Blockchain-Orchestrated Personalized Federated Learning with Synthetic Balancing for Non-IID Clinical IoT Data

---

Group Members:

Muhammad Ibrahim Iqbal 27085

Muhammad Maaz Siddiqui 27070

Muhammad Ibrahim Farid 27098

SUPERVISOR

Dr. S. M. Faisal Iradat

Assistant Professor, SMCS - IBA Karachi

CO-SUPERVISOR

Dr. Waseem Iqbal

Assistant Professor, Sultan Qaboos University

INDUSTRY MENTOR

Karim Jivani

Data & AI Partner, Bluetech Consulting

Letter to the Evaluator

Dear Evaluators,

On behalf of our team, I am pleased to present our proposal for the development of a blockchain-orchestrated personalized federated learning framework with synthetic balancing for non-IID clinical IoT data. We recognize the critical importance of trustworthy, fair, and privacy-preserving AI in healthcare, and we are excited to contribute toward advancing this field through our research.

Our project directly addresses well-documented challenges in healthcare federated learning, including data heterogeneity (non-IID distributions), rare-class imbalance, and the need for transparent governance. By combining federated learning, blockchain provenance, personalization strategies, and controlled synthetic data augmentation, our solution explores a novel pathway to improve fairness and performance in clinical predictive systems.

Key Features of Our Solution:

- **Federated Learning with Provenance:** Collaborative model training across simulated hospital clients, with blockchain-based logging of contributions for transparency and auditability.
- **Personalization Strategies:** Local fine-tuning and meta-learning approaches (e.g., PerFedAvg) to adapt models to site-specific data distributions.
- **Synthetic Data Balancing:** Controlled generation of rare-class clinical samples (e.g., rare arrhythmias) through generative models (GANs, TimeGAN, diffusion), authorized via smart contracts.
- **IoT Simulation:** Integration of heterogeneous clinical sensor streams (ECG, vitals) to emulate real-world hospital and ward-level data diversity.
- **Evaluation Metrics:** Comprehensive assessment of accuracy, fairness, auditability, and blockchain overhead.

We believe this research will provide valuable experimental insights into the trade-offs and potential benefits of combining blockchain governance with federated learning in healthcare. While exploratory in nature, the framework aims to contribute toward building fair, transparent, and personalized AI systems for sensitive domains.

We look forward to engaging in further discussion and receiving your feedback on how this research may be strengthened to maximize its impact.

Sincerely,

Ibrahim Iqbal
Group Lead

Disclaimer: We request, in the most humble way, that this proposal be evaluated by a domain expert to ensure an informed and accurate assessment of its technical and practical feasibility. This proposal is intended as a research exploration; the datasets, methods, and models described are indicative starting points and may evolve as the project progresses. Final outcomes will focus on validated insights and experimental contributions rather than production-level deployment.

Contents

1 Introduction

1.1 Problem Statement

Federated Learning (FL) has become a promising paradigm for enabling collaborative model training across distributed healthcare institutions without centralizing sensitive patient data. While FL preserves privacy under laws such as HIPAA, it still suffers from critical unsolved challenges. Traditional algorithms like FedAvg experience severe performance degradation under non-IID client data distributions, which are common in clinical IoT environments where hospitals and devices generate heterogeneous data. Moreover, rare-class events such as uncommon arrhythmias or climate-exacerbated conditions are underrepresented, leading to biased models.

Current FL systems also lack auditable governance mechanisms to track which clients contributed which updates, raising issues of accountability, fairness, and trust. While blockchain has been studied for provenance and Generative AI for data augmentation, no integrated framework exists that combines blockchain-based provenance, personalized FL, and controlled synthetic data generation to address fairness and non-IID challenges in healthcare IoT.

This project addresses these gaps by designing a Blockchain-Orchestrated Personalized Federated Learning framework with Synthetic Balancing for non-IID clinical IoT data.

1.2 Objectives

1.2.1 Primary Objectives

- To design and implement a blockchain-enabled federated learning system with on-chain provenance.
- To integrate controlled synthetic data generation for rare-class augmentation governed by smart contracts.
- To evaluate personalization strategies for non-IID clinical IoT data.

1.2.2 Secondary Objectives

- To simulate IoT-based clinical data streams using benchmark datasets.
- To compare centralized, vanilla FL, and proposed methods under varying non-IID conditions.
- To measure fairness, accountability, privacy, and computational overhead.
- To evaluate the security resilience of the proposed framework against adversarial attacks and model poisoning.

- To assess clinical safety by measuring explainability, reliability, and bias mitigation of AI-driven outputs.
- To embed ethical accountability through transparent governance, fairness audits, and traceability mechanisms.

1.3 Iterative Methodology for Project Execution

Given the exploratory and research-focused nature of this work, the project will follow an iterative milestone-driven methodology rather than strict software development practices. The execution plan is structured into research cycles, each delivering incremental progress:

Cycle 1: Establish baseline models (centralized training, FedAvg, FedProx).

Cycle 2: Integrate blockchain smart contracts for model provenance logging.

Cycle 3: Implement personalization strategies (fine-tuning, PerFedAvg).

Cycle 4: Incorporate controlled synthetic data generation (GANs, TimeGAN).

Cycle 5: Conduct evaluation on fairness, performance, and auditability.

Cycle 6: Perform ablation studies and finalize prototype with comparative results.

This methodology ensures the project progresses in well-defined stages while remaining flexible to adapt to research findings and supervisory feedback.

2 Methodology and Architecture

This section outlines the methodological framework and architectural design for the proposed blockchain-orchestrated personalized federated learning system with synthetic balancing in clinical IoT settings. The methodology is structured into well-defined components, each addressing a critical aspect of the problem.

2.1 Methodology

2.1.1 Data Collection and Simulation (MIMIC, PhysioNet, IoT streams)

The project will use publicly available healthcare datasets such as MIMIC-III and PhysioNet, which contain ECG signals, vitals, and clinical records. To simulate IoT hospital ward environments, these datasets will be partitioned into multiple client nodes, each representing a hospital or device with unique non-IID data distributions. Additionally, synthetic IoT data streams (e.g., ECG monitoring, blood pressure sensors) will be generated to simulate real-time data arrival.

