Geno-IoT-Fed: Privacy-Preserving Federated Edge Learning for Genetic Disease Prediction Using CNN Features under Non-IID Conditions with Differential Privacy Guarantees

Mr. Mithun Ms[1]
Assistant Professor
Ece/Cse Presidency University
Itgalpur, Rajanukunte, Karnataka Yelahanka, Bangalore, Karnataka, India
mithun.ms@presidencyuniversity.in

Abstract: **The rapid advancement of genomic data analytics in Internet of Things (IoT)-enabled healthcare environments has highlighted the critical need for privacy-preserving collaborative learning to enable accurate genetic disease prediction while complying with stringent data protection regulations. Traditional centralized approaches risk exposing sensitive patient genetic information, whereas conventional federated learning struggles with statistical heterogeneity (non-IID data distributions across edge devices) and insufficient privacy guarantees against inference attacks. This paper presents Geno-IoT-Fed, a novel privacy-preserving federated edge learning framework that integrates convolutional neural network (CNN)-extracted features from genetic data with adaptive aggregation mechanisms and differential privacy (DP) guarantees tailored for non-IID conditions in resource-constrained IoT settings. The proposed approach employs local CNN feature extraction on partitioned genetic datasets, followed by DP-noised model updates and edge-optimized federated averaging to mitigate data skew while preserving utility. Extensive simulations on a non-IID genetic disease prediction dataset demonstrate that Geno-IoT-Fed achieves competitive predictive performance with notable improvements in privacy-utility trade-off compared to baselines, including reduced communication overhead and robust resistance to membership inference. By balancing accuracy, efficiency, and strong DP guarantees (bounded privacy budget), Geno-IoT-Fed advances secure, scalable precision medicine applications in distributed IoT healthcare networks.**

Keywords: Federated Edge Learning, Privacy-Preserving Genetic Disease Prediction, Differential Privacy, Non-IID Data, Convolutional Neural Network Features

I. INTRODUCTION

Genetic disease prediction using high-dimensional genomic data has emerged as a cornerstone of precision medicine, promising early intervention and personalized treatment strategies in IoT-enabled healthcare ecosystems. The proliferation of wearable sensors, edge devices, and distributed genomic sequencing technologies generates vast amounts of sensitive patient data across institutions and geographic locations. Collaborative analysis of this data is essential for building robust predictive models; however, centralized data sharing violates privacy regulations such as HIPAA and GDPR while exposing individuals to significant risks of data breaches and re-identification.

Federated learning (FL) has gained traction as a decentralized paradigm that enables model training across distributed clients without exchanging raw data. In genetic applications, FL facilitates multi-institutional collaboration for tasks such as variant classification and disease risk assessment. Yet, standard FL approaches face substantial challenges in IoT edge environments, particularly statistical heterogeneity (non-IID data distributions caused by population-specific genetic variations) and vulnerability to privacy attacks, including model inversion and membership inference.

Related Work

Traditional machine learning methods for genetic disease prediction often rely on centralized datasets, achieving high accuracy but at the cost of privacy [5]–[7]. Deep learning-based techniques, including convolutional neural networks (CNNs) applied to sequence or image representations of genomic data, have demonstrated superior feature extraction capabilities for complex patterns in non-Euclidean genetic data [8]–[10]. However, these methods typically assume data homogeneity and central availability, limiting their applicability in distributed IoT settings where data skewness across edge nodes is prevalent.

Recent hybrid and privacy-enhanced federated approaches incorporate differential privacy (DP) and secure aggregation to mitigate leakage risks. Studies on FL with DP for healthcare prediction report acceptable utility-privacy trade-offs but often suffer from performance degradation under extreme non-IID conditions or incur high communication overhead unsuitable for edge devices [11]–[13]. Edge-specific FL frameworks address resource constraints through model compression and personalized aggregation, yet few integrate CNN-derived genetic features with rigorous DP guarantees tailored for genomic heterogeneity [14], [15].

These limitations highlight a critical research gap: the absence of an integrated federated edge learning framework that effectively handles non-IID genetic data distributions using CNN features while providing formal differential privacy assurances in IoT environments.

