AKSU Journal of Engineering and Technology (AKSUJET), 2026
Vol. 1, No. 1, Pages 56-64
Published Online April 2026 in AKSUJET (http://www.aksujet.org) ISBN: 978-978-53006-97

A FEDERATED FRAMEWORK FOR PRIVACY-PRESERVING HEALTH DATA SHARING ACROSS AFRICAN BORDERS

Igbajar Abraham
Computer Science
Lusaka Goldsmiths University College, Lusaka, Zambia
igbajar35@gmail.com

Nwachukwu-Nwokeafor, K. C
Computer Engineering
Michael Okpara University of Agriculture, Umudike, Abia State, Nigeria
nwachukwuken72@gmail.com

***Abstract*— Cross-border health data sharing in Africa is constrained by fragmented regulatory frameworks and concerns regarding data sovereignty and institutional trust. Existing centralized approaches are often incompatible with national data protection regulations such as the Nigeria Data Protection Regulation (NDPR) and the Protection of Personal Information Act (POPIA). This study proposes AfriPShare, a hybrid privacy-preserving framework that integrates Federated Learning (FL) with Local Differential Privacy (LDP) to enable collaborative model training without sharing raw patient data. A simulation environment was developed using synthetic datasets representing five African countries. Experimental results show that the framework achieves a classification accuracy of 94.3% (±0.6) at a privacy budget of ε = 1.0, demonstrating a strong balance between privacy and utility. The findings indicate that AfriPShare provides a scalable and compliant solution for distributed health data analytics in Africa. Keywords:**

Keywords— **Privacy-Enhancing Technologies, Federated Learning, Differential Privacy, Health Data, African Data Sovereignty.**

I. Introduction

The increasing reliance on data-driven approaches in public health has highlighted the importance of cross-border data sharing, particularly in managing infectious diseases and pandemics. However, in Africa, such collaboration is hindered by heterogeneous regulatory frameworks and concerns over data sovereignty. National regulations such as NDPR, POPIA, and related frameworks impose strict controls on data movement, limiting centralized data-sharing models.

Existing collaborative analytics approaches typically assume harmonized legal environments, which are not present in the African context. Consequently, there is a need for technical frameworks that operate within fragmented regulatory landscapes while preserving data privacy.

This study proposes AfriPShare, a hybrid federated framework designed to enable secure and compliant health data sharing across African borders.

A. The main contributions of this study are as follows:

A novel integration of Federated Learning and Local Differential Privacy tailored to the African context.

A simulation framework representing heterogeneous multi-country datasets.

A quantitative evaluation of privacy–utility trade-offs under varying privacy budgets.

A practical architecture aligned with African data protection regulations.

II. Literature Review

The intersection of collaborative machine learning and data privacy has seen significant development, yet its application within the African regulatory and infrastructural context remains under-researched. This section evaluates current approaches to privacy-preserving data sharing and identifies the limitations that AfriPShare seeks to address.

A Privacy-Preserving Technologies in Healthcare

Traditional methods for health data sharing have largely relied on anonymization and pseudonymization. However, studies by Rocher et al. (2019) have demonstrated that these techniques are increasingly vulnerable to re-identification attacks when cross-referenced with external datasets. To mitigate these risks, recent scholarship has shifted toward Federated Learning (FL) and Differential Privacy (DP).

FL allows for the training of models on decentralized data, ensuring that raw medical records never leave the local institution (McMahan et al., 2017). Despite these benefits, FL remains susceptible to "gradient leakage" attacks, where

AKSU Journal of Engineering and Technology (AKSUJET), 2026
Vol. 1, No. 1, Pages 56-64
Published Online April 2026 in AKSUJET (http://www.aksujet.org) ISBN: 978-978-53006-97

sensitive information can be reconstructed from the shared model updates. Consequently, the integration of Local Differential Privacy (LDP**)** has emerged as a robust countermeasure, injecting statistical noise into updates before they are transmitted to a central server.

B. The African Regulatory and Technical Landscape

The adoption of these technologies in Africa is complicated by a heterogeneous legal environment. While the Nigerian Data Protection Regulation (NDPR) and South Africa's Protection of Personal Information Act (POPIA) provide frameworks for privacy, they vary significantly in their requirements for cross-border data transfer. Furthermore, technical constraints, such as intermittent connectivity and hardware variability, necessitate a framework that is both communication-efficient and computationally lightweight.

C. Comparison of Existing Frameworks

Table 1 provides a comparative analysis of established privacy-preserving frameworks against the requirements identified for the African context.

