FedPrivEngine: A Federated Distributed Data Engineering Framework for Privacy-Preserving Analytics in Healthcare and Finance

Srinivas Lakkireddy

Independent Researcher, Buffalo Grove, USA
Email: reachlakkireddy@gmail.com

***Abstract*—Privacy-preserving analytics is essential in healthcare and financial domains where sensitive data must comply with regulations such as the General Data Protection Regulation (GDPR), the Health Insurance Portability and Accountability Act (HIPAA), and the Payment Card Industry Data Security Standard (PCI-DSS). Although Federated Learning (FL) enables collaborative model training without sharing raw data, existing frameworks often lack compliance-aware orchestration, efficient encrypted query execution, domain flexibility, and scalability under Homomorphic Encryption (HE). This paper presents FedPrivEngine, a federated and distributed data engineering framework designed for regulation-aware analytics across institutional silos. The framework integrates Homomorphic Encryption (HE), optional Differential Privacy (DP), Spark-based distributed orchestration, and a compliance-aware query planner to enforce policy constraints during federated execution. Two domain-specific models are implemented: HealthPrivNet, a Gated Recurrent Unit (GRU)-based model evaluated on the Open Safely healthcare dataset, and FinanceRiskNet, a Multi-Layer Perceptron (MLP)-based model assessed on the FICO HELOC credit risk dataset. Experimental results show that FedPrivEngine achieves high classification accuracy (93.8% in healthcare and 91.2% in finance), ensures zero policy violations, and supports encrypted query execution with success rates above 98%, while maintaining sublinear scalability and moderate system overhead. These results demonstrate the feasibility of secure, accurate, and regulation-aware federated analytics in policy-sensitive environments.**

***Keywords*—Federated Learning (FL), privacy-preserving analytics, Homomorphic Encryption (HE), compliance-aware orchestration, healthcare and finance**

I. Introduction

Data-driven decision-making is the norm nowadays, and privacy-preserving analytics has become a significant concern across domains such as healthcare and finance, where regulatory transparency requirements affect sensitive information. The presence of traditional centralized Machine Learning (ML) systems poses considerable risks, e.g., data leaks, data breaches, and non-compliance with General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA). A promising solution to this dilemma is federated learning, which enables collaborative model training and testing over little or no raw data sharing [1]. Yet, most existing frameworks either target a single domain, such as oblivious random-access memory, fail to support runtime compliance guarantees, or exhibit limited scalability under encrypted execution constraints, prohibiting widespread real-world adoption.

Several privacy-enhancing techniques have been recently proposed in the literature, such as homomorphic encryption [2], differential privacy [3], and blockchain-based access control [4]. Although these approaches provide theoretical guarantees, their limited integration with scalable orchestration engines and in-situ policy validation significantly limits their practical applicability, especially in a multi-institution setting. Also, many frameworks are either limited to training models or restrict query functionality, but do not provide an integrated system for training and executing queries over a private channel.

Recent work shows that federated learning is increasingly evident in the development of innovative healthcare, with partnerships combining IoT for data gathering and distributed clinical infrastructures. These systems emphasize the promise and the fundamental problems of performing privacy-preserving analytics across disparate biomedical data sources. Simultaneously, explainable AI has emerged as an essential component of clinical decision support systems to ensure that clinical model predictions are interpretable and trustworthy, particularly in federated environments where data locality prevents centralized interpretability of the models. This example explains why federated analytics frameworks that combine formal privacy guarantees with regulatory compliance and interpretable model execution are precisely what is needed.

In this work, we aim to fill these gaps by introducing a new framework, FedPrivEngine, for federated distributed data engineering that enables encrypted, compliant, and scalable analytics in healthcare and finance. It combines

Manuscript received September 10, 2025; revised September 15, 2025; accepted January 23, 2026; published May 22, 2026.

Spark-based orchestration with both homomorphic encryption and a compliance-aware query planner. We develop two domain-specific models: HealthPrivNet, for temporal healthcare analytics based on GRUs, and FinanceRiskNet, for credit risk prediction based on multi-layer perceptrons. It provides high prediction accuracy, enables secure fusion of multiple models, query execution over models in encrypted form, and includes jurisdiction-aware data governance mechanisms.

In summary, this paper offers a perspective on a system-based federated analytics framework rather than novel learning algorithms. FedPrivEngine serves as an example of how disparate technologies—federated learning, homomorphic encryption, optional differential privacy, and Spark-based orchestration—can be brought together within a cohesive architecture to enable encrypted analytics targeting regulated industries. It is built to support policy-constrained execution and encrypted model sharing between federated silos, while preserving domain-specific learning flows. The healthcare and financial models introduced are two golden use cases to demonstrate the applicability of the framework in dealing with heterogeneous analytical workloads. Experimental results on the benchmark database show the rationale and performance trade-offs underlying the proposed system design, rather than providing algorithmic or regulatory completeness.

The rest of the paper is organized as follows. Section II presents a literature review of federated learning and privacy-preserving analytics frameworks. Section III describes the proposed FedPrivEngine methodology, including the system architecture, model configurations, and algorithms; Section IV details the experimental results, covering model performance, privacy metrics, and scalability analysis. Section V discusses key findings and outlines the limitations of the current study. Finally, Section VI concludes the paper and highlights potential directions for future research.

II. Related Work

The convergence of federated learning and privacy-preserving techniques has emerged as a critical research area across multiple domains, driven by the need to balance collaborative analytics with stringent data protection requirements. This section examines recent advances in privacy-preserving federated analytics, organised by technical approach and application domain, and highlights both achievements and persistent challenges that motivate the present work.

A. Federated Learning in IoT and Edge Computing Environments

The proliferation of Internet of Things (IoT) devices has created unprecedented opportunities for distributed learning while simultaneously intensifying privacy concerns. Recent research has demonstrated that federated learning can effectively address these challenges across diverse IoT contexts. Privacy-enhanced frameworks for IoT systems have successfully integrated cryptographic techniques with fault-tolerant mechanisms, as shown by Awan *et al.* [1] through FedSim simulations. However, future scalability and trust automation remain open challenges. Building on these foundations, bandwidth-efficient approaches have been developed for intelligent building energy forecasting. Khalil *et al.* [5] achieved high accuracy using differential privacy mechanisms, tested on Pecan Street data, while maintaining communication efficiency. However, these solutions typically target specific IoT subdomains and have not been extensively validated in large-scale heterogeneous infrastructures.

In industrial IoT settings, intrusion detection systems have leveraged federated learning with differential privacy to preserve accuracy while protecting sensitive operational data. Ruzafa-Alcázar *et al.* [6] introduced the Fed+ framework, demonstrating that gradient compression combined with customized privacy mechanisms could maintain detection performance on benchmark datasets like ToN_IoT. However, gradient compression strategies and adaptive privacy tuning were identified as areas requiring further investigation. Complementing these advances, the importance of protecting model intellectual property has been recognized. Yang *et al.* [2] redefined Secure Federated Learning (SFL), introduced FedIPR for model ownership protection, and surveyed threats and defences while highlighting upcoming issues with IP rights and deployment scalability that require theoretical and practical resolution.

Edge computing environments have benefited from secure federated pipelines that combine distributed communication frameworks with encryption and differential privacy. Yan *et al.* [7] developed the FedLabX framework using Kafka, encryption, and differential privacy, demonstrating increased privacy protection and improved accuracy in simulated environments. However, limitations in handling non-IID data distributions and the need for strengthened security mechanisms against sophisticated attacks were acknowledged as critical areas for future development. Addressing similar challenges in distributed settings, Wang *et al.* [8] presented a differential privacy-based federated XGBoost framework with a novel two-stage protection technique that achieves substantial accuracy while maintaining privacy guarantees. However, indirect leakage avoidance and multi-node synchronization remain focal points for future work.

Advanced cryptographic approaches for IoT security have also been proposed. Jalali and Chen [9] introduced a federated learning framework that integrates Cheon–Kim–Kim–Song (CKKS)–based homomorphic encryption with the improved Transport Layer Security (iTLS) protocol to enable secure encrypted model aggregation while improving communication efficiency in IoT environments. Expanding the application scope, Hwang *et al.* [10] explored privacy-preserving personal identification using federated learning with multimodal vital signs data, achieving competitive accuracy while improving privacy. Their work highlighted opportunities to incorporate additional biometrics, larger datasets, and more advanced federated learning customization for practical deployment scenarios.

The intersection of IoT and anomaly detection has received significant attention through multi-layered security approaches. Arazzi *et al.* [11] proposed a fully privacy-preserving solution for IoT anomaly detection that combined federated behavioral fingerprinting with blockchain and homomorphic encryption, achieving 85% accuracy while identifying scalability and workload optimization as critical areas for improvement. Similarly, wireless sensor network security has advanced through novel architectural designs. Bukhari *et al.* [12] presented an FL-SCNN-Bi-LSTM model for WSN intrusion detection, tested on WSN-DS and CIC-IDS-2017 datasets, achieving approximately 99.9% accuracy while acknowledging that future work must concentrate on scalability and enhanced privacy mechanisms. Furthermore, Truong and Le [13] introduced MetaCIDS, a federated intrusion-detection framework for the metaverse leveraging blockchain technology, achieving up to 99% accuracy while noting that future research must focus on understanding how different attack patterns across devices affect detection performance.

Aggregation efficiency in federated systems has also been explored to reduce computational costs and improve resilience. Eltaras *et al.* [14] proposed an efficient verifiable protocol for privacy-preserving aggregation that reduces costs and enhances dropout resilience by leveraging auxiliary nodes. However, security enhancements and scalability improvements remain priorities for future research. These collective efforts in IoT and edge computing have made substantial progress in privacy-preserving federated learning but have also revealed persistent challenges in scalability, handling non-IID data, and integrating multiple privacy mechanisms within unified frameworks.

B. Healthcare Applications and Clinical Decision Support Systems

Healthcare has emerged as a particularly critical domain for privacy-preserving federated analytics, given stringent regulatory requirements and the sensitivity of patient data. Federated deep learning architectures have been specifically designed to address medical imaging challenges while preserving patient privacy. Riedel *et al.* [3] introduced ResNetFed, a federated model for COVID-19 detection from chest radiographs that preserves privacy through secure aggregation. Tested using COVIDx8 data, the model outperformed local training approaches while identifying Non-IID optimization, interpretability, and scalability as key areas for future development. The importance of interpretability in federated healthcare settings has been further emphasized by meta-analytical research demonstrating that transparency, usability, and trustworthiness constitute central success factors for the clinical adoption of AI-based decision systems, particularly when models operate across institutional boundaries in restricted data-access environments [15].

The telemedicine domain has witnessed growing interest in combining federated learning with blockchain technology for secure data management. Hiwale *et al.* [16] conducted systematic reviews of the integration of federated learning and blockchain for safe telemedicine across multiple publications, identified primary privacy issues, proposed unified frameworks, and emphasized the need for more sophisticated, scalable privacy-preserving solutions. This dual exploration underscores the persistent challenges in deploying federated telemedicine systems that balance security, privacy, and operational efficiency. The complexity of healthcare data protection has motivated research into complementary privacy-preserving approaches that operate at different system layers.

Medical data security has been enhanced through steganography-based techniques that provide additional confidentiality during storage and transfer by embedding sensitive health information into carrier signals. Riaz *et al.* [17] demonstrated robust steganographic techniques for protecting medical records in healthcare informatics. While such methods enhance data-level privacy, they primarily address storage and transmission security rather than mitigating privacy risks during collaborative model training or distributed analytics. These approaches are complementary to computation-level protections like homomorphic encryption and federated learning, suggesting that future hybrid systems may combine steganographic data protection with encrypted federated analytics to achieve multi-layered privacy preservation across institutional boundaries.

Privacy-preserving data mining frameworks have also been developed specifically for healthcare informatics. Darwish *et al.* [18] proposed a genetic-Tabu-based metaheuristic method for data sanitization to protect the privacy of medical records, demonstrating superior performance compared to existing techniques while identifying other privacy mechanisms and dataset variations as targets for future investigation. Domain adaptation techniques have emerged to enable knowledge transfer across healthcare institutions without compromising privacy. Song *et al.* [19] introduced a privacy-preserving unsupervised domain adaptation technique using homomorphic encryption and federated learning, demonstrating secure knowledge transfer without sacrificing accuracy on unsupervised tasks. However, efficiency improvements remain a focus for future research.

The genomic medicine domain presents unique privacy challenges due to the highly sensitive and permanent nature of genetic information. Kuo *et al.* [20] investigated genome privacy through safe protocols for genome analysis combined with blockchain-based access recording, demonstrating workable, secure techniques for genetic data management while identifying scalability and efficiency as primary areas for future research. Healthcare record linkage across institutions has also been addressed through privacy-preserving frameworks. Kiernan *et al.* [21] established a framework for privacy-preserving record linkage between electronic health records and administrative claims within a national clinical research network, achieving effective patient deduplication while noting that future work should focus

on matching accuracy, token flexibility, and integration with broader networks.

Homomorphic encryption has received particular attention for its potential to enable computation on encrypted healthcare data. Munjal and Bhatia [22] examined the development of homomorphic encryption, particularly Fully Homomorphic Encryption (FHE), classifying various schemes and assessing them using medical data such as electronic health records. Their systematic review indicated that improvements in privacy, scalability, and computational efficiency are priorities for future healthcare research. These healthcare-focused efforts collectively demonstrate significant progress toward privacy-preserving clinical analytics, yet highlight persistent challenges in achieving practical deployment that balances regulatory compliance, computational efficiency, and clinical utility across diverse institutional settings.