2.1.2 Data Pre-processing (Normalization, Feature Extraction, Sequence Formatting)

Collected data will undergo preprocessing to ensure compatibility with the federated learning pipeline. This includes:

- **Normalization:** Scaling features such as heart rate and blood pressure to standard ranges.
- **Feature Extraction:** Deriving relevant features (e.g., R–R intervals from ECG).
- **Sequence Formatting:** Structuring data into time-series windows for LSTM/CNN models.
- **Class Distribution Statistics:** Computed per client for blockchain logging and imbalance detection.

2.1.3 Model Development (Centralized, FL Baseline, Proposed System)

Three stages of model development will be implemented. **Centralized Model:** A benchmark model trained on pooled data will serve as a reference point. **Federated Baselines:** Standard FL algorithms such as FedAvg and FedProx will be implemented for distributed training. **Proposed System:** A blockchain-orchestrated federated learning framework will be developed, incorporating personalization and synthetic balancing. The global model will employ CNN/LSTM architectures suitable for clinical time-series prediction tasks (e.g., arrhythmia detection).

2.1.4 Blockchain Provenance and Governance (Smart Contracts, Audit Logs)

A blockchain layer will ensure transparency, accountability, and governance in the federated learning process. **Smart Contracts** will record model update hashes, metadata such as client contribution size and performance, and synthetic data requests. **Audit Logs** will provide immutable on-chain records detailing which clients contributed to each global model version. **Governance Rules** will enforce quotas on synthetic data generation and access permissions via contracts. The implementation will rely on technologies such as Ethereum testnet or Hyperledger Fabric, Solidity, and Web3.py.

2.1.5 Synthetic Data Generation and Balancing (GANs, TimeGAN, Diffusion)

Rare clinical events, such as uncommon arrhythmias, are often underrepresented in client datasets. To address this imbalance, clients may request rare-class augmentation through the blockchain. Upon approval, a generative AI module (e.g., GANs, TimeGAN, or diffusion models) will produce synthetic samples. These generated data will be added locally without exposing private raw data. All requests and generated outputs will be logged on-chain to ensure fairness and auditability.

2.1.6 Personalization Strategies (Fine-tuning, PerFedAvg)

Since client data distributions are non-IID, personalization is essential. Two strategies will be evaluated. **Local Fine-tuning:** Clients will fine-tune the aggregated global model on their respective local datasets. **PerFedAvg:** A meta-learning approach that enables the global model to adapt quickly to individual client distributions will also be tested.

2.1.7 Evaluation Metrics (Accuracy, AUC, Fairness, Overhead)

Model performance will be assessed across multiple dimensions. **Classification Performance** will be measured using accuracy, precision, recall, F1-score, and AUC. **Fairness** will be evaluated based on the variance in performance across different clients. **Blockchain Overhead** will be quantified by measuring latency and computational costs introduced by provenance logging. **Synthetic Data Impact** will be analyzed by examining improvements in rare-class recall after augmentation.

2.1.8 Validation and Continuous Learning

Validation will involve a comparative analysis of three baselines: centralized training (ideal reference), vanilla federated learning (FedAvg/FedProx), and the proposed system

(with blockchain, personalization, and synthetic balancing). Continuous learning will be simulated by periodically introducing new IoT data streams and updating models iteratively. Furthermore, ablation studies will isolate the impact of each component—blockchain, personalization, and synthetic balancing—to better understand their contributions to the overall system performance.

2.2 System Architecture Diagram

The system architecture is designed as a layered pipeline to separate functional components while illustrating their interactions. Clinical IoT data is collected and preprocessed at the Data Layer, processed locally at the Federated Learning Layer, and logged via blockchain for provenance and governance. Personalization and synthetic balancing layers adapt the model to client-specific data and augment rare classes. Finally, the Evaluation Layer assesses performance, fairness, and blockchain overhead, providing a modular and systematic framework for experimentation.

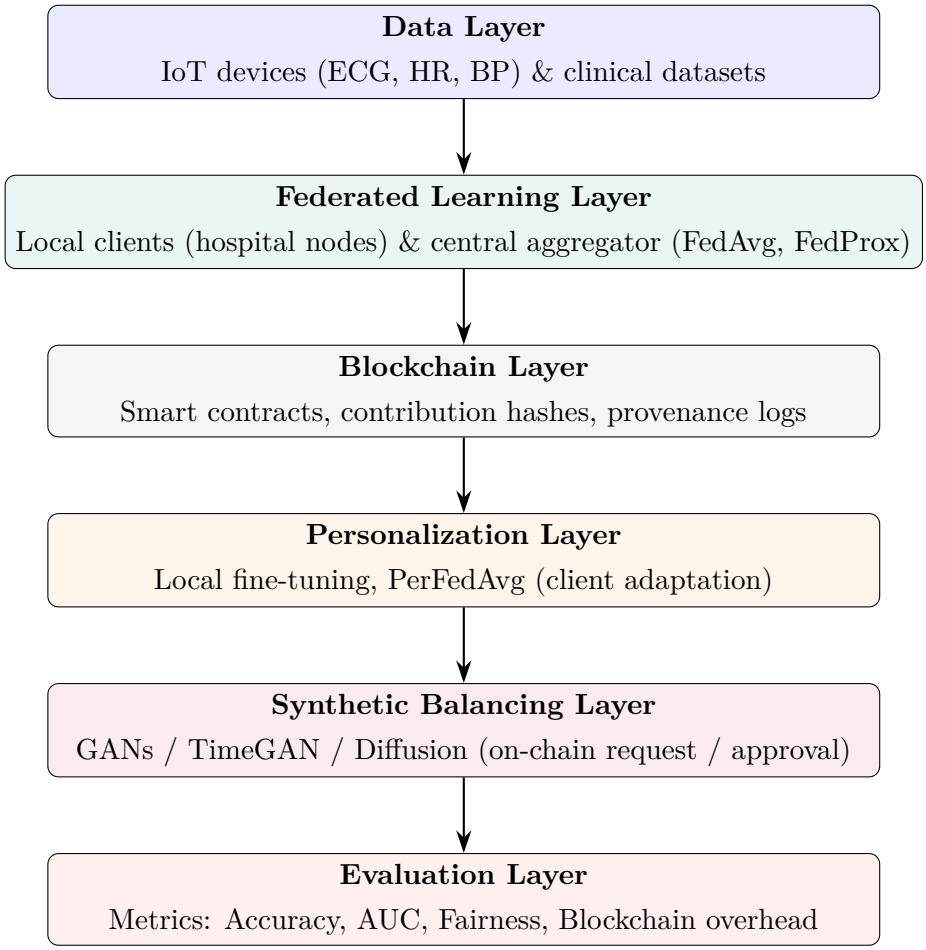

Figure 1: Top-down system architecture: data flows from IoT devices through federated learning, blockchain logging, personalization, synthetic balancing, and evaluation.

