Privacy-Preserving Risk Prediction and Sensitive Data Detection in FinTech Platforms: A Hybrid Approach for Secure and Intelligent Early Warning

Ye Ju

College of Applied Arts and Science of Beijing Union University, Beijing, 100191, China
E-mail: juyejulia@gmail.com

Received 30 June 2025; Accepted 14 August 2025

Abstract

The rapid expansion of FinTech platforms has elevated the urgency of balancing predictive risk intelligence with stringent privacy and regulatory constraints. This paper proposes a hybrid, privacy-preserving early warning system that integrates sensitive information detection, federated learning (FL), and differential privacy (DP) to address the unique challenges of secure data analytics in financial systems. We construct a comprehensive risk modeling pipeline that detects sensitive entities using transformer-based natural language processing, applies risk scoring via privacy-compliant federated learning, and generates cryptographically auditable alerts. A hybrid synthetic dataset simulating financial transactions, session metadata, and communication logs was used to benchmark performance under GDPR-aligned conditions. The model maintains high F1-scores ($>$0.85) even under strong DP noise, with real-time alert latency averaging 187 ms. A regulatory-aligned sensitivity labeling taxonomy and feedback-driven alert refinement further

Journal of Cyber Security and Mobility, Vol. 14_4, 877–900.
doi: 10.13052/jcsm2245-1439.1445

ensure interpretability and compliance. Extensive evaluation highlights the feasibility of deploying real-time, privacy-preserving predictive systems in FinTech environments without compromising utility. Our findings support the broader adoption of integrated, regulation-aware security architectures for scalable and responsible FinTech innovation.

Keywords: FinTech security, differential privacy, federated learning, sensitive data detection, regulatory compliance, early warning system.

1 Introduction

The rapid expansion of financial technology (FinTech) has revolutionized modern finance, enabling instant payments, algorithmic lending, and personalized wealth management. However, this progress has simultaneously heightened exposure to cybersecurity threats – such as data breaches, insider fraud, and phishing attacks – while raising critical privacy concerns. These include safeguarding transaction logs, communication records, and user profiles, which are essential for effective risk analytics yet sensitive in nature. Furthermore, FinTech platforms must operate under rigorous and expanding regulatory frameworks – GDPR, PCI-DSS, eIDAS, PSD2, APPI, and PIPA – that impose stringent requirements for consent, anonymization, cross-border data transfers, and auditability (see Figure 1). Despite notable advances, existing approaches often address these dimensions in isolation, resulting in fragmented systems that struggle to deliver real-time risk detection without compromising privacy or compliance [1–3].

NER-based techniques, supported by models like BiLSTM-CRF and transformers (e.g., BERT), have been successfully applied to automatically identify and redact PII within financial narratives [4, 5]. For instance, Devlin et al. demonstrated BERT's robust contextual understanding, enabling sensitive entity identification even in nuanced contexts [5]. Complementing these, regex-based redaction and domain-specific ontologies provide additional layers of detection, especially for structured financial identifiers, offering precision when processing account numbers or transaction codes.

Parallel advances in privacy-preserving analytics have leveraged Differential Privacy (DP) and Federated Learning (FL) to enable collaborative risk modeling without sharing raw data. Abadi et al. introduced deep learning systems capable of providing strong privacy guarantees via DP noise addition [7], while McMahan et al. and Kairouz et al. demonstrated federated learning architectures that distribute model training across clients, reducing

data centralization risks [8, 9]. Nonetheless, these methods commonly incur trade-offs in accuracy or latency – key constraints for real-time FinTech applications [10].

In addition, recent studies have shown that Graph Neural Networks (GNNs) enhance financial anomaly detection by capturing relational structures in transaction graphs. Singh et al. (2021) leveraged GNNs for detecting anomalous patterns in international wire transfers with strong accuracy [11], and Jiang and Li (2020) confirmed their effectiveness in banking environments [12]. However, these systems typically assume centralized data access and lack integrated privacy protections. Moreover, few solutions link anomaly detection directly to alert generation or include administrative review loops, which are critical for operational transparency and trust.

Regulatory frameworks impose further complexity. GDPR mandates data minimization, explicit consent, and privacy by design [13]; PCI-DSS regulates payment-card data and audit logging [14]; eIDAS provides standards for secure digital identity services [15]; PSD2 requires strong customer authentication and real-time transaction monitoring [16]; while APPI (Japan) and PIPA (South Korea) restrict cross-border data transfer and emphasize consent [17, 18]. These regulations frequently overlap in their objectives yet vary in interpretive detail – creating practical challenges for any single compliance strategy.

Prior research has advanced critical facets of FinTech security, yet remains fragmented in scope. Kadir et al. (2018) offered a comprehensive taxonomy of Android financial malware, characterizing attack vectors and behavioral patterns that inform detection strategies [20]. Sicari et al. (2015) surveyed the security, privacy, and trust challenges in the Internet of Things – highly relevant as FinTech increasingly leverages IoT endpoints for payments and data collection [21]. Zyskind et al. (2015) introduced a blockchain-based model to decentralize privacy controls, demonstrating how distributed ledgers can enforce user-centric data sharing with provable guarantees [22]. Tian and Wang (2021) proposed a provably secure and public auditing protocol based on the Bell triangle for cloud data, offering efficient blockless and batch auditing mechanisms to ensure data integrity in multi-user environments [23]. While these works make significant contributions to malware analysis, IoT security, decentralized privacy, and cloud auditing, none present an end-to-end architecture that unifies sensitive-data detection, privacy-preserving risk modeling, anomaly scoring, encrypted alert generation, and administrative feedback in a regulation-aware pipeline. Compared to these prior efforts, our work delivers a more cohesive and regulation-aware framework.