C. Smart Grid and Energy System Applications

Energy systems and smart grids represent another critical domain where federated learning intersects with privacy preservation, driven by the need to protect sensitive consumption data while enabling system optimization. Comprehensive reviews have examined federated learning applications in power systems, covering algorithms, applications, and implementation challenges. Yan *et al.* [23] proposed the FedLabX framework, which integrates Kafka-based communication, encryption, and differential privacy for federated learning in edge and Internet of Things environments, emphasizing federated learning's privacy and efficiency advantages for power system services, and identified aggregation optimization, communication efficiency, and security enhancements as focal areas for further research. Building on these foundations, we have systematically surveyed deep learning applications in smart grids to identify emerging trends and gaps. Massaoudi *et al.* [24] highlighted distributed deep learning, edge intelligence, and federated learning in their review of deep learning applications in innovative grid technology, noting existing drawbacks and recommending future research on explainable deep learning models to enhance trust and interpretability.

Renewable energy forecasting has particularly benefited from federated and distributed learning approaches. Massaoudi *et al.* [25] examined deep learning approaches for Photovoltaic (PV) power forecasting, emphasizing hybrid models and cutting-edge methods such as transfer learning and federated learning. Their review identified accuracy improvements and real-world deployment capabilities as primary objectives for future research, recognizing that practical implementation across distributed generation sites remains challenging. Anomaly detection in smart grids has leveraged blockchain-based reinforcement learning to enhance security while preserving privacy. Belhadi *et al.* [26] introduced ITSA, a blockchain-based reinforcement learning framework for intelligent grid anomaly detection that outperformed baseline approaches on the CASAS dataset. However, optimization for heterogeneous operational contexts was identified as a priority for future work.

Privacy-preserving distributed optimization has emerged as a fundamental requirement for multi-party cooperation in power systems. Tian *et al.* [27] proposed a privacy-preserving distributed optimization framework based on Secure Multi-Party Computation (SMPC) for power systems, demonstrating scalability while providing optimal solutions and protecting participant privacy. Hiwale *et al.* [16] enhancing convergence speed and scalability for massive interconnected systems were identified as primary goals for future research. Shen *et al.* [28] these innovative grid applications illustrate how federated learning can enable collaborative optimization in critical infrastructure while maintaining data sovereignty. However, practical deployment at utility scale continues to face challenges related to communication overhead, heterogeneous system integration, and real-time operational requirements.

D. Financial Services and Fraud Detection

Financial services have increasingly adopted privacy-preserving federated learning to enable collaboration while protecting sensitive customer information and maintaining regulatory compliance. Credit card fraud detection systems have particularly benefited from federated approaches that allow financial institutions to collaborate without sharing transaction data. Tang and Liu [29] developed a federated Structured Data Transformer (SDT) for credit card fraud detection, achieving strong AUC scores across two datasets and noting that real-world validation across multiple financial institutions is critical future work to assess practical deployment feasibility. Recent investigations have explored deep hybrid models to enhance fraud detection by leveraging sophisticated temporal and behavioral pattern learning. Jabeen *et al.* [30] demonstrated that deep hybrid CLST structures can significantly improve credit card fraud detection by learning complicated behavioral and temporal patterns. While these centralized hybrid models have shown superior predictive quality, they typically operate on raw financial data without federated aggregation, highlighting opportunities for future architectures that combine representational richness with privacy-preserving federation and regulatory compliance.

Privacy-preserving data mining techniques have also been developed for heterogeneous financial and governmental data sources. Lee and Jun [31] introduced a method that balances privacy and utility when mining heterogeneous open government data through micro-aggregation and distance-based record linkage without revealing individual identities. Their work identified model training integration and decision-maker support as areas requiring further development. These financial domain applications demonstrate that federated learning can enable secure cross-institutional collaboration while preserving customer privacy. However, practical deployment continues to face challenges related to regulatory compliance, model interpretability requirements, and the integration of sophisticated architectures within privacy-constrained frameworks.

E. Cryptographic Primitives and Secure Computation

The foundational cryptographic techniques underlying privacy-preserving federated learning have advanced significantly, particularly in homomorphic encryption and secure computation protocols. FHE has been extensively studied for its potential to enable arbitrary computation on encrypted data. Pulido-Gaytan *et al.* [32] and Shen *et al.* [33] examined developments in FHE for privacy-preserving neural networks, contrasting implementations like CryptoNets and LoLa while discussing prospects, challenges, and potential future research directions. Their analysis revealed that, despite theoretical elegance, practical deployment incurs substantial computational overhead, limiting real-world applicability. Addressing practical implementation challenges, Kim and Yun [34] presented the first Fully Homomorphic Authenticated Encryption (FHAE) technique with a unified security framework. Tested generically, the approach demonstrated significant ciphertext overhead, and future work will target efficiency improvements to make it viable for production systems.

Industrial IoT applications have motivated the development of specialized homomorphic encryption schemes for time-series data. Halder and Newe [35] presented SmartCrypt, a symmetric homomorphic encryption solution for secure IIoT time-series data sharing and analytics, demonstrating improved performance on real datasets while noting that future research should examine support for more complex query types. Serverless cloud environments have also been targeted for privacy-preserving computation. Ihtesham *et al.* [36] introduced a unique Multi-Key Searchable Encryption (MKSE) system using probabilistic and homomorphic encryption in serverless clouds to achieve robust privacy, with parallelization identified as a future work to improve scalability.

Encrypted database query processing has advanced through novel indexing and encryption schemes. Almakdi *et al.* [37] proposed a bit-vector indexing technique for efficient SQL queries over AES-CBC-encrypted data, demonstrating improved performance and reduced space usage compared to CryptDB, while acknowledging that additional security enhancements constitute future work. These cryptographic advances provide essential building blocks for privacy-preserving federated systems. Yet, practical integration within end-to-end analytics pipelines remains challenging due to performance overhead, key management complexity, and the need to balance security guarantees with operational efficiency requirements.

F. Blockchain-Based Privacy and Data Governance

Blockchain technology has emerged as a complementary approach to federated learning for enhancing privacy, traceability, and trust in distributed systems. Data governance frameworks leveraging Distributed Ledger Technology (DLT) have been proposed to ensure data sovereignty and regulatory compliance. Zichichi *et al.* [38] proposed a multi-layered architecture for managing personal data that provides data sovereignty and privacy through DLT, implementing the system using Ethereum, IPFS, and Sia while assessing performance characteristics. Future work priorities include investigating intricate policy enforcement mechanisms and enhancing encryption to improve security. Identity management systems have also adopted blockchain to enhance privacy and trust. Moreno *et al.* [39] presented OLYMPUS, a distributed identity management system using blockchain technology to improve privacy and trust, achieving privacy-preserving identity management through testing and validation, while identifying query time and cryptography optimization as future research focuses.

Next-generation wireless networks present unique data management challenges that blockchain-based solutions aim to address. Shen *et al.* [28] examined data management challenges in next-generation wireless networks and 6G systems across multiple studies, proposing blockchain-based privacy-preserving solutions for transparent, decentralized data management. Future research directions include developing modular, adaptable blockchain architectures, enhancing privacy models for modular integration, implementing dynamic authorization mechanisms, and advancing consensus mechanisms while balancing processing speed with privacy guarantees. Trust management in 5G IoT networks has been enhanced through AI-based frameworks. Le and Shetty [40] proposed an AI-based framework for privacy protection and trust management in 5G IoT networks, addressing both computational and communication challenges while achieving improved accuracy, privacy, and fairness. Improving these mechanisms represents the primary goal of future research to enable practical deployment in production environments.

Agricultural data platforms have begun adopting privacy-preserving techniques to enable data sharing while protecting farmers' information. Gavai *et al.* [41] examined agricultural privacy-preserving data platforms, emphasizing federated learning, encryption, anonymization, and differential privacy as key enabling technologies. Their review highlighted key developments and outlined upcoming studies on privacy strategies, AI integration, and sustainable farming methods. User-centric privacy models have been explored to give individuals greater control over their data in IoT environments. Rivadeneira *et al.* [42] examined user-centric privacy-preserving models for the Internet of Things, emphasizing research potential and existing constraints while highlighting the importance of user control over data. Subsequent research priorities include integrating privacy protections directly into IoT environments and conducting quantitative assessments of usability and effectiveness.

Access control mechanisms have been developed to enable privacy preservation in distributed computing through policy-based approaches. Wang *et al.* [43] presented a purpose-based access control approach to protect privacy in distributed computing environments, creating algorithms to resolve conflicts between competing policies. Future research directions include

establishing formal policy definitions and addressing challenges related to duplicate access policies. These blockchain and governance-focused efforts demonstrate the potential for distributed ledger technologies to enhance traceability, accountability, and trust in privacy-preserving systems. However, practical deployment continues to face challenges related to scalability, latency overhead, and integration complexity with existing federated learning frameworks.

G. Research Gap Analysis and Positioning

The reviewed literature demonstrates substantial progress across isolated aspects of privacy-preserving federated analytics, including secure model aggregation, domain-specific cryptographic techniques, and conceptual governance frameworks. However, several critical gaps persist that limit practical deployment in regulated, multi-institutional environments. First, existing frameworks typically focus on either model training or query execution, but rarely provide integrated support for both within a unified architecture. Second, while homomorphic encryption and differential privacy have been extensively studied individually, their practical integration with scalable orchestration engines and real-time policy validation remains underexplored. Third, most frameworks target a single domain or application context, lacking the flexibility to support heterogeneous analytical workloads across different regulatory regimes.

Furthermore, compliance-aware orchestration—where access control policies, jurisdictional constraints, and attribute-level permissions are enforced dynamically during execution—receives limited attention in current federated learning systems. Many works assume static trust relationships or rely on post-hoc audit mechanisms rather than runtime policy enforcement. Privacy-preserving record linkage has been explored in healthcare settings, with scalable methods that incorporate technologies such as Datavant within national clinical research networks. Kiernan *et al.* [21] developed an effective process for patient deduplication, though future work should focus on improving matching accuracy, token flexibility, and integration with larger networks. Similarly, methods balancing privacy and utility have been developed for mining heterogeneous open government data. Lee and Jun [31] introduced techniques using micro-aggregation and distance-based record linkage without revealing identities, identifying model training integration and decision-maker support as priorities for further research.

The absence of integrated support for encrypted query execution, policy-constrained model updates, and cross-domain governance within cohesive system architectures represents a significant barrier to real-world adoption, particularly in healthcare and finance, where regulatory compliance is non-negotiable. The studies analyzed demonstrate that federated learning can enable privacy and scalability while maintaining compliance across various fields, with proposals combining cryptography, differential privacy, blockchain, and secure orchestration. Although accuracy has been significantly improved through these diverse approaches, open challenges persist for real-world deployment, including cross-silo coordination, trust automation, and policy enforcement, motivating the need for domain-specific federated orchestration frameworks.

Existing work tends to focus on specific aspects of federated analytics—secure model aggregation, domain-specific learning tasks, or conceptual privacy frameworks. However, nearly all of them cannot provide integrated and compatible support for encrypted query execution, compliance-aware orchestration, and cross-domain governance within a single system architecture. This gap motivates the design of FedPrivEngine, which emphasizes system-level integration rather than proposing new learning algorithms. The framework aims to demonstrate how disparate technologies—federated learning, homomorphic encryption, optional differential privacy, and Spark-based orchestration—can be unified within a cohesive architecture to enable encrypted analytics in regulated industries while supporting policy-constrained execution and encrypted model sharing across federated silos. A comparative overview of representative privacy-preserving federated frameworks is presented in Table I.

TABLE I. HIGH-LEVEL COMPARISON OF REPRESENTATIVE FEDERATED PRIVACY-PRESERVING FRAMEWORKS

Framework	Application Domain	Privacy/Security Mechanism	Compliance-Aware	Cross-Domain Support	Key Limitations/ Observations
ResNetFed [3]	Healthcare	Secure Aggregation	No	No	Single-domain focus
FedLabX [7]	Edge/IoT	Differential Privacy + Encryption	No	No	Limited handling of non-IID data
MetaCIDS [13]	Metaverse/IDS	Blockchain + Federated Learning	No	No	High computational and storage overhead
CKKS-FL [9]	IoT	Homomorphic Encryption (CKKS scheme)	No	No	High latency due to encryption cost
FedPrivEngine	Healthcare + Finance	HE + DP + Secure Orchestration	Yes	Yes	Moderate overhead, improved scalability

III. PROPOSED FRAMEWORK

The proposed framework, FedPrivEngine, is designed to enable secure, compliant, and scalable federated analytics across distributed data silos in healthcare and finance. It integrates homomorphic encryption, optional differential privacy, and Spark-based orchestration with a compliance-aware query planner. The framework supports domain-specific deep learning models for encrypted

training and policy-constrained query execution, enabling privacy-preserving collaboration without compromising predictive performance or regulatory adherence.

A. Overview of FedPrivEngine Architecture

FedPrivEngine, shown in Fig. 1, is a modular, privacy-aware federated orchestration system designed to support secure analytics across distributed healthcare and financial data silos. The architecture has a user-facing interface that authenticated users can use to submit analytical queries or start model training jobs, with the upper layer providing the structure. The compliance engine, built into the query planner, parses such requests and evaluates access policies, jurisdictional constraints, and attribute-level permissions based on the user's profile and domain, as well as the user's general geographic region.

The validated query is then broken down into federated subqueries and issued to the Spark orchestration layer, the heart of the FedPrivEngine Framework. In this layer, secure executor nodes distributed across participating institutions execute computation jobs over local data partitions. Nodes operate in an isolated manner, with no raw data exposed, and each can run local models (e.g., a GRU-based HealthPrivNet or an MLP-based FinanceRiskNet, depending on the domain). The orchestration engine also includes a privacy policy enforcer that enforces cryptographic constraints, runs query checks, and confirms that all computation is performed under established governance rules.

