Journal of Health Informatics in Africa

Deployment-Oriented Evaluation of an Explainable Tabular Transformer for Tuberculosis Outcome Prediction Using Synthetic Low-Resource Healthcare Data

Shumirai S Sibanda [a], Belinda Ndlovu [b,*]

[a] Department of Informatics and Analytics, National University of Science and Technology, Zimbabwe
[b] School of Computing, University of South Africa, Pretoria, South Africa

Background and Purpose: Healthcare systems use EHRs and predictive analytics in clinical decision-making; however, many current healthcare-based AI applications prioritize predictive accuracy over explainability, reliability, subgroup consistency, and deployment feasibility in limited healthcare settings. This study assesses an explainable transformer architecture for predicting TB treatment outcomes by using a synthetic healthcare dataset to conduct replicable experiments without compromising patient privacy.
Methods: The proposed framework combines tabular machine learning with transformers, SHAP-based explainability, calibration tests, subgroup consistency tests, and evaluation for implementation into one AI pipeline in healthcare. Comparative experiments have been performed against Random Forest, XGBoost, and multilayer perceptron baselines on Accuracy, Precision, Recall, F1-score, and AUC-ROC measures. The transformer achieved the highest performance across all experiments, with Accuracy 82.1%, F1-score 0.741, and AUC-ROC 0.872, while retaining decent calibration and subgroup consistency.
Results: Explainability analysis showed that HIV status, drug resistance, treatment adherence, nutrition, and healthcare access measures were significant contributors to the model's predictive performance, demonstrating that the model identified clinically meaningful interactions in the synthetic environment. These results should be viewed solely as an indication of technical feasibility, as all experiments were conducted in a controlled, synthetic setting.
Conclusions: Beyond predictive performance, the study contributes a deployment-oriented healthcare AI evaluation framework that integrates explainability, calibration, subgroup inspection, and operational feasibility assessment. This paper highlights the need for healthcare AI models that not only have predictive ability but are also trustworthy and feasible.

Keywords: Tuberculosis (TB), Treatment Outcomes, Tabular Transformer Models, Explainable Artificial Intelligence, SHAP, Low Resource Healthcare, Synthetic Data, Clinical Decision Support, Deployment-Oriented

1 Introduction

Tuberculosis (TB) continues to be among the major causes of death due to infections worldwide, particularly in developing countries, where there is a lack of health facilities, access to information, and clinical decision support systems [1]. To increase treatment success in such settings, early detection of patients with a high risk of failure, recurrence, or death is crucial. However, predictive modelling in TB management is limited by data fragmentation, insufficient sample sizes, and poor use of complex analytical methods.[2].

Machine learning approaches have increasingly been explored for tuberculosis outcome prediction because of their ability to model complex nonlinear relationships across heterogeneous clinical variables [3]. Classical approaches, including Logistic Regression, Support Vector Machines, Random Forests, and

*Corresponding author address: University of South Africa, Preller St, Muckleneuk, Pretoria, 0002, South Africa. Email: 69970777@mylife.unisa.ac.za
 J Health Inform Afr. 2026;13(1):228-245. DOI: 10.12856/JHIA-2026-v13-i1-700

gradient boosting techniques, have demonstrated moderate-to-strong predictive capability across structured healthcare datasets [4]. More recent deep learning approaches, including recurrent and transformer-based architectures, further introduced contextual representation learning for longitudinal and heterogeneous healthcare information [5],[6],[7]. Nevertheless, many existing models remain decentralized by limited interpretability, weak deployment feasibility and insufficient evaluation of calibration, subgroup consistency and operational trustworthiness within resource-constrained healthcare environments, a gap that broader work on the ethics and governance of trustworthy medical artificial intelligence has similarly identified as a barrier to safe clinical translation [8], [9]

Tuberculosis treatment outcome prediction largely depends on heterogeneous tabular clinical variables, including demographic information, comorbidities, microbiological findings, treatment history and social determinants of health [5]. Unlike image-based or sequential biomedical data, these structured variables often exhibit complex nonlinear interactions that may not be fully captured by conventional linear or tree-based models [10], [11]. Tabular transformer architectures can model contextual feature interactions dynamically via self-attention mechanisms, making them particularly relevant for structured clinical prediction tasks involving heterogeneous patient factors [12]. This trend mirrors the broader uptake of transformer architectures across biomedicine, where self-attention mechanisms have increasingly been applied beyond sequential data to a range of structured and multimodal clinical prediction tasks [13].

Further research on models that simultaneously achieve high predictive performance, interpretability, and efficient computation is another gap in the existing literature [14]. In such clinical settings, where decision-making is critical, the models used need to do more than perform well; they must also be interpretable and work within limited computational resources [15]. Recent transformer models have demonstrated strong capability for modelling contextual interactions among features using self-attention, especially for heterogeneous tabular predictions [16],[17],[6]. However, note that their feasibility, interpretability, and deployability in resource-constrained settings have not yet been explored.

Another challenge is the insufficient amount of high-quality real-world data on TB, especially in low-resource settings. This has been a limiting factor in developing advanced prediction algorithms[18]. Synthetic data generation can be used to explore methodologies, as it allows one to test models under specific conditions. Synthetic data generation raises concerns about validity, generalisability and the risk of biasing a model with prior knowledge [5].

Despite growing research on the use of machine learning to predict the treatment outcome of TB, three main deficiencies remain inadequately explored. Firstly, several studies rely on datasets collected from the real world, which are inaccessible, incomplete, or contextualized, making it hard to conduct further research and experiments [19]. Secondly, many high-performing models lack sufficient levels of explainability for review purposes, especially in resource-constrained environments where transparency is crucial [15]. Lastly, only a few studies consider predictive modelling, interpretability, and subgroup consistency analysis within a single research approach, especially in resource-limited settings [5]. This study addresses these gaps by developing and evaluating an explainable tabular transformer that uses synthetic TB data as a methodological testbed.

For prediction performance alone to warrant the implementation of an AI system in resource-limited medical facilities, several other aspects, including interpretability, efficiency, feasibility, and the capability to deliver clinical insights, are equally essential. This study therefore investigates whether lightweight transformer-based architectures can support balanced healthcare AI evaluation by integrating predictive performance, interpretability, calibration reliability, subgroup consistency inspection, and deployment feasibility under simulated low-resource tuberculosis conditions.

This study makes four critical contributions to the AI for healthcare literature. First, it highlights the potential utility of lightweight transformer models for explainable TB outcome prediction in low-resource healthcare settings using structured tabular clinical data. Second, the paper advances healthcare AI evaluation for application purposes by integrating predictive modelling, calibration testing, subgroup consistency analysis, interpretability, and computational feasibility into an integrative methodology. Third, this study contributes to ethical AI research on healthcare by distinguishing between methodological feasibility and clinical validity in synthetic data experiments, thereby encouraging more conservative reporting of AI performance outcomes in healthcare. Fourth, the study provides a more holistic perspective on healthcare AI evaluation where predictive power, interpretability, calibration, and feasibility are seen as interrelated design elements. Overall, these contributions position deployment-oriented evaluation, rather

 J Health Inform Afr. 2026;13(1):228-245. DOI: 10.12856/JHIA-2026-v13-i1-700

than predictive accuracy alone, as a necessary criterion for healthcare AI research conducted under low-resource constraints, a reframing intended to be of direct relevance to future work in this field.

The study is guided by three research questions concerning predictive competitiveness, explainability-oriented healthcare AI evaluation, and the feasibility of deployment under simulated, infrastructure-limited tuberculosis conditions.

RQ1: Are transformer architecture systems competitive with classical machine learning models when predicting TB treatment outcomes in tabular, synthetic healthcare data?

RQ2: To what degree do methods such as SHAP explainability analysis, calibration checking, and subgroup screening assist in building an interpretable healthcare AI system under simulated healthcare conditions with limited TB?

RQ3: Can a transformer-based solution for TB treatment outcomes prediction be implemented within resource-constrained healthcare infrastructure without specialized hardware needs?

The remainder of this paper is structured as follows. Section 2 outlines the complete methodology employed. Section 3 presents the results of the model's performance evaluation, including details on accuracy, feature correlation, training curve, calibration, subgroup performance analysis, ablation analysis, and SHAP explanations. Section 4 analyses the implications of our findings for clinical practice, compares them with classical machine learning methods, identifies the limitations of our current study, and offers suggestions for future research. Section 5 draws conclusions from the results and provides recommendations for incorporating interpretable transformer-based systems into clinical TB treatment processes.