2.3 Workflow Diagram

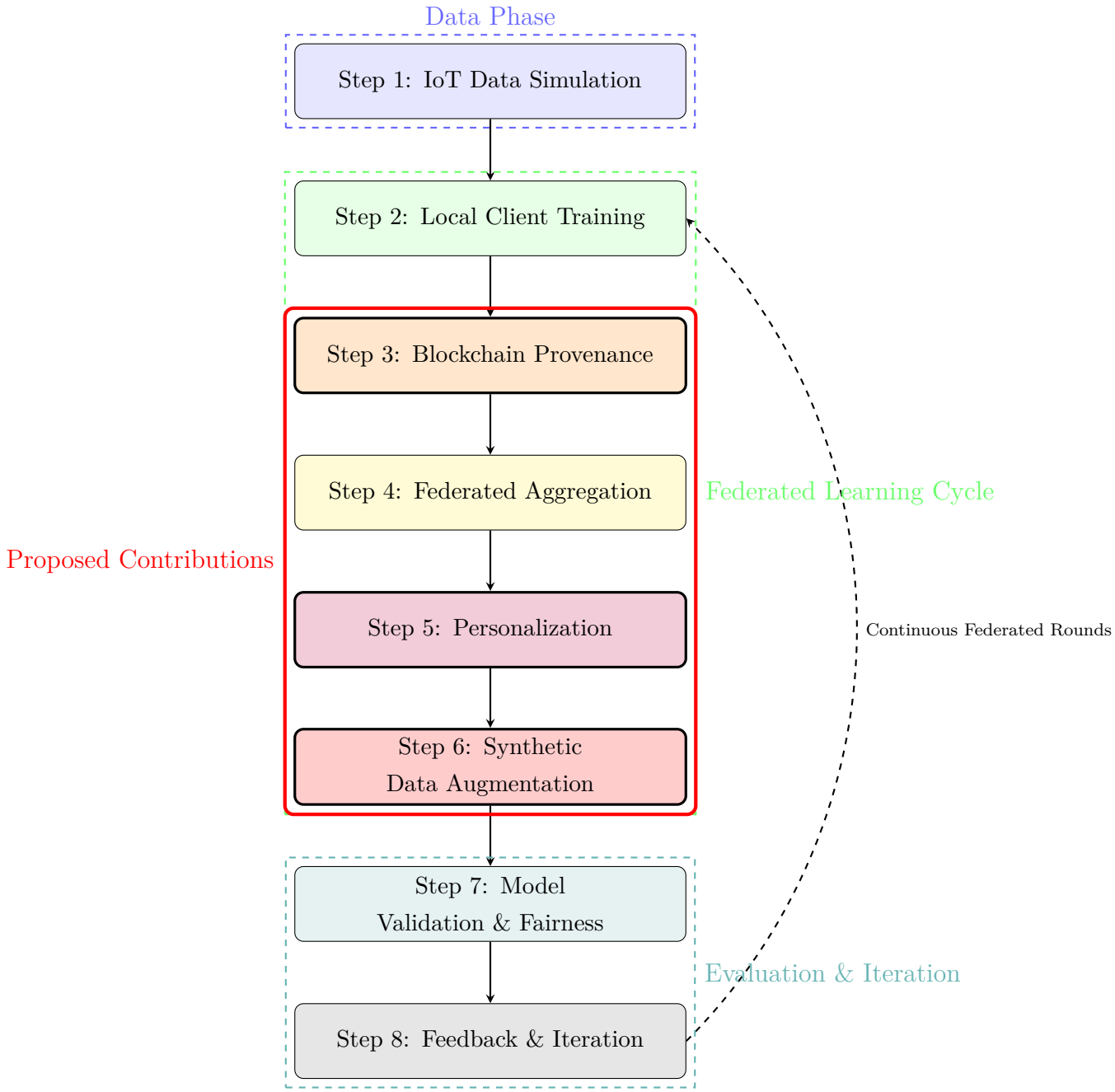

Figure 2: Vertical Workflow Diagram of Proposed System (Compact and Grouped Phases)

2.4 Conceptual Design & Interaction Model

The conceptual design and interaction model outlines a blockchain-orchestrated federated learning system for non-IID clinical IoT data. IoT clients collect and preprocess data (e.g., ECG, vitals), sending it to federated learning layers for training and aggregation

using FedAvg or PerFedAvg to manage heterogeneity. Blockchain governance logs model updates and metadata via smart contracts, ensuring provenance and controlling synthetic data generation for rare-class balancing. Evaluation modules assess accuracy, fairness, and overhead, feeding back to optimize the system, while IoT clients request governed synthetic data (e.g., GANs/TimeGAN) to enhance fairness, forming a closed-loop framework for healthcare AI.