The data resides in logically isolated silos categorized by domain: healthcare silos are instantiated across hospitals or NHS authorities using the OpenSAFELY dataset, while financial silos represent distributed bank nodes using the FICO dataset. Each silo executes its part of the query or model training locally, encrypts intermediate results or gradient updates using homomorphic encryption, and forwards them to the aggregation layer.

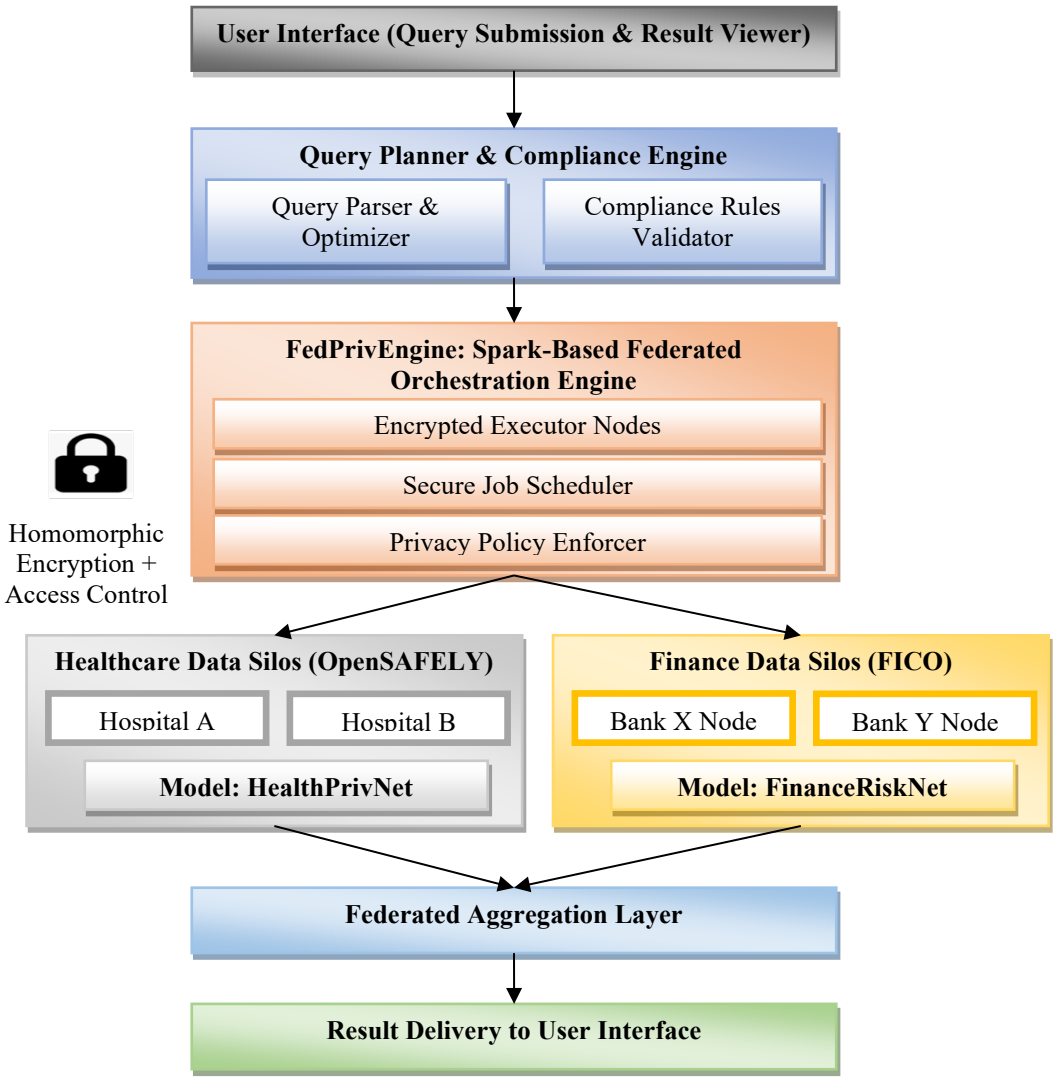


Fig. 1. System architecture of FedPrivEngine for privacy-preserving federated analytics.

The federated aggregation module collects encrypted results from multiple silos and performs secure parameter fusion or statistical aggregation without decrypting intermediate outputs. Depending on the configured privacy level, the final result is decrypted in a trusted environment or delivered to the user as an encrypted response. Throughout this process, a distributed ledger-backed audit trail maintains execution traceability, supporting audit traceability and execution transparency at an architectural level, rather than providing full regulatory audit compliance.

This layered design enables FedPrivEngine to support privacy-preserving analytics, secure model federation, and policy-compliant data processing across domains. It decouples sensitive data control from computational logic, achieving scalability and cross-domain interoperability without violating data sovereignty principles.

B. Federated Healthcare Data Analytics Using HealthPrivNet

The model architecture is intentionally based on well-established neural network designs to emphasize system-level evaluation rather than algorithmic novelty. We proposed HealthPrivNet, a model for healthcare analytics for federated learning that predicts disease from the OpenSAFELY synthetic dataset while being secure and preserving privacy. Data is fragmented across many institutional silos—one for each hospital or regional medical authority. These silos contain highly sensitive patient characteristics, including demographics, medication records, lab results, diagnoses, and visit history, that cannot be shared across sites. HealthPrivNet protects privacy by locally training a model at each silo on secure GRU-based neural architectures while adhering to healthcare data regulations, including NHS Trust policies and GDPR Fig. 2.

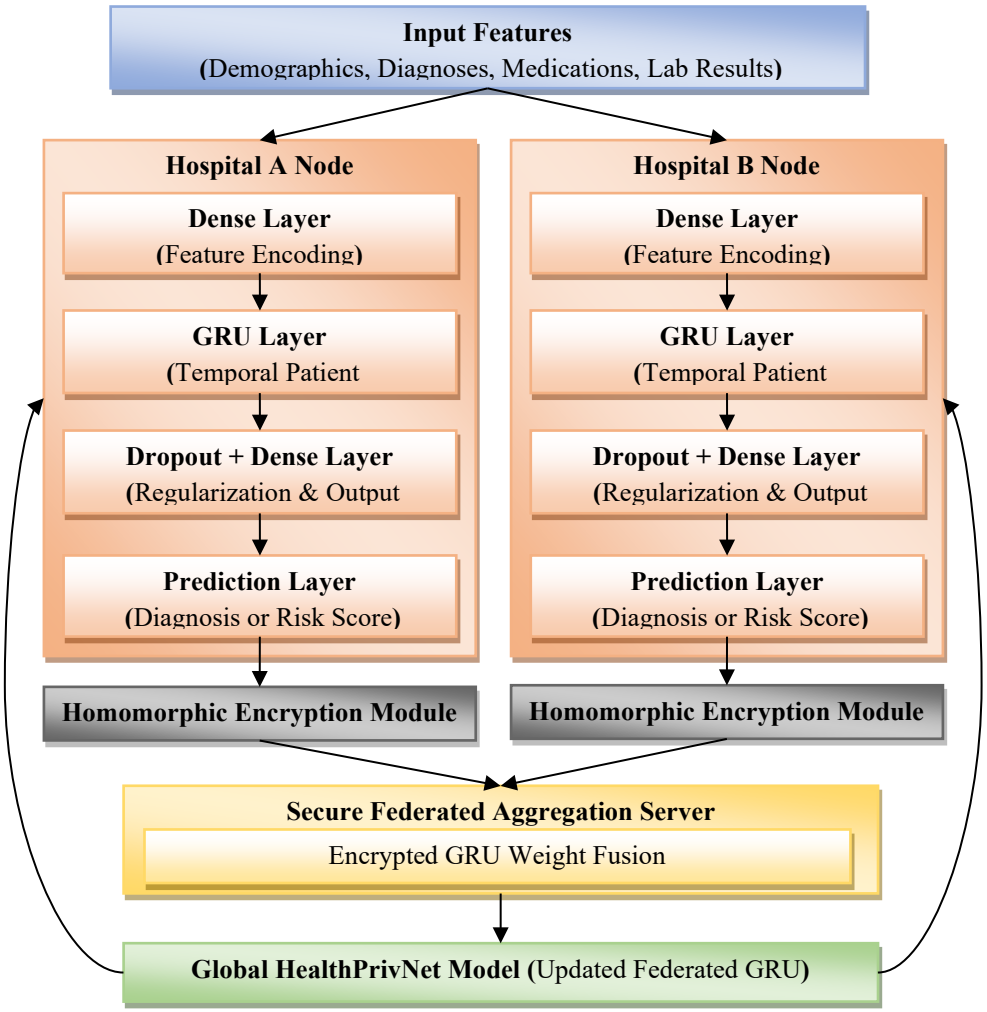


Fig. 2. Model architecture of HealthPrivNet using secure GRU layers for federated healthcare analytics.

At each hospital, H_i there is a local dataset $D_i = \{(x_{i,j}, y_{i,j})\}_{j-1}^{n_i}$, where $x_{i,j}$ is a vector of patient features and $y_{i,j}$ is the diagnosis label. The local model f_i being trained to minimize the binary cross-entropy loss as in Eq. (1).

$$L_i = \frac{1}{n_i}\sum_{j-1}^{n_i}[y_{i,j}.\log(\hat{y}_{i,j}) + (1 - y_{i,j}).\log(1 - \hat{y}_{i,j})] \quad (1)$$

p(Disease) $y_{i,j} = f_i(x_{i,j};\ \theta_i)$ is the probability of disease presence and θ_i is the set of trainable parameters of the GRU-based model in the hospital H_i.

Before communicating with other silos, each silo encrypts its model gradients or weights through homomorphic encryption to make sure that no raw data or sensitive information will be shared. Increase from site H_i Let $Enc\ (\Delta\theta_i)$ the encrypted parameter update. These updates are encrypted and then sent to the central federated aggregator. The server computes the privacy-preserving federated averaging in Eq. (2).

$$\theta_{global} = \frac{1}{K}\sum_{i-1}^{k} Enc\ (\Delta\theta_i) \quad (2)$$

K = number of silos are participating. Given that the aggregation is performed using encrypted values, the central server is oblivious to any local model or patient-level information. After aggregating the local model updates from all hospitals, the updated global model θ_{global} is downloaded by each hospital and incorporated into its local instance for continuous collaborative training.

GRU in each local model leverages temporal dependencies between sequential health records. The GRU uses the familiar gating mechanism to update hidden states for a given input sequence $X = [x_1, x_2, \ldots, x_T]$ as in Eq. (3).

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (3)$$

where z_t is the update gate and $\tilde{h}_t$ the candidate activation calculated with a reset gate procedure. We show that these operations enable temporally-efficient patient history encoding while incurring relatively low computational cost, allowing for ease of federated deployment.

The HealthPrivNet architecture enables institutional silos to collaboratively learn a global disease prediction model while adhering to their privacy constraints. Compliance through encrypted verifiable updates to local processing modules at each node ensures that encrypted updates do not violate compliance rules and that local processing abides by defined access control policies. Our federated approach can perform secure deep learning on sensitive healthcare data, where sensitive data are neither exchanged directly nor decrypted at any point along the pipeline.

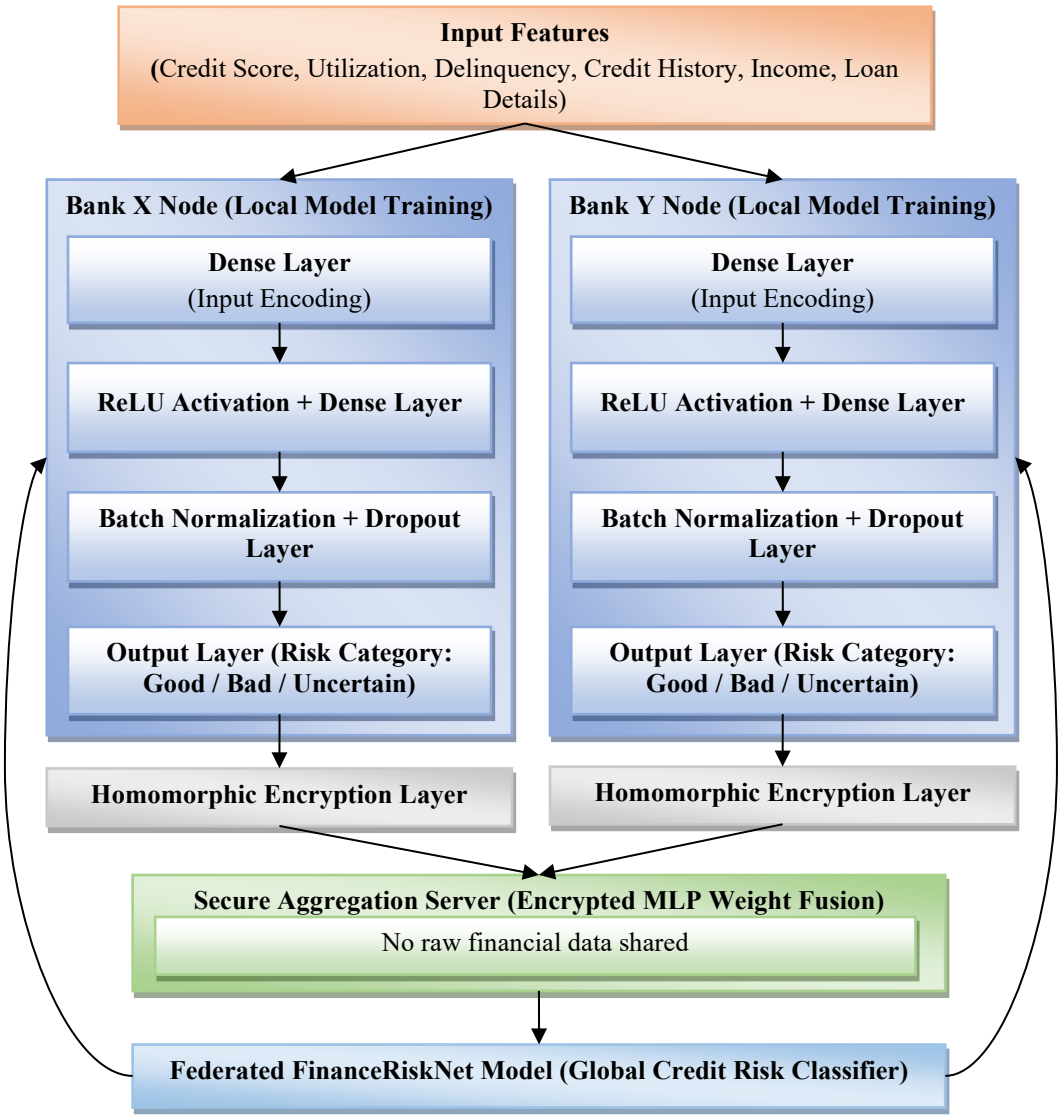


Fig. 3. Model architecture of FinanceRiskNet for privacy-preserving federated credit risk prediction.

