EJSIT
European Journal of Science, Innovation and Technology
ISSN: 2786-4936
www.ejsit-journal.com
Volume 5 | Number 6 | 2025

Privacy-First Explainable Federated Learning with Zero-Trust AI Infrastructure for Proactive Healthcare Decision Support

Dr. Rohith Vangalla[1] and Kalyan Chakravarthy[2]
[1]Optum Technologies, USA
[2]ITU, San Jose, CA, USA
ORCID: 0009-0001-5130-6697

ABSTRACT

Healthcare AI is increasingly deployed in decision support and patient-facing workflows, but scale-up is limited by privacy constraints on protected health information (PHI), expanding attack surfaces in hybrid delivery models, and weak transparency of complex models. This paper articulates ZT-XFL, a reference architecture that integrates federated learning (FL) with differential privacy (DP) and explainable AI (XAI) under a zero trust architecture (ZTA). The control plane enforces explicit verification, least-privilege authorization, and comprehensive audit logging as defined in ZTA guidance [1], while the training plane coordinates institution-local optimization (FedAvg-style aggregation) [3] and optionally applies DP-SGD and privacy budgeting to bound leakage from updates [6]. Secure aggregation is incorporated to prevent the coordinator from learning individual client updates [7]. The governance plane binds post-hoc explanations (LIME/SHAP) to immutable model versions and inference events [12], [13], reflecting the view that clinical explainability requirements are context-dependent and must be assessed alongside validation and system role [16]. We formalize a threat model spanning endpoint compromise, insider misuse, gradient-based inference, supply-chain risks, and poisoning of federated updates; and we map each threat to enforceable controls across identity, device posture, workload attestation, update screening, and explanation-access policy [1], [2]. Rather than presenting institution-specific results, we provide a reproducible evaluation protocol that jointly measures utility and calibration, privacy loss (ε, δ), security control coverage, explanation stability, and robustness against adversarial updates and inference attacks [9], [10], [11], enabling benchmarking on multi-site healthcare tasks without centralizing PHI.

Keywords: Federated Learning; Differential Privacy; Explainable AI; Zero Trust Architecture; Secure Aggregation; Healthcare AI Governance

1. INTRODUCTION

Clinical and operational AI systems are now routinely proposed for triage, risk stratification, care coordination, and patient engagement. Yet, healthcare data remains siloed, legally sensitive, and operationally costly to centralize, which constrains both model development and external validation [4], [5], [15]. Concurrently, hybrid cloud architectures, telehealth, and third-party integrations erode the practical value of perimeter security, motivating security postures that assume breach and enforce per-request access decisions [1].

Federated learning offers a principled alternative to centralized training by keeping raw data at the source and aggregating local updates into a shared model [3]. In healthcare, FL is viewed as a pragmatic response to fragmentation and governance barriers, but it introduces non-IID statistics, systems constraints, and new attack surfaces through updates and orchestration [4], [15]. Differential privacy provides formal guarantees by bounding the influence of any individual record on a model or released statistic, typically by gradient clipping and noise addition under a tracked privacy budget [6], [8].

Finally, clinical adoption hinges on accountability. Explanations can support calibrated trust, contestability, and auditing, but their utility depends on validation rigor, workflow integration, and the designated role of the system in decision-making [16]. We therefore treat security, privacy, and interpretability as a coupled design problem rather than independent modules.

Contributions: (i) a healthcare-specific threat model for distributed learning and inference; (ii) ZT-XFL, a three-plane reference architecture combining ZTA, FL, DP and XAI; (iii) a control-to-threat mapping suitable for audits; and (iv) an evaluation protocol that enables reproducible assessment without centralizing PHI.

2. RELATED WORK

Federated learning was popularized as an approach for training models on decentralized data through iterative model averaging (FedAvg), significantly reducing communication relative to centralized SGD in heterogeneous settings [3]. Healthcare surveys and perspectives identify FL as a route to unlock multi-institution learning while respecting governance constraints, but emphasize privacy leakage from updates and the need for secure protocols and robust aggregation [4], [15].

Differential privacy provides a mathematical definition of privacy loss and mechanisms for calibrating perturbations to sensitivity [8]. DP-SGD extends these ideas to deep learning through per-example clipping and noise with refined accounting [6]. In practice, DP is often positioned as a mitigation for membership inference and related leakage modes [10].

Zero trust architecture shifts defenses from implicit network trust to explicit verification and resource-centric protection, requiring authentication and authorization prior to establishing sessions and recommending continuous policy evaluation [1]. Planning guidance further ties zero trust deployments to risk management processes and stakeholder alignment [2].