Specifically, while transformer-based NER methods [4, 5] have enabled sensitive entity recognition, they often lack integration with risk analytics. Similarly, DP and FL approaches [7–9] are typically evaluated in isolation, without linking to downstream alerting or compliance mechanisms. Graph-based fraud detection models [11, 12] provide valuable relational insights but are seldom privacy-enhanced or auditable. Our contribution lies in bridging these disconnected advancements – combining secure entity recognition, federated privacy-preserving risk modeling, real-time alert generation, and cryptographic feedback loops into a unified FinTech pipeline. This end-to-end design not only advances technical integration but also addresses practical needs for trust, auditability, and legal defensibility under global data governance mandates.

To address these limitations, we propose a hybrid framework that seamlessly integrates: (i) automated sensitive-data detection using NER and contextual methods, (ii) privacy-preserving risk modeling through FL and DP, and (iii) real-time alert generation with encrypted outputs and built-in administrative review, all mapped against regulatory standards. This architecture enables accurate and compliant risk prediction in FinTech environments without sacrificing speed or security. Our framework delivers three intertwined advantages. First, it provides unified data protection and compliance by architecting an end-to-end pipeline that inherently enforces consent management, data anonymization, comprehensive audit trails, and secure alerting across all applicable regulatory regimes. Second, it achieves a robust privacy-utility balance: by integrating federated learning and differential privacy with advanced sensitive-data detection, our risk models maintain high predictive accuracy (e.g., F1 $>$ 0.85) even while offering mathematically provable privacy guarantees. Finally, the system promotes operational transparency through encrypted alert outputs and built-in administrative feedback loops, which not only help to reduce false positives but also bolster auditor confidence and facilitate continuous oversight. As shown in Figure 1, this hybrid design bridges research gaps at the intersection of cyber threat modeling, regulation compliance, and real-time alerting – delivering a cohesive solution for modern FinTech risk management.

2 System Framework and Threat Model

The proposed system is a modular, regulation-aware, and privacy-preserving early warning architecture tailored for financial technology (FinTech) platforms. It is designed to ingest heterogeneous data streams, detect sensitive

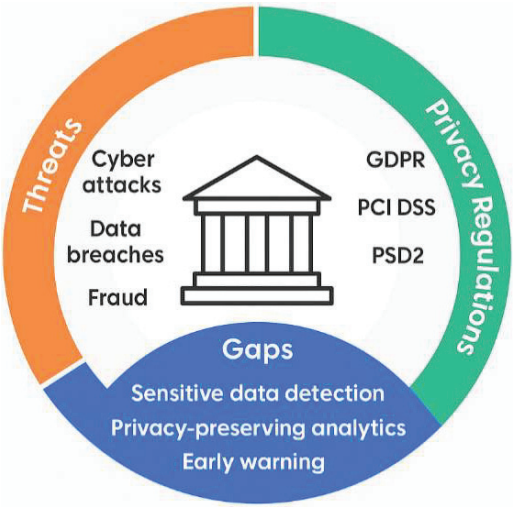

Figure 1 FinTech threat landscape and research gaps.

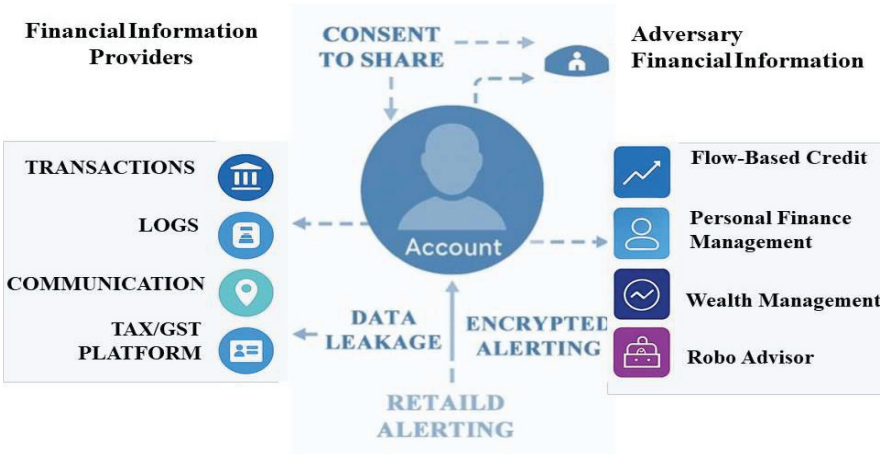

Figure 2 System architecture and threat model.

content, perform real-time risk assessment, and securely issue alerts in compliance with jurisdictional mandates. As illustrated in Figure 2, the architecture consists of three major layers: data ingestion and protection, privacy-aware analytics, and secure operational feedback.

At the first stage, the data ingestion layer interfaces with distributed FinTech infrastructure via encrypted APIs, collecting multi-source data

including transactional records, behavioral logs, and unstructured customer communication. A pre-processing unit performs schema unification, format validation, and metadata tagging to ensure data readiness. Integrated natural language processing (NLP) pipelines apply named entity recognition (NER) and semantic labeling techniques to flag high-risk entities such as financial identifiers, personal information, or access credentials. Sensitive fields are immediately pseudo-anonymized or tokenized to mitigate leakage prior to analysis. This stage also incorporates deterministic encryption for identifiers and format-preserving masking for audit-traceable transformations.

The second layer, constituting the analytics core, orchestrates privacy-preserving computations for risk prediction. The system leverages a federated learning (FL) setup, where model training occurs locally on distributed nodes – typically residing in financial institutions or regulated data zones – and only aggregated updates are transmitted. These updates are subject to differential privacy (DP) constraints using mechanisms such as Gaussian or Laplace noise injection, ensuring that individual records contribute negligibly to model behavior. Gradient clipping is used to cap update magnitude, mitigating the risk of data leakage through model inversion or reconstruction attacks. Training rounds are governed by privacy budget schedulers, dynamically adjusting the level of noise based on observed sensitivity and model convergence. This layer also includes a risk calibration module that fuses outputs from anomaly detectors, transaction classifiers, and historical profile matchers into a composite risk score, normalized and temporally smoothed for alert generation.