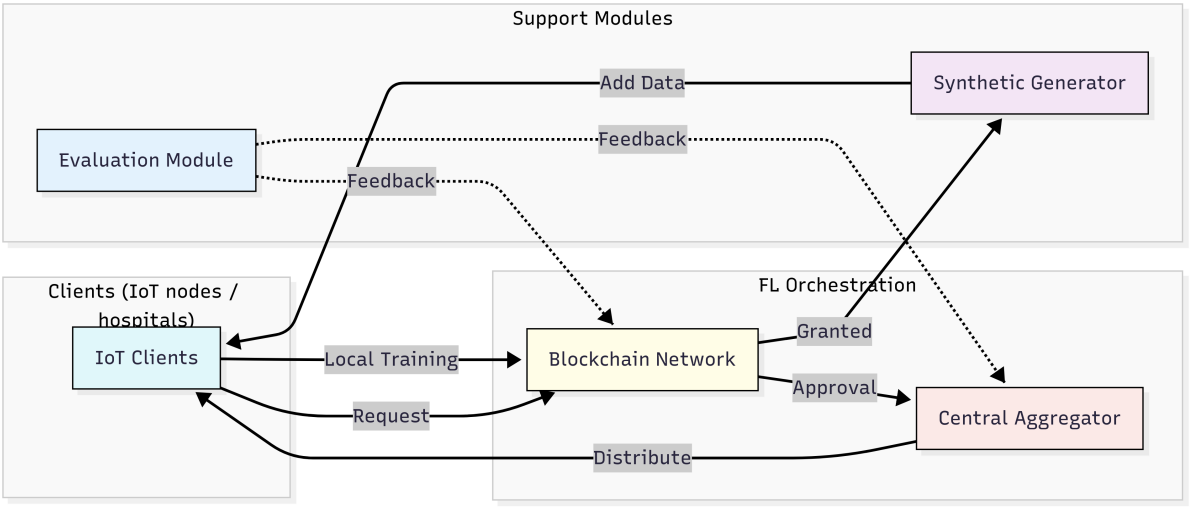

Figure 3: Conceptual design and interaction model.

2.5 Algorithmic Flowcharts

2.5.1 Flowchart 1: Federated Learning Cycle [Figure 4]

The federated learning cycle begins with each participating client (e.g., hospitals or IoT nodes) performing local training on its private dataset. After training, each client generates a model update which is sent to the federated server. The server performs aggregation (FedAvg or FedProx) to update the global model. To address non-IID distributions, a personalization step (e.g., fine-tuning or PerFedAvg) is performed for each client before redistributing the updated model. The cycle repeats until the model converges, at which point personalized models are available for all clients.

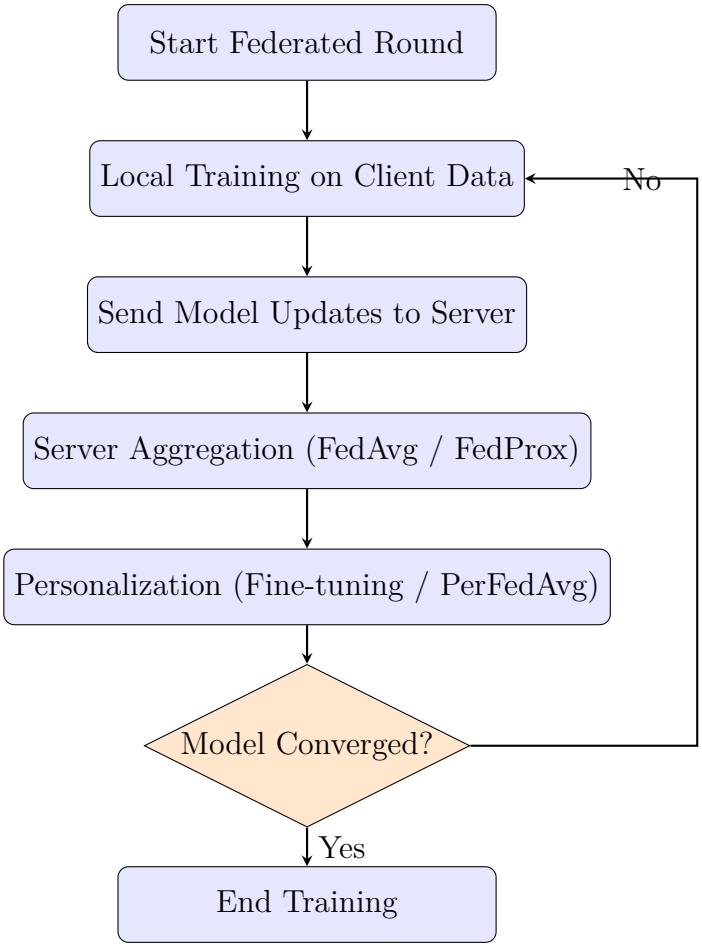

Figure 4: Federated Learning Cycle Flowchart

2.5.2 Flowchart 2: Blockchain Logging [Figure 5]

In the proposed framework, every model update received by the server undergoes blockchain-based provenance logging. When an update is submitted, the system computes a cryptographic hash of the update and associated metadata. This hash is then verified through a smart contract to ensure authenticity and integrity. Once verified, the information is recorded on the blockchain ledger, enabling auditability and governance of the training process. This mechanism not only prevents malicious contributions but also provides transparency in tracking which clients contributed to each round.

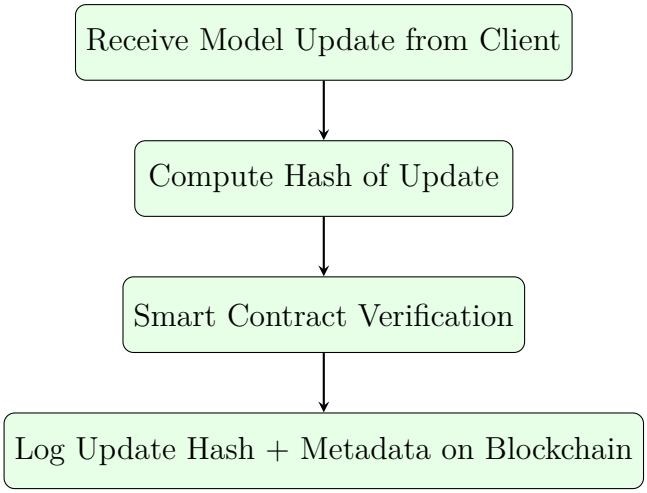

Figure 5: Blockchain Logging Flowchart

2.5.3 Flowchart 3: Synthetic Data Generation [Figure 6]

The synthetic data generation mechanism is designed to address the problem of rare-class imbalance in clinical datasets. When a client detects that certain classes (e.g., rare arrhythmias or extreme events) are underrepresented, it submits a request to the blockchain smart contract. The contract evaluates the request based on predefined governance policies. If approved, the request triggers the generative module (GANs, TimeGAN, or diffusion models) to synthesize realistic rare-class samples. These synthetic samples are then added to the client's local dataset to improve representation. If the request is rejected, no synthetic data is generated, thereby ensuring controlled and transparent usage of artificial data.

