilises cryptographic techniques, such as Advanced Encryption Standard (AES)-256, to ensure privacy, while DP introduces noise to balance privacy protection with model utility. Hybrid methods combine these approaches, though they face challenges related to scalability. The review also identifies current research gaps, including the need for adaptive defences against evolving adversarial tactics and lightweight protocols suitable for resource-constrained IoT environments. By synthesising the strengths and limitations of these defence mechanisms, this study provides a roadmap for developing secure, scalable FL systems, contributing to advancing privacy-preserving technologies in critical sectors.

Keywords: Federated Learning, Security Mechanisms, Privacy-Utility Trade-offs, Hybrid Frameworks

1. Introduction

The rapid development of machine learning (ML) has revolutionised industries ranging from healthcare and finance to autonomous systems and smart cities. By leveraging huge datasets, ML models achieve unprecedented accuracy in disease diagnosis, fraud detection, and personalized recommendations. However, the traditional centralised ML paradigm requires aggregating raw data into a single repository, which exposes sensitive information to risks such as unauthorised access, data breaches, and non-compliance. These challenges are exacerbated by inter-agency cooperation. To reconcile the tension between data utility and privacy, Federated Learning (FL) was developed as a groundbreaking framework, enabling collaborative model training across multiple devices or institutions without sharing raw data [1]. This method can potentially revolutionise fields where data privacy and security are paramount, such as healthcare and finance. FL addresses growing concerns around data privacy by allowing models to be trained across decentralised devices while ensuring private data never leaves its source. For instance, hospitals can use FL to collaboratively develop cancer diagnosis models while keeping patient records secure and complying with regulations like

DOI: 10.54254/2755-2721/155/2025.GL23426

the Health Insurance Portability and Accountability Act (HIPAA) or the General Data Protection Regulation (GDPR) [2].

FL achieves privacy through a combination of cryptographic and statistical techniques. Secure Aggregation (SA), a cornerstone of FL, encrypts data during transmission to ensure that only aggregated information is shared among participants. This method prevents isolated access to sensitive data, reducing leakage risks [3]. For example, each hospital encrypts its model updates using cryptographic protocols like homomorphic encryption in a multi-hospital collaboration. The server then aggregates these encrypted updates without decrypting individual contributions, ensuring no single institution's data can be reconstructed. Differential Privacy (DP) complements this by adding controlled noise to model gradients or outputs, making it harder to extract private information [4]. For instance, in an FL system for predicting patient outcomes, DP might inject Gaussian noise into gradient updates, masking individual patient contributions while preserving the overall model's utility. However, this noise introduces a trade-off: while privacy guarantees improve, model accuracy can degrade, particularly in tasks requiring high precision, such as medical image segmentation [5]. Despite these challenges, SA and DP remain critical tools for privacy preservation in FL, each addressing different aspects of the privacy-utility spectrum.

However, FL faces significant security challenges that threaten its viability. As the number of participants in an FL system grows, so does the attack surface for adversaries. Data poisoning attacks, where malicious actors inject false or misleading data into the training process, pose a severe risk. For example, an attacker might subtly alter labels in a medical dataset (e.g., mislabeling benign tumours as malignant), causing the global model to generate inaccurate diagnoses [6]. In financial applications, adversaries could manipulate transaction records to evade fraud detection algorithms. Another critical threat is model inversion attacks, where attackers reverse-engineer shared model updates to infer private training data. These attacks highlight the vulnerability of FL to sophisticated adversaries, particularly in high-stakes domains like healthcare and finance, where data breaches could have catastrophic consequences [7]. Current research focuses on two primary areas: attack identification and defence development.

This paper reviews the challenges faced by FL, evaluates existing defences, and highlights unresolved issues critical for its future. A significant challenge is adapting defences to evolving attack strategies, such as adaptive poisoning attacks and collusion attacks by malicious participants [8]. Even without malicious intent, privacy breaches, such as gradient inversion and member inference attacks, remain a concern. Compromised centralised servers can also undermine trust by exposing model updates or manipulating results. This review focuses on four defence methodologies—Security Aggregation, Differential Privacy, Byzantine Robust Aggregation, and Hybrid Security Framework—evaluating their advantages, limitations, and applicability across various FL scenarios.