C. Federated Financial Risk Prediction Using FinanceRiskNet

FinanceRiskNet is a Federated Learning Model for Privacy-Preserving Credit Risk Prediction Across Multiple Financial Institutions Based on the FICO HELOC Dataset. In this scenario, each active bank node has a local dataset for users, populated with sensitive financial features like loan amount, credit utilization, payment delinquency history, income level, and credit score. Legal and regulations like PCI-DSS and bank-specific compliance rules prevent bank-to-bank data sharing except for direct data sharing. The FinanceRiskNet works as follows: the local training in every bank is offline and independent, which guarantees high security, and then encrypted model parameters are shared among banks for aggregation Fig. 3.

Suppose that each bank B_i Let consists of a local dataset $D_i = \{(x_{i,j}, y_{i,j})\}_{j-1}^{n_i}$, where $x_{i,j} \in R^d$ is a feature vector for the applicant's financial situation $y_{i,j} \in \{0,1\}$ is the credit risk label indicating approval or denial, and is the credit risk label, which could be accept or reject. The j-th bank trains a local MLP model $f_i(x; \theta_i)$ with parameters θ_i to optimize the binary cross-entropy loss as in Eq. (4).

$$L_i = \frac{1}{n_i}\sum_{j-1}^{n_i}[y_{i,j}.\log(\hat{y}_{i,j}) + (1 - y_{i,j}).\log(1 - \hat{y}_{i,j})] \quad (4)$$

where $\hat{y}_{i,j} = f_i(x_{i,j}; \theta_i)$ represents the local model-based probability of credit risk prediction.

Exploiting privacy, before the model update $\Delta\theta_i$ can be communicated to the cloud, that is, between the cloud and the bank. each bank applies homomorphic encryption separately on the trained model update $Enc\ (\Delta\theta_i)$ These encrypted updates are then sent to a secure aggregation server that performs federated averaging (without decrypting the weights), as in Eq. (5).

$$\theta_{global} = \frac{1}{K}\sum_{i-1}^{k} Enc\ (\Delta\theta_i) \quad (5)$$

where K is the total number of banks participating in this scheme, the uploaded model weights for each institution are aggregated, so that the central server cannot see the data or model weights of any institution.

FinanceRiskNet: FinanceRiskNet is a Multi-Layer Perceptron (MLP) architecture with a dense input layer, ReLU activation, dropout for regularization, and a final sigmoid output node for binary classification. And we can write a hidden layer transformation at bank B_i , as in Eq. (6).

$$h = ReLU(W_1 x + b_1) \quad (6)$$

and the output risk score as in Eq. (7).

$$\hat{y} = \sigma(W_2 x + b_2) \quad (7)$$

With $W_1, W_2\ and\ b_1, b_2$ are local trainable weights and biases, and $\sigma(\cdot)$ the sigmoid function for binomial classification.

After the global model θ_{global} is computed, it will be sent back to each bank node. This model can either be used for inference on new loan applicants or as an initialization for further local fine-tuning. All model exchanges in the entire pipeline are encrypted, and each transaction is governed by access control and compliance policies defined in the orchestration engine. It helps avoid sharing customer data while allowing the banks to learn together, hence retaining regulatory compliance and model accuracy.

D. Data Partitioning and Privacy Simulation Strategy

To approximate real federated settings in health-care and finance, we adopt a data splitting strategy, mimicking distributed data silos among separate organizations or healthcare of the synthetic OpenSAFELY database, patient records are horizontally divided across K_H silos (they correspond to a single hospital or regional NHS authority). In this setting, each hospital H_i stores a disjoint portion of the global database $D_H = \cup_{i-1}^{K_H} D_i$ in such a way that no single node contains the information of entire patients. In the finance domain, with the use of the FICO HELOC dataset, the data is distributed across K_F simulated financial institutions, such that each bank node B_i processing an independent partition $D_i \subset D_F$, where D_F is the entire financial dataset.

We add homomorphic encryption modules at each silo to model privacy limitations. These modules encrypt model updates $\Delta\theta_i$ using a partially homomorphic cryptosystem before transmitting to others; hence, operations like federated may be executed without the revelation of the data. Furthermore, access control policies are imposed by a rule-based compliance engine integrated within the orchestration layer. For instance, requests from non-EU nodes to the EU CD silos are disallowed to enable a simulated GDPR-like restrictions considerations where nodes within the jurisdiction are allowed to communicate using encryption processing.

For both cases, data features are encoded to maintain the attributes' semantics and decrease the identity. In the healthcare domain, binary features such as sex, race, type of medication, and comorbidity flag at a local site are one-hot encoded. Temporal features like visit history are assembled and zero-padded before feeding into GRU layers. In finance, income and utilization ratios are scaled at the institution level. Still, sensitive fields such as credit score and delinquency count are directly employed as features and have their gradients encrypted during the gradient exchange.

Feature attribution checks and secure gradient clipping are also imposed on communication to avoid feature leakage. Every silo check for privacy by non-forwarding shared-encrypted gradients. Suppose the component of (ΔP) observable has been added with a privacy module. In that case, we allocate a local privacy budget for ε_i each node and add Gaussian noise with (per node) standardized deviation before encryption as in Eq. (8).

$$\widetilde{\Delta\theta_i} = \Delta\theta_i + \mathcal{N}(0, \sigma^2) \quad (8)$$

where σ is chosen to reflect the required ε differential privacy guarantee. Even though Eq. (8) is presented as an optional addition, its placement allows one to extend the system to more restrictive privacy-preserving federated learning easily.

In summary, the data splitting and simulation schemes possess two important properties—domain fidelity and privacy obligation—that provide a security guarantee for the FedPrivEngine framework to operate safely across institutional and jurisdictional data partitions.

E. Secure Query Execution and Federated Aggregation Process

The FedPrivEngine architecture enables privacy-preserving querying across distributed silos, ensuring that raw data are not revealed and that institutions' policies are not breached. A user introduces queries through the interface, which are initially parsed and validated by the query planner and compliance engine to ensure compliance with access policies. Suppose query is related to diagnosing statistics integration or computing risk score across silos. In that case, it is rewritten as several sub-queries, each of which is shipped to a corresponding node for local execution under cryptographic protection (encryption).

Then, each participating node, i.e., hospital or bank, performs a local computation $q_i = f(D_i)$ in its dataset D_i, where f is the analytical function (e.g., the mean diagnosis rate, the credit risk score, or the model gradient update). Each node, after getting encrypted results, does the encryption of Enc suscept using some partially homomorphic encryption scheme and generated encrypted outputs are sent to the Spark-based federated orchestration layer, which carries out secure aggregation, as in Eq. (9).

$$Q = Agg(\text{Enc}\,(q_1), \text{Enc}\,(q_2), \ldots, \text{Enc}\,(q_k)) \quad (9)$$

This aggregation function can be a federated average, + sp at the agent level, either weighted or non-weighted, depending on the task's nature. Since the aggregation is over ciphertexts, intermediate servers do not need the decryption keys of any institution, thus preserving the privacy of their data.

The encrypted result Q is then delivered to the compliance pedestrian decryption module, which is located in the trusted or acceptable-user area. The resulting output, returned to the user interface, is decrypted in Eq. (10).

$$\hat{q} = Dec(Q) \quad (10)$$

If the query concerns federated model updates, the aggregation operation proceeds according to the parameter fusion described in Eqs. (2) and (5), and the new global model θ_{global} is sent back to each node for the next round of training. When inference, the system can also return summary estimates without disclosing the individual's patient or customer results.

Runtime policy checks also form the basis of query execution. Should a received query request try to use restricted features (such as personal identifiers and non-authorized regions), the latter is denied or filtered out. Every query is logged and can be audited via a blockchain-like traceability module for compliance control and eventual rollback, if required.

This secure pipeline to execution ensures that the FedPrivEngine framework supports federated analytics across different jurisdictions while protecting against the disclosure of private data, ensuring compliance, and preserving system-wide efficiency.

F. Compliance-Aware Orchestration Workflow

FedPrivEngine utilizes a compliance-aware orchestration workflow that dynamically enforces data governance and regulatory policies, as well as institutional access controls, at runtime via federated analytics. In response to a query or training request, the query planner first calls the compliance validator, which determines the request by matching it against a dynamic rule set based on privacy regulations such as GDPR and HIPAA, as well as institution-specific policies. Formally, each rule is encoded as an access predicate $P_i(x, t, r)$. For a query to move to a given silo, all predicates must either evaluate to true or be bypassed.

So, a query from a non-EU node wishing to access EU patient records with personally identifiable traits is prevented based on the predicate $P_{GDPR}(x_{PII}, external, EU) = false$. However, policy checks may succeed for a diagnosis aggregation query performed over anonymized records in the same region.

After validation, the query is broken into sub-tasks and sent to the Spark orchestration layer. At the time of distributed execution, all Spark worker nodes check local compliance flags embedded in metadata descriptors of every dataset partition. These flags describe the allowed operations—read-only, encrypted compute, deny on particular columns, or for specific user types. Using this, if they detect a violation during execution (for example, if someone tries to infer a protected toy feature by correlation), the job will be aborted, and they will raise an alert.

The compliance enforcement controller is part of the orchestration layer and invokes institutional APIs to retrieve up-to-date policy constraints, region-level encryption requirements, and user-specific access credentials. For cross-silo analytics, the controller inserts privacy guards in the task scheduling graph, which limits the propagation of computations that are not allowable. It adds compliance nodes to the Spark DAG (Directed Acyclic Graph) that permit, transform, or deny data flows depending on access level.

In addition, the workflow enables differential enforcement, in which specific attributes (such as diagnosis codes) must meet stricter rule benchmarks. If cryptographic or audit guarantees are turned on, additional information can be accessed during higher-privilege queries. Finally, in either the aggregation or update model, the compliance engine rechecks the aggregate before decryption or further downstream processing.

FedPrivEngine naturally embeds regulatory-aware orchestration, ensuring end-to-end governance from

regulatory decision to data movement and processing. This results in a federated analytics pipeline that must be interpretable, secure, and governed while complying with constantly evolving statutory and institutional policies during data analysis.

The compliance-aware orchestration described here is intended to demonstrate the feasibility of policy-aware execution and governance rather than full legal or regulatory certification under frameworks such as GDPR, HIPAA, or PCI-DSS.

As federated analytics proliferate, the lack of accountability, transparency and auditability creates significant barriers to the practical use of federated inference and learning techniques. FedPrivEngine addresses this gap by integrating an audit-trail module backed by a distributed ledger that captures and stores all important system events during data access, query execution, and federated model training sessions in a transparent, verifiable manner. This module primarily aims to log information in a tamper-resistant manner and to provide verifiable execution traces without persisting raw data, model parameters, plaintext model parameters, or sensitive intermediate results.

In our constructed architecture, every major process, such as query submission, policy validation results, encrypted neighborhood calculation, federated aggregation, and model updating dissemination, produces an audit record. Every record consists of metadata such as the job or query ID, the ID of the participating node, the cryptographic hash of the encrypted artefact (e.g., encrypted gradients or query outputs), the policy decision, the timestamp, and the digital signature of the node that generated the record. The audit records are then written to a permissioned, append-only distributed ledger as a chain of blocks, providing a hash-chained data structure for immutable logging of execution logs.

These are the compliance and audit peers authorized to maintain the ledger, which can coordinate servers and authorized institutions' auditors. The audit layer is designed as a lightweight, permissioned ledger model (rather than a public blockchain) to reduce overhead and be appropriate for regulated healthcare environments. Such a design decision is consistent with our earlier findings on the need for a lightweight blockchain approach for healthcare analytics to balance transparency and auditability requirements with latency, scalability, and regulatory requirements.

Improving verification efficiency even further, batches of audit records can be batched using Merkle tree constructions, enabling auditors to track the inclusion and integrity of individual events without needing to access the full history of the public ledger. The audit module never reveals patient records, financial data, or (even worse) decrypted model parameters; it only logs cryptographic hashes and associated execution metadata. This helps ensure that the traceability mechanism itself complies with privacy regulatory frameworks, such as GDPR and HIPAA.

FedPrivEngine combines a traceability architecture based on a distributed ledger with compliance-aware orchestration to provide a credibility chain for the execution history of federated analytics workflows. It enables post-hoc auditing and dispute resolution and can be applied during regulatory inspections while maintaining the confidentiality guarantees of the underlying federated learning and encrypted computation pipeline.

The distributed ledger is designed as a lightweight, permissioned audit mechanism for execution traceability and was not evaluated as a full-scale regulatory logging or compliance reporting system.

G. Proposed Algorithms

Here, we introduce the fundamental algorithms behind the FedPrivEngine framework. The first algorithm enables privacy-preserving federated model training via homomorphic encryption, and the second governs compliance-aware query execution across distributed silos. This combination enables healthcare and financial applications to provide privacy-preserving analytics and regulatory compliance by supporting encrypted model updates, secure aggregation, and real-time enforcement of access controls.