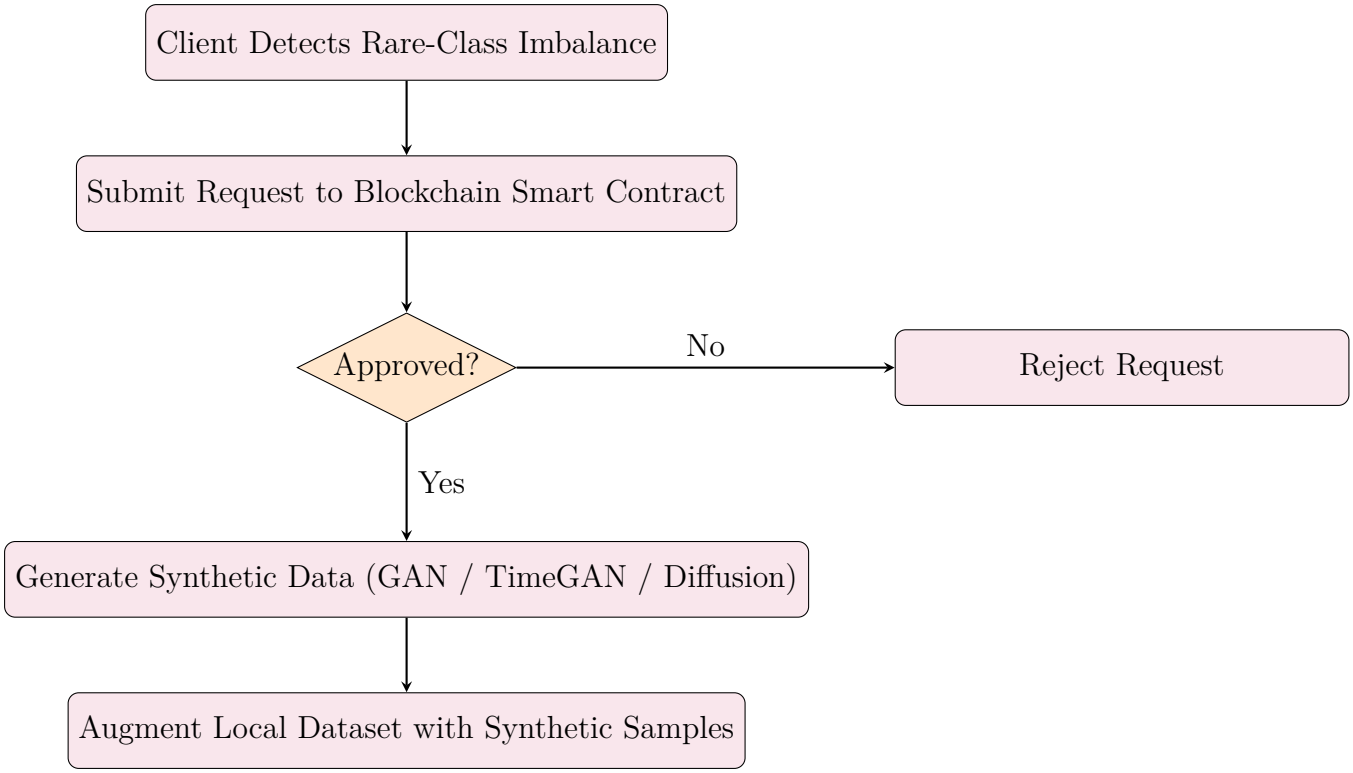

Figure 6: Synthetic Data Generation Flowchart (Compact Vertical Layout)

3 Data Structures and Algorithms

This chapter outlines the core data structures, algorithms, and libraries that will be utilized in the proposed blockchain-orchestrated personalized federated learning framework. These choices are guided by the requirements of handling distributed clinical IoT data, maintaining transparency via blockchain, and generating synthetic data for rare-class balancing.

3.1 Data Structures

3.1.1 Array List (Time-Series IoT Streams)

Time-series data from IoT devices such as ECG, heart rate, or blood pressure will be stored in array lists or equivalent tensor representations. These arrays will preserve the sequential ordering of events, enabling temporal models like LSTMs and CNNs to effectively process patterns over time.

Example:

$$X = [x_1, x_2, \ldots, x_n]$$

where each x_i is a vector of vitals at time i.

Usage: Local training at each client node, sequence modeling.

3.1.2 Maps / Dictionaries (Class Stats, Provenance, Quotas)

Dictionary-like structures will be employed to store key-value pairs for managing metadata and governance:

- **Class Statistics:** {class_label : sample_count} for imbalance detection.
- **Provenance Metadata:** {client_id : update_hash} linking clients to model contributions.
- **Synthetic Quotas:** {client_id : quota_remaining} ensuring controlled synthetic data generation.

3.2 Algorithms

3.2.1 Federated Learning Algorithms (FedAvg, FedProx, PerFedAvg)

- **FedAvg (Federated Averaging):** Clients train locally and send model weights to the server, which computes a weighted average.
- **FedProx:** Extends FedAvg with a proximal term to handle heterogeneous data distributions.

- **PerFedAvg:** A meta-learning approach where the global model is optimized for fast personalization at each client through fine-tuning.

These algorithms will form the baseline and personalized FL strategies for comparison.

3.2.2 Blockchain Consensus & Hashing (SHA-256, Smart Contracts)

Blockchain will provide immutable provenance of contributions:

- **Hashing (SHA-256):** Model updates will be hashed before being logged on-chain, ensuring lightweight and privacy-preserving provenance.
- **Smart Contracts:** Automated rules for update submission, synthetic data requests, and quotas.
- **Consensus Mechanism:** For experimentation, a permissioned blockchain (e.g., Hyperledger Fabric) or Ethereum testnet with Proof-of-Authority (PoA) will be used to avoid high energy costs.

3.2.3 Synthetic Data Algorithms (GANs, TimeGAN, Diffusion Models)

To address rare-class imbalance:

- **GANs (Generative Adversarial Networks):** For tabular clinical data augmentation.
- **TimeGAN:** Specifically for sequential/time-series IoT data like ECG.
- **Diffusion Models:** Potentially explored for generating more diverse, high-fidelity synthetic samples.

Synthetic generation will always be authorized via blockchain smart contracts, ensuring controlled use.