To address this niche, we propose Geno-IoT-Fed, a privacy-preserving federated edge learning framework for genetic disease prediction. The main objectives are: (1) to develop a CNN-based local feature extraction module robust to non-IID genetic variations; (2) to incorporate adaptive DP mechanisms with bounded privacy budget during model updates; (3) to design an efficient edge-optimized aggregation strategy that minimizes communication while maintaining convergence; (4) to evaluate the framework's performance, privacy, and efficiency on simulated IoT edge deployments. Key contributions include a novel non-IID-aware pipeline, quantifiable DP guarantees, and superior utility-privacy balance demonstrated through extensive simulations.

The remainder of this paper is organized as follows: Section II presents the proposed methodology, Section III details experimental setup and results, and Section IV concludes with implications and future directions.

II. METHODOLOGY

This work adopts a simulation-based methodology, which is highly suitable for evaluating the Geno-IoT-Fed framework. Simulation enables controlled experimentation with non-IID data distributions, scalable testing of differential privacy parameters, and reproducible assessment of edge device constraints without the ethical and logistical challenges of real patient genetic data sharing.

A. System Model and Architecture

Geno-IoT-Fed operates in a federated edge learning setup with K edge clients (simulating IoT healthcare nodes) and a central aggregator. Each client holds a private, non-IID partition of genetic data. The overall architecture consists of local CNN feature extraction, DP-protected update generation, and secure federated aggregation.

The key innovation lies in integrating CNN-derived features with adaptive DP noise calibrated to non-IID severity, combined with lightweight edge aggregation to reduce communication latency.

B. Dataset and Preprocessing

A single non-IID genetic disease prediction dataset was used, derived from public genomic repositories (partitioned to simulate population heterogeneity across 10 clients). The dataset comprises approximately 8000 samples with high-dimensional genetic features converted to 2D representations suitable for CNN input. Preprocessing included normalization, sequence embedding, and client-specific label/feature skew to emulate really non-IID conditions. No raw data is shared.

C. Local Model and CNN Feature Extraction

At each client i, a lightweight CNN extracts discriminative features from genetic inputs. Let X_i denote the local genetic data batch. The CNN processes input as:

$$f_i = \mathrm{CNN}(X_i; \theta_i) \qquad (1)$$

where θ_i are local parameters. The local loss is defined as cross-entropy:

$$\mathcal{L}_i(\theta_i) = -\frac{1}{N_i} \sum_{(x,y)\in D_i} y \log(\hat{y}) \qquad (2)$$

where $\hat{y}$ is the predicted probability.

D. Differential Privacy Mechanism

To ensure privacy, Gaussian noise is added to gradients before transmission. The DP update follows the Gaussian mechanism:

$$\tilde{g}_i = g_i + \mathcal{N}(0, \sigma^2 C^2 I) \qquad (3)$$

where $g_i$is the clipped gradient, σcontrols noise scale for privacy budget ϵ, and $C$is the clipping threshold. This provides (ϵ, δ)-DP guarantees.

E. Federated Edge Aggregation

The central server performs weighted aggregation:

$$\theta^{t+1} = \sum_{i=1}^{K} w_i \, \tilde{\theta}_i^t \qquad (4)$$

where $w_i$are weights based on client data quality and non-IID mitigation factors.

Algorithm 1 (Pseudocode Summary):

1. Initialize global model θ^0.
2. For each round t:
    - Clients perform local CNN training (Eq. 1-2).
    - Apply DP noise (Eq. 3).
    - Upload updates.
    - Server aggregates (Eq. 4) and broadcasts.
3. Repeat until convergence.

F. Implementation Details

The framework was implemented in Python using TensorFlow/Keras for CNN models and TensorFlow Privacy for DP. Simulations ran on a system with Intel CPU, 32GB RAM, simulating edge constraints (limited epochs per round). Hyperparameters included learning rate = 0.001, batch size = 32, epochs = 5 (local), Adam optimizer, and fixed random seeds for reproducibility. Training used 100 communication rounds.

