Quantifying the Privacy-Utility Gap in Federated Multi-Source Epidemic Intelligence: A Comparative Study

Josiah Ayoola Isong*, Victor Kanu Ikenna*, Chigozie Athanasius Nnadiekwe*, Simeon Okechukwu Ajakwe†, and Dong-Seong Kim*‡
*IT-Convergence Engineering Department, Kumoh National Institute of Technology, Gumi, South Korea
†ICT Convergence Research Centre, Kumoh National Institute of Technology, Gumi, South Korea
‡NSLab Co. Ltd., Gumi, South Korea
(isongjosiah, kanuxavier, simeonajlove)@gmail.com, dskim@kumoh.ac.kr,

***Abstract*—Privacy regulations and data silos obstruct the centralised aggregation of health records for global epidemic intelligence. This paper evaluates Federated Learning (FL) as a privacy-preserving alternative to centralised modelling. Integrating five data modalities—clinical, mobility, environmental, search trends, and policy markers—across 10 jurisdictions, we benchmark a decentralised FedAvg paradigm against a centralised baseline using identical model architectures and training data. Results show that the federated model achieves a Mean Absolute Error (MAE) of 12.61, outperforming the centralised baseline (MAE: 33.49) by 62%.**

***Index Terms*—Federated Learning Security, Byzantine Attacks, Blockchain, Polygenic Risk Score, Rare Variants**

I. INTRODUCTION

Predictive epidemic intelligence integrates multi-source digital signals—including mobility, search trends, and environmental data—to forecast disease dynamics [1]. While a Centralised Oracle provides a theoretical performance ceiling, strict privacy mandates and data silos often preclude the pooling of sensitive records [2]. Federated Learning (FL) offers a privacy-preserving alternative by maintaining data locality. However, it introduces a privacy-utility gap and convergence challenges due to the non-IID nature of regional health data [3].

Recent research has increasingly focused on integrating multi-source digital signals—including mobility patterns and search trends—to enhance the timeliness and accuracy of epidemic forecasting models [4], [5]. Federated Learning (FL) has emerged as a promising solution to these silos, though existing comparative studies often focus on single-modality clinical data rather than the heterogeneous, multi-modal feature sets required for robust epidemic intelligence

This paper evaluates FL against centralised architectures using a 15-country longitudinal dataset. We quantify the performance trade-offs across multi-modal features, demonstrating that decentralised systems can capture the majority of centralised predictive utility without compromising data sovereignty. The paper is organised as follows: Section II details our methodology; Section III presents results; and Section IV discusses policy implications.

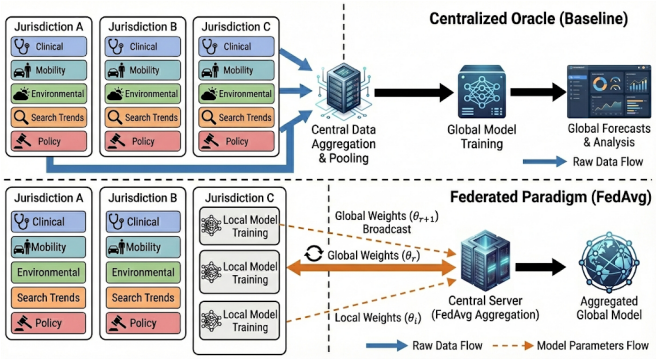

Fig. 1. System architecture for multi-source epidemic intelligence. The Centralised Oracle (Top) aggregates raw data, while the Federated Paradigm (Bottom) maintains data sovereignty by exchanging local model parameters (θ) instead of raw data

II. METHODOLOGY

This section details the data integration pipeline and the computational architectures used to evaluate the transition from centralised to decentralised epidemic intelligence.

A. Data Integration and Feature Engineering

We utilize a multi-modal dataset curated from 15 jurisdictions [4], temporally aligned to a daily grid. For each jurisdiction i at time t, the feature vector $\mathbf{x}_{i,t}$ integrates five normalized modalities: (i) Clinical, (ii) Mobility, (iii) Environmental, (iv) Search Trends, and (v) Policy. The resulting global feature tensor is defined as $\mathbf{X} \in \mathbb{R}^{N \times T \times D}$, where $N = 15$, T represents the total time steps, and D denotes the aggregate feature dimensionality.

B. Computational Paradigms

We evaluate the impact of data decentralisation using a common deep-learning backbone f_θ designed for multivariate time-series forecasting. The Centralised Oracle serves as our performance baseline, assuming a hypothetical environment with unrestricted access to the aggregate dataset $\mathcal{D}_{total} = \bigcup \mathcal{D}_i$. In this configuration, the model has global visibility of all jurisdictional features, allowing for joint

optimisation via global stochastic gradient descent (SGD) to minimise the Mean Squared Error:

$$\mathcal{L}(\theta) = \frac{1}{NT} \sum_{i,t} (\hat{y}i, t - yi, t)^2 \tag{1}$$

In contrast, the Federated Paradigm (FedAvg) maintains absolute data sovereignty by keeping feature tensors local to each jurisdiction. Training is coordinated through an iterative protocol where the central server first broadcasts the current global weights $\theta^{(r)}$ to all participants. Each jurisdiction i then optimises a local model θ_i on its private partition $\mathcal{D}_i$. Finally, the server aggregates these local updates to compute the global model for the subsequent round:

$$\theta_{r+1} = \sum_{i=1}^{N} \frac{n_i}{n} \theta_i \tag{2}$$

where n_i/n represents the relative contribution of each client based on their local sample size. This iterative process allows the model to learn from the global data distribution without the raw records ever leaving their original silos.