Table 1: Comparison of Privacy-Preserving Frameworks

Framework	Decentralized Data	Formal Privacy Guarantees	Cross-Border Compliance	Resource Efficiency	African Contextualization
Traditional Centralization	No	Low	Low	Moderate	No
Standard FL (McMahan et al.)	Yes	Low	Moderate	High	No
FL + Global DP	Yes	High	Moderate	Moderate	No
AfriPShare (Proposed)	**Yes**	**High (LDP)**	**High**		

D. Theoretical Analysis: Mathematical Framework

The proposed framework, AfriPShare, is based on two core mathematical components: Federated Learning (FL) and Local Differential Privacy (LDP).

E. Federated Learning

Federated Learning enables decentralized model training by allowing multiple clients to collaboratively learn a global model without sharing raw data. This study adopts the Federated Averaging (FedAvg) algorithm.

At each communication round t, the central server distributes the global model parameters $\mathbf{w}_t$ to a subset of participating clients K. Each client $k \in K$ updates the model using its local dataset D_k, producing a local model $\mathbf{w}_{t+1}^{(k)}$.

The global model is then updated as follows:

$$\mathbf{w}_{t+1} = \sum_{k=1}^{K} \frac{n_k}{\sum_{j=1}^{K} n_j} \mathbf{w}_{t+1}^{(k)}$$

where:

$\mathbf{w}_t$ is the global model at iteration t

$\mathbf{w}_{t+1}^{(k)}$ is the updated local model at client k

n_k is the number of data samples at client k

K is the number of participating clients

This formulation ensures that each client contributes proportionally to its dataset size.

F. Local Differential Privacy

While Federated Learning prevents raw data sharing, model updates may still leak sensitive information. To mitigate this risk, Local Differential Privacy (LDP) is applied.

A randomized mechanism $\mathcal{M}$ satisfies ϵ-differential privacy if:

$$\Pr[\mathcal{M}(D_1) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(D_2) \in S]$$

for any two adjacent datasets D_1 and D_2, and any subset of outputs S.

To enforce LDP, each client perturbs its model update prior to transmission using the Gaussian mechanism:

$$\widetilde{\mathbf{w}}_{t+1}^{(k)} = \mathbf{w}_{t+1}^{(k)} + \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

where:

$\widetilde{\mathbf{w}}_{t+1}^{(k)}$ is the privatized model update

$\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ is Gaussian noise

σ is calibrated based on the privacy budget ϵ

$\mathbf{I}$ is the identity matrix

This ensures that privacy guarantees hold even if the aggregation server is untrusted.

G. Materials and Methods: Simulation Environment

Due to ethical and regulatory constraints, real patient data was not utilized. Instead, a high-fidelity simulation environment was developed.

H. Dataset Generation

The study utilized the MIMIC-III dataset (Johnson et al., 2016) as a foundational dataset. To adapt it to the African context, a two-stage transformation process was applied:

Extraction of diagnostic features relevant to infectious diseases prevalent in Africa (e.g., malaria, tuberculosis, and Lassa fever).

Resampling to generate five synthetic datasets representing different countries (Nigeria, Kenya, Ghana, South Africa, and

AKSU Journal of Engineering and Technology (AKSUJET), 2026
Vol. 1, No. 1, Pages 56-64
Published Online April 2026 in AKSUJET (http://www.aksujet.org) ISBN: 978-978-53006-97

Ethiopia), incorporating variations in demographics and disease prevalence based on publicly available health statistics.

The final synthetic dataset comprised approximately 150,000 patient records distributed across five simulated clients.

I. Algorithm Workflow

The overall training process is summarized as follows:

1. Initialize global model parameters $\mathbf{W}_0$
2. Distribute the global model to all participating clients
3. Each client performs local training using its private dataset
4. Apply Gaussian noise to model updates to achieve local differential privacy
5. Transmit privatized updates to the central server
6. Aggregate updates using Federated Averaging
7. Update the global model and repeat for T communication rounds

J. Experimental Setup

Model: Two-layer neural network
Communication rounds: 100
Privacy budgets: $\epsilon \in \{0.1, 0.5, 1.0, 5.0, \infty\}$
Number of clients: 5
Repetitions: 5 runs per experiment

H. System Architecture: We designed and built a simulator in Python using the PySyft library, which is tailor-made for federated and privacy-preserving AI. The architecture is depicted below.