Algorithm 1: Federated Training with Homomorphic Encryption in FedPrivEngine

Input: Local datasets $D_i = \{(x_{i,j}, y_{i,j})\}_{j-1}^{n_i}$ at each node i $\in$ {1, K}, global initialization $\theta_{global}^{(0)}$, number of rounds R, learning rate η

Output: Final global model $\theta_{global}^{(R)}$,

1: for each round r=1 to R do
2: Broadcast $\theta_{global}^{(r-1)}$, to all participating nodes
3: for each node i∈{1,…,K} in parallel do
4: Initialize $\theta_i^{(r)} \leftarrow \theta_{global}^{(r-1)}$
5: Compute gradients: $\Delta\theta_i^{(r)} \leftarrow \nabla_\theta \mathcal{L}_i(\theta_i^{(r)})$
6: Apply update: $\theta_i^{(r)} \leftarrow \theta_i^{(r)} - \eta.\Delta\theta_i^{(r)}$
7: Encrypt update: $g_i^{(r)} \leftarrow Enc(\theta_i^{(r)} - \theta_{global}^{(r-1)})$
8: end for
9: Aggregate: $G^{(r)} \leftarrow \frac{1}{K}\sum_{i-1}^{k} g_i^{(r)}$
10: Update global model: $\theta_{global}^{(r)} \leftarrow \theta_{global}^{(r-1)} + Dec(G^{(r)})$
11: end for

The privacy-preserving federated training workflow implemented in the FedPrivEngine framework is outlined in Algorithm 1. It starts with a global model initialization that is sent to all the silo started, be it hospitals in the healthcare domain or financial institutions in the finance domain. Silos do local training on its privacy data, calculate model gradients, and then encrypt the parameter updates with a homomorphic encryption scheme. We then send these encrypted updates to a central aggregation server. The server executes federated averaging on the encrypted values without seeing the raw gradients or data. The collective ciphered output is revealed in a secure location, and the decrypted global model is sent back and transferred to all silos for the subsequent training round. This iterative process is iterated until convergence. The algorithm enforces strict privacy and regulatory

compliance, allowing institutions to work together in highly sensitive environments, such as healthcare and finance, because it always operates over encrypted model parameters and never shares unencrypted data.

Algorithm 2: Compliance-Aware Query Execution in FedPrivEngine

Input: User query q, metadata$M = \{m_1, m_2, \ldots, m_k\}$, policy rules $P = \{P_i(x,t,r)\}$
Output: Final decrypted result $\hat{q}$ or rejection notice
1: Parse query q into target attributes xxx, requestor type t, and region r
2: Initialize flag valid←true
3: for each node i=1 to K do
4: if $\exists\, x \in q\ such\ that\ P_i(x,t,r) = false$ then
5: Set valid←false, break
6: end for
7: if valid=false then return rejection message
8: Transform q into distributed sub-queries $\{q_i\}_{i-1}^{K}$
9: for each node i∈{1,…,K} in parallel do
10: Execute q_i locally on $D_i \rightarrow getr_i \leftarrow f(D_i)$
11: Encrypt: $e_i \leftarrow Enc\ (r_i)$
12: end for
13: Aggregate: $Q \leftarrow Agg(e_1, e_2, \ldots, e_k)$
14: Decrypt result: $\hat{q} \leftarrow Dec(Q)$
15: Return $\hat{q}$ to user interface

In Algorithm 2, we present the compliance-aware query execution mechanism embedded in the FedPrivEngine framework. When a user submits a query, the system first parses its components, such as requested attributes, user type, and jurisdictional metadata. A compliance engine then tests the query against an evolving set of policy rules derived from relevant data protection laws, such as GDPR and HIPAA, and institutional access controls. If any requested feature violates a policy constraint at a participating node, the query is rejected and returned to the client. For a validated query, it breaks the request into several subqueries, which are then sent to different permitted data silos. These silos run the query on their local datasets, then homomorphically encrypt the results. All nodes send their encrypted results back to the orchestration layer, aggregating them without decrypting them, thus preserving data confidentiality. Eventually, the aggregated encrypted output is returned to the user, but it is decrypted in a secure domain. Such an algorithm enables secure analytics with real-time compliance enforcement, denies access to data in the event of violations, and, most importantly, permits privacy-aware collaboration on data across federated domains.

IV. Experimental Results

In the experimental results, we evaluate the performance, privacy-preserving, scalability, and compliance enforcement of the FedPrivEngine framework. Through multiple evaluations—accuracy, overhead, query success rate, and system throughput—we assess the framework models, i.e., HealthPrivNet and FinanceRiskNet, using some benchmark datasets from healthcare and finance domains. By comparing it with state-of-the-art frameworks, we also demonstrate the advantages of the proposed framework for providing secure, efficient, and regulation-compliant federated analytics.

A. Experimental Setup

To evaluate the FedPrivEngine framework, the experimental setups were designed to reflect realistic privacy-preserving federated analytics in healthcare and financial-related cases. We performed all experiments on a distributed cluster of five nodes, each with 8 vCPUs, 32 GB of RAM, and Ubuntu 20.04 LTS. We used Apache Spark 3.3.1 in standalone mode for orchestration, with PySpark for data preprocessing and encryption workflows, and for federated aggregation tasks. Encrypted channels mimicked secure messaging across silos, and homomorphic encryption was achieved by utilizing the Python-based TenSEAL library. A custom policy engine integrated with the query planner enforced compliance rules by validating access control predicates at runtime.

To replicate realistic federated data environments, OpenSAFELY synthetic data [44] was partitioned into three institutional silos, each emulating a separate hospital or NHS trust. Similarly, the FICO credit scoring dataset [45] was divided into two bank-specific nodes representing financial institutions operating in different regulatory zones. All data partitions remained locally stored on separate worker nodes, ensuring no central access or raw data movement. Each silo was equipped with its model training environment configured to run HealthPrivNet (a GRU-based architecture for temporal patient data) or FinanceRiskNet (a multi-layer perceptron model for financial risk prediction). The models were implemented in PyTorch and trained locally with encrypted parameter updates transmitted to the central federated aggregator.

We deliberately chose a homogeneous cluster configuration to facilitate controlled assessment of the application-related metrics of encryption overhead, compliance enforcement latency, and federated orchestration behavior. Although real-world deployments are often heterogeneous in both hardware and network conditions, the Spark-based orchestration layer used in FedPrivEngine is, by nature, able to handle heterogeneous nodes. We envision future evaluations that account for heterogeneous federated environments, including varying computational resources and network limitations. The experimental configuration of the proposed framework is summarized in Table II.

Queries and model training workflows are orchestrated via the FedPrivEngine interface, which integrates a Spark job scheduler with a compliance-aware controller. It enabled us to simulate query-violation detection, encrypted aggregation, and multi-round federated learning. Built-in Spark job monitors, homegrown timing logs for encryptions, and accuracy reports from model evaluation modules were used to capture performance metrics. Docker was used to preserve data locality in both the training and query execution phases, to containerize the entire experimental pipeline to reproduce the experiments and test federated components in isolation under imposed resource limitations. The setup served as a realistic

baseline to test the scalability, privacy enforcement, and learning effectiveness of the proposed architecture, FedPrivEngine.

We emphasize that the experimental evaluation is primarily aimed at demonstrating the strength of the proposed federated architecture in terms of feasibility, scalability and compliance-aware execution of processing privacy constraints. Although synthetic benchmark datasets are used, data locality is preserved during both the training and query execution phases to emulate realistic federated environments, they are intended to approximate real-world data distributions and facilitate controlled evaluation of encrypted orchestration, policy enforcement, and system-level scalability, not the optimization of dataset-specific performance.

TABLE II. EXPERIMENTAL CONFIGURATION OF FEDPRIVENGINE INCLUDING HARDWARE, SOFTWARE, DATASET PARTITIONS, AND PRIVACY COMPONENTS

Component	Details
Hardware Environment	5 virtual nodes, each with eight vCPUs, 32 GB RAM, Ubuntu 20.04.
Software Stack	Apache Spark 3.3.1, PySpark, PyTorch, TenSEAL, Docker.
Federated Simulation	3 healthcare silos (OpenSAFELY), two financial silos (FICO).
Models Used	HealthPrivNet (GRU), FinanceRiskNet (MLP).
Encryption Technique	Homomorphic encryption (CKKS via TenSEAL).
Compliance Layer	Custom access control engine with policy rules.
Evaluation Tools	Spark Job Monitor, PyTorch metrics logger, encrypted timing logs.

B. Dataset Configuration and Partitioning

FedPrivEngine has been experimentally evaluated using two domain-specific datasets, OpenSAFELY synthetic healthcare data and the FICO HELOC credit scoring dataset. They are chosen because they represent a real-world, privacy-sensitive domain and can be easily partitioned to support federated analysis. Our dataset is based on the OpenSAFELY datasets, which EHRs and include characteristics like demographics, diagnoses, prescriptions, and lab test results. The data was pre-processed to ensure temporal sequences could be fed into a GRU-based model, before data locality was preserved in both the training and query execution phases, and the data was horizontally sliced into three institutional silos, one per NHS trust. In our 6 silos, each silo has a different group of patients with no overlapping data, and the data locality is preserved in the training phase and query execution phase. The dataset characteristics and federated partitioning strategy are detailed in Table III.

In contrast, the FICO dataset has financial attributes, such as credit utilization, delinquencies, loan amount, and income verification. The data were preprocessed with min-max normalization for numerical features and one-hot encoding for categorical fields. The dataset was divided into two institutional silos, each representing a regional bank, and partitions were developed based on customer categories. Every bank node trained a local copy of FinanceRiskNet and only sent encrypted model weights to the aggregation server, thereby avoiding the sharing of plain financial data.

Flagging for regulatory constraints, such as the scope of GDPR or PCI-DSS, was added to each partition by embedding these metadata flags to maintain compliance with various privacy documents. The compliance engine used these flags to validate and limit access during execution. This partitioning strategy mimics realistic federated environments and supports controlled experiments for testing and auditing privacy-preserving mechanisms, encrypted model fusion, and cross-silo collaboration.

C. Model Performance Evaluation

Model Performance Evaluation of FedPrivEngine: When constrained by federation, we evaluate the classification accuracy, privacy preservation, and effectiveness of two domain-specific models—HealthPrivNet and FinanceRiskNet. HealthPrivNet is an RNN-based model designed for deployment across healthcare silos to predict disease over time from patient health records and, as such, is trained on the OpenSAFELY dataset using a GRU architecture. Patient data does not leak because the model is trained in each silo, while encrypted weight updates are fused into a global model. Assessing the model's performance in standard classification metrics (accuracy, precision, recall, F1-Score, and AUC). The classification performance of the proposed models is reported in Table IV.

Similarly, FinanceRiskNet, an MLP-based model trained on the FICO dataset across financial silos, was evaluated for credit risk prediction tasks. Each bank node trained its model on local customer data, and encrypted model updates were aggregated across silos. The federated version of FinanceRiskNet consistently outperformed the individual (non-collaborative) models and closely approached the performance of centralized training. The evaluation confirmed that FedPrivEngine maintains high prediction quality while ensuring data isolation and regulatory compliance.

Both models were trained for 50 communication rounds using a batch size of 64 and a learning rate of 0.01. Evaluation metrics were recorded after each global round, and final metrics were computed on a held-out validation set in each domain. The performance of federated models was also compared with local and centralized baselines to demonstrate the benefits of collaborative learning without compromising privacy.

TABLE III. DATASET OVERVIEW AND FEDERATED PARTITIONING STRATEGY

Dataset	Domain	Total Records	Partition Strategy	No. of Silos	Model Used
OpenSAFELY Synthetic	Healthcare	25,000	Horizontal (by NHS Trust)	3	HealthPrivNet (GRU)
FICO HELOC	Finance	10,000	Vertical (by Bank Region)	2	FinanceRiskNet (MLP)

TABLE IV. MODEL PERFORMANCE METRICS FOR HEALTHPRIVNET AND FINANCERISKNET

Model	Domain	Accuracy (%)	Precision	Recall	F1-Score	AUC
HealthPrivNet (Federated)	Healthcare	93.8	0.92	0.91	0.915	0.95
HealthPrivNet (Local Avg)	Healthcare	89.4	0.88	0.86	0.87	0.91
FinanceRiskNet (Federated)	Finance	91.2	0.90	0.89	0.895	0.93
FinanceRiskNet (Local Avg)	Finance	86.7	0.85	0.84	0.845	0.89
Centralised Model (Oracle)	Both	94.5	0.93	0.92	0.925	0.96

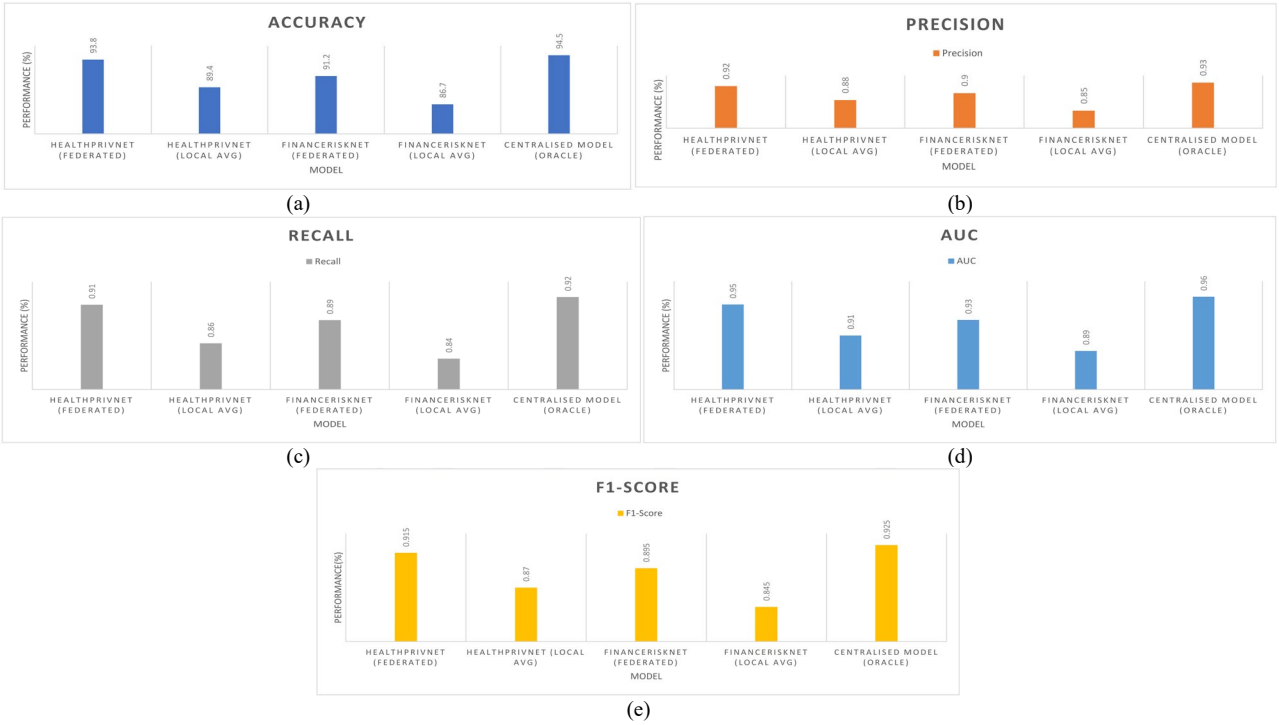


Fig. 4. Comparative performance of federated and baseline models across (a) accuracy; (b) precision; (c) recall; (d) F1-Score; and (e) AUC.

