Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics e-ISSN: 2656-8624
Homepage: https://ijeeemi.org/; Vol. 8, No. 3, pp. 405-430, August 2026

RESEARCH PAPER **OPEN ACCESS**

An Explainable XGBoost Framework for Mortality Prediction in People Living with HIV Receiving Antiretroviral Therapy

Ni Putu Likayuni Viona[1], Ngakan Nyoman Kutha Krisnawijaya[1], and Desak Nyoman Widyanthini [2]

[1] Department of Information Technology, Faculty of Engineering and Informatics, Universitas Pendidikan Nasional, Denpasar, Bali, Indonesia
[2] Department of Public Health and Preventive Medicine, Faculty of Medicine, Udayana University, Bali, Indonesia

Corresponding author: Ni Putu Likayuni Viona (e-mail: likayunii06@gmail.com), **Author(s) Email**: Ngakan Nyoman Kutha Krisnawijaya (e-mail: ngakankutha@undiknas.ac.id), Desak Nyoman Widyanthini (e-mail: desakwidyanthini@unud.ac.id)

Abstract **Mortality rates among people living with HIV (PLHIV) commencing antiretroviral therapy (ART) remain a pressing clinical challenge, underscoring the need to identify high-risk individuals early to deliver timely medical interventions. To address this, we developed an explainable machine learning framework built around an optimized Extreme Gradient Boosting (XGBoost) classifier to estimate mortality risk at ART initiation. The study evaluated a retrospective cohort containing 393 clinical records from 2006 to 2023, gathered at RSUD Sanjiwani in Bali, Indonesia. Data preprocessing involved imputing missing entries using median and mode strategies, along with one-hot encoding for categorical variables. To address class imbalance, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to the training partition (80:20 split). Hyperparameter tuning was carried out using Optuna paired with 5-fold cross-validation, while SHapley Additive exPlanations (SHAP) provided both global feature importance and patient-level local interpretability. To completely guard against target leakage, the primary baseline classifier was trained exclusively on pre-treatment predictors accessible at ART initiation (Day 0). This leakage-safe XGBoost variant yielded an accuracy of 59.49%, a weighted F1-score of 0.59, and an ROC-AUC of 0.5914, indicating that initial clinical parameters alone offer limited predictive capacity in this retrospective sample. Conversely, a sensitivity analysis that incorporated follow-up duration achieved an accuracy of 89.87%, a weighted F1-score of 0.90, and an ROC-AUC of 0.9546, demonstrating the substantial retrospective predictive information contributed by follow-up duration. SHAP interpretations confirmed that patient age and functional status represent important baseline predictors. Ultimately, these findings remind us that elevated retrospective performance driven by post-baseline metrics should not be equated with clinical readiness. Broad prospective validation, multicenter testing, and feature expansions specifically including baseline CD4 counts and viral load measurements remain necessary steps before clinical integration.**

Keywords: **HIV, Antiretroviral Therapy, Mortality Prediction, Explainable Machine Learning, XGBoost**

I. Introduction

Despite substantial progress in HIV prevention, diagnosis, and treatment, Human Immunodeficiency Virus (HIV) continues to impose a considerable global health burden. Data from the Joint United Nations Programme on HIV/AIDS (UNAIDS) indicate that approximately 40.8 million people were living with HIV worldwide in 2024. Although prevention and treatment coverage has expanded, HIV-related conditions remain an important cause of illness and death, particularly in low- and middle-income countries [1]. The introduction of antiretroviral therapy (ART) has markedly improved patient outcomes by suppressing viral replication, supporting immune recovery, and reducing HIV-related mortality [2]. Nevertheless, deaths among people living with HIV (PLHIV) receiving ART remain associated with delayed treatment initiation, opportunistic infections, poor adherence, and advanced clinical disease [3], [4]. Consequently, early identification of individuals at elevated risk is essential to facilitate rapid therapeutic management and foster more effective, data-driven clinical decision-making.

Mortality among PLHIV is a multifactorial outcome resulting from interactions among demographic, clinical, immunological, behavioral, and treatment-related factors. Previous studies have linked mortality among PLHIV to several factors, including older age, low CD4 cell counts, high viral load, opportunistic infections, poor ART adherence, advanced WHO clinical stage, and comorbid conditions [5], [6], [7]. In addition to clinical conditions, delayed HIV diagnosis, socioeconomic disadvantage, limited access to healthcare services, and interruptions in HIV care have been linked to compromised treatment efficacy and a heightened risk of mortality among PLHIV [8], [9]. Although conventional statistical approaches remain widely used, their predictive capability may be limited when relationships among clinical, demographic, and treatment-related variables are complex, nonlinear,

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

and interactive [10]. Furthermore, real-world HIV clinical datasets frequently contain missing values, heterogeneous patient characteristics, and imbalanced class distributions, which may reduce the reliability and predictive performance of traditional analytical models [11], [12]. These challenges underscore the need for more robust predictive approaches capable of effectively modeling complex clinical data and capturing nuanced patient risk profiles [13].

Machine learning has advanced clinical prediction by analyzing complex patient data and nonlinear relationships more effectively than conventional statistical methods [10]. Within the domain of HIV research, supervised machine learning algorithms are increasingly leveraged to model various clinical outcomes, notably mortality risk estimation, disease progression, ART adherence, and viral load suppression. Commonly used approaches include logistic regression, decision trees, random forests, support vector machines (SVM), artificial neural networks (ANN), and Extreme Gradient Boosting (XGBoost) [14], [15]. Because it is straightforward to implement and produces easily interpretable results, logistic regression often serves as the reference model in predictive studies. However, its predictive capability may be limited when relationships among multiple clinical variables are nonlinear or involve complex interactions [16]. Strong predictive results have been reported for random forest and XGBoost because these ensemble algorithms can capture nonlinear patterns and complex interactions within clinical data. Their ensemble and regularization mechanisms may also improve model generalization when appropriately configured and validated [17]. Several comparative studies have further reported that XGBoost consistently outperforms or achieves competitive performance compared with conventional statistical methods and other machine learning algorithms in predicting mortality and HIV-related clinical outcomes [18]. The increasing adoption of explainable artificial intelligence (XAI) has improved the interpretability of machine learning models. Among the available approaches, SHapley Additive exPlanations (SHAP) has become one of the most widely used methods for quantifying feature contributions and supporting clinical interpretation of model predictions [19]. The Synthetic Minority Over-sampling Technique (SMOTE) was used to rebalance the training data, while Optuna automated the search for suitable hyperparameter values and reduced the need for manual tuning [20]. Although machine learning applications in HIV prognosis have expanded significantly, the existing literature has a notable methodological vulnerability. A substantial portion of published prediction models inadvertently integrates post-baseline clinical variables—such as subsequent healthcare encounters or total observation duration—into initial risk estimation tasks. This practice causes severe target leakage (or look-ahead bias), which artificially inflates retrospective performance metrics while obscuring the authentic predictive value of indicators accessible at the exact time of antiretroviral therapy (ART) initiation (Day 0). Moreover, explainable machine learning frameworks validated on real-world clinical cohorts from low- and middle-income countries remain scarce, even though these settings routinely contend with incomplete baseline laboratory panels and elevated loss-to-follow-up rates. To address these gaps, this study presents an interpretable machine learning framework centered on an optimized Extreme Gradient Boosting (XGBoost) model designed to forecast