3.3 Libraries and Frameworks

3.3.1 PyTorch / TensorFlow

Used for model training and experimentation with CNNs, LSTMs, GANs, and diffusion models.

3.3.2 Flower / FedML

Frameworks for orchestrating federated learning simulations, client-server coordination, and custom algorithm implementation (FedAvg, FedProx, PerFedAvg).

3.3.3 Web3.py / Solidity / Hyperledger

- **Web3.py:** Python interface to interact with blockchain smart contracts.
- **Solidity:** Language for writing contracts (Ethereum).
- **Hyperledger Fabric:** Alternative framework for permissioned blockchain networks.

3.3.4 Scikit-learn, Pandas, Matplotlib

Supporting libraries for data preprocessing, evaluation, and visualization of results.

4 Project Timeline

4.1 Research Milestones (Indicative)

The project will be carried out in five structured phases, while maintaining flexibility to refine methods, datasets, and tools as research progresses.

Phase I – Exploration and Review. This phase will involve an in-depth literature review on federated learning, blockchain provenance, personalization, and synthetic data generation in healthcare. Suitable datasets such as MIMIC and PhysioNet will be identified, and IoT simulation strategies will be designed.

Phase II – Baseline Experiments. Centralized and federated learning models will be implemented to establish benchmarks under non-IID clinical data. Aggregation methods such as FedAvg and FedProx will be explored to understand their limitations and suitability.

Phase III – Framework Development. The blockchain layer will be introduced to enable provenance logging and governance of model updates. In parallel, personalization strategies and synthetic data generation methods will be incorporated to address fairness and rare-class imbalance challenges.

Phase IV – Evaluation and Analysis. Experiments will be evaluated in terms of accuracy, fairness, privacy, and auditability. Results will be compared against existing literature to highlight contributions and limitations of the proposed framework.

Phase V – Refinement and Documentation. The framework will be refined based on experimental insights, and final documentation will summarize results, challenges, and recommendations for future research.

4.2 Work Division

The project will be carried out collaboratively by three members, each taking lead responsibility for specific components while contributing across multiple areas.

Ibrahim Iqbal (Blockchain & Machine Learning). He will lead the development of the blockchain provenance and governance layer, including smart contracts, logging, and quotas. He will collaborate with Maaz on federated model design and training experiments, ensure blockchain-controlled access for synthetic data, and oversee overall system orchestration.

Maaz Siddiqui(Machine Learning & Synthetic Data Generation). He will lead the design and implementation of baseline and advanced federated learning models such as FedAvg, FedProx, and PerFedAvg. He will also develop CNN/LSTM architectures for time-series IoT data and work on synthetic data generation modules such as GANs, TimeGAN, and diffusion models. Collaborating with Ibrahim, he will ensure smooth integration of blockchain provenance with federated workflows.

Ibrahim Farid (Data Simulation & Evaluation). He will manage dataset collection, preprocessing, and IoT simulation, and partition the data into heterogeneous client subsets to replicate non-IID settings. He will lead evaluation and ablation studies, focusing on fairness, performance, and overhead. In addition, he will support personalization experiments and provide validation datasets.

5 Literature Review

5.1 Federated Learning in Healthcare (Challenges: non-IID, fairness, personalization)

Federated learning (FL) is widely recognized as a promising paradigm for privacy-preserving collaborative learning in healthcare, enabling model training across institutions without centralizing patient data. Reviews show FL's applicability to clinical tasks such as ICU prediction, medical imaging, and ECG analysis, but also document persistent challenges—most prominently data heterogeneity (non-IID client distributions), statistical bias, and degraded convergence under realistic conditions. Both theoretical and empirical analyses demonstrate that FedAvg's performance deteriorates as heterogeneity increases, motivating algorithmic modifications such as FedProx and the exploration of personalization methods. These issues are repeatedly highlighted in healthcare-focused FL reviews, which stress that personalization and fairness in non-IID clinical settings remain open research directions (Kairouz and McMahan 2021; Xu et al. 2021; Rana et al. 2024).

5.2 Blockchain for Federated Learning (Provenance, Governance, Auditability)

Blockchain and distributed ledger technologies have been proposed to provide immutable provenance, accountability, and decentralized governance for FL systems. Prior work describes architectures where on-chain smart contracts record model-update hashes, contribution statistics, and authorization events, thereby enabling audit trails, incentive mechanisms, and tamper-evident logs. Permissioned blockchains or proof-of-authority networks are commonly recommended for prototype deployments to avoid the energy and cost burdens of public chains, while still meeting privacy and participation requirements. Research also explores how blockchain can be combined with fair sampling and contribution scoring to mitigate free-riding and improve trustworthiness in cross-institutional healthcare collaborations (Kim et al. 2020; Li et al. 2024).

5.3 Synthetic Data in Clinical Settings (GANs, Rare-class Augmentation)

Generative models such as GANs, TimeGAN, conditional GANs, and more recently diffusion approaches, are increasingly employed to synthesize clinical signals (e.g., ECG and vitals) and tabular health records for augmentation and privacy-preserving research. Multiple studies demonstrate that synthetic samples can improve classifier performance for underrepresented classes when quality controls are applied. However, they also

highlight evaluation challenges, including the need to measure fidelity, utility, and potential membership leakage. Recent comparative work emphasizes the importance of quality control metrics (e.g., Wasserstein or KL distances, nearest-neighbor leakage tests, and downstream task utility) and cautions that poorly governed synthetic generation can introduce artifacts or privacy risks. These findings underscore the need for controlled and auditable synthetic data use in healthcare FL (Esteban, Hyland, and Rätsch 2017; Jordon et al. 2018; Azizi et al. 2023).

5.4 Personalization in Federated Learning (Meta-learning, Fine-tuning Approaches)

To address the challenges of non-IID client data, the literature presents two main personalization strategies: (1) local fine-tuning of the aggregated global model and (2) dedicated personalized FL algorithms. Examples of the latter include meta-learning approaches such as PerFedAvg, representation learning approaches like FedRep, and attention or adapter-based methods. Empirical studies indicate that personalization substantially improves client-level performance—especially for clients with skewed label distributions—and reduces variance across participating sites, contributing to fairness. Nevertheless, personalization introduces trade-offs in communication efficiency, system complexity, and the risk of overfitting, which require careful evaluation in clinical contexts (Fallah et al. 2020).