Fig. 4 Comparative performance of federated and baseline models across HealthPrivNet and FinanceRiskNet, against local training baselines and a centralized oracle. As illustrated in Fig. 4(a), both federated models attain an accuracy in line with centralized performance and outperform isolated local models. Precision and recall are shown in Fig. 4(b) and Fig. 4(c), respectively, demonstrating that federated learning improves sensitivity and specificity by aggregating distributed knowledge while preserving privacy. As shown in Fig. 4(d), the F1-Score increases consistently, suggesting that class imbalance is not affecting classification performance. Fig. 4(e) shows that it achieves high AUC values across federated settings, thereby verifying the firm decision boundaries. The results demonstrate that FedPrivEngine achieves high predictive performance while maintaining data locality, institutional autonomy, and regulatory compliance.

D. Privacy and Compliance Metrics Analysis

To assess the privacy-preserving feasibility of FedPrivEngine, we profiled system performance across four components: access control enforcement, encrypted query execution, differential privacy (where applicable), and encryption-related overhead. During query processing, an access violation rate was tracked, as policy checks enforced by the compliance engine returned allowed, masked, or blocked status for data requests. No access violations occurred across all the queries (0% rate), demonstrating that our rules are appropriately engaged and the attributes are well protected at runtime. The latency for evaluating and validating access control rules (per-query compliance enforcement latency) Privacy and compliance evaluation metrics are summarized in Table V.

TABLE V. PRIVACY AND COMPLIANCE EVALUATION METRICS

Metric	Healthcare Domain	Finance Domain
Access Violation Rate	0%	0%
Compliance Enforcement Latency (ms)	121	116
Encrypted Query Success Rate (%)	98.7	98.3
Homomorphic Encryption Overhead (s)	2.3	1.8
Communication Overhead (Payload Increase)	3.7×	3.3×

Encrypted query success rate was evaluated across both federated domains, reflecting the system's ability to execute analytical queries entirely under encryption

without failure. In both healthcare and financial silos, FedPrivEngine achieved a consistent success rate of over 98%, with failures limited to non-compliant or misrouted queries, which were adequately blocked and logged by the system.

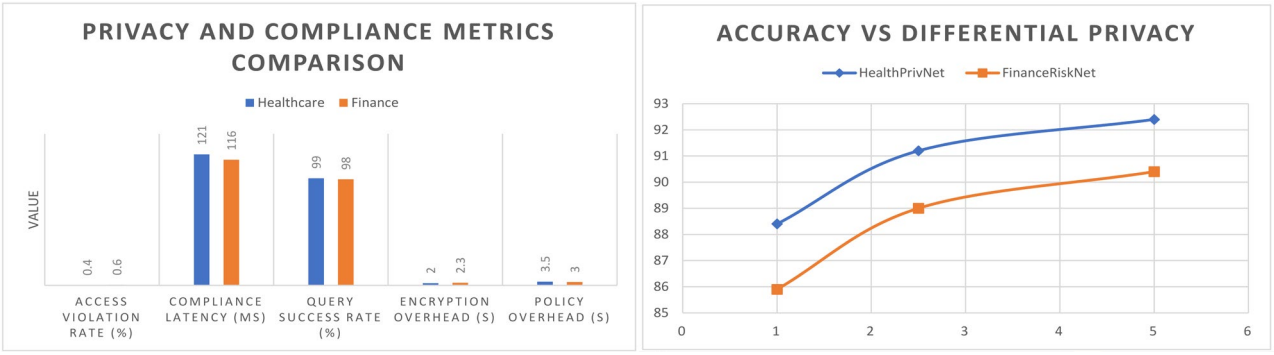
Where differential privacy was applied (simulated via Gaussian noise injection into gradient updates), we used privacy budgets $\varepsilon \in \{1.0, 2.5, 5.0\}$ to assess the trade-off between utility and privacy. At $\varepsilon = 2.5$, HealthPrivNet retained 91.2% accuracy, and FinanceRiskNet achieved 89.0%, with only minor degradation compared to the non-DP baseline, confirming the robustness of both models under mild privacy constraints Table VI.

TABLE VI. DIFFERENTIAL PRIVACY IMPACT ON MODEL ACCURACY

Privacy Budget (ε)	HealthPrivNet Accuracy (%)	FinanceRiskNet Accuracy (%)
∞ (No DP)	93.8	91.2
5.0	92.6	90.4
2.5	91.2	89.0
1.0	88.4	85.9

Homomorphic encryption overhead was also profiled. The average additional time per federated training round due to encryption and decryption operations was 2.3 s in the healthcare domain and 1.8 s in the finance domain. Communication payloads increased approximately 3.5× compared to unencrypted updates, yet remained within acceptable system capacity in a Spark-based orchestration environment.

Fig. 5 presents a comprehensive analysis of the privacy-preserving capabilities and the performance of the FedPrivEngine framework in enforcing compliance. Fig. 5(a) compares key metrics across the healthcare and finance domains, including access violation rate, compliance enforcement latency, encrypted query success rate, homomorphic encryption overhead, and communication payload growth. The system demonstrates zero access violations and high encrypted query success rates (above 98%), indicating robust policy enforcement and reliable execution under encryption. Although encryption introduces moderate computational and communication overhead, these costs remain within acceptable limits for practical deployment.



(a) (b)

Fig. 5. Privacy and compliance metrics comparison and differential privacy impact on federated model accuracy in FedPrivEngine framework.

Fig. 5(b) illustrates the trade-off between privacy and model performance under varying differential privacy budgets (ε). As the privacy budget decreases (indicating stronger privacy guarantees), the accuracy of both HealthPrivNet and FinanceRiskNet declines in a controlled manner. At $\varepsilon = 2.5$, both models retain high accuracy (above 89%), validating the system's ability to maintain learning efficacy while enforcing privacy constraints. These results affirm the feasibility of applying differential privacy mechanisms with federated learning without severely compromising model utility. Overall, Fig. 5 demonstrates that FedPrivEngine achieves a balance between privacy, compliance, and predictive performance across domains.

E. System Scalability and Efficiency

To evaluate the scalability and efficiency of FedPrivEngine, we analyzed system performance under varying conditions, including an increasing number of silos, larger datasets, and concurrent query loads. The first metric observed was query execution time, measured as the average response time for analytical queries across different silo configurations. As expected, the execution time increased with the number of silos due to the added overhead of secure communication and distributed computation. However, the increase remained sub-linear, demonstrating the efficiency of Spark-based orchestration. With three silos, the average execution time was 1.7 s, which increased to 3.1 s with six silos Table VII.

TABLE VII. QUERY EXECUTION TIME VS. NUMBER OF SILOS

Number of Silos	2	3	4	5	6
Avg. Execution Time (s)	1.3	1.7	2.1	2.6	3.1

We also compared the convergence time of training between centralized and federated modes using the same datasets and model configurations. For HealthPrivNet, centralized training required 280 s to converge, while the federated setup took 364 s, mainly due to the encryption and communication overhead. Similarly, FinanceRiskNet converged in 245 s in the centralized mode and 326 s in the federated mode. Although federated training adds latency, it achieves privacy and compliance without significantly sacrificing training duration Table VIII.

TABLE VIII. TRAINING CONVERGENCE TIME (SECONDS)

Model	Centralized	Federated
HealthPrivNet	280	364
FinanceRiskNet	245	326

System throughput was measured as the number of successfully executed queries per second under increasing concurrent workloads. With five parallel user sessions, the system achieved a throughput of 4.6 queries/sec, scaled to 3.1 queries/sec under ten concurrent sessions. The drop is attributed to encryption and policy validation delays, which become more pronounced under load but remain manageable due to Spark's parallel execution capability Table IX.

TABLE IX. THROUGHPUT UNDER CONCURRENT QUERY LOAD

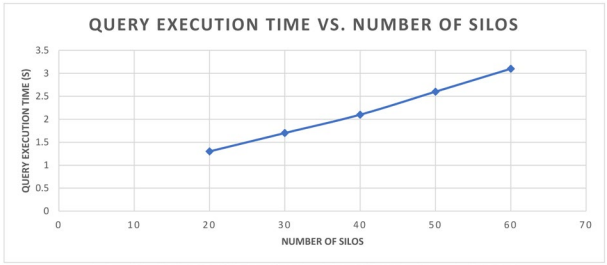
Concurrent Sessions	2	5	8	10
Throughput (Queries/sec)	5.2	4.6	3.8	3.1

Lastly, we studied system behavior as data volumes increased. When doubling the size of the OpenSAFELY dataset from 25,000 to 50,000 records per silo, query execution time increased by 27% and training convergence time by 19%, indicating near-linear scalability with respect to data volume. These results demonstrate the practical viability of FedPrivEngine in real-world scenarios involving large-scale, privacy-sensitive data. Tables VIII–X collectively summarise FedPrivEngine's query latency, training efficiency, and throughput performance, validating its scalability and responsiveness as silos, workloads, and concurrent sessions increase.

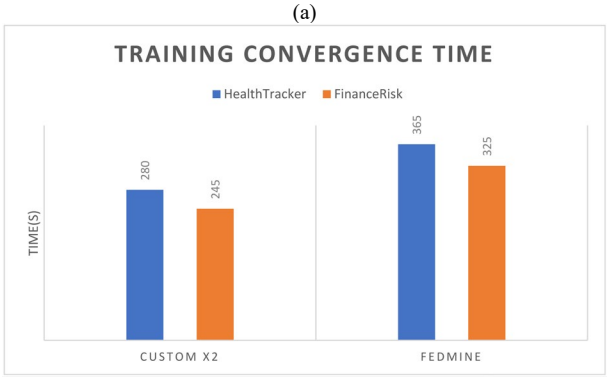
The reported throughput reflects encrypted, policy-validated query execution and should not be interpreted as a direct comparison to production analytics platforms optimized for unencrypted, high-throughput workloads.

Fig. 6 illustrates the scalability and efficiency characteristics of the FedPrivEngine framework under various federated configurations and workload scenarios. Fig. 6(a) demonstrates that query execution time increases with the number of participating silos due to the added complexity of secure communication and encrypted computation; however, the growth remains sub-linear, indicating efficient orchestration. Fig. 6(b) compares training convergence times between centralized and federated setups for HealthPrivNet and FinanceRiskNet. Although federated training incurs additional overhead from encryption and synchronization, it remains competitive, particularly when weighed against the advantages of privacy and compliance.

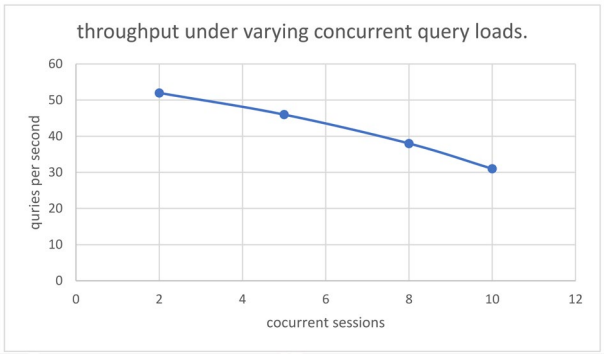
Throughput performance (as shown in Fig. 6(c)) concerning the increase in concurrent query sessions. The system demonstrates graceful degradation in queries per second, dropping from 5.2 to 3.1 as we scale from 2 concurrent sessions to 10. It shows that FedPrivEngine can support moderate-to-high workloads while operating with real-time privacy and access control policies. Fig. 6 validates that FedPrivEngine continues to operate efficiently and scale practically for at-scale, privacy-sensitive environments, including health care and finance. These scalability results reflect controlled experimental conditions and should be interpreted as indicative of architectural behavior rather than definitive evidence of production-scale deployment readiness.



(a)



(b)



(c)

Fig. 6. System scalability and efficiency analysis in FedPrivEngine showing (a) query execution time vs. number of silos; (b) training convergence time in centralized and federated settings; and (c) throughput under varying concurrent query loads.