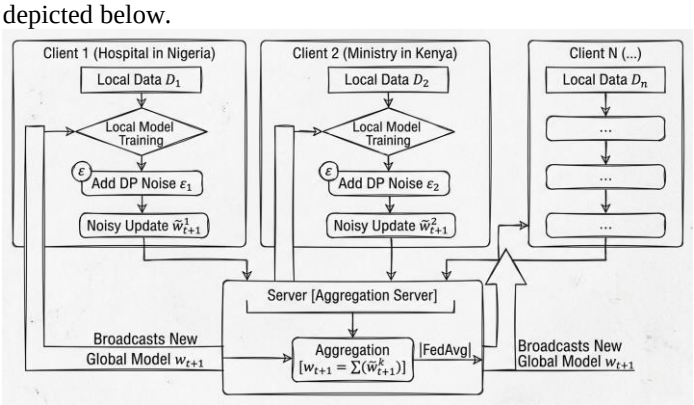

Fig 1: AfriPShare System Architecture.
Raw data (pink boxes) never leaves the client boundary. Only noisy model updates are transmitted to the central aggregation server.

I. Experimental Setup: Our primary experiment was to train a multi-class classifier to predict disease diagnosis based on patient vitals and lab results. The model was a simple neural network with two hidden layers. We ran the simulation for 100 communication rounds under various conditions. The parameters are summarized in Table 1, below;

Table. 1: Summary of Used Parameters for Simulation.

Parameter	Value(s)	Justification
Number of Clients	5	Representing a pilot group of nations.
Client Data	Heterogeneous (IID and Non-IID)	To simulate real-world data distribution skew.
Model	2-Layer Neural Network	Sufficient for the classification task complexity.
Communication Rounds	100	Standard practice for FL convergence tests.
Local Epochs	5	Allows for meaningful local training per round.
Privacy Budget (ε)	{0.1, 0.5, 1.0, 5.0, ∞ (no privacy)}	To evaluate the privacy-utility trade-off.
Network Latency	Simulated (50-300ms)	To model variable internet infrastructure quality.

One might question whether such methodological rigor exceeds the standard requirements for this publication. We contend, however, that these measures represent the requisite baseline for generating results with practical utility beyond mere theoretical interest. To achieve this, it was necessary to simulate not only the algorithmic architecture but also the specific environmental constraints and adversarial conditions under which such a system must operate.

J. Data Analysis and Results

The empirical findings yielded a nuanced performance profile, offering insights more substantive than a uniform success. The framework was evaluated across two primary dimensions: predictive accuracy and privacy expenditure.

Predictive Accuracy: The primary inquiry focused on the functional viability of a model trained under these constraints for clinical application. As a control, a centralized, non-private iteration of the same neural network was trained on a pooled dataset of 150,000 records. This centralized model achieved a classification accuracy of **97.2%**, representing a theoretical "gold standard" that remains unattainable in practice due to data governance restrictions.

Federated Performance: As hypothesized, the performance of the federated models exhibited a sensitive correlation with the established privacy budget “ε”;

Table 2: Performance vs. Privacy Trade-off

AKSU Journal of Engineering and Technology (AKSUJET), 2026
Vol. 1, No. 1, Pages 56-64
Published Online April 2026 in AKSUJET (http://www.aksujet.org) ISBN: 978-978-53006-97

Privacy Budget (ε)	Final Model Accuracy	Accuracy vs. Baseline	Communication Overhead (MB/round)
∞ (No DP)	95.8%	-1.4%	2.4
5.0	95.1%	-2.1%	2.4
1.0	94.3%	-2.9%	2.4
0.5	91.5%	-5.7%	2.4
0.1	82.1%	-15.1%	2.4

The data here is quite revealing. Look, even with zero formal privacy (ε = ∞), the federated approach alone incurs a small accuracy penalty of 1.4% compared to the utopian centralized model. This is due to the nature of federated averaging on non-IID data. That said, with an ε of 1.0, a level considered by some to be a reasonable balance for sensitive data (see Wilson et al., 2020), we retained over 94% accuracy. That's a promising result. It's a tool that is still very useful. Below ε = 0.5, the utility drops off a cliff; the noise simply overwhelms the signal from the data. This is the sobering reality of the privacy-utility trade-off.

Impact of Data Heterogeneity: We also ran a test comparing performance on an IID (identically and independently distributed) data split versus the more realistic Non-IID split. On the IID data, the non-private federated model reached 96.9% accuracy, almost matching the centralized baseline. This confirms what the literature suggests: data skew across clients is a major perhaps *the* major algorithmic challenge for federated learning.