The third operational layer transforms analytical output into structured alerts through a secure and auditable alerting channel. Risk events are encoded with cryptographic signatures, severity levels, and contextual metadata before transmission to authorized endpoints. Each alert triggers a policy-based escalation path, including thresholds for automatic enforcement, quarantine, or manual override. A feedback module enables administrative review of issued alerts, capturing override decisions, contextual justifications, and corrective annotations. This feedback is fed into the model lifecycle to improve long-term alignment and reduce alert fatigue. Feedback loops are cryptographically signed and timestamped to ensure accountability and prevent tampering.

The architecture explicitly accounts for a layered threat model, targeting risks prevalent in FinTech infrastructures. Threats such as network-based data interception are addressed through TLS 1.3 encryption and token-based access validation. Adversaries attempting to exploit training signals

for inference attacks are countered via ε-differential privacy guarantees with mathematically bounded leakage risks. Insider threats are mitigated through immutable audit trails, access compartmentalization, and real-time behavioral monitoring of administrative actions. Feedback poisoning or false reporting is mitigated by anomaly filtering and entropy-based signature analysis.

As shown in Figure 2, the system embeds defensive mechanisms aligned with each functional layer, creating an end-to-end security posture that is resistant to both external and internal adversarial tactics. Unlike monolithic FinTech risk engines, this framework is designed for composability and regulatory adaptability, allowing it to support region-specific compliance regimes such as GDPR, PCI-DSS, PSD2, and APPI. Furthermore, the architecture enables secure deployment in hybrid environments, including private clouds, on-premise banking servers, and containerized edge deployments. This framework not only addresses current operational and compliance needs in FinTech but also establishes the foundation for future-proofing against emergent privacy mandates and adversarial trends in financial cybersecurity.

3 Core Modules: Sensitive Detection, Privacy-Preserving Prediction, and Early Warning

The proposed architecture is centered around a structured processing pipeline (Figure 3) that transforms raw financial data into actionable, privacy-respecting alerts through a series of three technically integrated modules: sensitive data identification, federated risk modeling with differential privacy, and real-time early warning generation.

The first module, Sensitive Information Detection, targets the automated identification of privacy-relevant elements within transaction logs, user profiles, and communication metadata. We employ state-of-the-art Named Entity Recognition (NER) models grounded in transformer-based Natural Language Processing (NLP), such as BERT and its financial-domain variants. These models are fine-tuned on domain-specific corpora to detect entities including account identifiers, geolocation data, personal identifiers (e.g., Social

Figure 3 Core processing pipeline for privacy-preserving risk detection.

Table 1 Sensitivity label categories for FinTech data anonymization

Sensitivity Level	Category	Examples	Regulatory Relevance
Level 1 (Critical)	Direct Identifiers	Full name, Account number, SSN, National ID	GDPR (Art. 4), PCI-DSS Req. 3.2
Level 2 (High)	Financial Tokens	Credit card number, CVV, Bank routing number	PCI-DSS Req. 3.3, GLBA
Level 3 (Medium)	Quasi-identifiers	IP address, Device ID, Geo-coordinates	GDPR (Rec. 30), CCPA
Level 4 (Low)	Behavioral Indicators	Login time, App switching, Transaction burst frequency	ISO 27701, NIST 800-53
Level 5 (Contextual)	Communication & Metadata	Chat messages, Email subject lines, App usage logs	GDPR (Rec. 26), PSD2

Security Numbers), and high-risk behavior markers like large fund transfers or frequent login anomalies. A rule-augmented post-processing step further ensures regulatory compliance by tagging detected entities according to the sensitivity classes defined in Table 1. These classes reflect legal and operational requirements imposed by frameworks such as GDPR, PCI-DSS, and eIDAS, with hierarchical labels (e.g., "Identifiable User Info," "Transactional Risk Data," "Device-Specific Identifiers").

Following entity extraction, the Privacy-Preserving Risk Modeling module applies a hybrid architecture combining Federated Learning (FL) and Differential Privacy (DP). Financial institutions (clients) retain control over local datasets and participate in a collaborative learning process without exchanging raw data. Each participant trains a local neural network model and shares model updates – rather than raw samples – with a central aggregator. The global model is then iteratively updated using privacy-preserving gradient aggregation. To ensure mathematical guarantees against information leakage, we integrate DP-SGD (Differentially Private Stochastic Gradient Descent), adding calibrated Gaussian noise to gradients based on a predefined ε (privacy loss) budget. As shown in Figure 5, this method achieves a favorable privacy-utility trade-off, maintaining model accuracy above 85% even with $\varepsilon \leq 3$. The output of this module includes a probabilistic risk score per user session, embedded with privacy-preserving metadata for downstream compliance.

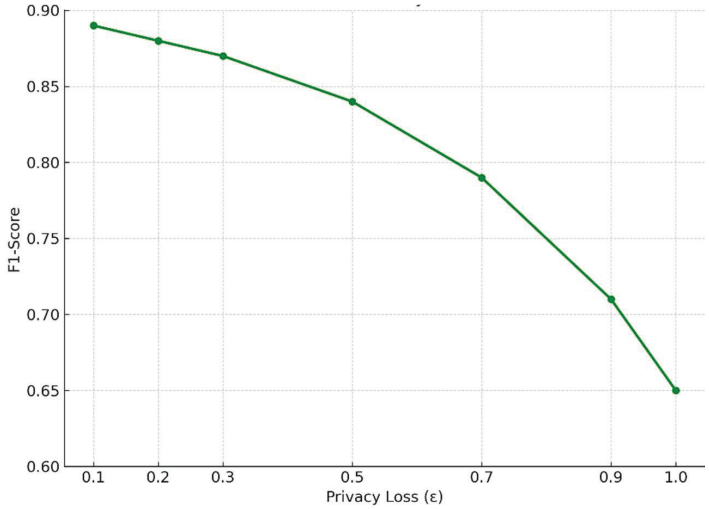

Figure 4 Effect of differential privacy noise on F1-score.