2 Materials and Methods

2.1 Methodological Overview

Figure 1 presents the end-to-end methodological pipeline, spanning synthetic data generation, model training, explainability analysis, evaluation, and deployment, which structures the remainder of this section.

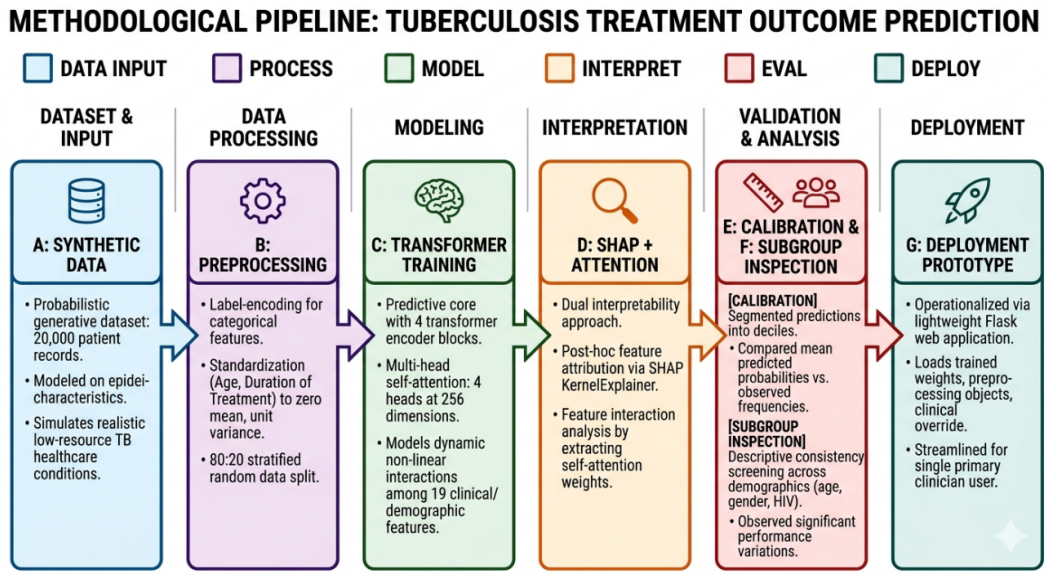

Figure 1: Methodological Pipeline Diagram

2.2 Study Design and Synthetic Data Generation

This study uses a controlled experiment to assess the feasibility of lightweight, explainable transformer models for predicting TB treatment outcomes in decentralized healthcare settings. The research methodology was designed to test predictive accuracy, interpretability, calibration, subgroup consistency,

 J Health Inform Afr. 2026;13(1):228-245. DOI: 10.12856/JHIA-2026-v13-i1-700

and deployability in a synthetic data evaluation setting, supporting reproducible healthcare AI experiments without compromising patients' data confidentiality.

Synthetic dataset was obtained from Kaggle (https://www.kaggle.com/datasets/nishatvasker/tb-patient-dataset-synthetic-bangladesh). These synthetic data were generated using a probability-based generation approach. The distribution of features was determined based on epidemiological characteristics from TB cohort studies conducted in Bangladesh and neighbouring countries. The dataset includes 20,000 patient records with 23 features, including demographic data, clinical variables, comorbidity information, treatment information, and social determinants. Because the dataset was generated through probabilistic epidemiological simulation rather than direct patient collection, the resulting feature distributions represent plausibility-oriented synthetic approximations rather than validated clinical population statistics. A correlation matrix was employed to ensure realistic relationships among comorbidities, demographic data, and patient outcomes. The simulation approach was evaluated by comparing the distributions of synthetic and real-world data using the Kolmogorov-Smirnov test and the chi-square goodness-of-fit test, where appropriate.

2.2.1 Synthetic Data Design

The synthetic dataset was constructed to support controlled methodological evaluation of explainable healthcare AI architectures under infrastructure-limited tuberculosis conditions rather than to replicate authentic clinical populations or enable direct clinical inference. Disease relationships, demographic distributions, treatment-risk interactions, and outcome patterns were intentionally structured using clinically informed assumptions to preserve plausible healthcare behaviour during model development. Consequently, all predictive, explainability, subgroup-consistency, and intervention-oriented scenario simulation findings should be interpreted as an evaluation of architectural and methodological behaviour under controlled synthetic conditions rather than evidence of validated clinical performance or epidemiological discovery.

2.2.2 Outcome Generation and Leakage Control.

To avoid target leakage, the treatment outcome was assigned probabilistically rather than deterministically. The probabilities were affected by risk factors such as HIV status, diabetes, malnutrition, medication resistance, length of time on the regimen, and health care accessibility, but some stochastic variation was built into the system so that the outcome was not completely predictable from any single factor. Instead of memorizing explicit rules, this architecture enables the model to learn probabilistic feature connections.

Instead of employing a deterministic rule assignment, outcome probabilities were generated from weighted probabilistic interactions among many clinical and social variables. Stochastic probability adjustment functions were used to construct conditional dependencies, guaranteeing that no single variable or feature combination would produce the desired result. To boost diversity in patient profiles and reduce overly deterministic patterns, random noise injection was also used.

2.2.3 Synthetic Data Plausibility Assessment.

When available, the generated distribution was compared with the current epidemiologic distribution to assess the plausibility of the synthetic data. The distribution tests included creating summary statistics, class balancing, correlation, and goodness-of-fit testing. Rather than comparing the data to actual TB groups, this procedure sought to assess the data's plausibility. Because tuberculosis treatment outcomes emerge from interactions among demographic, clinical, behavioural, and healthcare-access variables, the synthetic dataset was intentionally structured to preserve heterogeneous feature relationships relevant to resource-constrained treatment environments. Table 1 summarises the full feature set and variable types retained in the dataset.

Table 1:Features in the Synthetic Dataset

Feature Category	Features Included	Variable Type
Demographics	Age, Gender, Occupation, Region, City	Mixed (numerical / categorical)

 J Health Inform Afr. 2026;13(1):228-245. DOI: 10.12856/JHIA-2026-v13-i1-700

Feature Category	Features Included	Variable Type
Clinical Presentation	Symptoms, Sputum Smear Test, GeneXpert Test, Chest X-ray	Categorical / binary
Comorbidities	Diabetes, HIV, Chronic Lung Disease, Malnutrition	Binary
Treatment Information	Treatment Type, Duration, Drug Resistance, Relapse, Complications	Mixed
Social Determinants	Smoking, Alcohol, Living Conditions, Healthcare Access	Categorical
Outcome (target)	Cured (0) vs Poor Outcome (1)	Binary 70% cured / 30% poor

For model evaluation, Poor Outcome (1) was treated as the positive class, and Cured (0) as the negative class. For all evaluation criteria, poor outcome was considered the positive class, given the clinical focus on determining individuals who might not be successfully treated.

2.3 Data Preparation and Feature Correlation

Categorical features were label-encoded using scikit-learn LabelEncoder objects, with a separate encoder per column stored for inverse transformation during SHAP visualization. Numerical features (Age, Duration of Treatment, Region Code) were standardized using StandardScaler to zero mean and unit variance. The dataset was partitioned 80:20 using stratified random splitting, yielding 16,000 training and 4,000 test records.