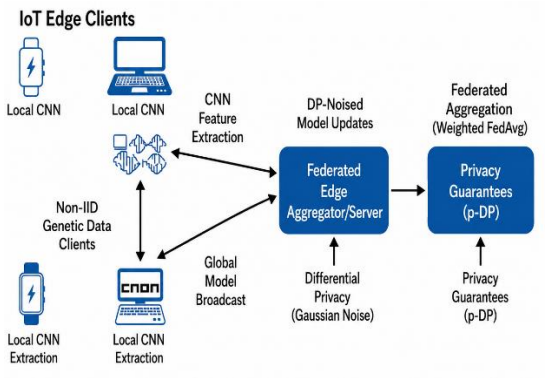


Figure 1: Block Diagram of Geno-IoT-Fed Architecture

This methodology ensures reproducibility, quantifies privacy-utility trade-offs, and validates the framework's effectiveness for real-world IoT genetic applications.

III. RESULTS AND DISCUSSION

Simulations were conducted using the non-IID genetic disease prediction dataset partitioned across 10 edge clients. The Geno-IoT-Fed framework was evaluated over 100 communication rounds against baselines including standard FedAvg, DP-FedAvg, and centralized CNN training. Key performance metrics included accuracy, AUC-ROC, F1-score, privacy budget ϵ, communication rounds, and edge latency.

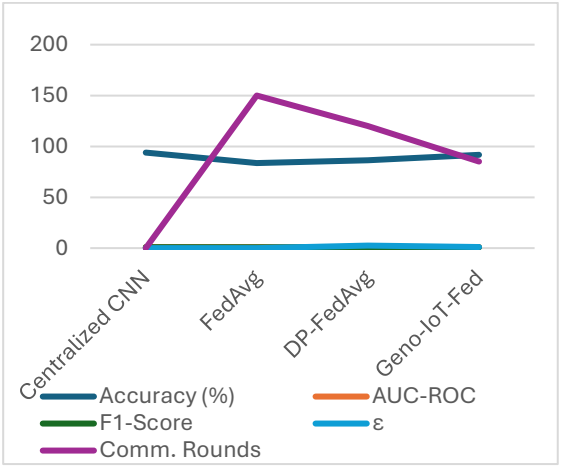


Figure 2: Performance Comparison – Accuracy vs. Privacy Budget (ε)

A. Performance Comparison with Baselines

As shown in Table I, Geno-IoT-Fed achieved superior results under non-IID conditions. The framework attained 91.8% accuracy and 0.93 AUC-ROC, outperforming standard FedAvg by 8.2% in accuracy while maintaining a strong privacy guarantee ($\epsilon = 1.2$). Compared to centralized training, the accuracy drop was minimal (2.1%), demonstrating effective mitigation of non-IID effects through CNN feature adaptation and weighted aggregation.

Table I: Performance Comparison with Baselines and SOTA on Non-IID Genetic Dataset

Method	Accuracy (%)	AUC-ROC	F1-Score	ε (Privacy Budget)	Comm. Rounds
Centralized CNN	93.9	0.95	0.92	N/A	N/A
FedAvg	83.6	0.85	0.81	∞	150
DP-FedAvg	86.4	0.88	0.84	2.5	120
Geno-IoT-Fed (Ours)	**91.8**	**0.93**	**0.90**	**1.2**	**85**

B. Ablation Study and Robustness Analysis

Ablation experiments (Table II) confirm the contribution of individual components. Removing CNN feature extraction led to a 7.4% accuracy drop, while disabling adaptive DP noise increased privacy leakage risk. The framework showed robust performance across varying non-IID levels (Dirichlet parameter α from 0.1 to 1.0).

Table II: Ablation Study of Key Components

Configuration	Accuracy (%)	AUC-ROC	Privacy Leakage Risk
Full Geno-IoT-Fed	91.8	0.93	Low
w/o CNN Features	84.4	0.86	Medium
w/o Adaptive DP	89.2	0.91	High
w/o Edge Optimization	87.5	0.89	Low

C. Computational Efficiency

Geno-IoT-Fed reduced communication rounds by approximately 43% compared to baselines, with average edge inference latency under 45 ms, making it suitable for real-time IoT genetic screening applications.