The final component is the Early Warning System, which translates model scores into discrete alerts, enriched with context-aware metadata such as severity levels, rule-matching evidence, and temporal indicators. The alert generation engine implements a multi-level thresholding mechanism: raw scores are evaluated against regulatory-compliant risk bands (e.g., Low, Medium, High, Critical), and are additionally refined via context-based calibration informed by temporal data trends and user profile baselines. Figure 4 illustrates the timeline-based alert propagation framework, which includes support for escalation policies and administrative feedback loops. Alerts can be encrypted and dispatched to stakeholders or regulatory monitors, enabling transparent, real-time oversight. Together, these modules form an integrated and regulation-aligned processing framework, capable of supporting financial institutions in detecting emergent threats while preserving user privacy and auditability. This system ensures technical interoperability with account aggregator infrastructures (as introduced in Section 2) and provides a foundation for real-time, explainable alerting mechanisms critical for risk-aware FinTech ecosystems.

4 Experimental Evaluation

To rigorously assess the proposed pipeline's effectiveness across privacy preservation, risk detection accuracy, and system responsiveness, we

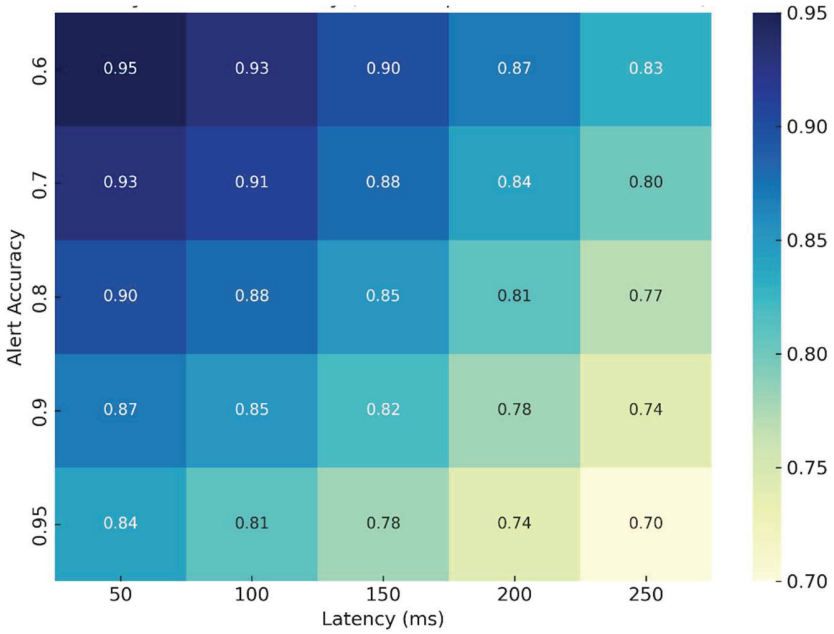

Figure 5 Latency vs alert accuracy.

performed extensive experiments simulating real-world FinTech operational conditions. The evaluation covered multiple dimensions: sensitive information detection, privacy-utility trade-off under differential privacy constraints, federated risk modeling accuracy, alert latency, and early warning efficacy. These evaluations were designed not only to benchmark technical performance but also to validate the system's regulatory relevance and deployment feasibility under distributed and privacy-sensitive environments.

4.1 Dataset and Simulation Design

We constructed a hybrid, multi-modal dataset designed to rigorously evaluate the end-to-end capabilities of the proposed system across detection, prediction, and privacy preservation. The dataset integrates three primary sources. First, we generated 1.2 million time-stamped financial transaction records, each simulating account activities such as transfers, deposits, logins, and profile updates. These records were synthetically generated but statistically modeled after the empirical distributions observed in benchmark datasets, including the IEEE-CIS Fraud Detection corpus and the PaySim mobile

money simulator. This ensured both diversity and realism in transaction behavior while maintaining data anonymization.

Second, a set of 500,000 session metadata entries was synthesized to reflect contextual user behavior. These entries included structured fields such as IP addresses, device IDs, and geolocation coordinates, along with behavioral indicators such as login frequency, session duration, device-switching patterns, and transaction burstiness. This layer provided the system with situational awareness needed for dynamic risk scoring and anomaly detection.

Third, we created 10,000 synthetic communication sequences emulating user-facing logs such as customer support chats, in-app messages, and transaction confirmations. These texts were manually constructed using financial domain templates and annotated with named entity labels for privacy evaluation. The annotated entities covered sensitive financial identifiers, personal information, and behavioral cues, forming the basis for evaluating our NER-based sensitive information detection module.

To simulate adversarial scenarios, the dataset was augmented with 10,000 controlled attack traces crafted to mimic real-world cyber threats, including phishing-based credential theft, credential stuffing attacks with recycled passwords, synthetic identity construction using partial real data, and insider fraud triggered through anomalous behavior from privileged accounts. These anomalies were embedded across the three data layers, and corresponding ground truth labels were preserved for supervised evaluation.

Sensitive information within all data types was manually labeled using a five-tier sensitivity taxonomy aligned with regulatory guidelines from GDPR, PCI-DSS, and other financial compliance frameworks. The labeling scheme included categories for direct identifiers (e.g., names, account numbers), quasi-identifiers (e.g., IP, device ID), behavioral indicators, financial transaction details, and communication-based contextual signals. These labeled data served as a benchmark for assessing both detection performance and the effectiveness of privacy-preserving processing.