Of the 23 attributes in the dataset, 19 served as predictor variables. Of the 23 variables in the dataset, 19 were retained as predictor variables for model training and interpretation after excluding the target outcome and auxiliary indexing attributes. A Pearson correlation matrix was computed across all 19 predictor features and the treatment outcome variable. The analysis revealed a strong positive correlation between Sputum Smear Test and GeneXpert result ($r = 0.52$), reflecting the natural co-occurrence of positive microbiological findings in active pulmonary TB, and between Sputum Smear and Chest X-ray findings ($r = 0.44$), consistent with active disease burden. Correlation analysis revealed a moderate positive association between drug resistance and treatment duration ($r = 0.38$), suggesting that resistant treatment profiles were linked to prolonged treatment exposure within the synthetic environment. Healthcare access variables also demonstrated a moderate negative association with adverse outcomes ($r = -0.28$), consistent with the intended relationship between limited healthcare access and poorer treatment trajectories. No predictor pairs exceeded an absolute correlation threshold of 0.60, suggesting the absence of severe multicollinearity across the retained features. These relationships are illustrated in the Pearson correlation matrix in Figure 2

2.4 Transformer Model Architecture

The Transformer model consisted of four stacked encoder blocks, with multi-head self-attention (4 heads, each with 256 dimensions), residual connections, layer normalization, and Conv1D feed-forward sublayers. Dropout was applied at rates of 0.25 in encoder blocks and 0.4 in classification layers to mitigate overfitting. The input features were treated as unstructured sequences of tokens, enabling the self-attention module to capture contextual interactions among structured clinical attributes – although this should be noted that attention values reflect interactions rather than causality. The output from the last encoder block was flattened by averaging over all elements and fed to a two-layer classifier with 128 hidden units, RELU activation, and sigmoid outputs. Optimization was conducted using the Adam optimizer (learning rate 1×10^{-4}) and binary cross-entropy loss, with class weights to account for imbalanced classes, for 20 epochs and a batch size of 64. The total number of trainable parameters in this model amounted to about 29,000.

 J Health Inform Afr. 2026;13(1):228-245. DOI: 10.12856/JHIA-2026-v13-i1-700

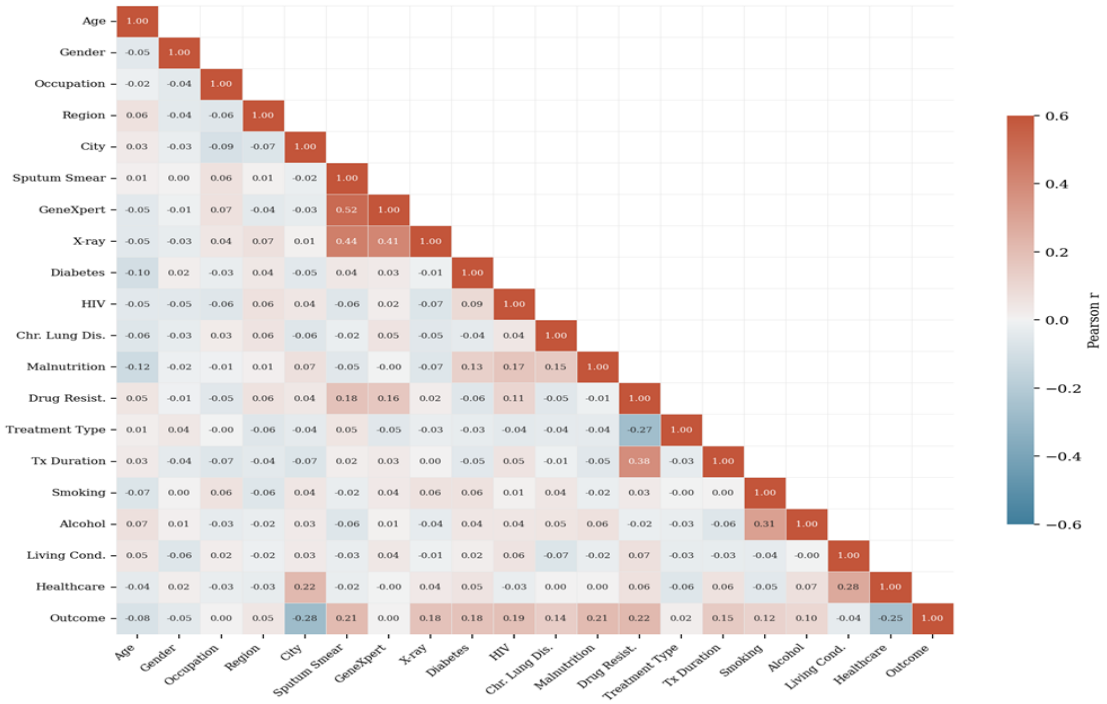

Figure 2: Pearson Correlation Matrix

2.5 Model Training and Self-Attention Weight Visualization

The training and validation curves shown in Figure 3 demonstrated stable convergence behaviour with limited evidence of severe overfitting. Training and validation losses remained closely aligned throughout optimization, while accuracy improved progressively across epochs before stabilizing in later training stages. These findings suggest comparatively stable learning behaviour under the controlled synthetic evaluation setting.

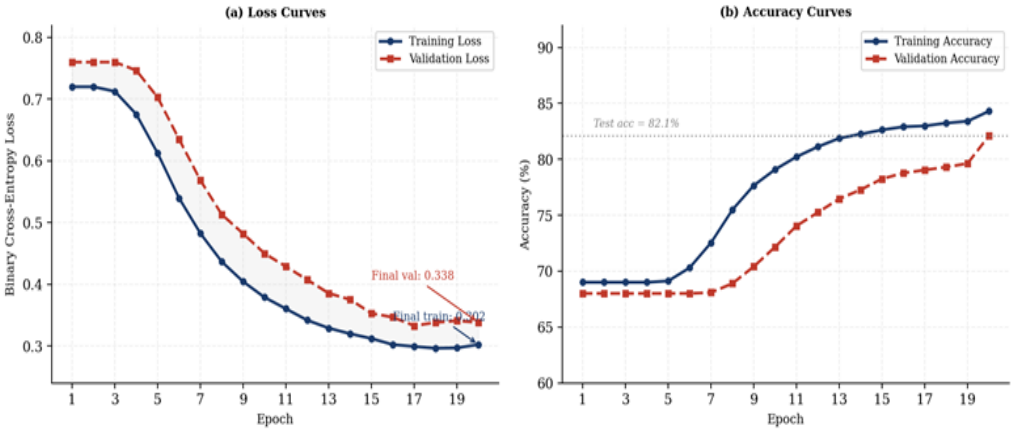

Figure 3: Training and Validation Performance Curves (20 epochs). (a) Loss curves (b) Accuracy curves

To characterize the transformer's intrinsic interpretability, self-attention weights were extracted from Encoder Block 1 and averaged across the four attention heads for 500 test-set predictions. The resulting mean (23×23) attention weight matrix shown in Figure 4 reveals the extent to which each feature contextualizes every other feature during inference, providing architectural-level evidence of feature-interaction learning that complements SHAP's post-hoc attributions. Attention-weight analysis was incorporated as an auxiliary architectural mechanism for contextual interaction analysis rather than as a definitive explanation of model reasoning behaviour or causal decision attribution.

 J Health Inform Afr. 2026;13(1):228-245. DOI: 10.12856/JHIA-2026-v13-i1-700

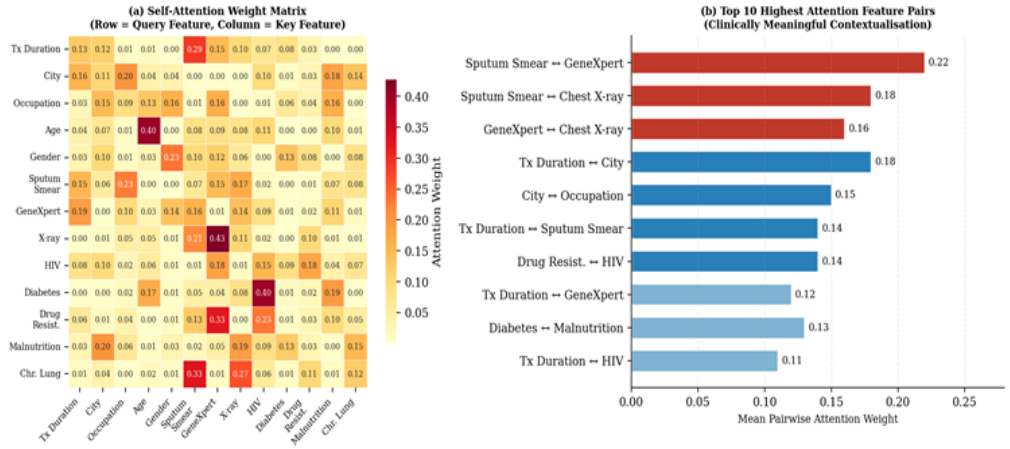

Figure 4: Transformer Self Attention Weight Analysis

2.6 Model Evaluation

The final model received its evaluation through the reserved test set, which contained 4,000 samples, using standard classification metrics. Accuracy, precision, recall, F1-score, and AUC-ROC were calculated using the scikit-learn library, together with precision and recall and F1-score and area under the ROC curve (AUC-ROC) statistics. The confusion matrix visualization shows the actual positive and negative results alongside incorrect positive and negative predictions, revealing the model's error distribution. All metrics were compared against the predefined clinical utility thresholds established during the business understanding phase. These thresholds were selected as heuristic, deployment-oriented reference criteria for controlled methodological evaluation and were not derived from validated tuberculosis treatment governance standards or prospective clinical decision-making protocols. Because the primary clinical objective is to flag patients at risk of unsuccessful treatment, Poor Outcome was treated as the positive class throughout all evaluation metrics

2.7 Baseline Model Comparison and Uncertainty Quantification

To assess the transformer model's relative performance, baselines were built using Logistic Regression, Random Forest, and XGBoost. The baselines were built on similar pre-processing and testing grounds to enable fair comparison. Comparison with baselines was conducted to determine whether the transformer model shows a statistically significant improvement over classical ML models.