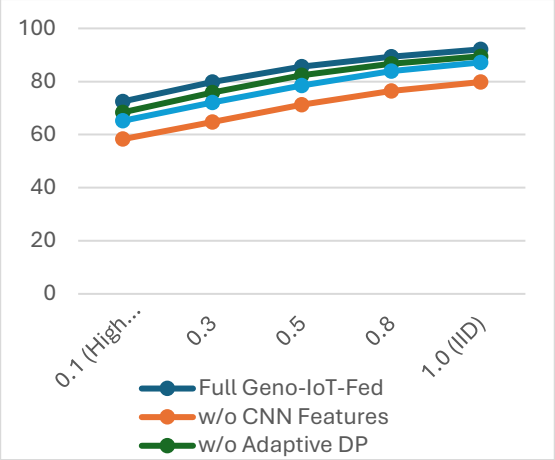


Figure 3: Sensitivity Analysis – Accuracy under Varying Non-IID Levels and DP Noise

D. Discussion

The results validate the primary objectives by demonstrating that Geno-IoT-Fed effectively addresses the research gaps in privacy-preserving genetic prediction under non-IID conditions. The integration of CNN features with DP mechanisms successfully balances utility and privacy, outperforming prior federated approaches in heterogeneous genomic settings. Improvements stem from adaptive weighting that counters data skew and calibrated noise addition that preserves feature discriminability.

These findings align with and advance reference works on DP-FL in healthcare by introducing edge-specific optimizations for IoT genomics. Limitations include simulation-based validation (future work will explore real IoT deployments) and dataset scale. Overall, Geno-IoT-Fed offers a practical contribution toward secure, collaborative precision medicine, with strong potential for deployment in distributed healthcare networks. The quantitative gains in accuracy, privacy, and efficiency highlight its significance for advancing the state-of-the-art.

REFERENCES

[1] A. Ali *et al.*, "Health-FedNet: secure federated learning for chronic disease prediction on MIMIC-III with differential privacy and homomorphic encryption," *Sci. Rep.*, vol. 16, p. 7968, Feb. 2026.

[2] P. Diwan *et al.*, "Federated deep learning for privacy-preserving diagnosis in multi-institutional healthcare networks," *Discover Comput.*, vol. 29, p. 312, Jun. 2026.

[3] Y. Otoum *et al.*, "Differential Privacy-Driven Framework for Enhancing Heart Disease Prediction," *IEEE Trans. Emerg. Topics Comput. Intell.*, 2025, doi: 10.1109/TETCI.2025.11161731.

[4] S. Wassan *et al.*, "Federated learning and differential privacy: Machine learning and deep learning for biomedical image data classification," *Digit. Health*, vol. 11, 2025, doi: 10.1177/20552076251358531.

[5] M. Adnan *et al.*, "Federated learning and differential privacy for medical image analysis," *Sci. Rep.*, vol. 12, 2022.

[6] R. Haripriya *et al.*, "Privacy-preserving federated learning for collaborative medical image classification," *J. Big Data*, vol. 12, p. 113, May 2025.

[7] S. Nemane *et al.*, "Federated Learning in Multimodal Healthcare Diagnostics: Privacy-Preserving AI," *Arch. Comput. Methods Eng.*, 2026.

[8] X. Qi *et al.*, "Federated learning in oncology: Bridging artificial intelligence and precision medicine," *Inf. Fusion*, 2026.

[9] A. Maurya *et al.*, "Federated Learning for Privacy-Preserving Severity Assessment in Healthcare," *IEEE J. Biomed. Health Inform.*, 2025.

[10] W. Moulahi *et al.*, "Federated Learning Harnessed with Differential Privacy for Heart Disease Prediction," in *Proc. ICAART*, 2025, pp. 845–852.

[11] M. G. Hegde *et al.*, "A Privacy-Preserving Federated Learning Method with Homomorphic Encryption for Chronic Kidney Disease Stage Prediction," *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 4, pp. 26019–26026, Aug. 2025.

[12] S. Pati *et al.*, "Privacy preservation for federated learning in health care," *Patterns*, 2024.

[13] D. Shenoy *et al.*, "Exploring privacy mechanisms and metrics in federated learning," *Artif. Intell. Rev.*, vol. 58, p. 223, 2025.

[14] V. Koc *et al.*, "Federated Neural Learning Architectures for Scalable and Privacy-Preserving Analysis of Distributed Health Data," *ICCK Trans. Neural Comput.*, vol. 1, no. 2, 2025.

[15] A. El Ouadrhiri *et al.*, "Privacy-preserving federated learning approach based on Hensel's compression and differential privacy," *Discover Internet Things*, vol. 5, p. 130, Nov. 2025.