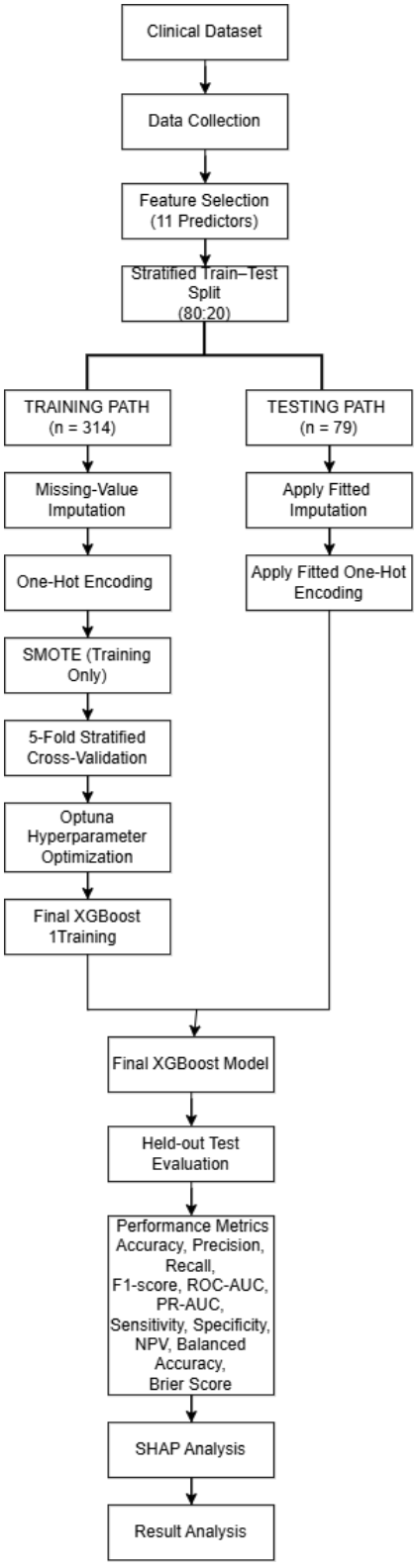


Fig. 1. Proposed Explainable Machine-Learning Workflow for Mortality Prediction Among PLHIV Receiving ART.

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

all-cause mortality among people living with HIV (PLHIV) who are commencing ART. Ultimately, this research strives to evaluate the true prognostic capacity of a strictly leakage-safe classifier built on Day 0 baseline records, systematically contrasting it with a sensitivity analysis model to illustrate how post-baseline temporal variables introduce confounding distortions.

Overall, this work advances the field through four distinct contributions. From a methodological standpoint, we establish a strictly leakage-safe primary baseline model driven exclusively by features available at ART initiation (Day 0), isolating post-baseline metrics like follow-up duration in a comparative sensitivity setup to quantify look-ahead distortion in HIV mortality modeling. Computationally, we deploy a standardized end-to-end framework that incorporates median and mode imputation, class rebalancing via SMOTE, hyperparameter tuning using the Optuna Tree-structured Parzen Estimator (TPE), and tree-based gradient boosting. To ensure model transparency and practical utility, we embed SHapley Additive exPlanations (SHAP), offering both broad feature importance rankings and patient-level risk attributions to facilitate clinical auditing. Finally, we provide empirical evidence gathered from a retrospective cohort at a regional referral hospital in Bali, Indonesia, delivering valuable context-specific insights tailored to resource-constrained clinical settings.

II. Materials and Method

The complete machine-learning workflow used in this study is illustrated in Fig. 1. After selecting the 11 predictors, the dataset was divided into training and held-out testing sets using a stratified 80:20 split. Missing-value imputation and one-hot encoding were fitted exclusively on the training data and subsequently applied to the held-out test set without refitting. SMOTE was applied only to the training data. Model development used stratified cross-validation and Optuna hyperparameter optimization, followed by final XGBoost training using the selected hyperparameters. The held-out test set was then used for independent performance evaluation, and SHAP analysis was performed to interpret the final model. Each stage of the workflow is described in the subsequent subsections.

A. Dataset

The dataset analyzed in this research comprised retrospective clinical records of PLHIV from RSUD Sanjiwani, Gianyar, Bali, Indonesia, covering 2006 to 2023. The intended clinical decision point is at ART initiation (Day 0). The model is designed to estimate mortality risk using predictor information available at or before ART initiation. The goal is to help clinicians identify high-risk PLHIV at ART initiation, enabling early targeted interventions such as closer monitoring or intensified prophylaxis. The records were generated during routine HIV care and included demographic, socioeconomic, clinical, and treatment-related information relevant to patient outcomes. This dataset was utilized to construct an explainable machine learning model aimed at forecasting mortality in PLHIV undergoing ART. To account for temporal changes in HIV treatment policies and clinical practice over the 2006–2023 study period, the calendar year of ART initiation was additionally grouped into three treatment-policy eras: Early Scale-up Era (2006–2012), SUFA Expansion Era (2013–2017), and Test and Treat Era (2018–2023). Mortality distribution was subsequently examined across these eras to assess temporal variation in patient outcomes. A total of 1,600 medical records were initially identified. After removing 21 duplicate records, 1,579 records remained and were assessed for eligibility. Based on the predefined eligibility criteria, 1,186 records were excluded: 557 patients were referred to another healthcare facility, 359 were lost to follow-up (LTFU), 264 had unknown treatment status, three discontinued ART, two had not initiated ART, and one had been transferred from another healthcare facility. Consequently, 393 eligible patients receiving ART were included in the final cohort. The participant screening and selection process is summarized in Fig. 2.

Table 1. Missing Values in the Predictor Variables of the Final Analytical Dataset

Variabel	Missing (n)	Missing (%)
Employment status	3	0.76
Marital status	62	15.78
Functional status	6	1.53
Clinical stage	35	8.91

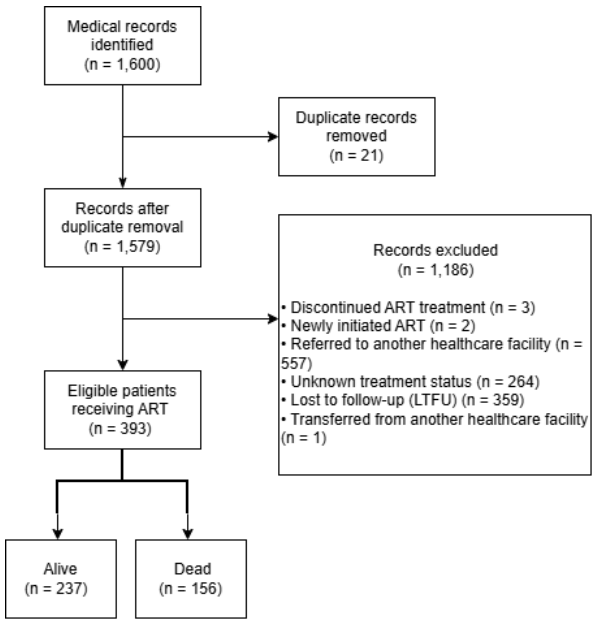


Fig. 2. Participant flow diagram showing record screening, exclusions, and selection of the final analytical cohort.