For explainability, LIME explains individual predictions through local surrogate models [12], while SHAP unifies additive feature attribution under Shapley values with desirable axioms [13]. In clinical decision support, the value of explainability depends on feasibility, validation, user groups, and how the system is positioned relative to human decision authority [16].

3. THREAT MODEL AND DESIGN OBJECTIVES

Threats considered include: (T1) external compromise of client endpoints or orchestration services; (T2) insider misuse and over-privileged access to data, models, or logs; (T3) inference attacks that recover information from model outputs or federated updates, including membership inference and model inversion [10], [11]; (T4) poisoning and backdoors introduced via malicious client updates, requiring robust aggregation [14]; and (T5) supply-chain compromise in dependencies, container images, and CI/CD artifacts.

Design objectives are: O1 Privacy—minimize exposure of PHI and bound leakage using DP and secure aggregation [6], [7]; O2 Security—enforce per-request authorization and continuous verification consistent with ZTA [1], [2]; O3 Integrity—detect anomalous or adversarial updates and maintain provenance; O4 Explainability & Governance—deliver validated, workflow-aligned explanations and audit artifacts [16]; and O5 Scalability—support multi-site training under non-IID distributions and operational constraints [3], [4].

4. PROPOSED ARCHITECTURE: ZT-XFL

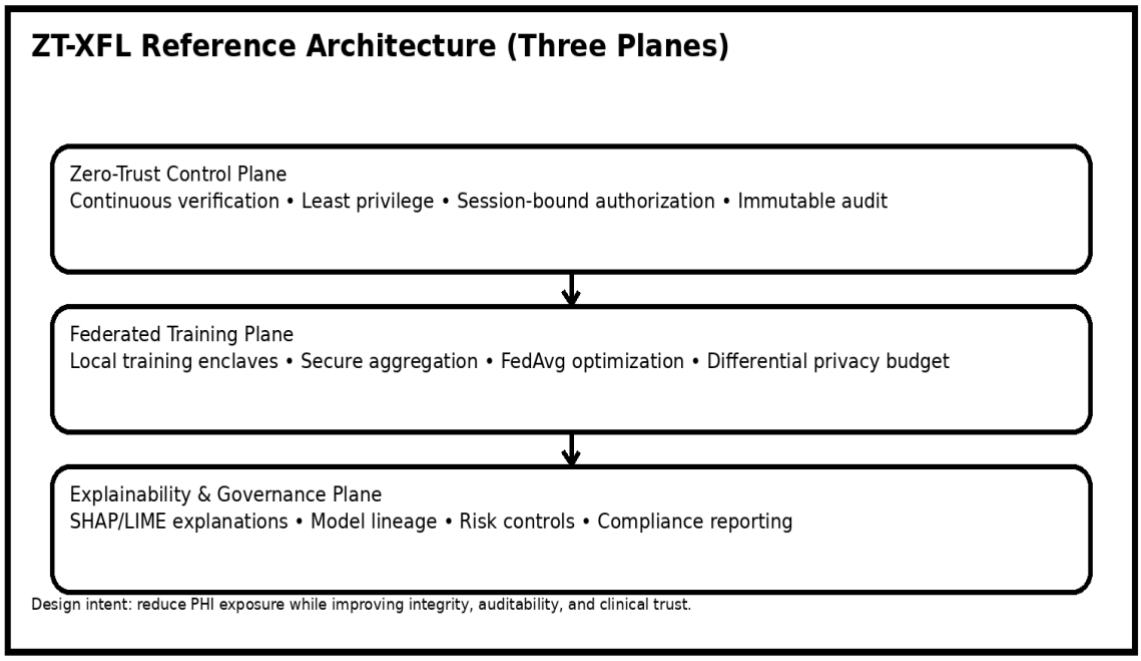

Figure 1. ZT-XFL three-plane reference architecture

Table 1. Threat-to-control mapping (illustrative)

Threat	Control family	Example enforcement
T1 Endpoint compromise	ZTA policy + workload attestation [1]	Deny if device posture fails; short-lived tokens
T3 Update/Output inference	DP + secure aggregation [6], [7]	DP-SGD; masked updates via secure aggregation
T4 Poisoning/backdoors	Robust aggregation [14]	Anomaly scoring; Krum-style selection
T2 Insider misuse	Least privilege + audit [1], [2]	Just-in-time access; immutable audit trail

ZT-XFL separates responsibilities into three planes (Fig. 1): a Zero-Trust Control Plane, a Federated Training Plane, and an Explainability & Governance Plane. The control plane implements policy enforcement points for every access to data, models, and inference services, ensuring authentication and authorization precede session establishment, and that logging is immutable and reviewable [1].