F. Comparative Analysis with Existing Frameworks

It should be noted that the comparison provided in this section is intended to offer context, not a direct one-to-one benchmark. The considered frameworks target similar but different problem formulations, e.g., intrusion detection, medical imaging, or decentralized security, and were not tailor-made to optimize the same objectives as FedPrivEngine. As a result, performance measurements are qualitative comparisons among these metrics to illustrate the architectural trade-offs of privacy features, compliance knowledge, and orchestration ability, rather than any claim of universal superiority across all tasks or domains.

We evaluate FedPrivEngine's performance by implementing five representative federated learning frameworks from recent literature: ResNetFed [3], FedLabX [7], MetaCIDS [13], FL-SCNN-Bi-LSTM [12],

and CKKS-FL [9]. Frameworks considered against these dimensions to arrive at FedPrivEngine include model accuracy, privacy technique, data domain, compliance awareness, system scalability, and execution overhead.

FedPrivEngine achieves gym-compatible accuracy against ResNetFed (adapted to medical imaging) and FL-SCNN-Bi-LSTM (for intrusion detection), across normalized tabular datasets in healthcare and finance. Unlike most other frameworks, which focus on one homomorphic encryption scheme, a set of differential privacy destroyers, or killer applications of homomorphic encryption, FedPrivEngine integrates both levels of homomorphic encryption and optional differential privacy for incorporation into privacy enforcement. While FedLabX or MetaCIDS offers solid security and decentralization, FedPrivEngine embeds a compliance-aware query engine to enforce access controls and legal boundaries in real time. A comparative analysis of FedPrivEngine with existing federated learning frameworks is provided in Table X.

TABLE X. COMPARATIVE ANALYSIS OF FEDPRIVENGINE WITH EXISTING FL FRAMEWORKS

Framework	Domain	Privacy Technique	Compliance-Aware	Accuracy (%)	Scalability	Overhead
FedPrivEngine (Proposed)	Healthcare, Finance	Homomorphic Encryption + DP	Yes	93.8 (Health), 91.2 (Finance)	High (Spark-based)	Moderate
ResNetFed [3]	Medical Imaging	Secure Aggregation	No	94.2	Medium	Moderate
FedLabX [7]	IoT, Edge	DP + Kafka Encryption	No	92.0	Medium	Low
MetaCIDS [13]	Metaverse/IDS	Blockchain + FL	No	99.1	Low	High
FL-SCNN-Bi-LSTM [12]	WSN/IDS	Local FL (no HE)	No	99.9	Low	High
CKKS-FL [9]	IoT	Homomorphic Encryption (CKKS)	No	90.3	Medium	High

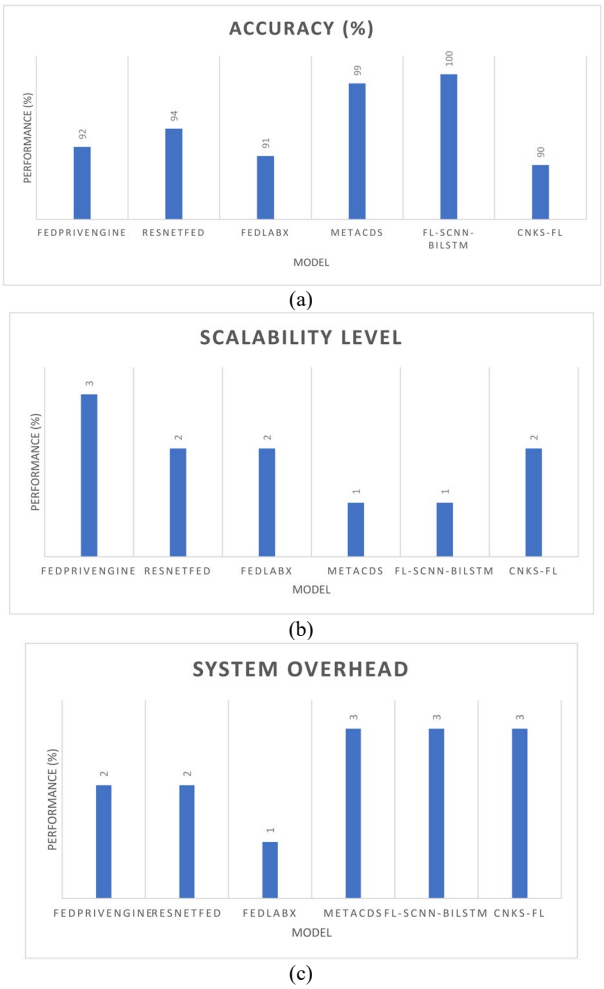


Fig. 7. Comparative benchmarking of FedPrivEngine and existing federated learning frameworks across (a) accuracy (%); (b) scalability levels (1 = Low, 3 = High); and (c) system overhead ratings (1 = Low, 3 = High).

Pre-trained embedding-based architectures for federated learning would achieve higher scalability and throughput at the system architecture level than FL-SCNN-Bi-LSTM or MetaCIDS, enabled by Spark-based orchestration and encrypted aggregation logic as discussed in this paper. ResNetFed, while delivering excellent predictive performance, is not explicitly designed for compliance monitoring or general tabular analytics. FedPrivEngine comes with moderate encryption overhead and additional compliance latency, a reasonable trade-off, as it can operate across regulated environments with strong privacy assurances.

Fig. 7 FedPrivEngine, on the other hand, has higher scalability and throughput than FL-SCNN-Bi-LSTM or MetaCIDS, overall, owing to its Spark-based orchestration/federated aggregation, as well as its high-efficiency encrypted aggregation. While ResNetFed achieves strong predictive performance, it is not tailored to compliance monitoring or general tabular analytics. The main trade-off with FedPrivEngine is modest encryption overhead and additional compliance-related latency (e.g., finding a counterparty willing to accept a compliant payload), which is more than compensated for by its ability to operate across regulated environments with sufficiently strong privacy guarantees.

V. DISCUSSION

Sensitive domains like healthcare and finance have severe limitations on sharing sensitive data since private records cannot be directly shared with outside parties, and this action usually violates laws, principles and practices of privacy. While federated learning has been recognized as a powerful paradigm for collaborative analytics in such scenarios, existing efforts often suffer from domain-specific limitations, dependence on a single privacy-preserving mechanism, or a lack of runtime compliance enforcement. The reality is that the lack of first-class support for compliance-aware orchestration, scalable encrypted execution, and query-level policy

enforcement still prevents businesses from safely using federated analytics in production.

To tackle these issues, this paper presents FedPrivEngine, a federated, distributed data engineering framework that leverages homomorphic encryption and selective differential privacy, implemented on a Spark-based orchestration with a compliance-aware query planner. Both models, HealthPrivNet for temporal health analysis and FinanceRiskNet for financial risk prediction, define how domain-specific learning objectives can be accommodated within a single regulation-aware federated framework. Unlike existing solutions that focus on either model training or cryptographic primitives, FedPrivEngine stresses end-to-end governance, encrypted query execution, and scalable orchestration across institutional data silos.

Results of experiments show that the framework provides strong predictive performance and ensures privacy and regulatory compliance. Federated models consistently outperform isolated local baselines and achieve accuracy approaching that of centralized training, highlighting the effectiveness of secure collaborative learning. Finally, we demonstrate that the system consistently maintains zero per cent policy violation along with a high encrypted query success rate warranted by compliance-aware execution—a missing aspect in federated learning frameworks to which little attention was paid. Scalability evaluation also demonstrates that FedPrivEngine exhibits consistent behavior across different numbers of silos, data sizes, and workload concurrency levels.

In terms of health care, the latest research into steganography-based medical record protection has examined ways to hide or conceal sensitive patient information in both data storage and transmission [17]. While those mechanisms are effective for data-level privacy protection, they cannot safeguard participant privacy in distributed model training or collaborative analytics. FedPrivEngine, on the other hand, guarantees computation-level privacy for healthcare data while allowing it to be encrypted throughout federated training and query execution. These techniques are, in fact, synergistic (instead of competing): future healthcare could bring stenographic data protection at the software/analytics layer into encrypted federated computation, so that multi-Anonymous Privacy-preserving is made possible: security-embedding and compliance analytics across institutions.

Earlier, in the financial sector, Deep Hybrid CLST architectures have been tested recently with great success for credit card fraud detection (where complex temporal and behavioral patterns are modelled) [45]. However, such architectures are commonly applied in relatively centralized institutions and assume unrestricted access to aggregated financial data, which limits their usefulness in regulated, multi-institute environments. In contrast, FinanceRiskNet integrates an ML-based, lightweight MLP model into a federated learning framework to focus on privacy-preserving properties, secure aggregation requirements, and regulatory compliance. This design sacrifices some architectural expressiveness for the ability to deploy in private setups. Still, we find encouraging evidence from hybrid deep learning models that suggest several fruitful directions for future extensions (e.g., adding temporal or representation-learning components while continuing to execute in encrypted and federated settings).

Explainability is also a key need in healthcare analytics federations, where doctors from different institutions must trust models' decisions without access to central patient records. A meta-analytic review of explainable AI in clinical decision support systems underscores the crucial requirements for transparency, interpretability, and usability for clinical adoption [42]. While FedPrivEngine is currently limited to privacy-preserving and compliance-aware analytics, the modular architecture of HealthPrivNet enables seamless inclusion of explainability mechanisms (e.g., feature attribution or temporal attention) at the local institutional level while still satisfying federated privacy constraints.

A. Limitations of the Study

The proposed model, however effective, has limitations that should be carefully noted. The most noticeable drawback is that the proposed framework can only be evaluated on synthetic and benchmark datasets. While the OpenSAFELY synthetic dataset and FICO benchmark dataset are commonly used and ideal for controlled experiments, they do not sufficiently represent the heterogeneity, noise properties, missing-data profiles, or institutional variation characteristic of actual healthcare and financial systems. Consequently, conclusions about real-world deployment behavior, resilience to operational constraints, and long-term system stability should be considered with caution.

Real-world healthcare and financial datasets exhibit nontrivial shifts in data distributions across institutions, driven by differences in population characteristics, data-collection procedures, measurement errors, and system-level noise. Such heterogeneity might affect federated convergence patterns, encrypted aggregation efficiency, and compliance enforcement in ways that are not apparent from synthetic or benchmark datasets. Overall, it implies that the current analysis shows feasibility and good architecture, but is not empirical in real deployment.

A second important limitation is related to the experimental facilities. All federated nodes in the evaluation were running on a homogeneous cluster with the same hardware and network configurations. This is not representative of what happens in real federated environments, where contributing hospitals or financial institutions usually have heterogeneous infrastructures with different computational power, storage capacity, and network latencies. While the Spark-based orchestration layer of FedPrivEngine is in principle capable of accommodating heterogeneous execution environments, this feature was not experimentally tested in the present work. As a result, statements about robustness to infrastructure heterogeneity should be considered architectural hopes, not empirically validated results.

In addition, the given experimental setting did not account for network variation, intermittent connections, and different resources—frequent components of real-world federated systems. These may add synchronization latency, straggler effects, and encryption overhead, which can impact scalability and throughput. A comprehensive evaluation in heterogeneous hardware and network environments is left for the future.

Solving such limitations will require the real or semi-realistic deployment and evaluation of FedPrivEngine in a federated environment, such as collaborations between healthcare organizations and banks, with operational data but anonymized. Subsequent work will focus on validating how the system behaves across heterogeneous infrastructures, under different network conditions, and with real-world data distributions, to provide stronger empirical evidence for scalability, robustness, and practical applicability claims.

B. Privacy Guarantees and System Overhead Limitations

Although FedPrivEngine uses several privacy-preserving technologies (homomorphic encryption, secure aggregation, access control reinforcement, and an optional layer of differential privacy), the existing system framework does not provide theoretical end-to-end privacy guarantees. Specifically, differential privacy is applied as an optional, emulated system, showing that the architecture of legal implications and the composability of combining homomorphic encryption and differential privacy can be used instead of a proper integrated privacy accounting mechanism. More advanced features, such as adaptive privacy budgeting, ε-accounting over the complete separation between rounds, and formal composition of privacy loss, are not available in this work.

Additionally, this work does not offer provable guarantees or a generic adversarial model. Interactions and the composability of combining homomorphic encryption, differential privacy, and access control policies are not formally investigated. Our construction is not parametrized, and its privacy assurance is thus at the architectural and empirical levels rather than as a formal (theoretical) guarantee. Formally providing privacy and security guarantees is an important next step.

Furthermore, the claimed "zero policy violations" reported in the experimental evaluation should also be taken with a pinch of salt. These experiments were conducted using synthetic, well-established benchmark datasets in a controlled experimental setting, with precise communication requirements that cannot be bypassed by adversarial behavior, resulting in a 3.5× increase in communication payload. A maximum per-round encryption overhead of 2.3 s was permissible at this scale. Still, it may multiply substantially in large-scale production settings due to access patterns and non-adversarial query loads. These are not realistic threat scenarios involving malicious users, adaptive attacks, or attempts to circumvent the policy. That is, the lack of policy violations shows that the compliance logic was implemented correctly under a controlled setup—not that it cannot be bypassed by adversarial behavior in actual deployment.

The computational and communication overhead due to the use of homomorphic encryption presents non-trivial practical deployment issues, which are not studied in detail in this paper. The observed approximately 3.5× growth in communication payload and maximum per-round encryption overhead of 2.3 s was permissible at the scale but may multiply substantially in large-scale production settings such as millions of patient records or financial transactions. Realities such as long network latency, straggler effects, and repeated rounds of encrypted aggregation can significantly impact throughput and system responsiveness.

