

A Tanzanian Maternal and Neonatal Healthcare Dataset Compliant with Federated Learning: Privacy, Fairness, and Compliance

Justin A. Mwakatobe¹✉, Kisangiri F. Michael²✉, Devotha Nyambo²✉

¹*Mbeya University of Science and Technology (MUST), Tanzania*

²*Nelson Mandela African Institution of Science and Technology (NMAIST), Tanzania*

Abstract

In the digital era, ensuring privacy, fairness, and security of patient data is of paramount importance in healthcare data sharing and research activities. Tanzania's fifth Health Sector Strategic Plan July 2021–June 2026 (HSSP V) underlines the necessity for the government to create a legislative framework to safeguard patient confidentiality, privacy, and data security. The existing literature addresses multifaceted challenges from classical Machine Learning (ML) and Deep Learning (DP) techniques in data sharing, security, privacy, regulatory compliance, data imbalance, overfitting, interpretability, fairness, and biases. This led to the gain in traction of the Federated Learning (FL) model for data sharing, fairness, privacy-preserving, and security between healthcare providers. Therefore, this paper presents a specialised maternal and neonatal dataset that can be used by researchers to develop a privacy-preserving FL model with fairness for healthcare data. The real dataset in Comma-Separated Value (CSV) format carries 20,333 patients' data from three hospitals (Zonal, regional, and district) in Tanzania. Cleaned and preprocessed datasets contain details such as encrypted ID, encrypted name, demographic, sponsor, vitals, diagnoses, procedures, medications, and outcome. The developed model will preserve data privacy with fairness by keeping data at its source and aggregating model updates instead of raw data. In conclusion, the dataset aims to support advancements in privacy and ethics of healthcare data analytics, application of FL in low-resource infrastructure like in Tanzania, where there is no need for new resources (trains on local infrastructure), fairness and bias mitigation, policy and regulatory impact, local contextualisation, and educational resources for researchers who wish to pursue FL technique. Furthermore, raw maternal and neonatal datasets will be encrypted with differential privacy, cryptographic protocols such as secure aggregation, and Secure Multi-Party Computation (SMPC) security techniques to ensure data privacy and fairness.

Keywords

Maternal, Neonatal, Privacy, Regulatory compliance, Data sharing, Health data

Current status of the research is: Work-in-progress

Potential collaboration with Authors

Continuing research in FL model, collaboration in research in the field of emerging technologies such as AI, privacy, data protection and cybersecurity.