The training plane orchestrates federated rounds using FedAvg-style aggregation [3]. For privacy, it optionally applies DP-SGD on clients or DP noise on aggregated updates with explicit privacy budgeting [6]. Secure aggregation prevents the coordinator from viewing any single client's update in the clear, limiting exposure even under honest-but-curious assumptions [7].

The governance plane binds explanations to model version identifiers and inference events to support traceability. Explanations are treated as governed outputs because they may reveal sensitive correlates or be misinterpreted if not validated and contextually appropriate [16].

5. IMPLEMENTATION BLUEPRINT

Federated optimization proceeds iteratively: the server selects eligible clients, distributes the current model, clients train locally for E epochs, and the server aggregates weighted updates. Selection and eligibility are driven by policy (device posture, workload attestation, least privilege) to align with zero-trust assumptions [1].

Privacy instrumentation includes: (i) per-example gradient clipping and Gaussian noise for DP-SGD with accounting [6]; (ii) secure aggregation to hide individual updates from the server [7]; and (iii) release policies for model checkpoints and explanations, so that governance artifacts are aligned with privacy budgets and purpose limitations [2].

Update integrity can be reinforced through robust aggregation rules that tolerate Byzantine behavior, such as Krum, when threat conditions justify the added computation [14].

6. EVALUATION PROTOCOL

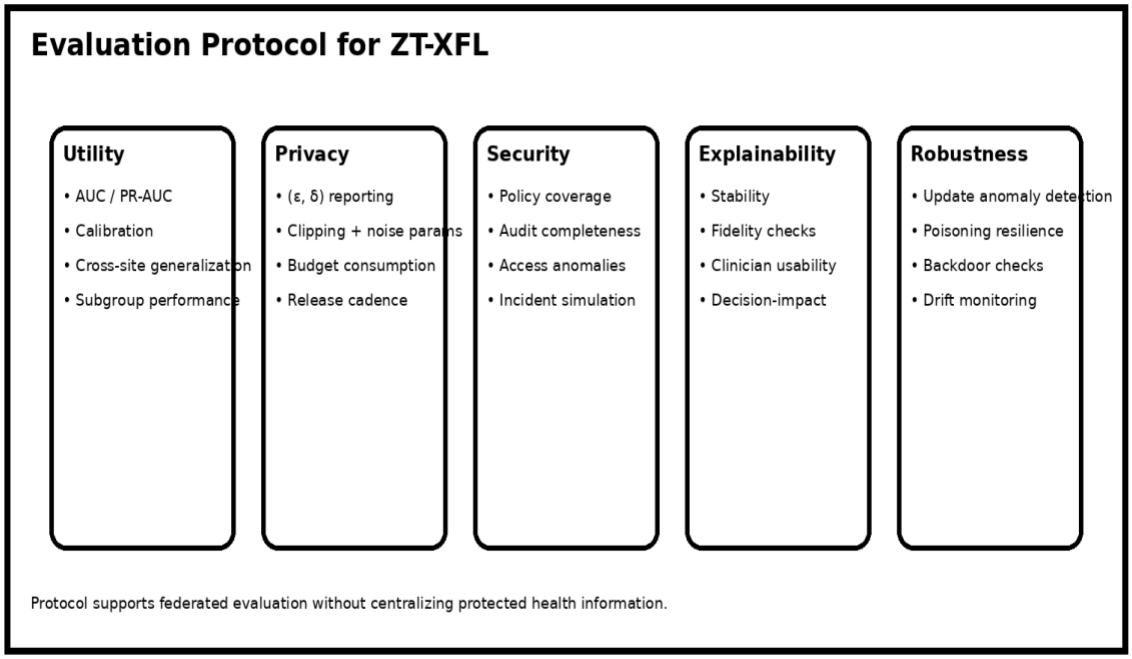

Figure 2. Evaluation dimensions and measurable criteria for ZT-XFL

Evaluation is designed to be reproducible without centralizing PHI. As summarized in Fig. 2, we assess: (E1) predictive utility and calibration; (E2) privacy loss via reported (ε, δ) budgets and DP configuration; (E3) security posture via policy coverage and audit completeness under ZTA principles [1], [2]; (E4) explainability via stability and usability measures grounded in clinical CDSS considerations [16]; and (E5) robustness against poisoning and inference attacks, referencing established attack models [10], [11], [14].