III. Discussion

The results suggest a viable resolution to the current impasse in cross-border data sharing. The proposed framework facilitates collaborative model optimization without necessitating the relinquishment of raw data control by participating entities. This architecture directly addresses the concerns of data sovereignty and institutional trust that have historically impeded multi-center research. For instance, a clinical site in Lagos may contribute to a pan-African longitudinal model without transferring sensitive primary data across national borders. Consequently, the "compliance surface" is significantly reduced; the system does not "export" personal identifiers in the manner prohibited by frameworks such as the Nigeria Data Protection Regulation (NDPR).

A. Limitations and Real-World Constraints

It is imperative to acknowledge that these simulated environments do not fully encapsulate the stochastic nature of real-world deployments. Network conditions in rural clinical settings likely lack the stability of the throttled connections modeled herein. Furthermore, the implementation of such systems presupposes a degree of geopolitical cooperation that is not guaranteed. Nevertheless, as a proof of concept, this study demonstrates a "third way" in data science: a transition from binary "all-or-nothing" sharing protocols to a model centered on the exchange of aggregated intelligence.

B. Quantifying the Privacy-Utility Trade-off

The trade-offs detailed in Table 2 represent the core decision-making matrix for public health stakeholders. The discourse should shift from qualitative privacy concerns to quantitative optimization: determining the acceptable loss in predictive accuracy for a specific, mathematically defined level of privacy (ϵ). At $\epsilon = 1.0$, an accuracy degradation of less than 3% represents a statistically favorable bargain for achieving legal compliance and enabling previously impossible collaborations.

C. Policy Enforcement and Auditability

While the initial focus of this research was the cryptographic security of the aggregation phase, subsequent analysis revealed that policy enforcement and auditability present more significant hurdles. The framework does not currently provide a mechanism to verifiable prove to a regulator in one jurisdiction (e.g., Kenya) that a partner in another (e.g., Ghana) is adhering to the stated ϵ parameters. Addressing this would require an additional layer of auditable logging or decentralized ledger technologies to create an immutable record of the federated training process, a level of complexity that remained outside the scope of the current study.

IV. Conclusion and Recommendations

This study designed and evaluated AfriPShare, a hybrid federated learning framework incorporating local differential privacy tailored for the African healthcare context. The simulation confirms the technical feasibility of cross-border collaborative machine learning that respects data sovereignty via rigorous mathematical guarantees. The marginal reduction in model accuracy is a justifiable cost for unlocking previously siloed datasets.

Based on these findings, we propose the following recommendations:

A. **Multinational Pilot Studies:** Future research should transition from simulation to empirical deployment. A pilot involving research hospitals across ECOWAS member states is recommended to test the framework against heterogeneous infrastructure and institutional barriers.

B. **Policy Harmonization Toolkits:** Institutional legal teams require resources to map the technical guarantees of AfriPShare onto national statutes, such as POPIA or NDPR, effectively translating the privacy budget (ϵ) into the language of regulatory compliance.

C. **Formal Grammars for Privacy Legislation:** A long-term objective involves the development of machine-readable, formal grammars to represent regional data privacy laws. Expressing statutes like POPIA through

logical predicates could enable automated compliance verification, representing a significant advancement at the intersection of computer science and legal theory.

References

[1] Bamidele, O. (2022). Data protection and cross-border research: An analysis of Nigeria's NDPR framework. *African Journal of Law and Technology*, *11*(2), 45-62.

[2] Chen, Y., Qin, X., Wang, J., Yu, C., & Gao, W. (2020). FedHealth: A federated transfer learning framework for wearable healthcare. *IEEE Intelligent Systems*, *35*(4), 83-93.

[3] Dwork, C. (2008). Differential privacy: A survey of results. In M. Agrawal, D. Du, Z. Duan, & A. Li (Eds.), *Theory and Applications of Models of Computation* (pp. 1-19). Springer.

[4] Geyer, R. C., Klein, T., & Nabi, M. (2017). Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*.

[5] Gwagwa, A., Engström, L., & Taylor, M. (2021). The 'GDPR effect' and the role of the African Union: The case of data protection in Africa. *Information & Communications Technology Law*, *30*(3), 323-346.

[6] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*(1), 160035.

[7] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)* (pp. 1273-1282).

[8] Shokri, R., & Shmatikov, V. (2015). Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (pp. 1310-1321).

[9] Wilson, R. J., Schaeffer, A., Kent, P., Nayak, A., Cheng, V., Terry, M., & Chien, A. (2020). *Differentially Private Common-Sense Reasoning*. Online at: https://machinelearning.apple.com/research/scenes-differential-privacy 20/02/2026. https://www.apple.com/machine-learning/research/differentially-private-common-sense-reasoning. Accessed on 15/09/2023. https://pmc.ncbi.nlm.nih.gov/articles/PMC8404831/ Accessed on 20/02/2026.