To estimate the level of uncertainty, 95% confidence intervals for all metrics were computed using the bootstrap method with 1,000 replications. The performance variation between different iterations of the model was studied.

2.8 Post-Hoc Interpretability Implementation with SHAP

To interpret model predictions and identify the most influential features, SHAP was employed. Based on cooperative game theory, SHAP values give a unified measure of feature relevance with both local and global interpretability [10]. To achieve a balance between efficiency and fidelity, KernelExplainer was initialized using the model's prediction function and a background dataset generated from the training data using k-means (50 clusters).

SHAP values were calculated for selected test examples, resulting in global plots that illustrate the importance of features for specific predictions. Explanation for every patient in terms of SHAP values was obtained using waterfall plots that showed how each feature contributed to the baseline prediction until the result prediction was made. The original names of categorical feature values were available through a label encoder, which enhanced clinical understanding. Log odds were used to calculate SHAP values, which were converted to percentage points to help clinicians better understand the data, even though they reflect

 J Health Inform Afr. 2026;13(1):228-245. DOI: 10.12856/JHIA-2026-v13-i1-700

only the importance of features relative to one another. SHAP values are used to determine importance of features but not their interactions.

The idea behind combining attention-weight visualization with SHAP analysis was to provide different perspectives on interpretability. SHAP analysis helps identify post-hoc feature importance in terms of their contribution to predictions, whereas attention analysis provides an architectural understanding of feature interactions within a transformer. Both these techniques help build a wider representation on interpretability without suggesting any causation whatsoever.

2.9 Deployment Prototype

The trained transformer model has been implemented using a Flask web application with the necessary security features suitable for a demo-level prototype. The implementation involves loading the model weights, pre-processing the objects, and using the SHAP KernelExplainer. The Rule-Based Safety Escalation Layer feature restricts predictions when a clinical-hazardous combination of risk factors is present. The rule-based safety escalation layer was intentionally designed as a bounded safety support mechanism operating independently from the transformer optimization process. Its role was not to artificially improve benchmark performance but to simulate conservative escalation safeguards commonly incorporated into safety-sensitive healthcare decision-support environments. The escalation rules and their triggering conditions are detailed in Table 2 and the associated deployment footprint and inference characteristics are reported in Table 3.

Table 2: Rule-Based Safety Escalation Layer

Condition	Escalation Action
HIV positive + MDR-TB + severe malnutrition	Escalate prediction to CRITICAL, risk regardless of model score
Missing adherence records (>7 days)	Flag for manual clinical review, suppress automated prediction
Drug Resistance confirmed + treatment duration > 12 months	Escalate to HIGH risk, alert clinician
Model confidence below 60% on any prediction	Flag as indeterminate, refer to specialist review

Table 3: System Deployment Characteristics

Metric	Value	Notes
Model size (Keras .h5)	0.34 MB	Full architecture + weights
TensorFlow Lite model size	0.12 MB	Post-quantization; mobile/edge deployment
Preprocessing objects (.pkl)	0.08 MB	StandardScaler + 23 LabelEncoders via joblib
Total deployment footprint	~0.54 MB	All inference artefacts combined
Average inference time	1.8 seconds	Per-patient prediction + SHAP on i7-8565U(no GPU)
SHAP explanation time	~1.2 seconds	KernelExplainer, k-means background k = 50
RAM footprint (inference)	~420 MB	Model + SHAP explainer loaded in memory
Full pipeline (train→SHAP)	~15 minutes	Consumer laptop; no GPU required
Offline capability	Full	No external API dependencies post-installation

 J Health Inform Afr. 2026;13(1):228-245. DOI: 10.12856/JHIA-2026-v13-i1-700

3 Results

3.1 Model Performance Evaluation

The model was assessed using a test dataset comprising 4,000 patient records: 2,800 with cured outcomes and 1,200 with poor outcomes.
Table 4 shows the results in relation to all set thresholds of clinical utility.

Table 4: Model comparison

Model	AUC-ROC	Accuracy	F1-Score	Recall	95% CI (AUC)
Logistic Regression	0.812	0.789	0.652	0.689	0.795-0.829
Random Forest	0.851	0.816	0.713	0.773	0.837-0.865
XGBoost	0.863	0.822	0.726	0.801	0.850-0.876
Transformer (proposed)	0.872	0.821	0.741	0.856	0.858-0.886

Collectively, the findings suggest that the transformer framework provided a stronger balance between discriminative performance, minority-class sensitivity, and deployment-oriented evaluation capability than the alternative baseline architectures.

Table 5: Model Performance Metrics

Metric	Score	Threshold	Status
AUC-ROC	0.872	>0.80	✓ PASS
Accuracy	82.1% (3,284 / 4,000 | TP = 1027, TN = 2257, FP = 543, FN = 173	>75%	✓ PASS
Precision	0.654	>0.60	✓ PASS
Recall (Sensitivity)	0.856	>0.75	✓ PASS
F1-Score	0.741	>0.75	✗Marginal (A=0.009)
Specificity	0.806	—	Reported
Brier Score	0.143	<0.20	✓ PASS

The transformer achieved the highest AUC-ROC, though the margin over XGBoost was modest (0.872 vs 0.863), as summarised in
Table 4. Notably, the 95% bootstrap confidence intervals for the two models overlap substantially (Transformer: 0.858-0.886; XGBoost: 0.850-0.876), indicating that this difference cannot be considered statistically significant based on interval comparison alone; a formal paired significance test would be required to establish superiority with confidence. This suggests the primary contribution lies not in raw predictive superiority but in the integration of explainability, feature interaction modelling, calibration assessment, and deployment-oriented design. It should also be noted that the F1-Score reported in Table 5 falls marginally below the pre-specified 0.75 clinical utility threshold.

3.2 Confusion Matrix

The confusion matrix demonstrates comparatively stable discrimination between cured and poor-outcome patients, although the remaining false-negative predictions remain clinically important because they represent potentially vulnerable patients incorrectly classified as lower risk within the synthetic evaluation environment. This distribution of errors is visualized in Figure 5.

 J Health Inform Afr. 2026;13(1):228-245. DOI: 10.12856/JHIA-2026-v13-i1-700

3.3 Ablation Analysis

To determine the role of each component in the framework under discussion, an ablation study was conducted by systematically eliminating or altering components of the model. Results are reported in Table 6.

