Privacy Preserving Federated Learning with Blockchain and Explainable AI: A Governance-Aware Framework for Equitable Healthcare Analytics

Dr. Avdhesh Gupta[1], Dr. Ashish Dixit[1,2]
[1]Post Doctoral Researcher, Lincoln University College, Malaysia
[2]Professor, Lincoln University College, Malaysia
[1]avvipersonal@gmail.com, [2]ashishdixit1984@gmail.com

Abstract: Various hospitals and other healthcare organizations are keeping data of their patients in the form of organized and unorganized data. These databases can be shared among other hospitals, diagnostic centres, and research laboratories for further investigation and research. But due to some privacy issues, regulations, hospital policies, and competitive concerns, the data cannot be shared among all healthcare industries. Therefore, a federated learning approach with some security like blockchain, etc., has emerged as a practical approach to share data with other healthcare institutions, keeping their records private and safe. Due to having different approaches to collecting data by different healthcare organizations, the data become heterogeneous and create some problems for others, like statistical heterogeneity across institutional datasets, vulnerability of model aggregation to adversarial or low-quality participants, and limited interpretability for all healthcare organizations to interpret results. Also, it can violate government regulations. In this paper, we are proposing a simple conceptual framework or model that integrates a federated learning approach with blockchain technology and explainable artificial intelligence, considering government regulations to address heterogeneity and security of data to have good data analytics and interpretation. The main objective in this paper is to provide federated learning researchers, security researchers, and the healthcare industry a formatted, structured architecture or framework for building federated healthcare systems that can be secure, fair, auditable, and explainable.

Keywords: Federated Learning, Blockchain, Explainable AI, Differential Privacy, Healthcare Data Privacy, Byzantine-Resilient Aggregation, Zero-Knowledge Proofs, Edge Computing, Regulatory Compliance.

Introduction

The practice of medicine is now generating massive quantities of data. Electronic health records (EHRs), medical imaging repositories, streaming wearable sensors, and laboratory information databases contain clinical data with the breadth and depth to dramatically improve clinical decision making if shared and analysed in aggregate. As it stands, healthcare data is not centralized. Patient data is segmented by the hospital, doctor's office, medical laboratory, or insurance company where it was collected. Healthcare providers typically have data governance policies which treat identifiable health information as sensitive. Many providers are legally barred from sharing information with outside entities. In the United States, HIPAA places limitations on the use of individually identifiable healthcare data. Similarly, the EU has created GDPR, a regulatory framework which imposes limitations on the use of personal data. Other countries have similar laws governing the privacy of healthcare data. India recently passed the Digital Personal Data Protection Act, 2023.

Federated learning was introduced to circumvent this limitation. Instead of pooling data together, FL "parsimoniously centralizes only model updates": each hospital fits a local model to their private dataset and communicates only gradient/parameter updates to a central server which averages these updates into a shared global model. Patient data never leaves the site of origin. Whilst this principle makes FL an enticing proposition for healthcare, there are challenges unique to healthcare that arise when moving from algorithmic concept to clinical implementation.

FL datasets often fail the independent and identically distributed (IID) assumption that many standard FL aggregation algorithms require to guarantee global model performance: even before collecting data, a specialised cardiac centre and a rural primary care clinic are going to see systematically different distributions of patient demographics, disease prevalence, and instrumentation. Due to various differences in data sharing over various healthcare institutions create issues in IID assumption and this leads to downgrade the quality of data and its interpretation by different healthcare institutions. But if we modify the data towards making it homogeneous leads to better data interpretation. Therefore, aggregator can solve this type of data heterogeneity.

Background and Problem Context

A. Data Fragmentation and Regulatory Constraints

Mir et al.'s systematic review summarizes the ethical aspects of FL deployments in healthcare organizations from 2024 to 2024, focusing on algorithmic fairness, governance, equitable access, and privacy preservation are recurrent and related issues rather than separate design criteria [1]. The integrated approach used in this study is directly motivated by their research, which emphasizes that technical privacy solutions alone are insufficient if governance and fairness are not addressed concurrently.