The prediction target was mortality status, formulated as a binary classification problem, with 237 records representing surviving patients (class 0) and 156 records representing deceased patients (class 1). This distribution showed a moderate class imbalance between the two outcome categories. After data preparation and feature selection, the final analysis included 12 variables: 11

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

predictors and 1 target variable. The predictors were selected based on the study objectives, data completeness, and clinical relevance. These variables represented demographic, socioeconomic, clinical, and treatment-related characteristics considered relevant to mortality prediction while maintaining data quality and consistency.

The predictor variables comprised age, gender, education level, employment status, marital status, functional status, clinical stage, HIV transmission risk factor, ART regimen, opportunistic infection, and ART duration. Age was treated as a numerical predictor representing chronological age at baseline in years, whereas gender was treated as a categorical predictor representing biological sex. Education level represented the highest completed formal education, employment status represented occupational condition at baseline, and marital status represented the patient's marital status category at baseline. Functional status was treated as a categorical predictor representing the patient's physical capacity and was categorized as working (normal activity), ambulatory (able to perform self-care and bedridden for less than 50% of the day), or bedridden (disabled and bedridden for more than 50% of the day). Clinical stage represented HIV disease severity and was categorized according to WHO clinical stages I, II, III, and IV, corresponding to increasing clinical severity. HIV transmission risk factor represented the primary documented route of HIV transmission. ART regimen represented regimen stability during follow-up, whereas opportunistic infection was treated as a binary predictor indicating the presence absence or of an active opportunistic infection documented prior to or at ART initiation. ART duration was treated as a numerical predictor and represented the total cumulative follow-up duration on ART from treatment initiation until death or the last recorded clinical observation, measured in months. Mortality status was the binary target variable, categorized as alive (class 0) or deceased (class 1) at the end of the observation period. Follow-up duration was defined as the elapsed time, in months, from ART initiation to death or the last recorded clinical observation. To eliminate target leakage and maintain a strictly baseline-oriented prediction model at ART initiation (Day 0), follow-up duration was omitted from the set of primary predictors. Nevertheless, this temporal metric was isolated and examined within a dedicated sensitivity analysis, allowing us to measure the precise magnitude of look-ahead bias introduced by post-baseline variables when modeling retrospective HIV mortality.

B. Data Collection

This study used retrospectively collected medical records from RSUD Sanjiwani, Gianyar, Bali, Indonesia. Data extraction was performed after obtaining institutional permission and ethical approval. The records contained demographic, socioeconomic, clinical, and treatment-related information on PLHIV receiving ART. To protect patient privacy and confidentiality, all personally identifiable information was removed before data processing and analysis.

Patient records were screened using predefined inclusion and exclusion criteria. Only records of PLHIV receiving ART with complete mortality outcome data were included. We removed duplicate entries and records that did not meet the eligibility criteria. After screening, 393 patient records remained for model development and evaluation. Ethical approval was obtained from the Research Ethics Committee of the Faculty of Medicine, Udayana University (Ethical Clearance No. 1567/UN14.2.2.VII.14/LT/2024). All records were anonymized before data processing and analysis.

C. Data Processing

Prior to model development, the clinical dataset underwent several preprocessing procedures to optimize data quality and compatibility with the machine learning model [21]. Missing data were assessed for each variable in the final analytical dataset (n = 393). Missing values were identified in employment status (0.76%), marital status (15.78%), functional status (1.53%), and clinical stage (8.91%), whereas no missing values were observed in the remaining variables. Variable-specific missingness is presented in Table 1. Numerical and categorical variables were handled separately within the preprocessing pipeline. Numerical variables were imputed using the median, whereas categorical variables were imputed using the most frequent category (mode). To prevent information leakage, we performed the train-test split before imputation, and estimated all imputation parameters exclusively from the training data. The fitted imputation transformations were subsequently applied to the held-out test set without refitting. Categorical predictors were subsequently transformed using one-hot encoding within the preprocessing pipeline. Following feature selection, the data were organized into predictor features (X) and a binary target variable (y), where the outcome represented survival (0) or mortality (1).

The 11 predictors were predefined before model development based on the study objectives, data availability, completeness, and clinical relevance. Although CD4 count, viral load (VL), tuberculosis (TB) status, adherence, and comorbidities are clinically important, these variables could not be consistently incorporated into the final predictor set because of structural limitations in the retrospective dataset. CD4, VL, and TB information was maintained in separate administrative or laboratory records that could not be reliably linked to the primary dataset. In addition, VL monitoring practices changed substantially during the 2006–2023 study period, resulting in structurally incomplete measurements across treatment eras. Adherence was not included because it represents a longitudinal behavior observed after ART initiation rather than information available at the baseline prediction point, while comorbidities were not systematically structured or coded at ART initiation. Categorical variables were reviewed according to their measurement scales. Sex, employment status, marital status, HIV risk factor, and ART regimen were treated as nominal categorical variables without an inherent ordinal structure, while opportunistic infection status was treated as a binary categorical variable. One-Hot Encoding (OHE) was used

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

to avoid introducing artificial numerical ordering among categories. OHE was implemented within the leakage-safe preprocessing pipeline after the train-test split, with the encoder fitted exclusively on the training data and subsequently applied to the held-out test set without refitting.

1. One-Hot Encoding

To represent nominal categorical variables without introducing artificial ordinal relationships, One-Hot Encoding (OHE) was applied within the preprocessing pipeline. Each category was converted into a separate binary indicator variable. The mathematical representation of OHE is expressed in Eq. (1) [22] .

$$x_{ij} = \begin{cases} 1, if\ sample\ i\ belongs\ to\ category\ j \\ 0, otherwise \end{cases} \quad (1)$$

Where x_{ij} denotes the binary indicator corresponding to observation i and category j. A value of 1 indicates membership of observation i in category j, whereas a value of 0 indicates that the observation does not belong to that category.

2. SMOTE

Following the train-test partition, class imbalance in the training dataset was addressed using the Synthetic Minority Over-sampling Technique (SMOTE). Importantly, oversampling was restricted to the training data so that synthetic observations did not influence the independent testing set. SMOTE generates new minority-class observations by interpolating between an existing minority observation and one of its neighboring minority observations. The synthetic sample generation process is expressed in Eq. (2) [23].

$$x_{new} = x_i + \delta \cdot (x_j - x_i), \quad (2)$$

Where x_{new} represents the generated synthetic minority-class observation x_i, denotes an original minority-class observation, x_j represents one of its K-nearest neighbors from the same minority class, and $\delta \in [0,1]$ is a randomly selected interpolation coefficient.