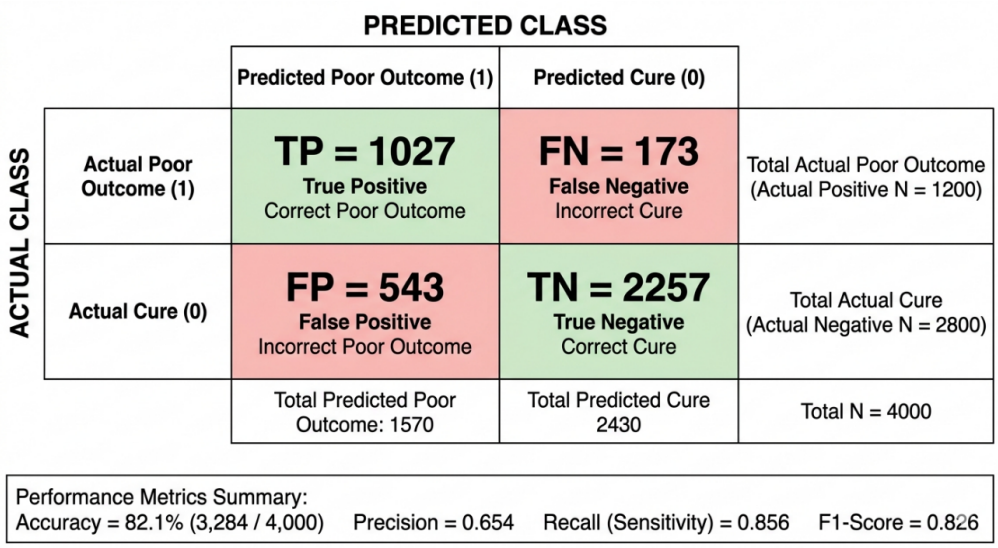

Figure 5: Confusion Matrix

Table 6: Ablation Analysis

Configuration	Description	AUC-ROC	F1-Score	Recall
Full Model (Proposed)	Transformer + full feature set	0.872 ± 0.006	0.741	0.856
Without Attention	No self attention	0.798 ± 0.010	0.644	0.718
Reduced Features	Top 10 Features only	0.843	0.712	0.827
No Class Weights	Transformer without imbalance handling	0.861	0.685	0.704
Logistic Regression	Linear baseline	0.812	0.652	0.689

The substantial decrease in AUC-ROC (from 0.872 to 0.798) and F1 score (from 0.741 to 0.644) due to the absence of the attention module indicates that modelling the interaction between contextual features played an important role in the transformer model's prediction process. In contrast, the decrease in performance was relatively smaller in the reduced-feature and no-class-weights scenarios, implying that contextualization via attention had a major architectural effect in this experiment. These findings suggest that contextual-feature interaction modelling was the principal architectural contribution of the transformer framework in the controlled synthetic healthcare environment.

3.4 Risk Stratification and Patient Analysis

The model produces probability results which get divided into four clinical risk groups through heuristically defined interpretability-oriented thresholds: LOW risk (70% cure probability), MEDIUM risk (50-70% cure probability), HIGH risk (30-50% cure probability), and CRITICAL risk (30% cure probability). Two representative patient cases were analyzed to demonstrate the model's behaviour across the risk spectrum. The proposed thresholds were incorporated solely for interpretability demonstration within the synthetic environment and should not be interpreted as clinically validated tuberculosis risk stratification criteria. The stratification boundaries were introduced exclusively for interpretability demonstration and workflow simulation within the synthetic environment and should not be interpreted as medically validated intervention thresholds for tuberculosis treatment management.

 J Health Inform Afr. 2026;13(1):228-245. DOI: 10.12856/JHIA-2026-v13-i1-700

3.5 Model Calibration and Reliability Assessment

Calibration performance was assessed by grouping predictions into deciles and comparing the mean predicted probability to the observed event frequency. The model demonstrated near-linear alignment with the ideal calibration line across mid-range probabilities, with minor overconfidence at extreme bins. A calibration score of 0.143 is considered to be acceptable for the internal evaluation (random = 0.25, perfect = 0.0). The calibration was calculated internally, using only synthetic data from the test set. An external calibration is needed because synthetic probabilities may not reflect the real-world probabilities of treatment outcomes. Figure 6 presents the corresponding reliability diagram.

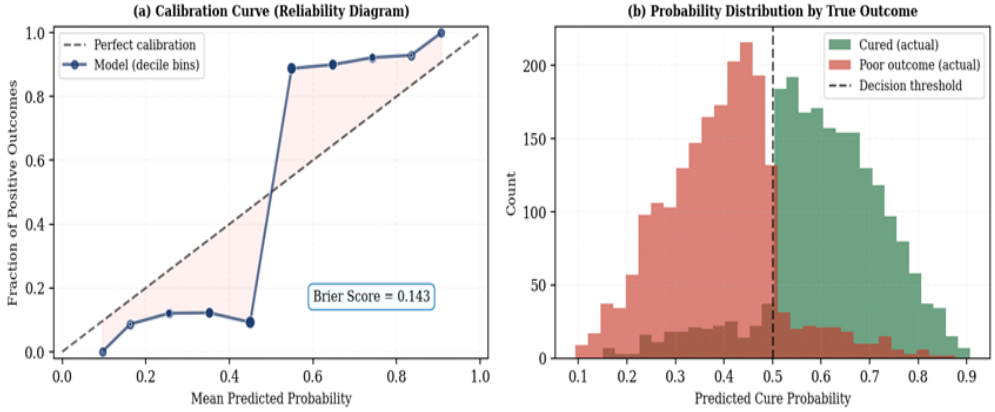

Figure 6: Model Calibration and Reliability Assessment

3.6 Sub-Group Consistency Inspection

The descriptive subgroup consistency test was used to determine whether there were substantial differences in predictive behaviour among synthetic demographic subgroups, such as age- and gender-based subgroups. The results indicated moderate differences in behaviour across some subgroups; however, these results cannot be considered a subgroup consistency analysis because of the artificial nature of the dataset used for testing. Figure 7 summarizes these subgroup comparisons.

The current analysis, then, can be considered an initial behavioural check procedure that might be helpful for future analyses of the consistency of real subgroups. Formal algorithmic fairness metrics such as equalized odds, predictive parity, or demographic parity were intentionally not interpreted inferentially because the synthetic demographic construction process could not reliably reproduce authentic structural healthcare inequalities or real population-level bias distributions. Consequently, the subgroup analysis should be interpreted primarily as behavioural stability inspection under controlled synthetic demographics rather than evidence of real-world fairness performance.

3.7 Illustrative Clinical Interpretation Scenario

To illustrate the interpretability approach, the framework was applied to a typical synthetic patient data sample, specifically Patient TB-2026-187, an 18-year-old female student from Rural Dhaka, diagnosed at Day 7 of her treatment regime. Based on the transformer model's output, the predicted probability of cure for this patient is 87.2%, categorizing the patient as LOW RISK in the over 70% cure category, with no rule-based safety escalations. In the clinical interpretation component, the lack of direct access to healthcare services is recognized as the only risk factor for the patient. At the same time, there are seven favourable factors that account for a positive diagnosis, including absence of drug resistance or previous relapse, non-smoker, healthy nutritional status, no presence of HIV disease, younger age, being enrolled in the DOTS program, and lack of current complications. While demonstrating the operation of the interpretability framework in practice, this example should not be viewed as a real-life clinical process. The corresponding output display is shown in Figure 8.