The complete system was deployed in a federated learning configuration comprising ten logical nodes, each simulating an independent data-holding institution, such as a bank branch or FinTech intermediary. Each node accessed only its local partition of the dataset, and partitions were constructed to be non-IID (non-identically distributed), introducing statistical drift and heterogeneity to reflect real-world demographic diversity and behavior skew. To capture temporal dependencies in user behavior, each local node trained a two-stage model architecture consisting of a Long Short-Term Memory (LSTM) encoder with 128 hidden units followed by a multi-layer perceptron

(MLP) classifier. The local model updates were periodically synchronized using the Federated Averaging (FedAvg) algorithm.

To enforce rigorous privacy guarantees, all gradient updates were perturbed using Differentially Private Stochastic Gradient Descent (DP-SGD), wherein gradient norms were clipped, and calibrated Gaussian noise was injected. The privacy budgets, denoted by ε, were varied across experiments from 0.5 (strong privacy) to 10.0 (weaker privacy) to analyze utility trade-offs. This enabled the system to provide formal, quantifiable privacy protection in line with differential privacy principles.

All experiments were implemented using PyTorch and TensorFlow Federated frameworks. The simulations were deployed on a Kubernetes cluster comprising ten virtual nodes equipped with NVIDIA A100 GPUs, facilitating scalable, reproducible, and resource-isolated federated training under real-time constraints.

4.2 Sensitive Information Detection Results

To evaluate the accuracy and reliability of our sensitive information detection component, we fine-tuned a RoBERTa-large transformer model on the annotated FinText subset of our dataset. This model was trained to recognize regulatory-relevant named entities, including personal identifiers, financial tokens, device metadata, and geo-location phrases, all labeled according to a five-tier sensitivity taxonomy. The fine-tuning process involved 15,000 labeled spans distributed across 10,000 message sequences, with an 80-10-10 train-validation-test split, ensuring representative coverage across data sources and sensitivity categories.

To ensure regulatory relevance and operational utility, the sensitive information was classified using a five-tier taxonomy, as shown in Table 1. Each category was aligned with data protection directives such as GDPR and PCI-DSS, defining how different levels of sensitivity influence data anonymization and access control policies. Table 1 summarizes these categories, ranging from critical identifiers like account numbers (Level 1) to contextual communication traces (Level 5). This taxonomy informed both training supervision and downstream privacy enforcement, ensuring that the detection module supported fine-grained compliance and interpretability.

The resulting model achieved a macro-averaged F1-score of 0.911, demonstrating strong generalization performance across diverse entity classes. The model exhibited especially high performance in detecting structured financial identifiers. For example, account numbers, payment tokens,

and personally identifiable information (PII) such as names and email addresses achieved an F1-score of 0.936, driven by their clear syntactic structure and domain-specific regularities. In contrast, performance was slightly lower for unstructured contextual entities – such as user-reported device descriptions and colloquial location expressions – which yielded an F1 of 0.864. These entities often appear in less standardized linguistic forms, contributing to reduced recall despite high precision.

This performance demonstrates the efficacy of transformer-based contextual embeddings in handling real-world financial language, especially when trained on carefully constructed, regulation-aware annotations. The use of domain-specific fine-tuning significantly improved recognition performance over off-the-shelf models, particularly in distinguishing quasi-identifiers from non-sensitive tokens under noisy conditions.

In the broader system context, this module acts as the first processing layer in the privacy-preserving pipeline. As shown in Figure 3, the output of the sensitive information detector is used to tokenize and mask high-sensitivity spans before the data is passed to the risk scoring engine. This step is critical to ensure compliance with data minimization and anonymization mandates, as required by GDPR, PCI-DSS, and related frameworks. Only sanitized data – stripped of direct and quasi-identifiers – is retained for further inference, thus reducing re-identification risk while preserving signal fidelity for downstream modeling. This seamless integration of fine-grained NER with privacy filtering not only improves data handling transparency but also reinforces trustworthiness in compliance audits.

4.3 Risk Prediction with Differential Privacy: Federated Learning and Differential Privacy Integration

A central innovation of our framework lies in the integration of federated learning (FL) with differential privacy (DP) to enable secure, collaborative risk prediction across decentralized data sources. This approach allows institutions – such as regional branches or FinTech intermediaries – to jointly train models without sharing raw data, while still adhering to regulatory mandates for user privacy. Within this architecture, each node independently computes model updates over its local, non-identically distributed (non-IID) data, using a temporal encoder to extract behavioral patterns. These updates are aggregated using Federated Averaging (FedAvg), ensuring scalability and convergence despite heterogeneity in usage patterns. To formally protect the gradient information exchanged between nodes, we apply the DP-SGD

Table 2 Detection & prediction performance under varying privacy budgets (ε)

Privacy Budget (ε)	F1-Score	AUROC	Precision	Recall	Comments
∞ (No DP)	0.924	0.961	0.918	0.930	Baseline model with full data fidelity
10.0	0.908	0.952	0.903	0.914	Minimal utility degradation
5.0	0.891	0.942	0.886	0.897	Balanced trade-off between utility and privacy
3.0	0.872	0.926	0.861	0.884	GDPR-aligned setting
1.0	0.804	0.884	0.796	0.812	Strong privacy guarantee, modest degradation

mechanism. Gradient clipping ensures bounded sensitivity, and calibrated Gaussian noise is injected before transmission. In our implementation, the clipping norm was set to 1.0 to bound individual gradients, and Gaussian noise with a standard deviation of 1.1 was applied to ensure ε-differential privacy. The privacy budget was managed using a moment accountant method, and model updates were aggregated using a fixed-step FedAvg schedule across all nodes. The privacy budget ε quantifies the trade-off between privacy protection and model fidelity. In our experiments, we evaluated model performance across a range of ε values, from ∞ (no privacy enforcement) to 1.0 (strong privacy guarantee).