2. Methodology

2.1. Design of federated learning security mechanism

The design of FL security mechanisms must align with the specific characteristics and threat models of application scenarios. In privacy-sensitive areas like healthcare, finance, and the Internet of Things (IoT), unique data characteristics and security requirements present challenges for FL deployment. In healthcare, multi-institution collaboration improves model generalisation, but non-independent and identically distributed (Non-IID) data, such as inconsistent image annotations and diverse patient populations, pose difficulties. Furthermore, sensitive patient data makes gradient inversion attacks a significant threat, allowing attackers to infer private information from shared model updates. Data poisoning attacks can also undermine diagnostic model accuracy by label tampering [9,10]. In finance, extreme class imbalances and strict regulatory standards complicate FL deployment. For example,

DOI: 10.54254/2755-2721/155/2025.GL23426

credit card fraud detection suffers from insufficient sensitivity to minority samples, while financial data is vulnerable to reverse model attacks [11]. IoT environments face additional challenges with resource constraints and dynamic network conditions, risking Byzantine node poisoning attacks [12].

In summary, the FL security mechanism needs to address three common challenges: 1) privacy-utility trade-off: although encryption or noise injection protects privacy, it will lead to a significant decrease in model performance (Noise injection leads to a significant decrease in model accuracy, especially in medical imaging tasks that require high accuracy) [3,10,12]; 2) Dynamic adversarial adaptability: The continuous evolution of attack strategies (such as gradient cracking driven by quantum computing) requires the defense mechanism to have real-time response and self-optimization capabilities [6]; 3) Heterogeneous device scalability: Resource-constrained IoT devices are difficult to carry the high computing overhead of hybrid frameworks, and lightweight protocol design is urgently needed [9,11]. These challenges provide a practice-oriented entry point for the theoretical optimisation and technical implementation of the follow-up defence mechanism.

2.2. Proposed analytical framework

This review adopts a structured approach to evaluate FL security, organised into four thematic pillars (shown in Figure 1).

The threat taxonomy of Artificial Intelligence (AI) involves classifying attack vectors, such as data poisoning and model inversion, across various stages of the lifecycle, including data collection, aggregation, and deployment, while considering adversarial capabilities in both white-box and black-box settings. A critical examination of defence mechanisms reveals various cryptographic, statistical, and hybrid strategies, focusing on their theoretical foundations and practical limitations. Evaluating the trade-offs between privacy guarantees, computational overhead, and model utility is crucial, as these factors vary across defence paradigms. Additionally, research gaps remain in addressing challenges like dynamic adversarial adaptation and scalability issues in IoT environments.

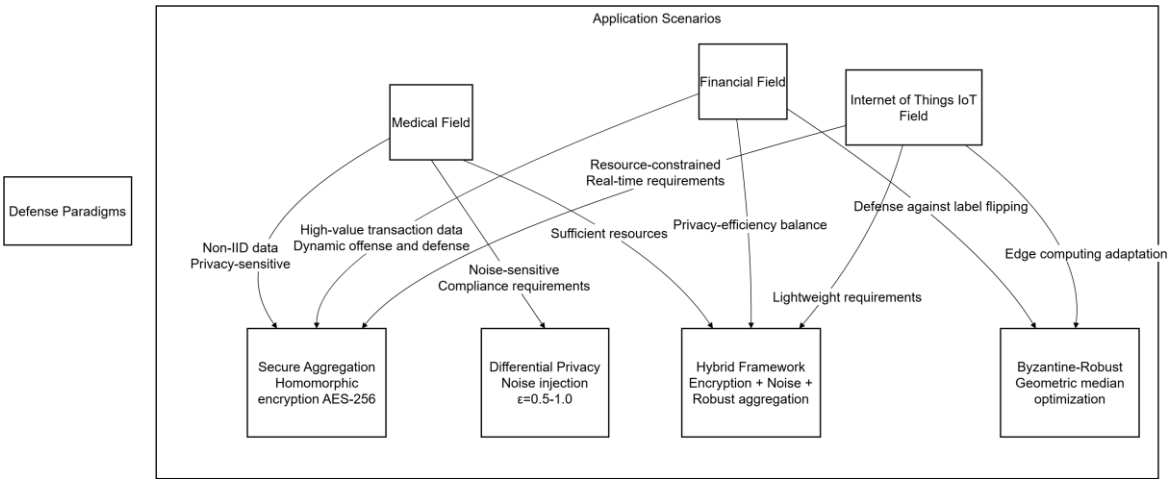

Figure 1: Framework for analysing FL security mechanisms (picture credit: original)