 J Health Inform Afr. 2026;13(1):228-245. DOI: 10.12856/JHIA-2026-v13-i1-700

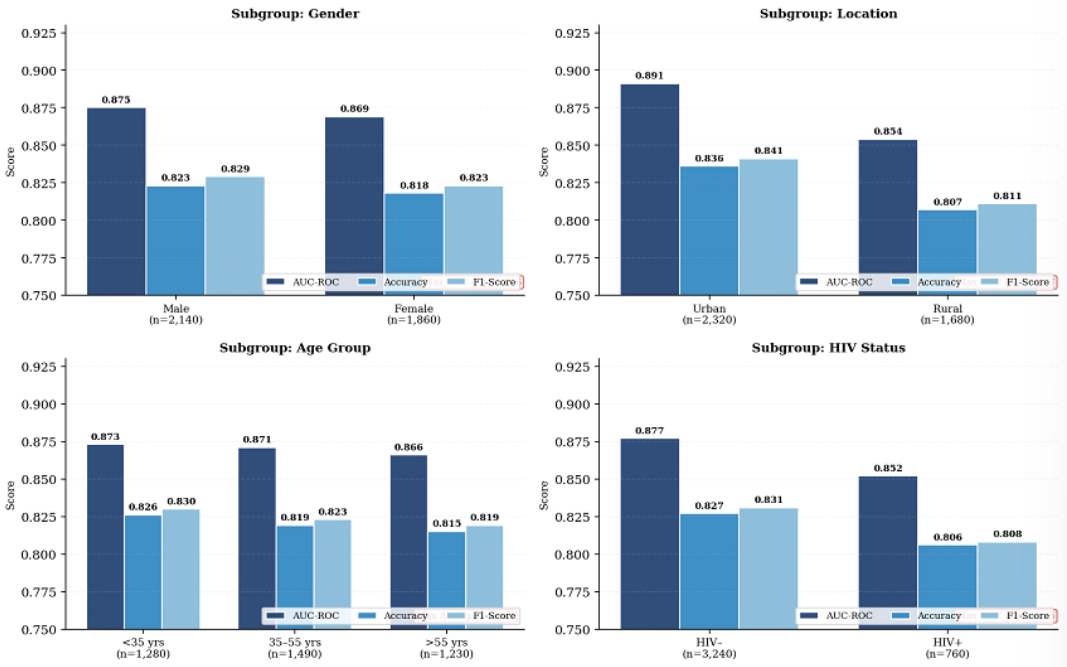

Figure 7: Sub-Group Fairness Analysis

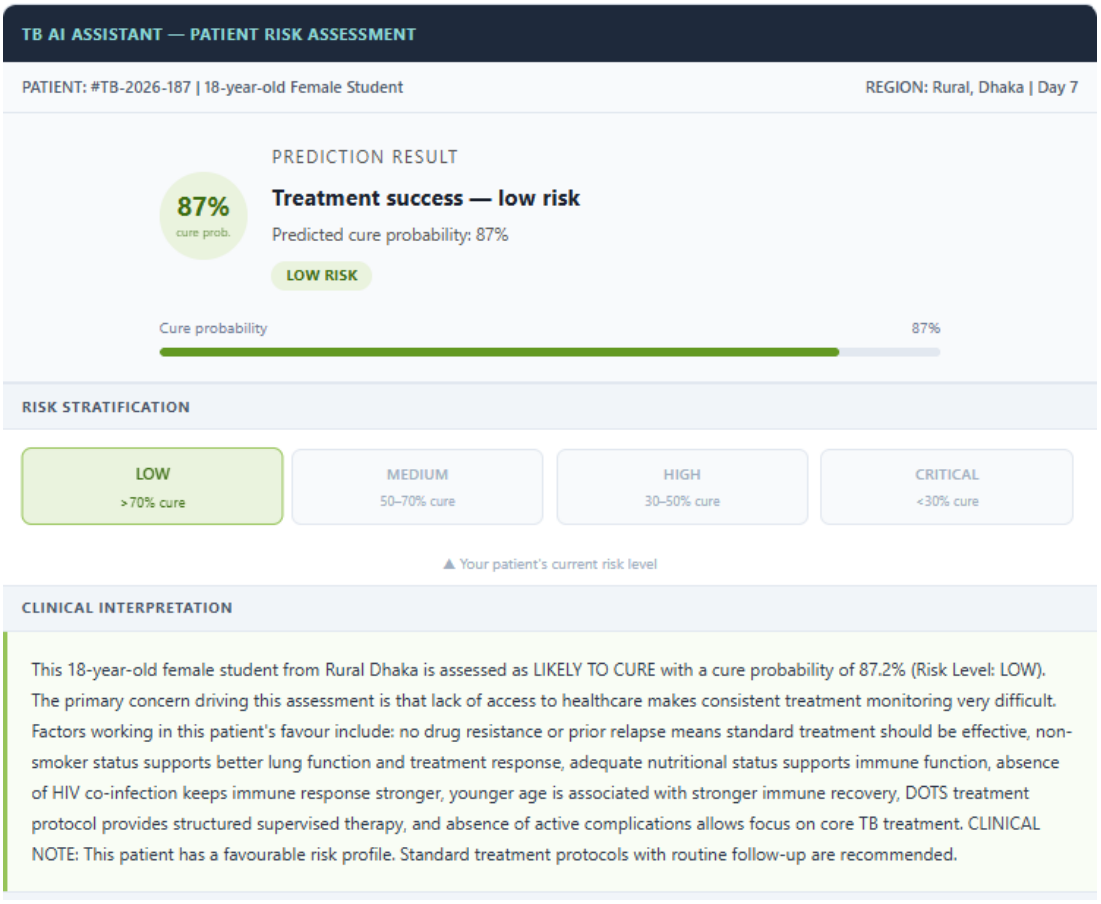

Figure 8: Patient Prediction

 J Health Inform Afr. 2026;13(1):228-245. DOI: 10.12856/JHIA-2026-v13-i1-700

4 Discussion

4.1 Deployment-Oriented Healthcare AI Perspective

The proposed transformer framework demonstrated methodological feasibility for deployment-oriented healthcare AI evaluation in controlled, synthetic, resource-constrained tuberculosis settings. Although the transformer achieved competitive predictive performance relative to classical machine learning baselines, the observed performance gains over XGBoost remained comparatively modest. The justification for transformer utilization in the present study therefore lies less in isolated predictive superiority and more in the architectural ability to support contextual feature interaction modelling, integrated interpretability inspection, calibration-oriented optimization behaviour, and future extensibility toward heterogeneous longitudinal healthcare representations beyond static tabular prediction alone. These findings suggest that the principal contribution of the framework lies less in raw predictive superiority and more in integrating explainability, calibration assessment, subgroup consistency inspection, deployment feasibility analysis, and operational trust considerations into a unified healthcare AI evaluation pipeline.

The transformer architecture offered advantages in contextual feature interaction modelling through self-attention mechanisms, enabling dynamic weighting of heterogeneous clinical and social determinants during prediction. This may explain the framework's comparatively improved ability to identify high-risk treatment profiles involving comorbidity interactions, treatment adherence patterns, microbiological findings, and healthcare access constraints. Nevertheless, the observed performance improvements should be interpreted cautiously, as all experiments were conducted under controlled, synthetic conditions rather than in operational clinical environments.

The explainability analysis further demonstrated that variables such as HIV status, drug resistance, treatment adherence, nutritional condition and healthcare accessibility consistently contributed to prediction behaviour across the synthetic evaluation setting. Although SHAP and attention-based interpretability mechanisms improved transparency and supported inspection of model behaviour, these outputs represent statistical associations learned during optimization rather than evidence of causal clinical relationships. Within tuberculosis treatment workflows characterized by prolonged treatment duration, uncertain adherence patterns, a high comorbidity burden, and limited specialist oversight, explainability may therefore serve as an operational trust-support mechanism rather than a definitive clinical reasoning system.

The deployment findings additionally highlight an important trade-off between predictive optimization and operational feasibility in decentralized healthcare environments. In low-resource settings, a marginally more accurate but computationally intensive and non-interpretable model may ultimately prove less useful than a slightly less accurate framework that remains computationally lightweight, explainable, offline-capable, and operationally deployable on existing infrastructure. The proposed framework demonstrated relatively small deployment requirements, rapid inference capability and offline operational feasibility without dependence on specialized GPU infrastructure. These characteristics may improve suitability for decentralized healthcare environments where computational resources, internet connectivity and advanced technical support remain limited. CPU-only inference under two seconds, and offline operability is intended to give future work in this space a concrete, reproducible benchmark against which lightweight healthcare AI deployments in comparable low-resource settings can be measured.

Importantly, the study suggests that healthcare AI optimization in constrained settings becomes inherently multi-dimensional. Effective deployment may depend not only on predictive performance but also on interpretability, calibration reliability, subgroup behavioural stability, infrastructure compatibility, workflow suitability, governance readiness and clinician trust. Consequently, predictive accuracy alone may be insufficient as the dominant evaluation criterion for healthcare AI systems intended for operational clinical environments.

Nevertheless, the current framework should not be interpreted as deployment-ready clinical intelligence. Real-world implementation would require substantially stricter governance procedures, including external clinical validation, prospective calibration assessment, subgroup equity auditing, clinician usability testing, workflow integration analysis, model drift monitoring, incident reporting systems, and institutional oversight for responsible AI deployment. The present findings therefore represent methodological feasibility under controlled synthetic conditions rather than evidence of clinically validated operational performance.