3. Hyperparameter Optimization Using Optuna

The XGBoost classifier was optimized through automated hyperparameter tuning using Optuna to improve predictive performance [24]. Appropriate hyperparameter selection is important because model settings can substantially affect predictive performance [25]. Compared with manual tuning and exhaustive grid search, Optuna can explore a predefined hyperparameter search space more efficiently [26]. Hyperparameter optimization was performed separately for the Primary Leakage-Safe Model and the Exploratory Sensitivity Model using a Tree-structured Parzen Estimator (TPE) sampler. A total of 300 trials were conducted for each model, with the optimization objective defined as maximizing the mean ROC-AUC obtained using 5-fold stratified cross-validation on the training data. The folds were shuffled using a fixed random seed of 42. The search ranges were 100–300 for n_estimators, 3–8 for max_depth, 0.01–0.30 for learning_rate, 0.70–1.00 for subsample, 0.70–1.00 for colsample_bytree, 0–5 for gamma, and 1–10 for min_child_weight. The TPE sampler and XGBoost classifier also used a random seed of 42 to support reproducibility. No pruning or early-stopping strategy was applied, and the optimization process terminated after completion of all 300 trials. Hyperparameter tuning was performed exclusively using the training data, while the held-out test set was kept separate and used only for final model evaluation. The search ranges and final selected hyperparameters for both models are presented in Table 2.

Table 2 XGBoost Hyperparameter Search Ranges and Final Selected Values for the Primary and Exploratory Models

Hyperparameter	Search Range	Primary Leakage-Safe Model	Exploratory Sensitivity Mode
n_estimators	100-300	116	190
max_depth	3-8	7	5
learning_rate	0.01-0.30	0.08701	0.07937
subsample	0.70-1.00	0.76285	0.93185
colsample_bytree	0.70-1.00	0.97404	0.74611
gamma	0-5	0.28366	0.78037
min_child_weight	1-10	1	5

The hyperparameter optimization produced different configurations for the Primary Leakage-Safe Model and the Exploratory Sensitivity Model, reflecting differences in the predictor sets used by the two models. For the Primary Leakage-Safe Model, which excluded Follow-up Duration, Optuna selected 116 estimators, a maximum tree depth of 7, a learning rate of 0.08701, a subsample ratio of 0.76285, a colsample_bytree ratio of 0.97404, a gamma value of 0.28366, and a min_child_weight of 1. In comparison, the Exploratory Sensitivity Model, which included Follow-up Duration, selected 190 estimators, a maximum tree depth of 5, a learning rate of 0.07937, a subsample ratio of 0.93185, a colsample_bytree ratio of 0.74611, a gamma value of 0.78037, and a min_child_weight of 5. These differences indicate that the optimal XGBoost configuration varied according to whether Follow-up Duration was included in the predictor set. Both configurations were selected independently by Optuna from the same predefined search space using the mean ROC-AUC from 5-fold stratified cross-validation on the training data, without using information from the held-out test set. To describe the optimization mechanism underlying the TPE sampler, the expected improvement criterion can be formulated as follows. In the conventional minimization formulation of TPE, the improvement associated with a candidate hyperparameter configuration can be represented by Eq. (3) [27].

4. TPE Density Function:

$$p(\theta|y) = \begin{cases} l(\theta), & y < y^* \\ g(\theta), & y \geq y^* \end{cases} \quad (3)$$

Where $p(\theta|y)$ denotes the conditional probability density of a hyperparameter configuration θ given the observed

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

objective value $y;\theta$ represents a candidate hyperparameter configuration; y denotes the objective value obtained from evaluating that configuration; y^* represents the threshold used to separate better-performing and remaining trials; $l(\theta)$ denotes the probability density of hyperparameter configurations associated with objective values below y^*, $g(\theta)$ and denotes the probability density of configurations associated with objective values equal to or greater than y^*.

5. Expected Improvement:

Based on this formulation, the expected improvement of a candidate hyperparameter configuration is obtained by integrating the improvement over the conditional distribution of the objective value, as expressed in Eq. (4) [27].

$$EI_{y^*}(x) \propto \left[\gamma + \frac{g(x)}{l(x)}(1-\gamma)\right] \quad (4)$$

where $EI_{y^*}(x)$ denotes the expected improvement associated with hyperparameter configuration x, y^* represents the selected objective-value threshold $l(x)$, denotes the probability density of hyperparameter configurations associated with objective values below $y^*, g(x)$ represents the probability density of the remaining configurations, and γ denotes the probability mass assigned to the region below y^*.

6. XGBoost Model Development

An XGBoost classifier was developed to predict mortality among PLHIV receiving ART. The XGBoost algorithm generates a sequence of decision trees, where each successive tree is trained to correct the residual errors of the preceding trees, thereby enhancing predictive performance [28]. XGBoost has been widely used in healthcare prediction because it can model nonlinear relationships, accommodate high-dimensional clinical data, and apply regularization to reduce overfitting [29]. The prediction generated by XGBoost is obtained by combining the outputs of multiple regression trees in an additive ensemble. The XGBoost prediction function is expressed in Eq. (5) [30].

$$\hat{y}_i = \phi(x_i) = \sum_{\kappa=1}^{\mathrm{K}} f_k(x_i), \quad f_k \in \mathcal{F} \quad (5)$$

where $\hat{y}_i$ represents the predicted output for observation i, $\mathcal{K}$ denotes the total number of trees in the ensemble $f_k(x_i)$, represents the prediction generated by the $\hbar$ -th tree for observation x_i, and $\mathcal{F}$ denotes the space of regression trees.

- Objective Function

Model training is performed by minimizing an objective function that combines prediction error with a penalty for model complexity. The regularized XGBoost objective function is defined in Eq. (6) [30].

$$\mathcal{L}(\phi) = \sum_{i=l}^{n} l\,(\hat{y}_i, y_i) + \sum_{\kappa=1}^{\mathrm{K}} \Omega(f_k) \quad (6)$$

where n represents the number of observations $l(\hat{y}_i, y_i)$, is the differentiable loss function measuring the discrepancy between the observed outcome y_i and predicted output $\hat{y}_i$ and $\Omega(f_k)$ represents the regularization term associated with tree f_k.

- Regularization

To control tree complexity and reduce overfitting, XGBoost incorporates a regularization penalty into its objective function. The regularization term is expressed in Eq. (7) [30].

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda||\omega||^2 \quad (7)$$

where T denotes the number of terminal leaves in the tree, ω represents the vector of leaf weights, γ controls the penalty associated with adding terminal leaves, and λ controls the regularization applied to the leaf weights.

Second-order approximation.

- Second-Order Taylor Expansion

At boosting iteration, XGBoost adds a new tree to the prediction obtained from the previous iteration. The objective function combines the prediction loss with a regularization term to control model complexity, as expressed in Eq. (8) [30].

$$\mathcal{L}^{(\mathrm{t})} = \sum_{\mathrm{i}=1}^{\mathrm{n}} l(\,y_i, \hat{y}_i{}^{(\mathrm{t}-1)} + f_{\mathrm{t}}(x_i)) + \Omega(f_{\mathrm{t}}) \quad (8)$$

where $\mathcal{L}^{(\mathrm{t})}$ denotes the objective function at boosting iteration $t, \hat{y}_i{}^{(\mathrm{t}-1)}$ represents the prediction obtained from the previous iteration, $f_{\mathrm{t}}(x_i))$ denotes the contribution of the newly added tree, and $\Omega(f_{\mathrm{t}})$ represents its regularization term.