2.2.1. Secure Aggregation

As a core privacy-preserving technology in FL, SA has rapidly evolved from theory to practice (Figure 2). Its fundamental principle involves encrypting local model updates on clients using cryptographic methods such as secret sharing or homomorphic encryption, ensuring the server can only access the aggregated global model result while remaining oblivious to individual client data. Specifically, each client's model update wi is encrypted as a ciphertext $cici$, and homomorphic

DOI: 10.54254/2755-2721/155/2025.GL23426

encryption enables direct additive operations on these ciphertexts, ultimately yielding the aggregated global model update. This approach safeguards data privacy and addresses practical challenges like client dropouts.

Regarding system robustness, 2025 research has advanced dynamic client participation mechanisms. To mitigate aggregation failures caused by client disconnections, new schemes introduce redundant sharding techniques based on threshold cryptography, enabling successful aggregation even when 30% of clients fail. Furthermore, domestic scholars in Computer Research and Development highlight that hybrid encryption strategies combining gradient sparsification and secret sharing reduce computational latency to 1.2 times that of plaintext schemes, significantly outperforming early DP approaches [11].

Future research on SA will emphasise multi-server architectures and post-quantum cryptography. As quantum computing threats escalate, lattice-based key exchange mechanisms have been integrated into the latest FL frameworks. These advancements underscore the rapid evolution of SA toward efficiency, verifiability, and quantum resistance.

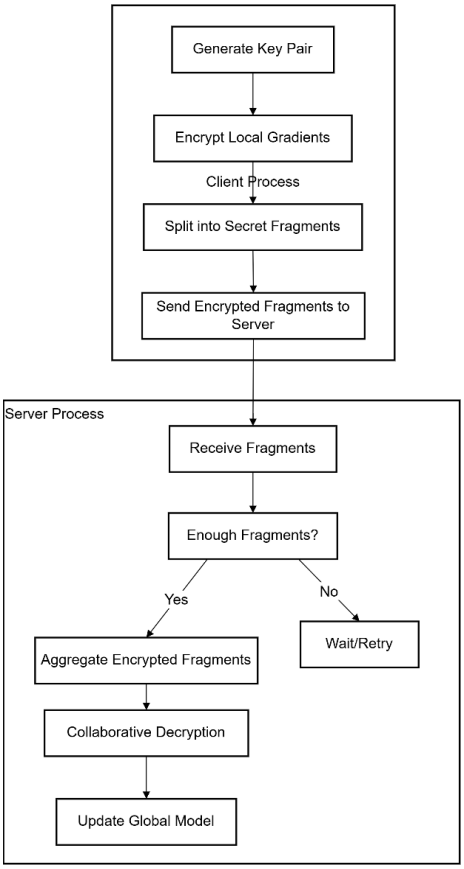

Figure 2: Secure Aggregation flow chart (picture credit: original)

2.2.2. Differential Privacy

DP, introduced by Dwork in 2006, has emerged as a cornerstone of privacy protection (Figure 3). Its core principle is rigorously ensuring that individual privacy cannot be inferred during data analysis through mathematically grounded noise injection mechanisms. Specifically, DP requires that an algorithm's output distribution remains statistically indistinguishable even when minor changes occur.

In recent years, DP has evolved through its integration with FL. A 2024 study proposed the Personalised Local Differential Privacy (PLDP) framework, which allows users to customise privacy

DOI: 10.54254/2755-2721/155/2025.GL23426

levels and balances privacy-utility tradeoffs via dynamic noise-threshold mechanisms [12]. For instance, noise injection is triggered only when client participation rates exceed predefined thresholds during fluctuations, significantly reducing computational overhead. Additionally, quantum-resistant DP has emerged as a frontier, leveraging lattice-based cryptographic protocols to defend against quantum computing attacks, offering future-proof privacy solutions.

Looking ahead, DP will deepen its integration with AI governance, focusing on automated privacy budget allocation, privacy-preserving generative models, and hybrid frameworks combining DP with secure multi-party computation. These advancements will further solidify DP's foundational role in data-driven societies.