 J Health Inform Afr. 2026;13(1):228-245. DOI: 10.12856/JHIA-2026-v13-i1-700

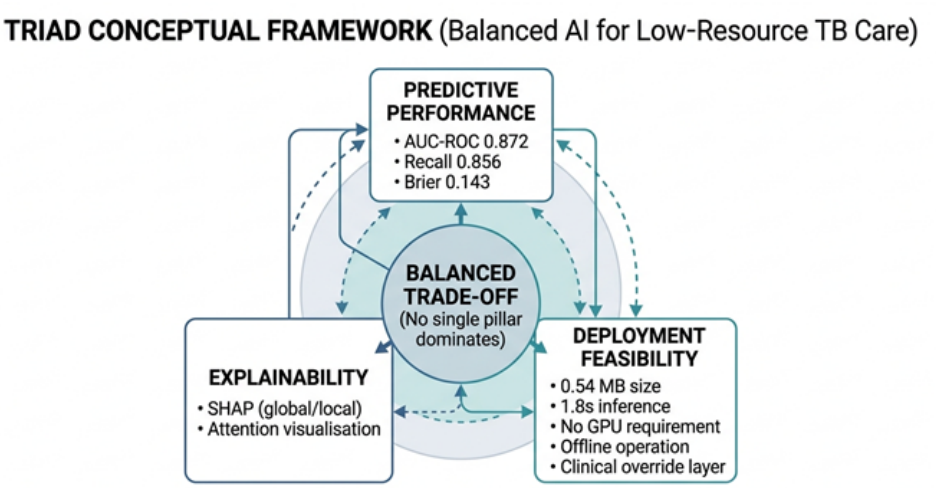

Figure 9: Conceptual Framework

The proposed conceptual framework, depicted in Figure 9, summarises how predictive performance interpretability and operational feasibility interact within resource-constrained deployment settings. It suggests that, in cases where AI systems for health care operate in resource-constrained settings, they should not be made to maximize predictive accuracy on their own terms. Rather, utility could be achieved by balancing predictive accuracy, interpretability, and practicality. The conceptual framework is proposed as an interpretive synthesis derived from controlled experimental findings rather than as a validated operational implementation model for real-world healthcare systems.

4.2 Subgroup Consistency Screening under Synthetic Data Constraints

As previously noted, this screening reflects internal consistency only. Regarding performance by gender, age category, geographical location, and HIV status, small differences were observed in the artificial data. Proper assessment using metrics such as equalized odds and demographic parity, as well as externally calibrated subgroup analysis, is needed before drawing any conclusions about equity in the real world.

Due to factors such as healthcare access, socioeconomic status, prevalence of HIV infection, and disparities based on region, there may be disparities in treatment outcomes across different populations due to structural inequality. Therefore, in future studies concerning subgroup consistency, researchers should investigate the ability of models to maintain stability within clinically vulnerable subgroups.

4.3 Deployment and Governance Implications

While the suggested framework proved computationally feasible for implementation in lightweight local settings, its use in actual healthcare settings would require much stricter validation processes and governance practices [8]. These include external clinical validation, calibration, clinician usability assessment, model drift, security assessment, incident reporting, and oversight frameworks for the responsible use of AI in healthcare facilities [9].

These results imply that methodological feasibility is more important than operational feasibility when evaluating healthcare artificial intelligence. In decentralized healthcare settings, the success of such implementation might depend not only on predictive performance but also on explainability, computational availability, workflow suitability, and trustworthiness.

4.4 Comparison with Prior Work

This study used synthetic data, which must be borne in mind when interpreting comparisons with previous studies. However, many other studies have used real patients, images, or populations affected by the disease.

 J Health Inform Afr. 2026;13(1):228-245. DOI: 10.12856/JHIA-2026-v13-i1-700

The performance of the proposed transformer architecture on synthetic data falls within the range of existing models for TB outcome prediction reported in the literature review, as shown comparatively in Table 7.

Using radiomics and longitudinal CT scans, researchers [20] developed a deep learning model that achieved AUC values of 0.764-0.867 on their test sets. This study differs from previous methods that address only parts of these criteria by combining transformer performance, SHAP explainability, attention-weight interpretability, deployment feasibility, calibration, subgroup-consistency screening, and a novel rule-based safety-escalation layer.

The author [21] employed XGBoost with SHAP explanations to predict treatment failure among patients with TB-diabetes comorbidity, achieving an AUC of 0.928 on their test set. While this exceeds the transformer's AUC, their model was developed on a more homogeneous population (patients with TB-diabetes comorbidity) and may not generalize as broadly to the general TB patient population represented in the synthetic dataset. The transformer model's ability to achieve strong performance across a diverse synthetic population suggests methodological feasibility for future evaluation under heterogeneous clinical conditions.

The author [22] introduced the Decoder Transformer for Temporal Health Data (DT-THRE), achieving 78.5% accuracy on sequential EHR data, a substantial improvement over baseline models (40.5%). The transformer model's accuracy of 82.1% is consistent with their reported performance, although their model was evaluated on general health data rather than specifically on TB cohorts, limiting direct comparability.

The author [13] reviewed the expanding application of transformer architectures across biomedical domains, noting their capacity to model complex, non-sequential feature interactions beyond their original sequence-to-sequence formulation. The present study extends this trajectory to tabular clinical data, suggesting that self-attention principles remain applicable to structured patient records when features are treated as a token sequence.

Table 7: Comparative Positioning of the Proposed Framework

Study	Data Type	Explainability	Subgroup Screening	Calibration	Deployment	Transformer
[20]	Real CT/imaging	Partial (radiomics)	Yes	No	Yes	No
[21]	Real World tabular	Shap only	No	No	No	No
[22]	Real EHR sequential	None reported	No	No	No	No
Proposed Study	Synthetic tabular	SHAP Attention	Descriptive	Yes	Prototype	Yes

Collectively, these comparisons imply that the value of this proposed framework lies not in absolute predictive power but rather in its integrative analysis of explainability, calibration, subgroup robustness testing, and implementational feasibility.

4.5 Reproducibility and Experimental Control

All preprocessing procedures, training-test partitions, random initialization settings, optimization parameters, and evaluation pipelines were fixed across experiments to improve reproducibility within the controlled synthetic environment. Multiple baseline models were trained under comparable preprocessing conditions to minimize experimental inconsistency.

5 Limitations and Threats to Validity

Several limitations need to be considered when interpreting the results of this study. First, the transformer architecture itself offers limited methodological novelty relative to existing tabular transformer approaches in the literature, and its predictive improvement over a strong gradient-boosted baseline (XGBoost) was

 J Health Inform Afr. 2026;13(1):228-245. DOI: 10.12856/JHIA-2026-v13-i1-700

modest and, as noted above, not confirmed by formal statistical significance testing. The primary contribution of this work, therefore, lies in the integrated evaluation pipeline rather than in architectural motivation. To start, all tests in this paper have been run on synthetic data generated under controlled conditions. Second, the analysis of subgroups' consistency was purely descriptive and did not involve any calculations of equity measures or performance differences. Although there were no significant performance degradation cases in the subgroups identified in the synthetic dataset, this cannot in any way be used as proof of equal performance in the real world. Third, the explainability techniques employed in the current investigation provide interpretability guidance rather than causation. SHAP values and attention visualizations reveal statistical patterns that were identified by the model during its optimization process and should not be taken as proof of any clinical causation. The current prototype evaluation focused more on computational feasibility and did not conduct usability testing with clinicians, workflow analysis, or operational healthcare implementation analysis.

Additionally, the framework was tested only in a retrospective manner and lacks external validation, prospective testing, or clinician evaluation.

The outputs from the SHAP analysis are explanatory and provide model-driven association interpretations rather than fundamental medical insight. While there was considerable correspondence between some of the highly weighted variables and the existing risk factors for tuberculosis, this does not equate to these being verified clinical determinants of outcome. The simulation exercises performed for intervention scenarios were based on simplified, hypothetical causal foundations and, as such, can only yield plausible intervention behaviour.

Deployment in practice will require much more stringent governance and validation steps, including external calibration assessment, institutional oversight, medical practitioner usability evaluations, model drift detection, security auditing, and ongoing post-deployment assessment.

This study demonstrates that epistemological modesty is necessary when implementing AI systems in healthcare. Uncertainty is generated by synthesized assumptions, insufficiently representative clinical data, and varying deployment conditions, even with excellent performance under tightly controlled conditions. Therefore, AI must be used as a probability model to inform decisions, not as a decision-maker itself.

5.1 Ethical and Responsible AI Considerations

The proposed framework was developed exclusively for controlled methodological evaluation and not for autonomous clinical deployment. The study intentionally incorporated bounded interpretability claims, explicit acknowledgement of synthetic data limitations, conservative subgroup interpretation and deployment-oriented safety constraints to reduce overstatement of healthcare AI capability. Any future translational implementation would require prospective clinical validation, external calibration assessment, clinician-centred usability evaluation, institutional governance oversight, and continuous post-deployment monitoring.

6 Conclusion

This study developed and tested a transformer-based framework for tuberculosis treatment outcome prediction in controlled, resource-constrained synthetic healthcare settings, incorporating explainability, calibration assessment, subgroup consistency inspection, and intervention scenario simulation into a unified healthcare artificial intelligence pipeline.

