# A Hybrid Machine Learning Model for TB/HIV Progression Prediction Using Resource-Constrained Electronic Health Record (EHR) Data in Zambia

Joe Phiri
*School of Computing, Technology and Applied Science*
ZCAS University
Lusaka, Zambia
phirijoe26@hotmail.com

Aaron Zimba
*School of Computing, Technology and Applied Science*
ZCAS University
Lusaka, Zambia
aaron.zimba@zcasu.edu.zm

Chiyaba Njovu
*School of Computing, Technology and Applied Science*
ZCAS University
Lusaka, Zambia
chiyaba.njovu@zcasu.edu.zm

***Abstract*—Tuberculosis (TB) remains a leading cause of mortality among people living with HIV (PLHIV) in Zambia, posing a major challenge to an already strained health system. Zambia's national electronic health record (EHR) systems, contains valuable longitudinal data that could support predictive tools for early TB intervention. However, issues such as data sparsity, limited analytical capacity, and poor interpretability of machine learning (ML) models have slowed clinical adoption. This study proposes a hybrid ML framework that integrates Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Long Short-Term Memory (LSTM) networks, enhanced with SHapley Additive exPlanations (SHAP) for transparency. The Design Science Research (DSR) methodology guides iterative model development, evaluation, and deployment. Preprocessing employs Multiple Imputation by Chained Equations (MICE) for missing data, Min-Max normalization for scaling, and SMOTE for class balancing. Data mapping from EHRs has been completed, and a preprocessing pipeline is under development. Initial training and validation are being conducted using synthetic EHR datasets, with performance measured by F1 Score and Area Under the Precision-Recall Curve (AUC-PR). Prototype models will be tested in simulated clinical workflows to assess feasibility and responsiveness.**

**The research contributes a novel ensemble-based approach that fuses static and temporal variables with explainable AI, supporting early HIV/TB progression prediction and clinician trust in low-resource settings. Future work will focus on real-world validation, stakeholder feedback, and integration into national digital health systems.**

***Keywords—Tuberculosis, HIV, Machine Learning, EHR, SHAP, LSTM, Analytics***

## I. INTRODUCTION

Zambia faces a dual burden of HIV and tuberculosis (TB), with TB contributing significantly to AIDS-related mortality [1]. Electronic health record (EHR) systems, particularly through the SmartCare platform, present a valuable resource for developing predictive analytics tools. However, challenges such as data sparsity, infrastructure limitations, and clinician distrust of black-box models hamper the adoption of artificial intelligence (AI) in healthcare. This paper presents a hybrid machine learning (ML) framework designed to predict TB/HIV progression in PLHIV using real-world EHR data from Zambia. The model architecture combines Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Long Short-Term Memory (LSTM) networks, supplemented by SHapley Additive exPlanations (SHAP) for interpretability. The approach is tailored to the constraints and requirements of low and middle income countries (LMIC) settings.

## II. LITERATURE REVIEW

Geldsetzer et al. [2] explored the potential of ML models for TB/HIV interventions in LMICs, identifying key limitations in data quality and model transparency. Rajkomar et al. [3] highlighted the risk of algorithmic bias and the importance of fairness and reproducibility in health-focused AI applications. Locally, Zambia's EHR systems holds longitudinal patient records but suffer from data inconsistency and limited analytical usage [4]. While XGBoost has shown success in structured health data prediction [5], LSTM models excel in capturing temporal trends, such as CD4 count dynamics. Lundberg and Lee [6] proposed SHAP to address model explainability—a critical requirement in healthcare adoption. Earlier approaches such as LIME provided instance-level explanations for model predictions [7], but SHAP offers more consistent and theoretically grounded interpretability. Recent reviews have also highlighted the growing role of large language models, foundation models, and digital twins in clinical data analysis [8], underscoring the importance of explainability and applicability in AI.

The proposed research addresses a gap by combining static and temporal ML models in a stackable ensemble while maintaining clinical transparency.

## III. METHODOLOGY

### A. Research Design

The research follows a Design Science Research (DSR) methodology to guide the iterative development and validation of the model.

### B. Conceptual Framework

The predictive system comprises interconnected modules designed to address the full pipeline from data acquisition to clinical deployment. It begins with data collection and harmonization, where diverse patient data from Zambia's EHR systems are aggregated and standardized. This is followed by a feature engineering and transformation phase, in which raw inputs are refined into analytically useful variables. The hybrid machine learning (ML) model development module integrates various algorithms, including tree-based and temporal models, to improve predictive performance. Subsequently, a model interpretation and trust-building component leverages SHAP techniques to generate transparent explanations for predictions, facilitating clinical acceptance as illustrated in Fig. 1. Finally, the system culminates in a deployment pipeline optimized for integration within resource-constrained healthcare settings.

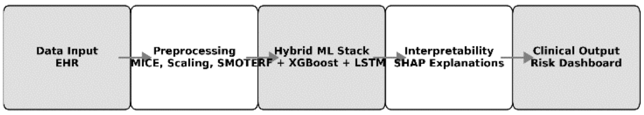

Fig. 1. Conceptual Framework of the Model

### C. Data Preparation

The study will utilize anonymized datasets extracted from Zambia's national EHR systems. These datasets encompass a range of relevant patient-level variables essential for modeling TB/HIV disease progression. The demographic attributes include patient age, gender, and geographical location. Clinical records comprise ART initiation dates, HIV staging information, and documented TB history. Laboratory data such as CD4 cell counts, viral load measurements, and hemoglobin levels are also included. Importantly, the dataset features longitudinal sequences of CD4 values over time, which are critical for modeling patient trajectories using temporal deep learning approaches. All data handling procedures will adhere to national data governance requirements, including compliance with the Zambia Data Protection Act [9].