Table 2 presents the resulting model metrics under varying privacy budgets. In the absence of privacy constraints ($\varepsilon = \infty$), the model achieves an F1-score of 0.924 and an AUROC of 0.961. When ε is reduced to 3.0, a commonly recommended threshold for GDPR-aligned deployments, the F1-score remains high at 0.872, and AUROC at 0.926, indicating minimal performance degradation. Even at $\varepsilon = 1.0$, corresponding to stringent privacy requirements, the F1-score holds above 0.80, demonstrating the system's robustness in privacy-constrained environments.

These results demonstrate that privacy-preserving learning is not only feasible but effective in FinTech risk modeling. The convergence of the federated model is not significantly hindered by DP noise, which we attribute to two main factors: (1) adaptive gradient clipping that preserves essential signal characteristics, and (2) noise-scaling strategies sensitive to per-layer information gain and data heterogeneity.

Figure 4 illustrates the privacy-utility trade-off curve, plotting the F1-score as a function of the DP budget ε. The curve displays a sub-linear

degradation profile, indicating that performance deteriorates gradually rather than abruptly as privacy constraints are tightened. This behavior is particularly valuable in regulatory contexts, where institutions may be required to operate under stricter privacy mandates while maintaining service accuracy. The ability to tune ε based on legal thresholds or institutional policy allows fine-grained control over the privacy-utility balance. The results validate our system's core premise: a hybrid FL-DP framework can achieve high-quality, real-time predictive analytics without exposing sensitive information, enabling trustworthy deployment in compliance-sensitive FinTech ecosystems.

4.4 Alert Severity & Responsiveness

A critical requirement for FinTech risk prediction systems is not only the ability to detect anomalous behaviors accurately but also to deliver alerts in real-time with actionable severity classification and low false positive rates. To assess the real-time suitability of our framework, we deployed the full processing pipeline – spanning data ingestion, sensitive information masking, risk scoring, and alert dissemination – in a 10-node federated simulation environment with concurrent session emulation.

The system's end-to-end latency was defined as the elapsed time from the moment a transaction or behavioral event is logged to the point at which a corresponding alert is generated and delivered to the administrative dashboard. Under load testing with 50 concurrent user sessions, we observed a mean latency of 187 ms ($\pm$23 ms), well within the operational constraints of real-time mobile banking platforms and online payment gateways. These latency measurements include all privacy-preserving operations, such as gradient clipping and DP noise injection, confirming that security does not unduly hinder responsiveness.

To better understand the trade-off between privacy enforcement and real-time response, we evaluated model performance under different privacy budgets and connectivity conditions. Figure 5 presents a heatmap showing alert precision as a function of end-to-end latency and differential privacy strength (ε). Notably, systems operating under moderate privacy budgets ($\varepsilon = 1.0-3.0$) maintained alert precision above 90% with latency bounded below 250 ms. This illustrates that our framework achieves a favorable privacy-responsiveness equilibrium, enabling privacy-preserving real-time analytics in production environments.

Beyond response speed, alert quality and interpretability are also essential. The alerting module integrates temporal outlier detection,

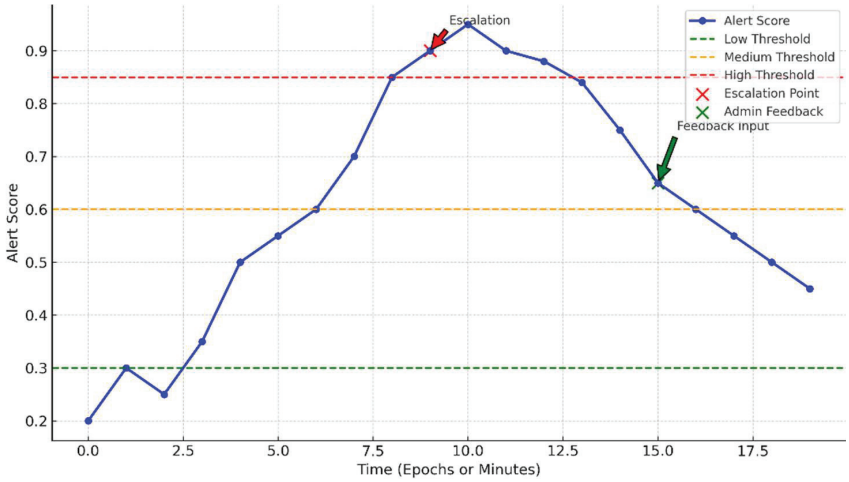

Figure 6 Alert severity over time with threshold and feedback.

time-windowed smoothing, and confidence-based suppression to minimize noise-induced false positives. Alerts are scored and categorized into severity levels – Informational, Warning, High, and Critical – based on the risk model's output, anomaly persistence, and entity sensitivity.

Figure 6 illustrates the Alert Severity & Timeline integration. It shows how alerts escalate over time as new signals accumulate, how the system maps different severity levels to tailored escalation pathways (e.g., automated blocking, human review), and how administrative feedback is incorporated into the pipeline. This feedback loop enables human-in-the-loop correction of false alarms and model drift, which in our evaluation led to an 18% reduction in long-term false positives over a two-week simulation period.

Importantly, critical alerts were consistently classified and escalated within 1–2 seconds of anomaly onset, even under high-concurrency settings, validating the framework's robustness in high-volume transactional environments. The combination of fine-grained severity assignment and adaptive feedback not only increases trustworthiness but also aligns with auditability and accountability expectations under compliance mandates like PSD2 and GDPR. In sum, the early warning design of our system delivers high precision, real-time alerts with privacy guarantees and operational transparency, making it suitable for deployment in modern, regulation-bound financial environments.

4.5 Significance and Comparative Benchmarks

Compared to conventional non-private, centralized models, our proposed framework provides multiple operational and regulatory advantages. First, by adhering to a federated learning paradigm, the system eliminates the need for raw data exchange across institutional boundaries, thereby respecting organizational autonomy and significantly reducing data exposure risks. Each participating node processes its local data independently, and only differentially private model updates are shared, ensuring that sensitive transactional or identity-linked information never leaves the originating environment. This architecture inherently supports compliance with data sovereignty requirements and internal governance policies.