Although the transformer model performed similarly well to other machine learning methods from a predictive standpoint, the main novelty of this research is the establishment of a methodological framework incorporating explainability, calibration assessment, subgroup inspection, and deployment feasibility.

It must be emphasized, however, that the results of this study must be understood in the context of the limitations of a synthetic evaluation approach rather than actual clinical validation, which has not occurred. Therefore, the paper does not propose a clinically applicable tuberculosis treatment predictor but rather an experimental framework for constrained AI application in healthcare. More broadly, the study highlights the growing importance of deployment-oriented healthcare AI evaluation frameworks in which predictive performance, explainability, calibration reliability, subgroup stability and operational feasibility are treated as interdependent design requirements rather than isolated optimization targets. In healthcare, limited

 J Health Inform Afr. 2026;13(1):228-245. DOI: 10.12856/JHIA-2026-v13-i1-700

environments, trustworthy and practically deployable AI systems may ultimately prove more valuable than narrowly optimized, high-performance prediction models lacking interpretability or operational realism. This study therefore proposes a deployment-oriented evaluation integrating explainability, calibration, subgroup screening, and operational feasibility alongside predictive accuracy as a methodological template that future healthcare AI research in low-resource settings can adopt and extend. In particular, the rule-based safety-escalation layer introduced here offers a reusable design pattern for pairing predictive models with human-oversight triggers in settings where full clinical validation remains pending, an approach that later studies working under similar resource and validation constraints may wish to adapt.

Acknowledgements

During manuscript development, the authors utilized generative AI systems and AI-assisted visual generation tools to support editorial refinement, conceptual and methodological overview, and pipeline figure development. Figure generation involved structured prompt engineering based on author-defined workflow specifications and iterative visual adjustments. The authors retained full responsibility for all scientific interpretations, methodological decisions, verification processes and final manuscript content. All AI-generated outputs were independently reviewed and validated prior to inclusion.

Conflicts of Interest

The authors declare no conflicts of interest relevant to this research.

References

[1] Global Tuberculosis Report 2023. World Health Organization, 2023.

[2] S. I. Nafisah and G. Muhammad, “Tuberculosis detection in chest radiograph using convolutional neural network architecture and explainable artificial intelligence,” Neural Comput. Appl., vol. 36, no. 1, pp. 111–131, Jan. 2024, doi: 10.1007/S00521-022-07258-6/METRICS.

[3] M. C. Hosu, L. M. Faye, and T. Apalata, “Comorbidities and Treatment Outcomes in Patients Diagnosed with Drug-Resistant Tuberculosis in Rural Eastern Cape Province, South Africa,” Diseases, vol. 12, no. 11, Nov. 2024, doi: 10.3390/diseases12110296.

[4] S. R. N. Kalhori and X.-J. Zeng, “Evaluation and Comparison of Different Machine Learning Methods to Predict Outcome of Tuberculosis Treatment Course,” J. Intell. Learn. Syst. Appl., vol. 05, no. 03, pp. 184–193, 2013, doi: 10.4236/jilsa.2013.53020.

[5] S. S. Sibanda and B. Ndlovu, “Explainable Transformer and Machine Learning Models in Predicting Tuberculosis Treatment Outcomes . A Systematic Review,” J. Appl. Informatics Comput., vol. 10, no. 1, pp. 150–164, 2026, doi: 10.30871/jaic.v10i1.11846.

[6] T. Govender et al., “Decentralized, Integrated Treatment of RR/MDR-TB and HIV Using a Bedaquiline-Based, Short-Course Regimen Is Effective and Associated With Improved HIV Disease Control,” J. Acquir. Immune Defic. Syndr., vol. 92, no. 5, pp. 385–392, Apr. 2023, doi: 10.1097/QAI.0000000000003150.

[7] J. Machemedze and B. Ndlovu, “Transformer-Based Models for Electronic Health Records and Omics in Healthcare : A Systematic Literature Review,” J. Appl. Informatics Comput., vol. 10, no. 1, pp. 90–105, 2026, doi: 10.30871/jaic.v10i1.11893.

[8] J. Herington et al., “Ethical Considerations for Artificial Intelligence in Medical Imaging : Deployment and Governance,” pp. 1509–1515, doi: 10.2967/jnumed.123.266110.

[9] J. Zhang and Z. Zhang, “Ethics and governance of trustworthy medical artificial intelligence,” BMC Med. Inform. Decis. Mak., vol. 9, pp. 1–15, 2023, doi: 10.1186/s12911-023-02103-9.

[10] A. K. Mondal, A. Bhattacharjee, P. Singla, and A. P. Prathosh, “XViTCOS: Explainable Vision Transformer Based COVID-19 Screening Using Radiography,” IEEE J. Transl. Eng. Heal. Med., vol. 10, 2022, doi: 10.1109/JTEHM.2021.3134096.

[11] S. Gupta et al., “Four Transformer-Based Deep Learning Classifiers Embedded with an Attention U-Net-Based Lung Segmenter and Layer-Wise Relevance Propagation-Based Heatmaps for COVID-19 X-ray Scans,” Diagnostics, vol. 14, no. 14, Jul. 2024, doi: 10.3390/diagnostics14141534.

[12] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "TabTransformer : Tabular Data Modeling Using Contextual Embeddings," no. Rong 2020, 2020.

[13] S. Madan, M. Lentzen, J. Brandt, D. Rueckert, M. Hofmann-Apitius, and H. Fröhlich, "Transformer models in biomedicine," BMC Medical Informatics and Decision Making, vol. 24, no. 1. BioMed Central Ltd, Dec. 2024, doi: 10.1186/s12911-024-02600-5.

[14] E. Chanda, "The clinical profile and outcomes of drug resistant tuberculosis in Central Province of Zambia," BMC Infect. Dis., vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12879-024-09238-8.

[15] L. Shichman et al., "Predictive Modeling and Machine Learning Insights into Multi-Drug Resistant Tuberculosis Treatment Outcomes," in 2025 IEEE Systems and Information Engineering Design Symposium, SIEDS 2025, 2025, pp. 185–190, doi: 10.1109/SIEDS65500.2025.11021160.

[16] N. Ndlovu and B. Ndlovu, "Explainable Artificial Intelligence in Multimodal Malaria Prediction : A Systematic Review and Roadmap Integrating Climate Change , Parasite Genomics , and Public Health Decision Support," J. Appl. Informatics Comput., vol. 10, no. 2, 2026, doi: 10.30871/jaic.v10i2.12347.

[17] O. T. Chikumo and B. Ndlovu, "Transformer-based Models for Cardiovascular Disease Predictions from Electronic Health Records: A Systematic Review," J. Appl. Informatics Comput., vol. 10, no. 1, pp. 75–89, Feb. 2026, doi: 10.30871/JAIC.V10I1.11899.

[18] M. C. Hosu, L. M. Faye, and and T. Apalata, "Predicting Treatment Outcomes in Patients with Drug-resistant Tuberculosis and HIV Co-infection Using a Supervised Machine Learning Algorithm." Sep. 2024, doi: 10.20944/preprints202409.1747.v1.

[19] Y. Yang, L. Xu, L. Sun, P. Zhang, and S. S. Farid, "Machine learning application in personalized lung cancer recurrence and survivability prediction," Comput. Struct. Biotechnol. J., vol. 20, pp. 1811–1820, Jan. 2022, doi: 10.1016/j.csbj.2022.03.035.

[20] M. Nijiati et al., "Deep learning and radiomics of longitudinal CT scans for early prediction of tuberculosis treatment outcomes," Eur. J. Radiol., vol. 169, Dec. 2023, doi: 10.1016/j.ejrad.2023.111180.

[21] A. Z. Peng et al., "Explainable machine learning for early predicting treatment failure risk among patients with TB-diabetes comorbidity," Sci. Rep., vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-57446-8.

[22] O. Boursalie, R. Samavi, T. E. Doyle, and O. Boursalie, "Decoder Transformer for Temporally-Embedded Health Outcome Predictions," in Proceedings - 20th IEEE International Conference on Machine Learning and Applications, ICMLA 2021, 2021, pp. 1461–1467, doi: 10.1109/ICMLA52953.2021.00235.

 J Health Inform Afr. 2026;13(1):228-245. DOI: 10.12856/JHIA-2026-v13-i1-700