International Journal of Recent Advances in Multidisciplinary Topics
Volume 7, Issue 11, November 2026
www.ijramt.com | E-ISSN: 2582-7839 | RESAIM Publishing | www.resaim.com

Federated and Explainable Learning for Trustworthy AI in Regulated Domains: A Compliance-Centric Framework for Secure Cross-Organizational Collaboration

Jami Kiran*
Department of Cyber Security & Data Sciences, Paris American International University, Almere, Netherlands

Abstract: **Federated learning enables organizations to train shared machine-learning models without transferring raw records, but deployment in regulated sectors remains constrained by privacy, explainability, accountability, and cross-domain governance requirements. This paper presents a compliance-centric framework that combines regulation-aware federated training, privacy-preserving aggregation, explainable AI, continuous compliance evidence, and trust management. The framework separates local data control from global model coordination and introduces policy checks, auditable explanations, and human oversight throughout the learning lifecycle. Healthcare, financial-services, and government scenarios illustrate how the design can support collaboration while maintaining organizational and regulatory boundaries. Because the supplied manuscript does not report reproducible datasets or numerical experiments, the evaluation section is framed as a validation protocol rather than an empirical performance claim. The resulting architecture offers a practical research blueprint for developing federated systems that are private, interpretable, governable, and suitable for independent assurance.**

Keywords: **federated learning, explainable AI, privacy, compliance, differential privacy, homomorphic encryption, trustworthy AI.**

1. Introduction

Federated learning (FL) allows multiple organizations to optimize a shared model while retaining sensitive data at its source. This property is attractive in healthcare, finance, and government, where centralized pooling can conflict with confidentiality duties, contractual restrictions, security policies, and data-protection law. Data locality alone, however, does not make an FL system compliant or trustworthy: model updates may leak information, participants may operate under different rules, and a distributed model may remain difficult to explain or audit.

The proposed framework addresses those issues as a single lifecycle. It couples privacy-preserving training with policy enforcement, explanation generation, evidence logging, and accountable human review. Its principal contribution is an integrated design in which privacy, explainability, and compliance are system properties rather than post-deployment additions.

2. Background and Related Work

A. Federated Learning

In a typical FL round, a coordinator distributes a global model to participating nodes. Each node trains locally and returns an update rather than raw observations. The coordinator aggregates eligible updates and redistributes the revised model. This reduces direct data movement but does not eliminate inference attacks, poisoning, unauthorized participation, or governance obligations.

B. Explainable Artificial Intelligence

Explainable AI (XAI) techniques help people understand the factors associated with a prediction. Local methods such as LIME approximate model behavior near a selected instance, while SHAP assigns feature contributions using a game-theoretic formulation. In regulated use cases, explanations must be linked to the model version, decision context, intended audience, and known limitations.

C. Privacy and Regulatory Context

Differential privacy limits the influence of individual records through calibrated randomness, while homomorphic encryption can enable computations on protected values. These techniques address different threat models and introduce accuracy, latency, and operational trade-offs. Legal compliance also depends on purpose, role allocation, lawful processing, data-subject rights, retention, security controls, and documented governance; no single technical control establishes compliance.

3. Compliance-Centric Framework

A. Architectural Overview

The framework contains five coordinated layers: local governance and training; protected update exchange; policy-aware orchestration; explainability and fairness analysis; and an

*Corresponding author: jklife2010@gmail.com

Table 1
Framework control layers and responsibilities

Control layer	Primary function
Local governance	Purpose, eligibility, local data control, consent or authority, feature restrictions
Protected coordination	Secure update exchange, aggregation, privacy accounting, update validation
Explanation and fairness	Local and global explanations, stability checks, subgroup analysis, uncertainty
Assurance	Provenance, policy decisions, approvals, audit records, incidents, remediation

assurance layer for evidence, approvals, and audit. Raw records remain within the participating organization. Only authorized, protected updates and approved explanation artifacts cross organizational boundaries.

B. Regulation-Aware Federated Training

Before a node participates, the policy engine evaluates its jurisdiction, processing purpose, consent or authority record, permitted feature set, retention rule, and model-use restrictions. Ineligible updates are rejected or quarantined. A signed policy decision and model version are recorded for later assurance.

C. Privacy-Preserving Aggregation

Secure aggregation, encryption, and differential privacy can be composed according to the deployment threat model. Each round should document the privacy parameters, participating nodes, key-management state, update validation, and dropout handling. Privacy budgets must be monitored across repeated training and analysis rather than evaluated one round at a time.

D. Explainability and Auditing

Explanation services operate at both local and global levels. Local organizations can generate case-level explanations without exposing source data; the coordinating service can publish aggregate behavior and drift indicators. Every explanation is associated with the model, data window, method, parameters, audience, and reviewer decision. Audit records distinguish evidence from interpretation.

E. Cross-domain Collaboration Policies

Collaborating organizations may be subject to different rules and risk tolerances. Machine-readable participation agreements define permitted purposes, minimum controls, update formats, validation thresholds, explanation duties, incident escalation, and withdrawal procedures. The strictest applicable constraint can be enforced for shared operations while local obligations remain under each participant's control.

4. Privacy, Security, and Trust Controls

Threats include membership and property inference, reconstruction of training records, malicious or low-quality updates, compromised coordinators, collusion, explanation leakage, and model drift. The framework therefore uses layered controls: authenticated nodes, least-privilege roles, secure aggregation, update clipping, anomaly screening, differential privacy, cryptographic key rotation, immutable event logs, and independent review.