B. Federated Learning Combined with Cryptographic and Blockchain Mechanisms

In order to eliminate the single point of failure present in a centralized aggregator, a number of recent studies have investigated combining FL with distributed ledgers and cryptographic privacy methods. Yang's blockchain-and-homomorphic encryption framework shows that accuracy comparable to standard FL can be maintained while reducing single-point-of-failure and inference-attack risks by using cross-silo FL in conjunction with smart contracts to decentralize aggregation and threshold homomorphic encryption to protect parameters in transit [2]. In a similar vein, Chang et al. suggest a blockchain-based FL approach for smart healthcare in which edge nodes keep track of the ledger and an adaptive differential privacy algorithm adjusts injected noise based on training progress in conjunction with a gradient-verification consensus protocol intended to identify poisoning attempts [3]. By framing the technical synergy between immutable ledgers and collaborative model training as a basis for trust in multi-institutional healthcare data ecosystems, Alsamhi et al. offer a more comprehensive conceptual treatment of how blockchain and FL together reshape decentralized, patient-centric data sharing [4].

C. Explainability and Optimisation in Federated Healthcare Models

Using SHA-256-backed model updates, Levy Grasshopper-based feature selection, and SHAP-based interpretability layers, Bhardwaj and Sumangali present a framework that combines blockchain, explainable AI, and metaheuristic optimization into a single FL pipeline to improve robustness and clinical transparency, reporting accuracy above 95% on benchmark clinical datasets [5]. However, the authors point out an open gap in their architecture: it does not create a verifiable feedback loop where explainability outputs are either fed back into the optimization process or directly anchored into the blockchain ledger.

D. Fairness-Aware and Privacy-Preserving Benchmarks

Work on fairness-aware federated learning (FAFL) for healthcare has developed benchmark approaches for assessing how FL schemes perform under data heterogeneity and privacy limitations outside of the blockchain-specific research. In comparison to regular FL, privacy-preserving operation can significantly increase communication overhead and delay, as demonstrated by benchmarking work that indicates significant variation in how various FAFL algorithms respond to varied clinical data distributions [6]. All of the design choices discussed in this work include this trade-off between privacy, fairness, and communication cost.

E. Foundational and Survey-Level Treatments

By framing the fragmentation and sensitivity of clinical data as the primary driving force for federated approaches and reviewing the statistical and systems issues that follow, Xu et al. offer an early and frequently cited treatment of FL specifically for healthcare informatics [7]. The IEEE Access survey by Nezhadsistani, Moayedian, and Stiller offers a systematic review of blockchain-based federated learning (BCFL) systems in healthcare, examining consensus protocols, cryptographic protocols, and storage topology decisions relevant to clinical deployment. It also serves as a structural reference for the governance layer [8]. Finally, Farrugia et al.'s methodological recommendations for formulating research questions, hypotheses, and objectives are used to establish the study plan for this paper[9].

Research Objectives

This work organizes its objectives around four specific, design-relevant aims, followed by the research questions each objective is meant to address, in accordance with the organized approach to creating research objectives given by Farrugia et al. [9].

1) Need to design a lightweight federated architecture that allow healthcare organizations to allow to share heterogeneity data with security.
2) Heterogeneity data conversion or to extract results or contribute same results as homogeneous data provide.
3) Design an aggregation mechanism that can detect unreliable data sharing or any other updates from individual participants.
4) Provide auditable Blockchain Governance to keep track of data sharing and data transactions securely, so that the stakeholders and regulators can verify compliance without needing to trust a single coordinating party.

Proposed Methodology and Architecture

The suggested framework is structured as a tiered architecture with a federated learning core surrounded by layers for fairness and privacy, aggregation-integrity, and governance. An explainability layer is added at the point where clinicians are shown model results. Below is a description of each layer.

A. Federated Learning Core

The fundamental training procedure is based on a cross-silo FL pattern that is appropriate for the healthcare industry. Each participant has a significant, high-value local dataset, but the number of participants is moderate (tens to low hundreds of institutions rather than millions of edge devices). Within this core, the framework supports various complementary strategies based on institutional requirements: federated transfer learning, which enables healthcare organization or institutions having low local data to bootstrap from a pretrained global model instead of training from scratch; multi-task learning formulations, which allow related clinical prediction tasks to share representations while maintaining task-specific output layers; and personalized FL approaches, which allow each institution to retain a locally adapted model variant alongside the shared global model.

B. Fairness-Aware, Privacy-Preserving Aggregation

To overcome the issue of heterogeneity of data, the aggregation step must be implemented, and this method can be named as bias-detection stage that identifies patient's latest information and statistics before forwarding to make aggregation. A good aggregation technique that has adaptive weighting to avoid large scale homogeneous data collection at a smaller cost that indicate Research Objective 2. Differential privacy has been added with this process by adding calibrated noise into shared gradients using the adaptive noise-scaling method described by Chang et al. to balance privacy guarantees against preserved model accuracy as training progresses [3].