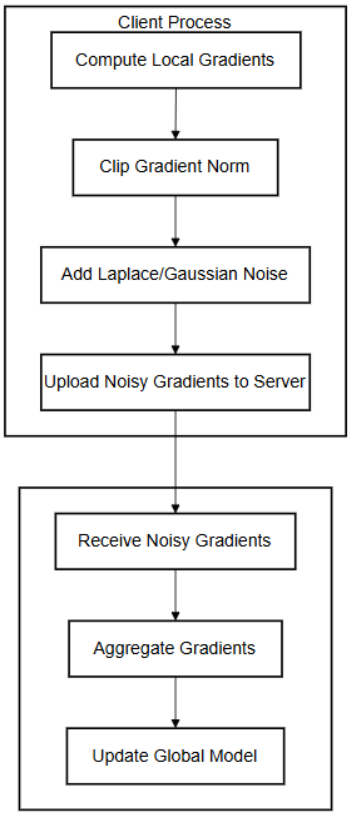

Figure 3: Differential Privacy flow chart (picture credit: original)

2.2.3. Byzantine-Robust aggregation

Byzantine-robust aggregation is a core technology for defending against malicious node attacks in FL (as shown in Figure 4). Early methods, such as the Krum algorithm, resist outliers by selecting candidate models with gradients closest to other nodes, albeit with lower computational efficiency. Subsequent research proposed optimisation strategies based on regularised objective functions, where constraint terms were introduced into global model updates to enforce parameter alignment with most nodes, thereby enhancing robustness. Recently, dynamic adaptive methods have become mainstream, exemplified by approaches that design robustness metrics and adaptive weight allocation mechanisms to address complex scenarios involving data heterogeneity and high proportions of Byzantine nodes.

Current research focuses on multidimensional collaborative optimisation: dynamic strategies adjust aggregation rules in real-time via reinforcement learning to counter external attacks, privacy-preserving mechanisms integrate DP with robust aggregation, and decentralised architectures reduce

DOI: 10.54254/2755-2721/155/2025.GL23426

reliance on central servers through distributed consensus protocols. For example, matrix mapping-based aggregation algorithms resist attacks by monitoring the consistency of user models with historical states. Such methods significantly improve system robustness by quantitatively analysing the behavioural patterns of Byzantine nodes. These directions collectively drive Byzantine-robust aggregation toward more efficient and secure goals.

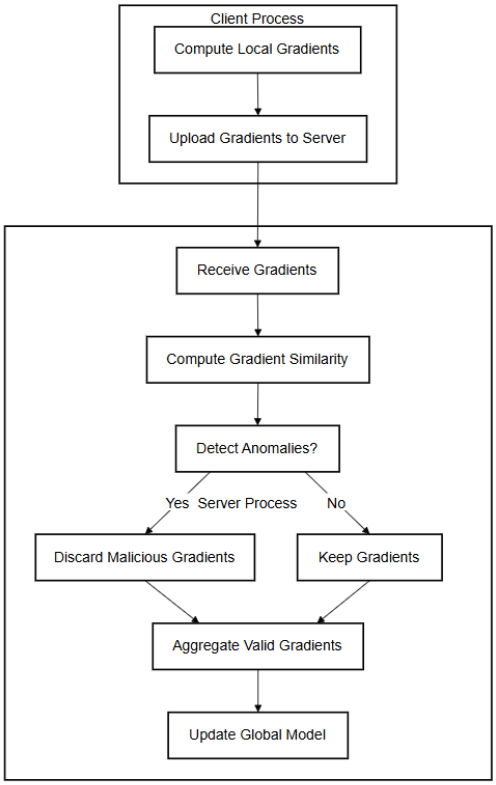

Figure 4: Byzantine-Robust aggregation flow chart (picture credit: original)

2.2.4. Hybrid defence paradigms

Hybrid defence paradigms in FL address privacy and security challenges by integrating cryptographic techniques, DP, and system-level strategies (as shown in Figure 5). As a decentralised machine learning framework, FL enables collaborative model training while keeping data localised. However, its distributed nature introduces risks such as model poisoning, gradient leakage, and backdoor attacks. Early defences relied on data isolation and basic encryption, whereas modern hybrid approaches balance privacy and model utility through dynamic noise injection, blockchain-audited aggregation, and robust optimisation.