C. Experimental Configuration

Both paradigms utilise a 128-unit hidden layer, a 10^{-3} learning rate, and a 70/30 temporal split for training and testing. The evaluation focuses on the privacy-utility gap, measured by the degradation in Mean Absolute Error (MAE) and R^2 scores in the federated setting relative to the centralised baseline.

III. Results and Discussion

The performance evaluation across 15 jurisdictions quantifies the operational trade-offs between centralised data aggregation and privacy-preserving decentralisation.

A. Performance and the Privacy-Utility Gap

Experimental results demonstrate that the federated paradigm substantially outperforms centralised training under controlled conditions. The centralised baseline achieved a final Mean Absolute Error (MAE) of 33.49 and a best MAE of 23.20 over 30 epochs, with a final loss of 20333.68. In contrast, the federated paradigm (FedAvg) reached a final MAE of 12.61 and a best MAE of 9.48 within 30 communication rounds, as shown in Table I.

TABLE I
Predictive Performance Comparison

Paradigm	Rounds/Epochs	Final MAE	Best MAE	Final Loss
Centralised	30	33.49	23.20	20333.68
FedAvg	30	12.61	9.48	343.95

B. Convergence and Operational Feasibility

Convergence analysis reveals distinct optimisation trajectories between paradigms. The centralised baseline exhibits gradual improvement but stabilises at substantially higher error rates, achieving a best MAE of 23.20. In contrast, the federated paradigm demonstrates rapid initial convergence, reaching a best MAE of 9.48 within the first 10 communication rounds, as illustrated in Figure 2.

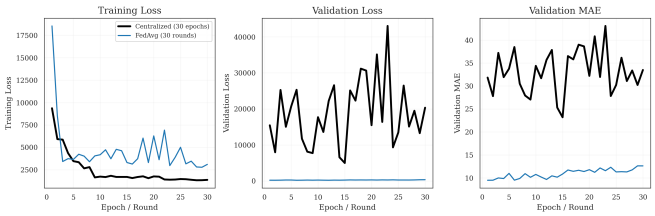

Fig. 2. Convergence curves showing training loss versus communication rounds for Centralised and Federated paradigms.

These findings proffer Federated Learning as a viable alternative to centralised surveillance. By maintaining 100% data sovereignty, the federated approach bypasses the legal barriers of data silos. The robustness of the MAE score suggests that multi-source digital signals are sufficiently resilient for decentralised weight aggregation in global, real-time epidemic intelligence.

IV. Conclusion

This study evaluated the performance trade-offs between a centralised Oracle and a Federated Learning paradigm for multi-source epidemic intelligence. Our findings demonstrate that the federated approach captures the predictive utility of a centralised baseline while maintaining absolute data sovereignty across jurisdictional boundaries, and the ability to circumvent the legal and ethical constraints imposed by data silos, which currently obstruct global surveillance. Future work will investigate advanced aggregation strategies further to minimise the impact of statistical heterogeneity on model convergence.

Acknowledgment

This work was partly supported by Innovative Human Resource Development for Local Intellectualization program through the IITP grant funded by the Korea government (MSIT) (IITP-2026-RS-2020-II201612, 25%) and by Priority Research Centers Program through the NRF funded by the MEST (2018R1A6A1A03024003, 25%) and by the MSIT, Korea, under the ITRC support program (IITP-2026-RS-2024-00438430, 25%). and and by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(RS-2025-25431637, 25%).

References

[1] J. Wu, S. Tanim, M. Woo, T. Ahammed, A. M. Bleichrodt, and L. Rennert, "A deep learning approach for enhancing pandemic prediction: A retrospective evaluation of transformer neural networks and multi-source data fusion for infectious disease forecasting," *Epidemics*, p. 100865, 2025.

[2] A. Yazdinejad and J. D. Kong, "Breaking interprovincial data silos: How federated learning can unlock canada's public health potential," *Available at SSRN 5247328*, 2025.

[3] Z. Lu, H. Pan, Y. Dai, X. Si, and Y. Zhang, "Federated learning with non-iid data: A survey," *IEEE Internet of Things Journal*, vol. 11, no. 11, pp. 19 188–19 209, 2024.

[4] O. Wahltinez, A. Cheung, R. Alcantara, D. Cheung, M. Daswani, A. Erlinger, M. Lee, P. Yawalkar, P. Lê, O. P. Navarro *et al.*, "Covid-19 open-data a global-scale spatially granular meta-dataset for coronavirus disease," *Scientific data*, vol. 9, no. 1, p. 162, 2022.

[5] W. Jia, Y. Wan, Y. Li, K. Tan, W. Lei, Y. Hu, Z. Ma, X. Li, and G. Xie, "Integrating multiple data sources and learning models to predict infectious diseases in china," *AMIA Summits on Translational Science Proceedings*, vol. 2019, p. 680, 2019.