Missing values are handled using Multiple Imputation by Chained Equations (MICE), class imbalance with SMOTE, and scaling with Min-Max normalization. CD4 slopes and trends are derived for LSTM input.

### D. Model Architecture

The model combines:

RF: interpretable tree-based learner for static features

XGBoost: optimized boosting on structured data

LSTM: time-series analysis of longitudinal variables

A meta-learner integrates the predictions. SHAP is used to provide interpretable insights for both individual and cohort-level predictions as illustrated in Fig. 2.

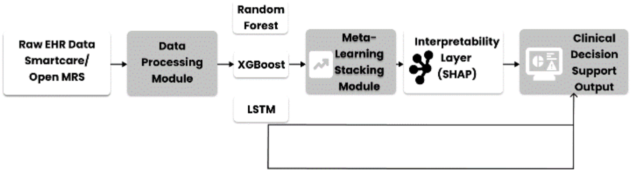

Fig. 2. Hybrid Model Architecture

### E. Evaluation Metrics

To assess the performance of the hybrid predictive model in the context of imbalanced health data, particularly where the accurate identification of TB progression is vital, two primary evaluation metrics are adopted: the F1 Score and the Area Under the Precision-Recall Curve (AUC-PR). The F1 Score serves as the harmonic mean of precision and recall, making it an ideal choice in situations with uneven class distributions, such as rare disease detection. It ensures that both false positives and false negatives are taken into account when evaluating the model. On the other hand, AUC-PR provides a holistic view of how well the model maintains high precision and recall across various threshold levels. Unlike ROC-AUC, AUC-PR is more informative in contexts like TB/HIV comorbidity prediction, where the positive cases are significantly fewer than the negatives. These metrics collectively ensure the robustness and reliability of the predictive system when deployed in real-world clinical environments. Table I summarizes the evaluation metrics used in this study and their respective purposes.

TABLE I. PERFORMANCE METRICS AND THEIR PURPOSE

| Metric | Formula | Purpose |
|---|---|---|
| F1 Score | $F_1 = \backslash frac\{2 \cdot P \cdot R\}\{P + R\}$ | Balance between precision and recall |
| AUC-PR | Integrated PR curve | Robust performance in imbalanced datasets |

[a] P = Precision, R = Recall; AUC-PR = Area Under the Precision-Recall Curve

### F. Algorithmic Workflow

Preprocessing Pipeline

Input: Raw EHR Data

Apply MICE for missing fields

Normalize numerical variables

Generate CD4 time windows

Use SMOTE for class rebalancing

Output: Feature matrix

Model Training

A. Train RF and XGBoost on structured data

B. Train LSTM on CD4 time-series

C. Feed base outputs into meta-classifier

D. Apply SHAP for interpretability

E. Evaluate and export final model

## IV. Discussion and Conclusion

The proposed hybrid model framework has the potential to overcome several technical and contextual limitations observed in the application of machine learning (ML) to electronic health records (EHRs) in low- and middle-income countries (LMICs). SHapley Additive exPlanations (SHAP) enhance transparency and may increase clinical adoption by showing why a patient is at high risk. Temporal modeling using long short-term memory (LSTM) networks enables capturing disease progression dynamics that static models often miss.

Implementation challenges include harmonizing variable definitions across provinces, managing missing time-stamped entries, and building lightweight deployment pipelines compatible with Zambian public health infrastructure. Further development will focus on improving LSTM stability, enhancing SMOTE realism, and validating SHAP usability in clinical contexts. As emphasized by Challen et al. [10], addressing bias and ensuring clinical safety remain critical when deploying AI in real-world healthcare settings.

Overall, this work demonstrates the feasibility of combining ensemble learning, temporal modeling, and explainable AI to strengthen predictive analytics in resource-constrained environments. By integrating interpretability with clinical utility, the proposed framework offers a pathway toward actionable decision support in Zambia's national EHR ecosystem. Future efforts will prioritize real-world validation, stakeholder engagement, and adaptation of the framework to other chronic disease domains, contributing to scalable and sustainable digital health strategies in LMIC settings.

## Acknowledgment

This work is supported by the ZCAS University as part of a doctoral research program.

## References

[1] World Health Organization, Global Tuberculosis Report. Geneva, Switzerland: WHO, 2023.

[2] J. Geldsetzer, K. Reinmuth, F. Ouma, H. Barnighausen, and T. Vollmer, “Impact of health information technology in low- and middle-income countries,” Int. J. Med. Inform., vol. 152, pp. 104–116, 2021.

[3] A. Rajkomar, J. Dean, and I. Kohane, “Scalable deep learning with electronic health records,” NPJ Digit. Med., vol. 1, no. 18, pp. 1–10, 2018.

[4] G. Churchyard, A. Mametja, L. Mvusi, F. Ndjeka, Y. Hesseling, R. Reid, and M. Babatunde, “Epidemiology of HIV and tuberculosis coinfection in South Africa,” Clin. Infect. Dis., vol. 59, no. 3, pp. 123–130, 2014.

[5] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794

[6] S. Lundberg and S. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 2017, pp. 4765–4774

[7] M. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?: Explaining the predictions of any classifier,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 1135–1144

[8] Y. Fuse, “Artificial intelligence in clinical data analysis: A review of large language models, foundation models, digital twins, and allergy applications,” J. Clin. Data Sci., vol. 13, 2025, Art. no. 100080. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1323893025000802

[9] Government of Zambia, Data Protection Act, 2021.

[10] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, and J. Tsaneva-Atanasova, “Artificial intelligence, bias and clinical safety,” BMJ Qual. Saf., vol. 28, no. 3, pp. 231–237, 201