C. Byzantine-Resilient Aggregation

To fulfil third research objective, the aggregator uses averaging after data statistics, to avoid or prevent a small number of malfunctioning users from substantially distorting the entire model. This technique is assessed in addition with the fairness layer, since naive outlier rejection might provide bias against legitimately anomalous institutions.

D. Blockchain Governance Layer

The final done aggregation needs to work on permissioned blockchain. This ledger of blockchain keeps track the participating healthcare organization, who shared data of its patients and all other data that are needed for further interpretation. According to government the data that need to be shared need to take consent to share and the data use agreement, and the minimum number of participants. Before a record of a training round is permanently added to the blockchain, smart contracts verify the record meets the stated requirements to be added to the ledger. This creates an immutable trail that can be audited independently, which allows data protection officials to review and focus on the evidence without having to trust the account of a central coordinator. This directly addresses Research Objective 4. The consensus and storage topology is constructed based on an analysis of BCFL design patterns, as documented by Nezhadsistani et al. [8]. This design is constructed following the guidelines and is in agreement with the ideas presented by Alsamhi et al. regarding the use of blockchain technology and FL in the sharing of healthcare data [4].

E. Explainability Layer

The local training model aggregates data from different locations, resulting in a black box for each of the locations. To address the challenge of black box systems and system output explainability, the model builds an additional layer of explainability at the inference phase. This additional layer employs SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) to provide explanation for each prediction in a way that the medical doctors can understand and that SHAP and LIME are other post-hoc explainability methods. The new approach proposed by the author stores a hash of each explanation on the blockchain as well as the associated prediction. This is different from the approach laid out by Bhardwaj and Sumangali [5], where both explainability and blockchain logging are performed, but are separate processes. This will make sure that the gap in feedback with the clinical decision-support and explanation process is solved.

Technologies that support the study

Beyond the core architectural layers, the framework depends on a small set of supporting technologies that operationalise privacy, trust, and efficiency in a clinical deployment context.

1) With Zero Knowledge Proofs (ZKPs), an organization can cryptographically show a property of its data/computation without disclosing the patient data, like showing that a model update was done with legitimate local data. Here, ZKPs show an authentication method for participant updates and can be used with differential privacy.
2) Reducing the amount of data that needs to be sent to and from the central servers during federated rounds is achieved through local data preprocessing. This is done at the point of data capture (for example, departmental servers, hospital workstations, etc.). For organizations that lack sufficient networking infrastructure, such as hospitals, this is done to reduce the challenges of bandwidth and latency.
3) Since these algorithms enable the discrimination of fraudulent and invalid contributions without relying on the privacy and fairness mechanisms discussed above, they are treated as a separate layer of supporting technology.
4) In the context of a regulated clinical environment, the governing and operational requirements of a regulated clinical environment are met by employing a permissioned, rather than wholly public,

ledger. Only validated healthcare organizations that are allowed to act as validating nodes are able to participate. This compromises between the leverage of auditability that blockchain technology brings and the concerns of a regulated clinical environment.

Healthcare Integration Roadmap

It will be mandatory to adopt a staged approach towards implementation as a one-time implementation would be impossible given the regulatory and trust requirements of the healthcare sector. There are three phases in the proposed solution.

Phase 1: Governance Compliant Ecosystem Design: Define the federated architecture, smart contracts, and data management processes in a way that they are compliant with HIPAA, GDPR, India's Digital Personal Data Protection Act 2023, and the data governance practices of the participating institutions before starting any inter-institutional training.

Phase 2: Multi-Institutional Testing: Rather than focusing solely on optimizing for accuracy, run intensive testing at several participating hospitals or clinics in a controlled environment, measuring the federation across three dimensions: predictive accuracy, fairness between larger and smaller participating institutions and privacy leakage.

Phase 3: Real-World Deployment Pilot: Run a real-world pilot test at a few partner institutions where we measure the performance of the model, adoption of the clinical personnel, additional overhead of the blockchain governance layer and regulatory compliance.

Discussion: Trade-Offs And Broader Alignment

A. Privacy, Fairness, and Efficiency as Competing Constraints

Each of the design decisions outlined in has a cost that must be stated clearly rather than taken for granted. While differential privacy enhances disclosure protection, benchmarking work on fairness-aware FL has shown that it tends to raise communication overhead and potentially lower model accuracy if noise is not well adjusted [6]. The framework clearly distinguishes the fairness layer from the robustness layer rather than confusing them because Byzantine-resilient aggregation increases robustness to adversarial participants but runs the risk of penalizing institutions whose data is legitimately, rather than maliciously, unusual.