7. DISCUSSION

ZT-XFL emphasizes that FL alone is not a privacy guarantee: update leakage and output leakage can still expose sensitive membership or attributes, motivating DP and secure aggregation [6], [7], [10]. Similarly, security controls must extend beyond network boundaries to identity, device, and workload trust signals evaluated continuously, as described in ZTA guidance [1], [2].

Explainability is not treated as an aesthetic overlay but as a governed interface. Clinical literature argues that explainability requirements vary with the context, the system's role, and

validation maturity; therefore, explanations must be coupled with model lineage, limitations, and monitoring [16].

A limitation of this manuscript is that it presents a deployable pattern and evaluation criteria rather than reporting institution-specific performance. This is deliberate to avoid disclosure of PHI and proprietary deployment details while enabling reproducible benchmarking on public or federated simulation datasets [4], [15].

8. CONCLUSION

This work presented ZT-XFL, a privacy-first pattern for healthcare AI that composes zero trust controls with federated learning, differential privacy, secure aggregation, and explanation governance. The architecture reduces PHI centralization pressure, strengthens security through per-request authorization, and supports accountable deployment through traceable explanations and audit artifacts.

DATA AVAILABILITY STATEMENT

No protected health information is shared. The evaluation protocol can be instantiated on federated simulations or approved multi-site studies without centralizing raw data.

REFERENCES

[1] S. Rose, O. Borchert, S. Mitchell, and S. Connelly, "Zero Trust Architecture," NIST Special Publication 800-207, Aug. 2020. doi:10.6028/NIST.SP.800-207.

[2] S. Rose, "Planning for a Zero Trust Architecture: A Planning Guide for Federal Administrators," NIST CSWP 20, May 2022. doi:10.6028/NIST.CSWP.20.

[3] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," Proc. AISTATS, PMLR 54, pp. 1273–1282, 2017.

[4] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated Learning for Healthcare Informatics," Journal of Healthcare Informatics Research, vol. 5, no. 1, pp. 1–19, 2021. doi:10.1007/s41666-020-00082-4.

[5] N. Rieke et al., "The Future of Digital Health with Federated Learning," npj Digital Medicine, vol. 3, art. 119, 2020. doi:10.1038/s41746-020-00323-1.

[6] M. Abadi et al., "Deep Learning with Differential Privacy," Proc. ACM CCS, pp. 308–318, 2016. doi:10.1145/2976749.2978318.

[7] K. Bonawitz et al., "Practical Secure Aggregation for Privacy-Preserving Machine Learning," Proc. ACM CCS, 2017. ePrint:2017/281.

[8] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," Theory of Cryptography Conference (TCC), LNCS 3876, pp. 265–284, 2006. doi:10.1007/11681878_14.

[9] U.S. Department of Health & Human Services, "HIPAA Guidance Materials," accessed Dec. 2025. https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/index.html.

[10] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," IEEE S&P, 2017. doi:10.1109/SP.2017.41.

[11] M. Fredrikson, S. Jha, and T. Ristenpart, "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures," Proc. ACM CCS, 2015. doi:10.1145/2810103.2813677.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," Proc. ACM SIGKDD, 2016. doi:10.1145/2939672.2939778.

[13] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," NeurIPS, 2017.

[14] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Byzantine-Tolerant Machine Learning (Krum)," NeurIPS, 2017. arXiv:1703.02757.
[15] P. Kairouz et al., "Advances and Open Problems in Federated Learning," Foundations and Trends in Machine Learning, vol. 14, no. 1–2, pp. 1–210, 2021. doi:10.1561/2200000083.
[16] J. Amann et al., "To Explain or Not to Explain?—AI Explainability in Clinical Decision Support Systems," PLOS Digital Health, 2022. doi:10.1371/journal.pdig.0000016.
[17] World Health Organization, "Ethics and Governance of Artificial Intelligence for Health: WHO Guidance," 2021. ISBN:9789240029200.

AUTHOR BIOGRAPHY

Dr. Rohith Vangalla is a Senior IEEE Member, AI Researcher, and Lead Software Engineer at Optum Technologies, specializing in AI-driven solutions, cybersecurity, and enterprise AI infrastructure. His work focuses on privacy-preserving learning, explainable AI, and zero-trust architectures for regulated healthcare environments. He has served as a reviewer and judge for academic venues and contributes to responsible AI governance and secure AI platform engineering.