Trust is represented as evidence rather than as a single score. Relevant evidence includes participant identity, software and model versions, policy decisions, attestation where available, training-round logs, explanation quality checks, fairness analysis, security incidents, and remediation status. A decision is trustworthy only to the extent that its evidence is current, complete, and appropriate to the intended use.

5. Explainability, Fairness, and Human Oversight

The framework integrates model-level summaries, local feature attribution, subgroup performance analysis, counterfactual testing, and explanation stability checks. Explanation outputs must not reveal protected information or allow reconstruction of local records. Where privacy noise materially changes an explanation, the system should communicate the uncertainty rather than present false precision.

Human oversight is required at defined decision gates: approving the training purpose, admitting participants, accepting privacy parameters, releasing a model, reviewing high-impact decisions, responding to drift, and suspending collaboration. Reviewers need sufficient authority, domain knowledge, and time to challenge both the model and its explanation.

6. Operational Workflow

The lifecycle begins with a documented purpose and participant agreement. The global model is initialized and distributed to eligible nodes; local training runs on approved data; updates are protected and validated; eligible updates are aggregated; explanation and fairness analyses are generated; and compliance evidence is reviewed before release. Monitoring continues after deployment, with triggers for retraining, investigation, withdrawal, and incident response.

Data never needs to leave the local node in raw form, but model updates and explanations remain information-bearing artifacts. They require classification, encryption, access control, retention limits, and monitoring. The framework records provenance for each artifact and prevents a model from advancing when mandatory evidence or approval is absent.



Fig. 1. Compliance-centric federated learning architecture.

7. Sector Applications

A. Healthcare

Hospitals can collaborate on predictive models without pooling patient records. Local governance can restrict eligible cohorts and features, while clinical reviewers assess performance, explanation quality, and safety before use. Technical privacy measures support but do not replace healthcare privacy, security, ethics, and medical-device obligations.

B. Financial Services

Financial institutions can train fraud or risk models across organizational boundaries while retaining transaction data locally. Explanations, fairness controls, model-risk management, and adverse-decision duties must be addressed for the actual product and jurisdiction.

C. Government and Public Services

Public bodies can analyze service or policy outcomes while preserving agency boundaries. Participation rules should reflect statutory purpose, proportionality, records management, transparency, procurement, equality, and public accountability.

8. Evaluation Protocol and Research Questions

A credible evaluation should use predefined datasets or approved institutional data, reproducible partitions, documented node heterogeneity, baseline models, repeated trials, and uncertainty intervals. The study should compare predictive utility, communication cost, convergence, privacy loss, explanation fidelity and stability, subgroup performance, policy-check coverage, audit completeness, and resilience to malicious updates.

The central research questions are: (1) how much utility is lost at specified privacy budgets; (2) whether explanation quality remains stable across nodes and rounds; (3) whether automated policy checks correctly block non-compliant operations; (4) how heterogeneous data and regulation affect convergence and fairness; and (5) whether auditors can reconstruct material decisions from retained evidence. Quantitative claims should be added only after this protocol is executed and independently reviewed.

9. Limitations and Future Work

Privacy controls can reduce model utility and explanation fidelity; homomorphic operations can add latency; heterogeneous nodes can impair convergence; and post-hoc explanations may not represent causal reasoning. Regulations can overlap or conflict, and machine-readable rules require legal and organizational interpretation. The framework is therefore a design blueprint, not a certification method or guarantee of compliance.

Future work should implement a reference architecture, formalize the policy language, evaluate compositional privacy loss, test secure aggregation under node dropout, measure explanation leakage, and conduct domain-specific studies with qualified reviewers. Cross-domain governance should also be evaluated for withdrawal, model unlearning, incident response, and changes in legal requirements.

10. Conclusion

Federated learning can support collaboration in regulated environments only when privacy, explainability, compliance evidence, and accountable oversight are engineered together. The proposed framework combines these controls across the full model lifecycle and makes uncertainty and policy decisions auditable. Its practical value should be established through reproducible experiments and independent assurance before deployment in high-impact settings.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Stat. (AISTATS)*, Fort Lauderdale, FL, USA, Apr. 2017, pp. 1273–1282.

[2] C. Dwork and A. Roth, *The Algorithmic Foundations of Differential Privacy*. Hanover, MA, USA: Now Publishers, 2014.

[3] C. Gentry, "Fully homomorphic encryption using ideal lattices," in *Proc. 41st Annu. ACM Symp. Theory Comput. (STOC '09)*, Bethesda, MD, USA, May/Jun. 2009, pp. 169–178.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD '16)*, San Francisco, CA, USA, Aug. 2016, pp. 1135–1144.

[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Long Beach, CA, USA, Dec. 2017, pp. 4765–4774.

[6] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, Barcelona, Spain, Dec. 2016, pp. 3315–3323.

[7] P. Kairouz *et al.*, "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, nos. 1–2, pp. 1–210, Jun. 2021.

[8] European Parliament and Council of the European Union, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), Off. J. Eur. Union, vol. L 119, pp. 1–88, May 4, 2016.

[9] U.S. Department of Health and Human Services (HHS), *Health Insurance Portability and Accountability Act of 1996 (HIPAA) and Implementing Regulations*, Public Law 104-191, Washington, DC, USA, Aug. 21, 1996.