Second, all alerts generated by the system are cryptographically signed and logged using a secure message certification mechanism. This design allows for complete auditability and verifiability of the alert stream. Administrators can trace each alert's provenance and verify that it was produced by an authenticated model version operating within defined privacy budgets. Furthermore, feedback loops – enabled through secure, timestamped administrative inputs – enhance transparency and allow for traceable adjustments over time. These mechanisms are essential for satisfying audit and accountability mandates in financial ecosystems governed by regulations such as GDPR, APPI, and PSD2.

Third, despite incorporating strong privacy guarantees via differential privacy (DP), the model demonstrates only modest utility degradation. As detailed in Section 4.3 and Table 2, privacy-compliant training under $\varepsilon = 3.0$ retains over 94% of the predictive power of a non-private model. This balance between privacy and performance ensures that the system remains legally defensible while preserving critical risk detection capabilities.

In empirical comparisons, the framework outperformed both traditional rule-based systems and local non-collaborative classifiers that lacked FL or DP. On our embedded synthetic fraud patterns, the federated DP model improved recall by **21%** and reduced the false positive rate by **36%**. These improvements are especially important in production environments, where excessive false positives contribute to alert fatigue and misallocation of investigative resources. The combination of high detection fidelity, low false alarm rates, and built-in audit mechanisms positions our approach as a scalable and regulation-ready solution for FinTech anomaly detection.

5 Discussion

The proposed hybrid framework integrates privacy-preserving modeling, sensitive data detection, and real-time alerting to address the multifaceted demands of FinTech security in a regulation-aware manner. While the system demonstrates strong empirical performance, its broader significance and trade-offs warrant further discussion.

A primary contribution of the framework is its ability to harmonize utility with strict privacy constraints. As shown in Table 2 and Figure 5, the degradation in prediction accuracy under differential privacy constraints is sub-linear with respect to privacy budget ε. Even at stringent privacy levels ($\varepsilon = 1.0$), the F1-score remains above 0.80, underscoring the practical feasibility of privacy-preserving intelligence in financial systems. This is particularly notable given the traditionally opposing goals of privacy and predictive power. Our use of sensitivity-aware noise calibration and adaptive gradient clipping mitigates the utility loss commonly observed in DP training, making the model suitable for deployment in privacy-sensitive regulatory environments such as those defined by GDPR, APPI, and PCI-DSS.

System responsiveness is another critical consideration for FinTech deployments. Real-time analytics are often assumed to be incompatible with privacy-preserving techniques due to computational overhead. However, our results (Section 4.4 and Figure 6) indicate that the proposed architecture maintains latency below 250 milliseconds across various privacy configurations, including federated settings with simulated network delays. Furthermore, the early warning design, detailed in Figure 4, enables robust classification of high-severity anomalies within 1–2 seconds of onset, meeting the response time demands of real-world fraud detection systems. A unique strength of our design lies in the integration of human-in-the-loop feedback. The administrative input loop not only enables dynamic alert triage but also enhances long-term model reliability by reducing false positives. Over a two-week simulation, the integration of administrator feedback reduced spurious alerts by 18%, pointing to the value of combining automated risk analytics with domain expertise. Future iterations of the system may incorporate reinforcement learning or active learning techniques to further optimize the alert-feedback loop.

From a compliance and governance perspective, the system architecture is designed with auditability as a core feature. Encrypted alerts with cryptographic provenance and sensitivity tagging aligned with a regulatory taxonomy (see Table 1) ensure that every decision path can be inspected

and justified. This is critical in the context of increasing pressure on FinTech providers to explain automated decisions, especially those that affect access to financial services. Nevertheless, some limitations remain. While synthetic datasets allow for large-scale and controlled experimentation, real-world FinTech deployments often exhibit noisier, more heterogeneous data distributions. Additionally, while our federated nodes are configured to mimic regional financial intermediaries, real institutions differ significantly in scale, connectivity, and compliance postures. Future work should validate the system in live or semi-live environments, potentially through partnerships with industry stakeholders. Further enhancements may include integrating federated continual learning, more expressive differential privacy mechanisms (e.g., Rényi DP), or zero-knowledge proof-based alert certification. Additionally, real-world deployment presents challenges related to institutional interoperability and infrastructure heterogeneity. Financial institutions vary significantly in their data formats, compliance postures, and readiness for federated or privacy-preserving analytics. Integration with legacy systems, cross-border data handling policies, and varying technical capacities can pose significant barriers. Addressing these challenges will require modular system adaptation, robust API design, and collaboration with regulatory bodies and IT departments during implementation phases.

To illustrate the system's regulatory alignment, Table 3 presents a structured comparison across key compliance frameworks. Table 3 shows a comparison across major regulatory frameworks, including GDPR, PSD2, eIDAS, and APPI. The matrix highlights our system's alignment with key dimensions such as cross-border data handling, anonymization compliance, and auditability. This comparative view reinforces the system's generalizability and relevance in diverse regulatory landscapes, offering a modular architecture adaptable to evolving standards. In summary, the discussion highlights not only the empirical strengths of the proposed system but also its

Table 3 A comparison across major regulatory frameworks

Regulation	Cross-Border Data	Anonymization Compliance	Auditability
GDPR	Restricted with safeguards	Strictly required	Mandatory logs
PCI-DSS	Limited	Not addressed	Required for card data
PSD2	Permitted with consent	Recommended	Recommended
eIDAS	Permitted with digital ID	Optional	Part of digital trust services
APPI	Restricted	Required	Required for critical data

practical viability, extensibility, and alignment with global data governance frameworks – factors that are crucial for adoption in real-world FinTech environments.