Recent advancements include blockchain-audited aggregation 99, which enhances trust by recording model updates in an immutable ledger, alongside contrastive learning-based methods to detect backdoor patterns by comparing model embeddings. Cross-domain governance frameworks integrate IoT device security with FL model protection to achieve end-to-end resilience. Future research directions include quantum-resistant encryption to counter post-quantum threats and meta-learning-driven noise scaling to optimise privacy-utility trade-offs dynamically [9]. These innovations aim to unify the scalability of decentralized learning with robustness against evolving attacks, promoting FL adoption in sensitive domains like healthcare and finance.

DOI: 10.54254/2755-2721/155/2025.GL23426

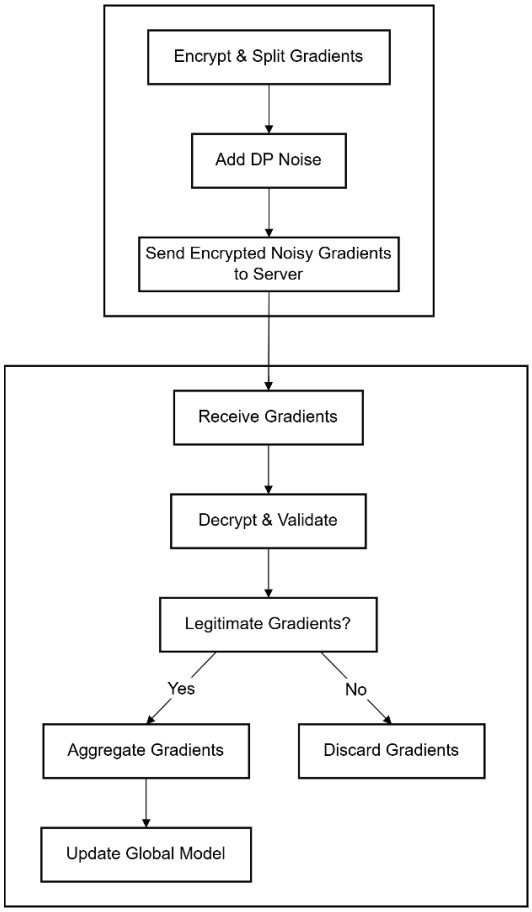

Figure 5: Hybrid defense paradigm flow chart (picture credit: original)

3. Results and discussion

3.1. Results analysis

Overall, FL's four major security mechanisms show significant differences in privacy protection strength, anti-attack ability, and computing efficiency, and their applicability needs to be evaluated in combination with the scenario's characteristics.

SA: Homomorphic encryption is used to ensure the privacy of gradient transmission of medical cross-institutional collaboration, but encrypted computing may cause a delay, which makes it challenging to meet the real-time needs of the Internet of Things, and relying on trusted servers may lead to the risk of single point of failure [2,11];

DP: mathematically provable privacy budget (ε=0.5-1) Defend against member inference attacks in financial scenarios, but noise injection reduces the accuracy of medical image segmentation, and privacy budget allocation relies on human experience [3];

Byzantine Robust Aggregation: 30% of malicious nodes are tolerated through geometric median optimization, which is suitable for financial collaborative defence against poisoning attacks, but the false positive rate increases under medical non-independent identical distribution data, and the computational complexity increases exponentially with node size [7];

Hybrid framework: Combining encryption and noise technology to balance privacy-utility (e.g., 78% accuracy + ε=1.0 privacy protection in medical scenarios), the problem is that the computational overhead is 1.2 times that of the plaintext solution, and it needs to rely on hardware acceleration to adapt to IoT devices [9,11].

DOI: 10.54254/2755-2721/155/2025.GL23426

In summary, the performance boundary of each mechanism is determined by the needs of the scenario: medical priority privacy intensity, financial focus on dynamic offensive and defensive confrontation, and the Internet of Things needs to balance security and efficiency.

3.2. Discussion

The core challenge in FL security lies in balancing privacy, utility, and efficiency, which creates multi-dimensional trade-offs across applications. For instance, while SA prevents data leakage via homomorphic encryption, it fails to counter model-level reverse attacks (e.g., gradient inversion) [3,12]. DP mitigates such attacks through noise injection but sacrifices model accuracy, particularly in precision-critical tasks like medical imaging [3,12]. Hybrid frameworks partially address these issues via gradient sparsity, yet remain dependent on scenario-specific tuning rather than universal solutions [11].