5.5 IoT in Healthcare (Streaming Vitals, ECG, Clinical Sensors)

The growth of clinical IoT, encompassing bedside monitors, wearables, and implantable devices, has enabled continuous collection of vital signals such as ECG, heart rate, oxygen saturation, and blood pressure. These time-series data streams are valuable for predictive modeling but exacerbate heterogeneity due to differences in device types, sampling rates, and clinical practices across institutions. Literature on IoT in healthcare highlights opportunities such as real-time monitoring and longitudinal modeling, while also acknowledging challenges including connectivity issues, missing data, and difficulties in labeling. These realities motivate the use of simulated IoT streams and robust preprocessing strategies for realistic FL experiments, including windowing and methods to handle noisy or incomplete signals (Islam et al. 2015; Khan et al. 2024).

6 Novelty and Contributions

Recent work has explored blockchain-based federated learning, personalization for non-IID data, and synthetic data generation in healthcare. However, most approaches treat these areas separately. Our project aims to unify these components into a single auditable framework for clinical IoT data.

Key Contributions

- **Blockchain Governance** – Smart contracts for logging updates, authorizing synthetic data requests, and ensuring transparent provenance.
- **Personalized Federated Learning** – Fine-tuning and PerFedAvg to adapt global models to heterogeneous clinical clients.
- **Synthetic Data Balancing** – Blockchain-controlled augmentation using GANs/TimeGAN/diffusion for rare-class fairness.
- **Integrated Orchestration** – A modular framework combining IoT simulation, FL, blockchain, personalization, and augmentation with evaluation on accuracy, fairness, and overhead.

Novelty

Our contribution lies in *integration*. While prior work often focuses on one of these areas in isolation, we combine blockchain governance, personalization, and synthetic balancing into one system tailored for non-IID clinical IoT data. This integration produces an auditable, adaptive, and governed federated learning pipeline that addresses both distributional heterogeneity and rare-class imbalance. As evident in Fig. 7.

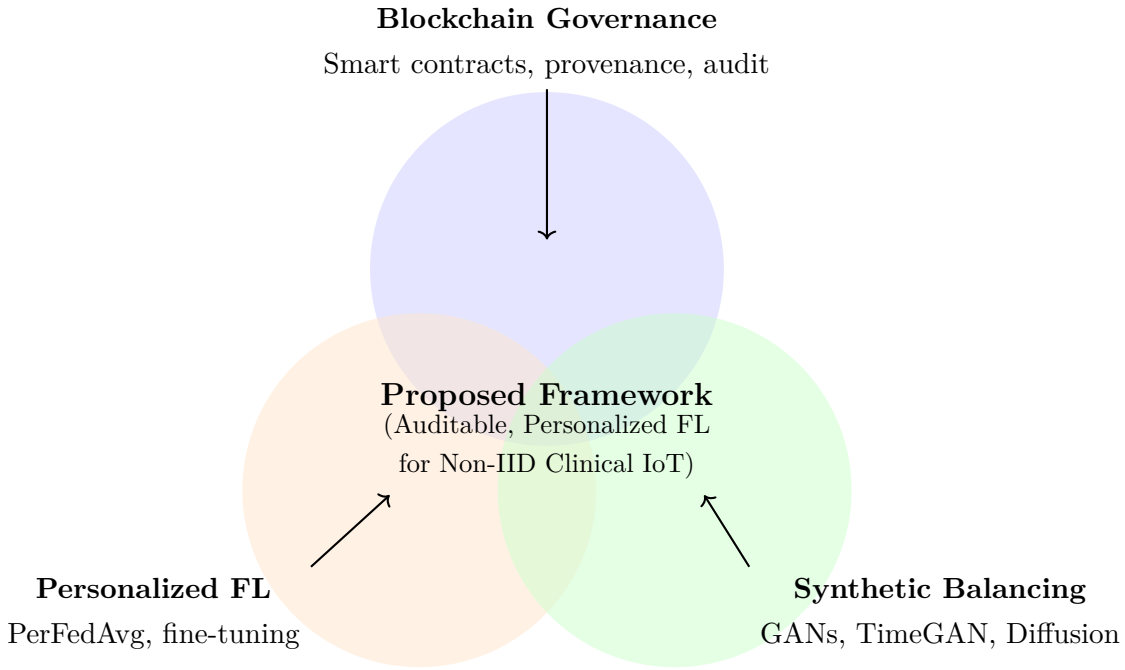

Figure 7: Novelty summary: integration of blockchain governance, personalized FL, and synthetic balancing.

7 Limitations

The study relies on publicly available datasets such as MIMIC-III and PhysioNet; however, these resources may not fully represent the diversity and complexity of real-world clinical IoT environments. Validation of the proposed system will be conducted within a simulated multi-client setting rather than in live hospital deployments, as regulatory and ethical considerations place such implementations beyond the project's scope.

Another important consideration is the computational and communication overhead introduced by the blockchain layer. While this overhead will be measured and analyzed, it may impose limitations on scalability, particularly in high-frequency IoT scenarios. Similarly, synthetic data generation offers potential to mitigate class imbalance, but maintaining fidelity and privacy remains a challenge, as there is always the risk of sensitive pattern leakage.

Finally, personalization in federated learning may enhance performance for individual clients, yet it also brings trade-offs such as increased computational requirements and a higher risk of overfitting. These constraints collectively highlight the boundaries of this research while guiding the interpretation of its findings.

8 Scope of the Project

This project will investigate the feasibility of a blockchain-orchestrated federated learning system enriched with personalization techniques and synthetic balancing mechanisms for clinical IoT data. The scope encompasses a comprehensive literature review covering federated learning, blockchain-based provenance, synthetic augmentation, and healthcare IoT applications. It also involves simulating heterogeneous clinical IoT data through benchmark datasets, implementing baseline federated learning methods such as FedAvg and FedProx, and exploring personalization strategies including fine-tuning and PerFedAvg.

The system will incorporate blockchain-driven provenance logging to ensure transparency and accountability, alongside controlled synthetic data augmentation to address rare-class imbalance. Evaluation will focus on key dimensions such as accuracy, fairness, and governance, thereby assessing both technical performance and ethical considerations.