- Second-Order Taylor Approximation

To facilitate efficient optimization, the objective function in Eq. (7) is approximated using a second-order Taylor expansion around the prediction obtained from the previous boosting iteration. After omitting constant terms, the approximation is expressed in Eq. (9) [30].

$$\mathcal{L}^{(\mathrm{t})} \approx \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i) + \Omega(f_t)\right] \quad (9)$$

where g_i and represent the first-order gradient and second-order Hessian of the loss function, respectively, while $f_t(x_i)$ denotes the contribution of the tree added at iteration t . This approximation enables XGBoost to use both gradient and curvature information during model optimization.Where, First-Order Gradient and Second-Order Hessian The first-order gradient and second-order Hessian used in the Taylor approximation are calculated from the loss function evaluated at the prediction from the previous boosting iteration, as defined in Eq. (10) [30].

$$g_i = \partial_{\hat{y}} l(y_i, \hat{y})|_{\hat{y}_i{}^{(t-1)}} \; and \; h_i = \partial^2_{\hat{y}} l(y_i, \hat{y})|_{\hat{y}_i{}^{(t-1)}} \quad (10)$$

where g_i measures the first-order change in the loss with respect to the model prediction, whereas h_i represents the corresponding second-order curvature information.

7. Optimal leaf weight

For a fixed tree structure, the optimal weight assigned to each terminal leaf is determined from the accumulated first- and second-order gradient statistics. The optimal leaf weight is calculated using Eq. (11) [30].

$$\omega_j^* = -\frac{G_j}{H_j + \lambda} \quad (11)$$

where ω_j^* denotes the optimal weight of leaf $j, G_j = \sum i \in l_j G_i$ represents the set of observations assigned to that leaf $j, H_j = \sum i \in l_j h_i$ represents the corresponding sum of second-order Hessian values Ij, denotes the set of

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

observations assigned to leaf j, and λ is the regularization parameter applied to the leaf weights.

8. Split Gain

During tree construction, candidate splits are evaluated according to the reduction in the regularized objective function. The gain obtained by dividing a parent node into left and right child nodes is expressed in Eq. (12) [31].

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \quad (12)$$

where G_L and G_R denote the sums of first-order gradients for observations assigned to the left and right child nodes, respectively H_L and H_R denote the corresponding sums of second-order gradients λ is the regularization parameter and γ represents the complexity penalty associated with introducing an additional split.

9. SHAP

- Additive Feature Attribution

To interpret XGBoost predictions, SHapley Additive exPlanations (SHAP) were used to quantify the contribution of individual predictors. SHAP represents an explanation as an additive combination of feature contributions. The additive feature-attribution model is expressed in Eq. (13) [32].

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z_j' \quad (13)$$

where $g(z')$ denotes the explanation model, ϕ_0 represents the baseline or expected model output, M is the number of input features z_j', indicates the presence of a feature j in the simplified input representation, and represents the contribution assigned to that feature.

- SHAP value

The contribution of an individual feature is calculated using the Shapley-value formulation, which considers its marginal contribution across possible subsets of the remaining features. The SHAP value for feature is expressed in Eq. (14) [33].

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(M-|S|-1)!}{M!}\left[f(S \cup \{j\}) - f(S)\right] \quad (14)$$

where $\mathcal{F}$ denotes the complete set of features, $\mathcal{S}$ represents a subset of features that does not contain feature j, $\mathcal{M}$ is the total number of features, and $f(S \cup \{j\}) - f(S)$ represents the marginal contribution of a feature when it is added to a subset. The magnitude of ϕ_j reflects the strength of the feature contribution, while its sign indicates the direction of its influence on the model output.

10. Evaluation Metrics

- Accuracy

Classification performance was evaluated using complementary metrics derived from the confusion matrix. Overall classification correctness was quantified using accuracy, as expressed in Eq. (15) [34].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (15)$$

where $TP, TN, FP, and\ FN$ denotes true positives, true negatives, false positives, and false negatives. In this study, mortality was treated as the positive class and survival as the negative class.

- Precision

The reliability of positive mortality predictions was assessed using precision, which represents the proportion of predicted positive cases that were correctly classified. Precision is defined in Eq. (16) [34].

$$Precision = \frac{TP}{TP + FP} \quad (16)$$

where TP represents mortality cases correctly classified as mortality and FP represents surviving patients incorrectly classified as mortality cases.

- Recall / Sensitivity

The ability of the classifier to identify actual mortality cases was evaluated using recall, which is equivalent to sensitivity for the positive class. Sensitivity is expressed in Eq. (17) [34].

$$Recall\ /\ Sensitivity = \frac{TP}{TP + FN} \quad (17)$$

where TP denotes correctly identified mortality cases and FN represents mortality cases incorrectly classified as survival.

- Specificity

The ability of the model to correctly identify patients who survived was quantified using specificity, as defined in Eq. (18) [34].

$$Specificity = \frac{TN}{TN + FP} \quad (18)$$

where TN denotes surviving patients correctly classified as survival and FP represents surviving patients incorrectly classified as mortality.

- Negative Predictive Value (NPV)

The reliability of negative predictions was additionally assessed using the negative predictive value (NPV), which is expressed in Eq. (19) [35].

$$NPV = \frac{TN}{TN + FN} \quad (19)$$

where TN represents correctly predicted survival cases and denotes mortality cases incorrectly predicted as survival.

- F1-score

Because precision and recall capture different aspects of positive-class performance, the F1-score was used to summarize both measures through their harmonic mean, as expressed in Eq. (20) [34].

$$F1 = 2\frac{Precision \times Recall}{Precision + Recall} \quad (20)$$

where $Precision$ represents the correctness of positive predictions and $Recall$ represents the ability to identify actual positive cases. A higher F1-score indicates a better balance between these two measures.