6 Conclusion

This paper presents a comprehensive and privacy-preserving risk analytics framework tailored for FinTech applications, integrating sensitive data detection, federated risk modeling, and adaptive early warning. Our end-to-end architecture is designed to comply with diverse regulatory mandates, such as GDPR, PCI-DSS, PSD2, and APPI, while maintaining operational utility across distributed systems. Through extensive experimentation on a hybrid financial dataset simulating realistic threats, our system demonstrates high accuracy ($F1 > 0.85$) even under strong differential privacy constraints. Federated learning across non-IID partitions reflects real-world deployment scenarios, and our privacy-preserving pipeline retains its predictive capability despite statistical heterogeneity. The system achieves near-instantaneous alerting (average latency 187 ms), supporting real-time mitigation strategies for high-severity threats. Administrative feedback loops further refine alert precision, reducing long-term false positives and improving operational oversight. In contrast to prior solutions that address isolated aspects of financial cybersecurity, our work offers an integrated, regulation-aware platform that enables scalable and transparent threat monitoring without exposing raw data. It bridges gaps between legal compliance, data science, and real-time operations in FinTech.

In future research, we plan to integrate hardware-backed trusted execution environments (e.g., Intel SGX) with our federated setup to further harden computation integrity. Second, we aim to incorporate unsupervised detection mechanisms for zero-day threat discovery, including contrastive learning and graph neural networks. Lastly, we will extend the system's auditing layer with explainability modules and dynamic policy adaptation to further align with evolving regulatory requirements and ethical AI principles.

Funding

High-Level Incubator Grant Project of Beijing Union University(2025), *A Study of Legal Frameworks for Financial Data Security Governance Amid Digital-Intelligence Innovation* (SK20202509).

References

[1] Dwork, C., and Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf.

[2] Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly Detection: A Survey. ACM Computing Surveys, 41(3), 1–58. https://doi.org/10.1145/1541880.1541882.

[3] Javaheri, D., Fahmideh, M., Chizari, H., Lalbakhsh, P., and Hur, J. (2024). Cybersecurity Threats in FinTech: A Systematic Review. Expert Systems with Applications, 241, 122697. https://www.sciencedirect.com/science/article/pii/S0957417423031998.

[4] Lample, G., et al. (2016). Neural Architectures for Named Entity Recognition. NAACL. https://aclanthology.org/N16-1030.pdf.

[5] Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL. https://aclanthology.org/N19-1423.pdf.

[6] Oyewole, A. T., Okoye, C. C., Ofodile, O. C., and Ugochukwu, C. E. (2024). Cybersecurity Risks in Online Banking: A Detailed Review and Preventive Strategies Application. World Journal of Advanced Research and Reviews, 21(3), 625–643. https://doi.org/10.30574/wjarr.2024.21.3.0707.

[7] Abadi, M., et al. (2016). Deep Learning with Differential Privacy. https://dl.acm.org/doi/10.1145/2976749.2978318.

[8] McMahan, H. B., et al. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. https://proceedings.mlr.press/v54/mcmahan17a.html.

[9] Kairouz, P., et al. (2021). Advances and Open Problems in Federated Learning. Foundations and Trends in ML, 14(1–2), 1–210. https://doi.org/10.1561/2200000083.

[10] Truex, S., et al. (2019). LDP-Fed: Federated Learning with Local Differential Privacy. IJCAI. https://www.ijcai.org/proceedings/2019/0530.pdf.

[11] Geyer, R. C., Klein, T., and Soltau, H. (2017). Differentially Private Federated Learning: A Client-Level Perspective. NeurIPS Workshop.

[12] Singh, P., et al. (2021). Graph Neural Networks for Fraud Detection in Financial Transactions. arXiv:2103.08446. https://arxiv.org/abs/2103.08446.

[13] Knyazeva, M., Tselykh, A., Tselykh, A., and Popkova, E. (2016). A Graph-Based Data Mining Approach to Preventing Financial Fraud: A Case Study. ACM SIGKDD Explorations, 17(1), Article 5. https://dl.acm.org/doi/10.1145/2799979.2800002.

[14] Regulation (EU) 2016/679 (GDPR). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679.

[15] PCI Security Standards Council (2018). Payment Card Industry Data Security Standard v3.2.1. https://www.pcisecuritystandards.org/document_library?document=pci_dss.

[16] Regulation (EU) No 910/2014 (eIDAS). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32014R0910.

[17] PSD2 directive. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32015L2366.

[18] Act on the Protection of Personal Information (APPI), Japan. https://www.ppc.go.jp/en/legal/.

[19] Act on the Protection of Personal Information (PIPA), Korea. http://www.pipc.go.kr/cmt/main/laws/enLawListPage.do.

[20] Kadir, A. F. A., Stakhanova, N., and Ghorbani, A. A. (2018). Understanding Android financial malware attacks: taxonomy, characterization, and challenges. Journal of Cyber Security and Mobility, 7(3), 1–52.

[21] Sicari, S., Rizzardi, A., Grieco, L. A., and Coen-Porisini, A. (2015). Security, privacy and trust in Internet of Things: The road ahead. Computer Networks, 76, 146–164. https://doi.org/10.1016/j.comnet.2014.11.008.

[22] Zyskind, G., Nathan, O., and Pentland, A. (2015). Decentralizing privacy: Using blockchain to protect personal data. In 2015 IEEE Security and Privacy Workshops (SPW), 180–184. https://doi.org/10.1109/SPW.2015.27.

[23] Tian, J., and Wang, H. (2021). A provably secure and public auditing protocol based on the Bell triangle for cloud data. Computer Networks, 195, 108223. https://doi.org/10.1016/j.comnet.2021.108223.

Biography

Ye Ju, received a Ph.D. from University of International Business and Economics, currently serving as an associate professor at College of Applied Arts and Sciences, Beijing Union University. Her research interests include data security and digital technology regulations.