This project is positioned as a research exploration rather than a production-ready deployment. The techniques, models, and datasets outlined here are indicative starting points; alternative approaches may be pursued if they better align with the objectives or yield stronger insights during experimentation.

9 Future Scope

This work creates several avenues for future research and development:

- **Clinical Deployment:** Extend the framework to real-world hospital consortia operating under HIPAA and GDPR compliance.
- **Advanced Personalization:** Investigate federated transfer learning, adapter modules, and attention-based methods for personalization across multi-modal healthcare data.
- **Scalable Blockchain:** Explore lightweight consensus mechanisms (e.g., DAG-based ledgers) to reduce the computational and communication overhead of blockchain.
- **Secure Multi-party Computation & Differential Privacy:** Integrate blockchain with techniques such as homomorphic encryption and differential privacy to strengthen privacy and security guarantees.
- **Cross-domain Applications:** Adapt the framework for other sensitive domains, including finance, smart grids, and climate-health modeling.

10 Expected Outcomes & Deliverables

10.1 Expected Outcomes

- A research model demonstrating provenance-aware federated learning with personalized adaptation and synthetic data balancing.
- Clear evidence that the system improves performance on non-IID and rare-class scenarios compared to baseline federated learning approaches.
- An auditable blockchain ledger of contributions and synthetic data usage, ensuring trust and accountability in clinical IoT environments.

10.2 Deliverables

- A comparison with existing approaches, highlighting the advantages of the proposed framework.
- A comprehensive study of the field, documenting the state-of-the-art in federated learning, blockchain, and synthetic data generation.
- A framework for addressing the challenges of non-IID data, rare-class imbalance, and transparent governance in clinical IoT environments.

11 Conclusion

This project proposes a novel framework that integrates federated learning, blockchain provenance, personalization, and synthetic data balancing to address the pressing challenges of non-IID data, fairness, and rare-class representation in clinical IoT. Unlike conventional federated approaches, the system introduces on-chain auditability for governance and controlled synthetic augmentation to mitigate data imbalance.

Through simulation experiments with benchmark datasets, the project aims to provide evidence-based insights into the feasibility and trade-offs of such a design. While the prototype will remain exploratory and limited in scale, it is expected to highlight both the potential benefits and practical limitations of combining blockchain and generative AI within federated healthcare ecosystems.

Ultimately, this work aspires to contribute toward trustworthy, personalized, and fair AI in healthcare, and to serve as a foundation for future extensions in real-world clinical collaborations.

12 Appendix: List of Abbreviations

.

AI Artificial Intelligence

AUC Area Under the Receiver Operating Characteristic Curve

BP Blood Pressure

CNN Convolutional Neural Network

ECG Electrocardiogram

FedAvg Federated Averaging (baseline FL algorithm)

FedProx Federated Proximal (robust FL algorithm for heterogeneous data)

FL Federated Learning

F1-score Harmonic mean of precision and recall

GAN Generative Adversarial Network

HIPAA Health Insurance Portability and Accountability Act

HR Heart Rate

IoT Internet of Things

LSTM Long Short-Term Memory (neural network)

MIMIC-III Medical Information Mart for Intensive Care III (clinical dataset)

non-IID Non-Independent and Identically Distributed (heterogeneous data)

PerFedAvg Personalized Federated Averaging (meta-learning approach)

R–R Interval Time interval between successive R-peaks in ECG signals

SHA-256 Secure Hash Algorithm 256-bit

Web3.py Python library for interacting with Ethereum blockchain

References

[1] P. Kairouz, H. B. McMahan, et al., "Advances and Open Problems in Federated Learning," *Foundations and Trends in Machine Learning*, 2021.

[2] J. Xu, et al., "Federated learning for healthcare informatics," *Journal of Biomedical Informatics*, 2021.

[3] A. Fallah, et al., "Personalized federated learning: A meta-learning approach," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[4] S. M. R. Islam, et al., "The Internet of Things for Health Care: A Comprehensive Survey," 2015.

[5] R. Khan, et al. "Advanced Federated Ensemble Internet of Learning for IoT-Enabled Medical Systems." *Scientific Reports* 14 (October 2024)

[6] Y. Li, C. Xia, D. Huang, L. Sun, and T. Wang. "BFLN: A Blockchain-based Federated Learning Model for Non-IID Data." *arXiv preprint* (2024).

[7] Z. Azizi, et al. "A Comparison of Synthetic Data Generation and Federated Learning for International Healthcare." *Scientific Reports* 13, no. 1 (July 2023)

[8] N. Rana, et al. "Role of Federated Learning in Healthcare Systems: A Survey." *Mathematical and Computational Applications* 29, no. 2. June 2024.

[9] X. Yang, D. Wan, G. Han, and W. Zhang. "Personalized Federated Learning with Hierarchical Reweighting for Multi-Center Clinical Prediction." *Computer Methods and Programs in Biomedicine* 271 (August 2025)

[10] R. Cowlishaw, N. Longépé, and A. Riccardi. "Balancing Centralisation and Decentralisation in Federated Learning for Earth Observation-Based Agricultural Predictions." Published March 26, 2025.

[11] M. Arafeh, M. Wazzeh, H. Sami, H. Ould-Slimane, C. Talhi, A. Mourad, and H. Otrok. "Efficient Privacy-Preserving ML for IoT: Cluster-Based Split Federated Learning Scheme for Non-IID Data." *Future Generation Computer Systems* 236 (April 2025)

[12] T. Bhardwaj and K. Sumangali. "An Explainable Federated Blockchain Framework with Privacy-Preserving AI Optimization for Securing Healthcare Data." July 2025.

[13] N. T. Madathil, F. K. Dankar, M. Gergely, A. N. Belkacem, and S. Alrabaee. "Revolutionizing Healthcare Data Analytics with Federated Learning: A Comprehensive Survey of Applications, Systems, and Future Directions." 2025.

[14] X. Yang, D. Wan, G. Han, W. Zhang, and W. Tang. “Personalized Federated Learning with Hierarchical Reweighting for Multi-Center Clinical Prediction.” 2025.

[15] M. Firdaus, H. T. Larasati, and K. Hyune-Rhee. “Blockchain-Based Federated Learning with Homomorphic Encryption for Privacy-Preserving Healthcare Data Sharing.” May 2025.