- ROC-AUC

Model discrimination across classification thresholds was assessed using the receiver operating characteristic (ROC) curve and the area under the ROC curve (ROC-AUC). The ROC curve relates the true-positive rate to the false-positive rate across decision thresholds, while ROC-AUC summarizes the overall discriminatory ability of the model. The area under the ROC curve can be expressed in Eq. (21) [34].

$$ROC\text{–}AUC = \int_0^1 TPR(FPR)\,d(FPR) \quad (21)$$

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

Algorithm 1: Complete computation workflow

Input: Clinical dataset D = {(xi, yi)}i=1,...,N

Train-test ratio r = 80:20

Number of Optuna trials T = 300

Number of cross-validation folds K = 5

- Output: Optimized XGBoost model M*

Predicted mortality outcomes ŷ

Predicted probabilities p

Performance metrics E

SHAP explanations Φ

6. Define predictor matrix X and binary mortality outcome y, where y = 0 represents survival and y = 1 represents mortality.
7. Split (X, y) into stratified training and held-out testing sets (Xtrain, ytrain) and (Xtest, ytest) using an 80:20 ratio.
8. Identify numerical and categorical predictors.
9. Fit median imputation to numerical predictors using Xtrain only.
10. Fit mode imputation to categorical predictors using Xtrain only.
11. Apply the fitted imputers to Xtrain and Xtest without refitting.
12. Fit One-Hot Encoding (OHE) to categorical predictors using Xtrain only.
13. Transform Xtrain and Xtest using the fitted encoder.
14. Apply SMOTE exclusively to the preprocessed training data to obtain the balanced training set (Xtrain_bal, ytrain_bal).
15. Initialize Optuna using the Tree-structured Parzen Estimator (TPE) sampler.
16. For trial t = 1 to T do
17. Sample an XGBoost hyperparameter configuration θt.
18. Initialize K-fold stratified cross-validation.
19. For fold k = 1 to K do
20. Train XGBoost using θt on the training portion of fold k.
21. Predict mortality probabilities on the validation portion.
22. Compute ROC-AUCk.
23. end for
24. Compute the mean validation ROC-AUC across K folds.
25. end for
26. Select θ* that maximizes the mean cross-validation ROC-AUC.
27. Train the final XGBoost classifier M* using θ*.
28. Generate class predictions ŷ and mortality probabilities p on the held-out testing set.
29. Evaluate M* using Accuracy, Precision, Recall/Sensitivity, Specificity, NPV, F1-score, ROC-AUC, and Brier Score.
1. Initialize SHAP TreeExplainer using M*.
2. Compute SHAP values Φ for the held-out testing observations.
3. Determine global predictor importance and generate
4. Local SHAP explanations.
5. Return M*, ŷ, p, E, Φ.

where TPR represents the true-positive rate (sensitivity), denotes the false-positive rate, defined as $FPR = FP/(FP + TN)$, and $d(FPR)$ denotes integration with respect to the false-positive-rate scale. Larger ROC-AUC values indicate a greater ability to discriminate between mortality and survival across classification thresholds.

- Brier Score Calibration

In addition to discrimination, the agreement between predicted probabilities and observed outcomes was assessed using probability calibration. The Brier score quantifies the mean squared difference between the predicted probability and the observed binary outcome and is expressed in Eq. (22) [36].

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \quad (22)$$

Where $\mathcal{N}$ denotes the number of evaluated observations (p_i, represents the predicted probability of mortality for observation , and denotes the corresponding observed binary outcome, coded as 1 for mortality and 0 for survival. Lower Brier scores indicate better agreement between predicted probabilities and observed outcomes.

11. Algorithm End-to-End Explainable XGBoost Framework for Mortality Prediction

To provide a systematic overview of the proposed methodology, the complete computational workflow is summarized in algorithm 1. The pseudocode describes the sequential implementation of data preprocessing, stratified train-test splitting, SMOTE-based class rebalancing, Optuna-TPE hyperparameter optimization with cross-validation, XGBoost model training and testing, performance evaluation, and SHAP-based model interpretation **Algorithm 1** End-to-End Explainable XGBoost Framework for Mortality Prediction. As shown in Algorithm 1, the held-out testing dataset was kept separate from model development and hyperparameter optimization and was used only for final performance evaluation and SHAP-based interpretation.

B. Statistical Analysis

Five classification metrics were used to evaluate the predictive capability of the proposed model: accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC). Accuracy quantifies the share of all observations that the model classified correctly. Precision indicates how many instances predicted as positive were actually positive, whereas recall reflects the model's ability to detect positive instances from all true positive cases. The F1-score combines precision and recall into a single indicator by calculating their harmonic mean, thereby providing an overall measure of classification performance [37]. ROC-AUC summarizes how well the classifier separates

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

mortality from survival over a range of classification thresholds, with larger values reflecting stronger discrimination. To obtain a more robust estimate of model performance, repeated stratified cross-validation was performed on the training data using five folds and ten repetitions with a fixed random seed of 42. Stratification was used to preserve the class distribution across folds. Performance was summarized using the mean, standard deviation (SD), and 95% confidence interval for accuracy, precision, recall, F1-score, and ROC-AUC. The held-out test set (20%, n = 79) remained separate from the cross-validation procedure and was used only for final model evaluation.

Additionally, an ablation analysis was performed to assess the individual contributions of Optuna and SMOTE by comparing baseline XGBoost, XGBoost with Optuna, XGBoost with SMOTE, and XGBoost with both Optuna and SMOTE. Logistic regression, random forest, and gradient boosting were also evaluated as benchmark models using the same evaluation setting. Additional clinical performance measures were calculated on the held-out test set, including precision–recall area under the curve (PR-AUC), sensitivity, specificity, negative predictive value (NPV), balanced accuracy, and Brier score. Bootstrap-based 95% confidence intervals were estimated for these performance measures. An exploratory threshold analysis was performed across probability thresholds from 0.30 to 0.70 to assess the sensitivity–specificity trade-off. Because no clinically validated decision threshold was prespecified, the threshold analysis was considered exploratory. Model calibration was additionally evaluated using a calibration plot.

To improve the interpretability of the optimized XGBoost model, SHAP was applied to explain how individual features contributed to its predictions [33]. SHAP is an XAI method derived from cooperative game theory that assigns each feature a value representing its contribution to a specific prediction. SHAP values were calculated using TreeExplainer on the held-out test set comprising 79 samples. Global predictor importance was quantified using the mean absolute SHAP value across all test samples, where larger values indicate a greater overall contribution of a predictor to the model output. For one-hot encoded categorical predictors, SHAP values corresponding to the encoded categories were interpreted in relation to their original predictor. In addition to the global SHAP summary plot, dependence and category-level analyses were used to further examine the effects of the major predictors. In this study, SHAP was applied to interpret the model at both the overall and individual-prediction levels. Global interpretation was performed using a SHAP summary plot to identify the overall importance, magnitude, and direction of predictor contributions across the testing dataset. For local explainability, SHAP waterfall plots were generated for correctly classified, false-positive, and false-negative predictions. To avoid preferential selection of favorable cases, representative observations from the correctly classified and false-positive groups were selected as those with predicted probabilities closest to the median probability within each respective group. Because only one false-negative prediction occurred in the held-out test set, that case was necessarily included. Waterfall plots were used to illustrate the direction and magnitude of individual predictor contributions to each prediction. This approach improved model transparency and facilitated the interpretation of mortality predictions among PLHIV receiving ART.