B. Limitations

Instead of a deployed and empirically proven system, this paper describes a conceptual and architectural framework; no accuracy, fairness, or privacy-leakage figures are promised for the integrated framework itself, and the staged roadmap has not yet been implemented. Empirical data from individual component technologies (such as the accuracy results reported for particular blockchain-FL-XAI systems) should not be interpreted as performance claims for the combined architecture suggested here, but rather as evidence for the viability of individual components. In order to get such figures under controlled, multi-institutional conditions, future work should prioritize implementing the roadmap's Phase 1 and Phase 2 stages.

C. Alignment with Sustainable Development Goals

The motivation for this work also aligns with several United Nations Sustainable Development Goals. By improving the ability of healthcare systems to learn collaboratively without compromising patient privacy, the framework supports SDG 3 (Good Health and Well-Being). The integration of blockchain, federated learning, and explainable AI as a coordinated technical stack contributes to SDG 9 (Industry, Innovation and Infrastructure). Fairness-aware aggregation that protects smaller and under-resourced institutions from being marginalised in the global model directly supports SDG 10 (Reduced Inequalities). Finally,

because the framework is explicitly designed for multi-institutional participation under shared governance rules, it is structurally aligned with SDG 17 (Partnerships for the Goals).

Conclusion

Federated learning offers a structurally sound answer to the problem of fragmented, privacy-sensitive healthcare data, but a federated learning pipeline alone does not resolve the fairness, robustness, transparency, and governance requirements that clinical deployment demands. This paper has proposed a layered framework in which fairness-aware, differentially private aggregation is combined with Byzantine-resilient robustness checks, a permissioned blockchain governance layer, and a SHAP/LIME-based explainability layer whose outputs are anchored back into the same ledger used for training governance. Three research questions concerning security and privacy, fairness under heterogeneity, and edge-computing efficiency were used to structure the architecture, and a three-phase roadmap, moving from governance-compliant design through multi-institutional validation to real-world pilot deployment, was proposed as the path from this conceptual framework to a deployable clinical system. The principal contribution of this paper is therefore not a new algorithm in isolation, but an integrated architectural reference and research agenda that treats privacy, fairness, robustness, and explainability as jointly optimised design constraints rather than as separate add-ons to a federated learning system. Future work should implement and empirically evaluate the proposed roadmap on real multi-institutional healthcare data to validate the trade-offs.

References

1. M. B. R. A. Mir, S. R. Abbas, and S. W. Lee, “Federated Learning in Healthcare Ethics: A Systematic Review of Privacy-Preserving and Equitable Medical AI,” Healthcare, 2026.
2. C. Yang, “Federated Medical Learning Framework Based on Blockchain and Homomorphic Encryption,” Wireless Communications and Mobile Computing, vol. 2024, Art. no. 8138644, 2024, https://doi.org/10.1155/2024/8138644 .
3. Y. Chang, C. Fang, and W. Sun, “A Blockchain-Based Federated Learning Method for Smart Healthcare,” Computational Intelligence and Neuroscience, vol. 2021, Art. no. 4376418, 2021, https://doi.org/10.1155/2021/4376418
4. S. H. Alsamhi, R. Myrzashova, A. Hawbani, S. Kumar, S. Srivastava, L. Zhao, X. Wei, M. Guizani, and E. Curry, “Federated Learning Meets Blockchain in Decentralized Data Sharing: Healthcare Use Case,” IEEE Internet of Things Journal, vol. 11, no. 11, pp. 19602–19615, Jun. 2024, https://doi.org/10.1109/JIOT.2024.3367249 .
5. T. Bhardwaj and K. Sumangali, “An Explainable Federated Blockchain Framework with Privacy-Preserving AI Optimization for Securing Healthcare Data,” Scientific Reports, vol. 15, Art. no. 21799, Jul. 2025, https://doi.org/10.1038/s41598-025-04083-4
6. Towards Fair and Privacy Preserving Federated Learning for the Healthcare Domain, arXiv preprint arXiv:2308.01529, Aug. 2023.
7. 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, https://doi.org/10.1007/s41666-020-00082-4
8. N. Nezhadsistani, N. S. Moayedian, and B. Stiller, “Blockchain-Enabled Federated Learning in Healthcare: Survey and State-of-the-Art,” IEEE Access, 2025, https://doi.org/10.1109/ACCESS.2025.3587345
9. P. Farrugia, B. A. Petrisor, F. Farrokhyar, and M. Bhandari, “Research Questions, Hypotheses and Objectives,” Canadian Journal of Surgery, vol. 53, no. 4, pp. 278–281, Aug. 2010.