This work does not simulate or measure the overheads under extreme data volumes or real-time operational pressures. In the future, we plan to improve encryption pipelines, study hybrid or partial homomorphic schemes, assess batching and compression methods, and perform bulk stress-testing of the feasibility of systems in production-sized federated environments Table XI.

TABLE XI. NOTATIONS FOR FEDPRIVENGINE FEDERATED ANALYTICS FRAMEWORK

Notation	Description
H_i	The i^{th} hospital node or healthcare data silo
B_i	The i^{th} bank node or financial institution silo
D_i	Local dataset at node i, such that $D_H = \cup_{i-1}^{K_H} D_i$
$x_{i,j}$	Feature vector of the j^{th} instance in dataset D_i
$y_{i,j}$	Target label for the j^{th} instance in dataset D_i
θ_i	Local model parameters at node i
$\Delta\theta_i$	Parameter update from local training at node i
$\hat{y}_{i,j}$	Predicted output for input $x_{i,j}$ using model f_i
$\mathcal{L}_i$	Local loss function at node ii (cross-entropy for classification)
Enc (·)	Homomorphic encryption function
Dec (·)	Decryption function applied after secure aggregation
θ_{global}	Aggregated global model parameters from all participating nodes
q_i	Query result computed locally at node i
Q	Aggregated encrypted query result across all silos
ε	Privacy budget in differential privacy
σ	Standard deviation of Gaussian noise in differential privacy mechanism
K_H, K_F	Number of healthcare silos and financial silos respectively
$P_i(x,t,r)$	Compliance policy predicate based on attribute x, user type t, and region r
$f_i(x;\theta_i)$	Local model (GRU or MLP) prediction function at node i
Agg(.)	Federated aggregation operation over encrypted data

VI. Conclusion

This study presented FedPrivEngine, a systems-oriented federated data engineering framework designed to support privacy-aware analytics across healthcare and financial domains. By integrating homomorphic encryption, optional differential privacy, and Spark-based orchestration under a policy-aware query planner, the framework demonstrates how encrypted model training and query execution can be coordinated across distributed, institutionally governed silos. The dual-model configuration—HealthPrivNet and FinanceRiskNet—serves as a representative use case to evaluate the framework's applicability to heterogeneous analytical workloads, rather than as a novel modelling contribution.

Experimental results on synthetic and benchmark datasets demonstrate the feasibility of the proposed architecture and highlight key performance and privacy trade-offs in encrypted federated analytics. The evaluation demonstrates that secure collaboration can be achieved with acceptable overhead under controlled experimental conditions. However, these results should be interpreted as architectural validation rather than evidence of real-world deployment readiness or regulatory certification.

Several limitations remain and motivate future work. First, validation on real-world institutional datasets is required to assess robustness under realistic data heterogeneity, noise, and operational constraints. Second, while differential privacy is supported as an optional component, future work will focus on fully integrating adaptive privacy budgeting and formal privacy accounting mechanisms. Third, although CKKS-based homomorphic encryption provides strong practical confidentiality, further optimization is necessary to reduce communication and computation overhead. Future efforts will investigate hybrid cryptographic schemes, partial homomorphic alternatives, batching strategies, and hardware acceleration to improve latency and throughput.

While the current evaluation focuses on healthcare and finance, FedPrivEngine is designed with modularity in mind and can be extended to other settings such as IoT analytics, multimodal data processing, and vertical federated learning. Future studies will include broader benchmarking, ablation analyses, and evaluations across heterogeneous infrastructure configurations to characterize scalability and generalizability better.

Finally, we plan to release an open-source version of FedPrivEngine with containerized deployment, standardized APIs, and documentation to support reproducibility and community adoption. Usability studies with domain practitioners will also be conducted to assess deployment complexity and operational workflows in compliance-sensitive environments.

Conflict of Interest

The author declares no conflict of interest.

References

[1] K. A. Awan, I. U. Din, A. Almogren *et al.,* "Privacy-preserving big data security for IoT with federated learning and cryptography," *IEEE Access*, vol. 11, pp. 120918–120934, 2023.

[2] Q. Yang, A. Huang, L. Fan *et al.,* "Federated learning with privacy-preserving and model IP-right-protection," *Machine Intelligence Research*, vol. 20, no. 1, pp. 19–37, 2023.

[3] P. Riedel, R. von Schwerin, and D. Schaudt *et al.,* "ResNetFed: Federated deep learning architecture for privacy-preserving pneumonia detection from COVID-19 chest radiograph," *Journal of Healthcare Informatics Research*, vol. 7, no. 2, pp. 203–224, 2023.

[4] K. Sabiri, F. Sousa, and T. Rocha, "A systematic review of privacy-preserving blockchain applications in healthcare," *Multimed Tools Appl*., vol. 84, no. 32, pp. 39925–39980, 2025. https://doi.org/10.1007/s11042-024-20541-z.

[5] M. Khalila, M. Esseghira, and L. M. Boulahia, "Privacy-preserving federated learning: An application for big data load forecast in buildings," *Computers & Security*, vol. 131, 103211, 2023.

[6] P. Ruzafa-Alcázar, P. Fernández-Saura, E. Mármol-Campos *et al.,* "Intrusion detection based on privacy-preserving federated learning for the industrial IoT," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 1145–1154, 2023.

[7] Y. Yan, M. B. Alshawki, M. Zoltay *et al.,* "Fedlabx: A practical and privacy-preserving framework for federated learning," *Complex & Intelligent Systems*, vol. 10, no. 1, pp. 677–690, 2024.

[8] X. Wang, W. Fan, and X. Hu, "Differential privacy-preserving of multi-party collaboration under federated learning in data center networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 8, no. 2, pp. 1223–1237, 2024.

[9] N. A. Jalali and H. Chen, "Federated learning security and privacy-preserving algorithm and experiments research under Internet of Things critical infrastructure," *Tsinghua Science and Technology*, vol. 29, no. 2, pp. 400–414, 2024.

[10] T. H. Hwang, J. Shi, and K. Lee, "Enhancing privacy-preserving personal identification through federated learning with multimodal vital signs data," *IEEE Access*, vol. 11, pp. 121556–121566, 2023.

[11] M. Arazzi, S. Nicolazzo, and A. Nocera, "A fully privacy-preserving solution for anomaly detection in IoT using federated learning and homomorphic encryption," *Information Systems Frontiers*, vol. 27, no. 1, pp. 367–390, 2023.

[12] S. M. S. Bukhari, M. H. Zafar, M. Abou Hou *et al.,* "Secure and privacy-preserving intrusion detection in wireless sensor networks: Federated learning with SCNN-Bi-LSTM for enhanced reliability," *Ad Hoc Networks*, vol. 155, 103407, 2024.

[13] V. T. Truong and L. B. Le, "MetaCIDS: Privacy-preserving collaborative intrusion detection for metaverse based on blockchain and online federated learning," *IEEE Open Journal of the Computer Society*, vol. 4, pp. 253–266, 2023.

[14] T. Eltaras, F. Sabry, W. Labda *et al.,* "Efficient verifiable protocol for privacy-preserving aggregation in federated learning," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2977–2990, 2023.

[15] S. R. Abbas, Z. Abbas, A. Zahir *et al.,* "Federated learning in smart healthcare: A comprehensive review on privacy, security, and predictive analytics with IoT integration," *Healthcare*, vol. 12, 2587, 2024. doi: 10.3390/healthcare12242587

[16] M. Hiwale, R. Walambe, V. Potdar *et al.,* "A systematic review of privacy-preserving methods deployed with blockchain and federated learning for the telemedicine," *Healthcare Analytics*, vol. 3, 100192, 2023.

[17] H. Riaz, R. A. Naqvi, M. Ellahi *et al.,* "Robust steganography technique for enhancing the protection of medical records in healthcare informatics," *IEEE Journal of Biomedical and Health Informatics*, 2025.

[18] S. M. Darwish, R. M. Essa, M. A. Osman *et al.,* "Privacy preserving data mining framework for negative association rules: An application to healthcare informatics," *IEEE Access*, vol. 10, pp. 76268–76280, 2022.

[19] L. Song, C. Ma, G. Zhang *et al.,* "Privacy-preserving unsupervised domain adaptation in federated setting," *IEEE Access,* vol. 8, pp. 143233–143240, 2020.

[20] T. T. Kuo, X. Jiang, H. Tang *et al.,* "iDASH secure genome analysis competition 2018: Blockchain genomic data access logging, homomorphic encryption on GWAS, and DNA segment searching," *BMC Medical Genomics*, vol. 13, 98, 2020.

[21] D. Kiernan, T. Carton, S. Toh *et al.,* "Establishing a framework for privacy-preserving record linkage among electronic health record and administrative claims databases within PCORnet®, the national

patient-centered clinical research network," *BMC Research Notes*, vol. 15, no. 1, 337, 2022.
[22] K. Munjal and R. Bhatia, "A systematic review of homomorphic encryption and its contributions in healthcare industry," *Complex & Intelligent Systems*, vol. 9, no. 4, pp. 3759–3786, 2023.
[23] R. Zheng, A. Sumper, M. Aragüés-Peñalba *et al.,* "Advancing power system services with privacy-preserving federated learning techniques: A review," *IEEE Access*, vol. 12, pp. 76753–76780, 2024.
[24] M. Massaoudi, H. Abu-Rub, S. S. Refaat *et al.,* "Deep learning in smart grid technology: A review of recent advancements and future prospects," *IEEE Access,* vol. 9, pp. 54558–54578, 2021.
[25] M. Massaoudi, I. Chihi, H. Abu-Rub *et al.,* "Convergence of photovoltaic power forecasting and deep learning: State-of-art review," *IEEE Access*, vol. 9, pp. 136593–136615, 2021.
[26] A. Belhadi, Y. Djenouri, G. Srivastava *et al.,* "Privacy reinforcement learning for faults detection in the smart grid," *Ad Hoc Networks*, vol. 119, 102541, 2021.
[27] N. Tian, Q. Guo, H. Sun *et al.,* "Fully privacy-preserving distributed optimization in power systems based on secret sharing," *IEnergy*, vol. 1, no. 3, pp. 351–362, 2022.
[28] X. S. Shen, D. Liu, C. Huang *et al.,* "Blockchain for transparent data management toward 6G," *Engineering*, vol. 8, pp. 74–85, 2022.
[29] Y. Tang and Z. Liu, "A credit card fraud detection algorithm based on SDT and federated learning," *IEEE Access*, vol. 12, pp. 182547–182560, 2024.
[30] M. Jabeen, S. Ramzan, A. Raza *et al.,* "Enhanced credit card fraud detection using deep hybrid CLST model," *Mathematics*, vol. 13, no. 12, 1950, 2025.
[31] J. S. Lee and S. P. Jun, "Privacy-preserving data mining for open government data from heterogeneous sources," *Government Information Quarterly*, vol. 38, no. 1, 101544, 2021.
[32] B. Pulido-Gaytan, A. Tchernykh, J. M. Cortés-Mendoza *et al.,* "Privacy-preserving neural networks with Homomorphic encryption: Challenges and opportunities," *Peer-to-Peer Networking and Applications*, vol. 14, no. 3, pp. 1666–1691, 2021.
[33] X. S. Shen, C. Huang, D. Liu *et al.,* "Data management for future wireless networks: Architecture, privacy preservation, and regulation," *IEEE Network*, vol. 35, no. 1, pp. 8–15, 2021.
[34] J. Kim and A. Yun, "Secure fully homomorphic authenticated encryption," *IEEE Access*, vol. 9, pp. 107279–107297, 2021.
[35] S. Halder and T. Newe, "Enabling secure time-series data sharing via homomorphic encryption in cloud-assisted IIoT," *Future Generation Computer Systems*, vol. 133, pp. 351–363, 2022.
[36] M. Ihtesham, S. Tahir, H. Tahir *et al.,* "Privacy preserving and serverless homomorphic-based searchable encryption as a service (SEaaS)," *IEEE Access*, vol. 11, pp. 115204–115218, 2023.
[37] S. Almakdi, B. Panda, M. S. Alshehr *et al.,* "An efficient secure system for fetching data from the outsourced encrypted databases," *IEEE Access*, vol. 9, pp. 78474–78494, 2021.
[38] M. Zichichi1, S. Ferretti, G. D'Angelo *et al.,* "Data governance through a multi-DLT architecture in view of the GDPR," *Cluster Computing*, vol. 25, no. 6, pp. 4515–4542, 2022.
[39] R. T. Moreno, J. García-Rodríguez, J. B. Bernabé *et al.,* "A trusted approach for decentralised and privacy-preserving identity management," *IEEE Access*, vol. 9, pp. 105788–105804, 2021.
[40] T. Le and S. Shetty, "Artificial intelligence-aided privacy preserving trustworthy computation and communication in 5G-based IoT networks," *Ad Hoc Networks*, 126, 102752, 2022.
[41] A. K. Gavai, Y. Bouzembrak, D. Xhani *et al.,* "Agricultural data privacy: Emerging platforms & strategies," *Food and Humanity*, vol. 4, 100542, 2025.
[42] J. E. Rivadeneira, J. S. Silva, R. Colomo-Palacios *et al.,* "User-centric privacy preserving models for a new era of the internet of things," *Journal of Network and Computer Applications*, vol. 217, 103695, 2023.
[43] H. Wang, L. Sun, and E. Bertino "Building access control policy model for privacy preserving and testing policy conflicting problems," *Journal of Computer and System Sciences,* vol. 80, no. 8, pp. 1493–1503, 2014.
[44] D. M. Bean, J. Teo, R. Bendayan *et al.* (2020). OpenSAFELY: A secure analytics platform for electronic health records in the NHS. medRxiv. [Online]. Available: https://www.opensafely.org/
[45] FICO. (2018). Explainable machine learning challenge: Home Equity Line of Credit (HELOC) risk data. *FICO Community*. [Online]. Available: https://community.fico.com/s/explainable-machine-learning-challenge