III. Results

A. Accuracy

Over the 2006–2023 study period, overall mortality exhibited a steady downward trajectory across successive treatment policy eras. During the initial Early Scale-up Era (2006–2012), the fatality rate reached its peak at 42.55% (20/47). This figure saw a modest decline to 41.25% (99/240) in the subsequent SUFA Expansion Era (2013–2017), before eventually reaching a minimum of 34.91% (37/106) throughout the Test and Treat Era (2018–2023). Notably, this step-down in mortality aligns with the nationwide policy evolution that favored earlier ART initiation without relying on baseline CD4 count thresholds. Baseline characteristics were examined according to mortality outcome in the original cohort before SMOTE. The mean age was similar between surviving and deceased patients, whereas differences were more apparent in functional status and clinical stage. Advanced clinical disease was more frequent among deceased patients, while HIV risk factor and opportunistic infection status showed relatively similar distributions between the two outcome groups. Baseline

Table 3. Baseline Characteristics of the Study Cohort Stratified by Mortality Outcome

Characteristic	Alive (n=237)	Dead (n=156)
Age, years, mean ± SD	35.91 ± 11.15	35.55 ± 11.23
Functional status, Ambulatory	35 (14.77%)	26 (16.67%)
Functional status, Bedridden	7 (2.95%)	23 (14.74%)
Functional status, Working	195 (82.28%)	107 (68.59%)
WHO clinical stage I	74 (31.22%)	44 (28.21%)
WHO clinical stage II	111 (46.84%)	56 (35.90%)
WHO clinical stage III	44 (18.57%)	34 (21.79%)
WHO clinical stage IV	8 (3.38%)	22 (14.10%)
Opportunistic infection, category 0	21 (8.86%)	12 (7.69%)
Opportunistic infection, category 1	216 (91.14%)	144 (92.31%)

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

characteristics stratified by mortality outcome are summarized in Table 3.

Table 4. Classification Performance of the Primary Leakage-Safe and Exploratory Sensitivity XGBoost Models on the Held-Out Testing Dataset

Outcome Class	Precision	Recall	F1-score	Support
Primary Leakage-Safe Model				
Alive (0)	0.66	0.69	0.67	48
Dead (1)	0.48	0.45	0.47	31
Macro Average	0.57	0.57	0.57	79
Weighted Average	0.59	0.59	0.59	79
Exploratory Sensitivity Model				
Alive (0)	0.98	0.85	0.91	48
Dead (1)	0.81	0.97	0.88	31
Macro Average	0.89	0.91	0.90	79
Weighted Average	0.91	0.90	0.90	79

A held-out set of 79 patient records was used to assess the predictive capability of the explainable machine learning model. Before SMOTE, the training set consisted of 189 surviving patients and 125 deceased patients. After SMOTE was applied exclusively to the training set, the minority mortality class increased from 125 to 189 samples, resulting in a balanced training distribution of 189 surviving and 189 deceased samples. The held-out test set was not subjected to SMOTE and remained unchanged for independent model evaluation. Evaluation of the exploratory sensitivity XGBoost model incorporating follow-up duration yielded an accuracy of 89.87% and an ROC-AUC of 0.9546 on the independent testing set. Precision, recall, and F1-score were also computed to assess how effectively the model differentiated mortality from survival among PLHIV receiving ART. The class-specific performance results are presented in Fig. 3. Baseline characteristics stratified by mortality outcome are summarized in Table 3. These descriptive characteristics were calculated from the original cohort before SMOTE. The mean age was comparable between surviving and deceased patients, with values of 35.91 ± 11.15 years and 35.55 ± 11.23 years, respectively. More pronounced differences were observed in functional status. Bedridden functional status was more frequent among deceased patients than among surviving patients (14.74% vs. 2.95%), whereas working functional status was more common among surviving patients (82.28%) than among deceased patients (68.59%). Differences were also evident across WHO clinical stages. Stage IV was substantially more frequent among deceased patients than among surviving patients (14.10% vs. 3.38%), while stages I and II were proportionally more common among survivors. Stage III was also slightly more frequent in the mortality group (21.79%) than in the survival group (18.57%). In contrast, opportunistic infection status showed relatively similar distributions between the two outcome groups, with category 1 observed in 91.14% of surviving patients and 92.31% of deceased patients. Overall, the descriptive findings suggest that functional status and WHO clinical stage showed clearer differences between mortality groups than age or opportunistic infection status.

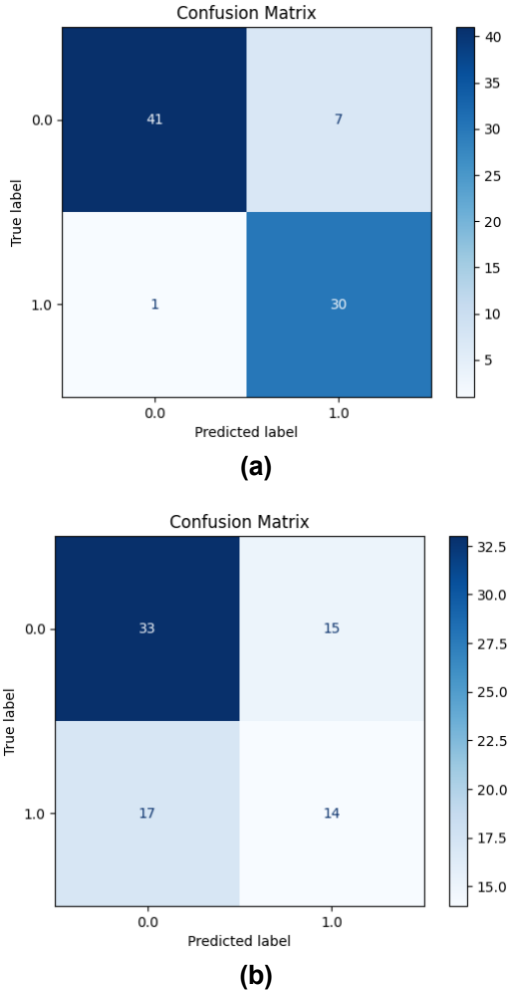


Fig. 3. Confusion Matrices of the (a) Exploratory Sensitivity Model and (b) Primary Leakage-Safe XGBoost Model on the Held-Out Test Set

As shown in Table 4, differences in classification performance were observed between the Primary Leakage-Safe Model and the Exploratory Sensitivity Model. The Primary Leakage-Safe Model, which excluded Follow-up Duration and was restricted to predictors available at ART initiation, achieved a precision of 0.66, recall of 0.69, and F1-score of 0.67 for the Alive class. For the Dead class, precision was 0.48, recall was 0.45, and the F1-score was 0.47, indicating comparatively limited ability to identify mortality cases using baseline predictors alone. The macro-average precision, recall, and F1-score were all 0.57, while the weighted-average precision, recall, and F1-score were all 0.59. In contrast, the Exploratory Sensitivity Model, which included Follow-up Duration, demonstrated substantially higher classification performance. For the Alive class, precision, recall, and

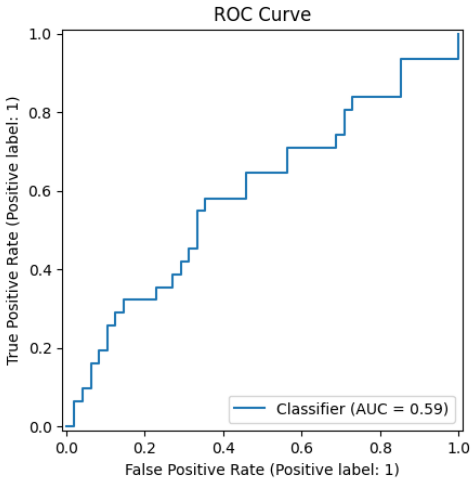
Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

F1-score were 0.98, 0.85, and 0.91, respectively. For the Dead class, the corresponding values were 0.81, 0.97, and 0.88. The macro-average precision, recall, and F1-score were 0.89, 0.91, and 0.90, respectively, while the weighted-average values were 0.91, 0.90, and 0.90. The increase in mortality-class recall from 0.45 in the Primary Leakage-Safe Model to 0.97 in the Exploratory Sensitivity Model illustrates the substantial influence of Follow-up Duration on retrospective classification performance. However, because Follow-up Duration represents post-baseline information unavailable at ART initiation, the duration-inclusive model was retained only as an exploratory sensitivity analysis, whereas the leakage-safe model remained the primary Day-0 prediction model.