Efficiency-security conflicts further complicate FL deployment. Due to exponential computational complexity, Byzantine-robust aggregation filters malicious nodes but scale poorly in large Iot networks. Lightweight protocols (e.g., gradient compression) reduce overhead but inadvertently expose transaction patterns in financial fraud detection [9,11], underscoring the tension between resource constraints and security.

Key unresolved challenges persist:

Dynamic adversarial adaptation: Due to incompatible cryptography, existing defences struggle against evolving attacks, such as quantum computing-enhanced gradient inversion [6].

Cross-domain governance: Heterogeneous regulations lack interoperable frameworks for healthcare, finance, and IoT [9,11].

Lightweight-strong security compatibility: Optimising encryption and aggregation for resource-limited Iot requires hardware acceleration or neural architecture search (NAS) [9,11].

Addressing these challenges demands a shift from isolated optimizations to systematic innovation. Dynamic defence models must be integrated with hardware-aware protocol engineering to enable FL's scalable adoption in privacy-sensitive domains.

4. Conclusion

This study provides a comprehensive evaluation of security mechanisms in FL, emphasising the critical balance between data privacy and model utility in sensitive domains like healthcare, finance, and Iot. By systematically analysing defence strategies against emerging threats, such as data poisoning, model inversion, and Byzantine attacks, the research offers valuable insights for developing secure and regulatory-compliant FL systems. Future research will focus on adaptive defence strategies to counter evolving adversarial tactics, including collusion attacks and threats posed by quantum computing.

References

[1] *Sun, Y., Ochiai, H., Esaki, H. (2021). Decentralized deep learning for multi-access edge computing: A survey on communication efficiency and trustworthiness. IEEE Transactions on Artificial Intelligence, 3(6), 963-972.*

[2] *Zhu, L., Tang, X., Shen, M., et al. (2021). Privacy-preserving machine learning training in IoT aggregation scenarios. IEEE Internet of Things Journal, 8(15): 12106-12118.*

[3] *Noble, M., Bellet, A., Dieuleveut, A. (2022). Differentially private federated learning on heterogeneous data. International conference on artificial intelligence and statistics. PMLR, 10110-10145.*

[4] *Gong, X., Chen, Y., Wang, Q., et al. (2022). Backdoor attacks and defenses in federated learning: State-of-the-art, taxonomy, and future directions. IEEE Wireless Communications, 30(2), 114-121.*

[5] *Wei, W., Liu, L. (2021). Gradient leakage attack resilient deep learning. IEEE Transactions on Information Forensics and Security, 17, 303-316.*

DOI: 10.54254/2755-2721/155/2025.GL23426

[6] *Qammar, A., Ding, J., Ning, H. (2022). Federated learning attack surface: taxonomy, cyber defences, challenges, and future directions. Artificial Intelligence Review, 55(5), 3569-3606.*

[7] *Park, S., Choi, W. (2022). Byzantine fault tolerant distributed stochastic gradient descent based on over-the-air computation. IEEE Transactions on Communications, 70(5), 3204-3219.*

[8] *Tripathy, P. K., Shrivastava, A., Agarwal, V., et al. (2024). Federated learning algorithm based on matrix mapping for data privacy over edge computing. International Journal of Pervasive Computing and Communications, 20(5), 633-647.*

[9] *Li, X., and Wang, Y. (2022). Blockchain-audited hybrid defense for financial federated learning. IEEE Trans. Dependable Secure Comput, 19(3), 1124–1136.*

[10] *Zhao, Y., Chen, J. (2022). A survey on differential privacy for unstructured data content. ACM Computing Surveys (CSUR), 54(10s), 1-28.*

[11] *Yang, Q., and Liu, Y. (2024). VFLAIR: A vertical federated learning framework with robustness evaluation. ACM SIGSAC Conf. Secure Mach. Learn. (SecML), 45–58.*

[12] *Guo, S., Wang, X., Long, S., et al. (2023). A federated learning scheme meets dynamic differential privacy. CAAI Transactions on Intelligence Technology, 8(3), 1087-1100.*