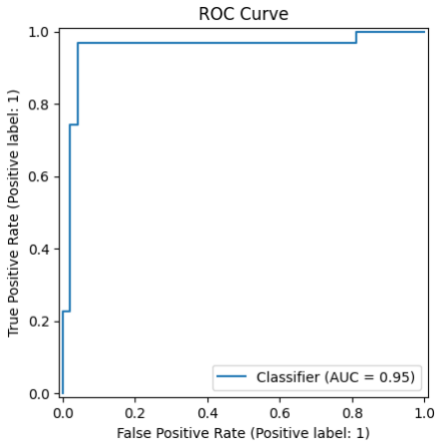
Fig. 3. Compares the classification performance of the Primary Leakage-Safe Model and the Exploratory Sensitivity Model on the held-out testing dataset. The Primary Leakage-Safe Model correctly classified 33 of 48 survival cases and 14 of 31 mortality cases; however, it misclassified 15 survival cases as mortality and 17 mortality cases as survival. This corresponds to a mortality-class recall of 0.45 and indicates limited sensitivity for mortality detection when restricted to predictors available at ART initiation. In contrast, the Exploratory Sensitivity Model correctly classified 41 of 48 survival cases and 30 of 31 mortality cases, with seven false-positive and only one false-negative prediction. The reduction in false-negative mortality predictions from 17 in the Primary Leakage-Safe Model to one in the Exploratory Sensitivity Model is consistent with the increase in mortality-class recall from 0.45 to 0.97 reported in Table 4. Overall, the comparison demonstrates the substantial increase in retrospective classification performance when Follow-up Duration is incorporated. However, because Follow-up Duration represents post-baseline information that is unavailable at ART initiation, the higher-performing duration-inclusive model is interpreted only as an exploratory sensitivity analysis.

Fig. 4. Compares the ROC curves of the Primary Leakage-Safe Model and the Exploratory Sensitivity Model on the held-out testing dataset. The Primary Leakage-Safe Model, which excluded Follow-up Duration and was restricted to predictors available at ART initiation, achieved a ROC-AUC of 0.591, indicating limited discriminative ability in distinguishing mortality from survival. Although its ROC curve generally remained above the diagonal reference pattern, the relatively modest ROC-AUC indicates that baseline predictors alone provided limited ability to consistently rank mortality cases above survival cases across classification thresholds.

In contrast, the Exploratory Sensitivity Model, which included Follow-up Duration, achieved a ROC-AUC of 0.9546, indicating excellent discriminative performance. Its ROC curve rose sharply toward the upper-left region and remained close to that region across most thresholds, demonstrating substantially stronger separation between mortality and survival. This finding is consistent with the classification results presented in Table 4, where mortality-class recall increased from 0.45 in the Primary Leakage-Safe Model to 0.97 in the Exploratory Sensitivity Model. The marked difference in ROC-AUC between the two models further demonstrates the substantial influence of Follow-up Duration on retrospective predictive performance. However, because Follow-up Duration represents post-baseline information that is unavailable at ART initiation, the higher-performing duration-inclusive model is retained only as an exploratory sensitivity analysis, whereas the Primary Leakage-Safe Model remains the principal model for the intended Day-0 prediction setting.



a. Primary Leakage-Safe Model



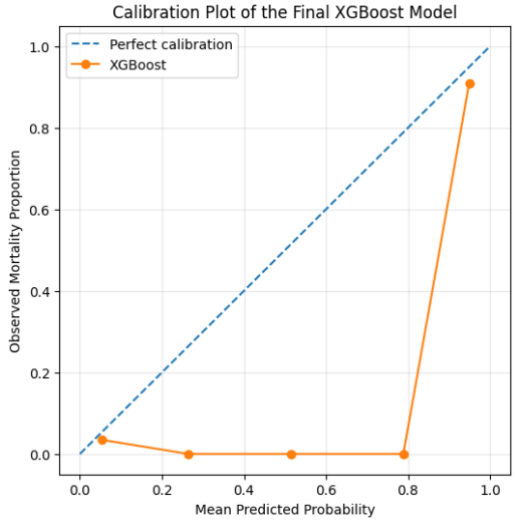
b. Exploratory Sensitivity Model

Fig. 4. ROC Curves of the (a) Primary Leakage-Safe and (b) Exploratory Sensitivity XGBoost Models on the Held-Out Test Set

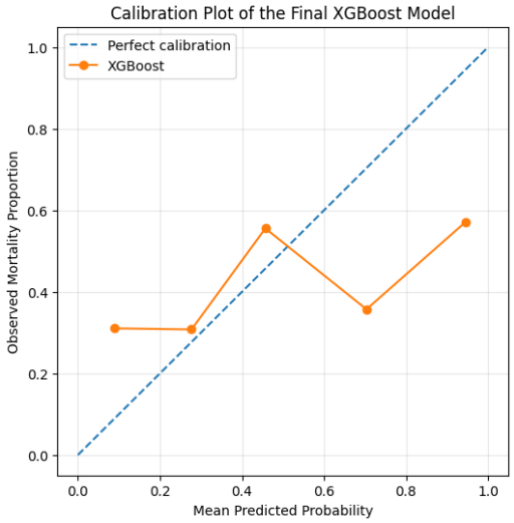
B. Additional Clinical Performance Evaluation

To provide a more comprehensive evaluation of model performance in the imbalanced clinical prediction setting, additional clinical performance measures were calculated on the held-out test set for both the Primary Leakage-Safe Model and the Exploratory Sensitivity Model. These complementary measures were used to assess discrimination, class-specific performance, reliability of negative predictions, and probabilistic prediction error. As summarized in Table 5, the Primary Leakage-Safe Model achieved a ROC-AUC of 0.5914 (95% CI: 0.4618–0.7231) and a PR-AUC of 0.4968 (95% CI: 0.3490–0.6792), indicating limited discriminative performance when prediction was restricted to baseline predictors available at ART initiation. Sensitivity was 0.4516 (95% CI: 0.2812–

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383



a. Exploratory Sensitivity Model



b. Primary Leakage-Safe Model

Fig. 5. Calibration Plots of the (a) Exploratory Sensitivity and (b) Primary Leakage-Safe XGBoost Models on the Held-Out Test Set

Table 5. Additional Clinical Performance of the Primary Leakage-Safe and Exploratory Sensitivity XGBoost Models on the Held-Out Test Set.

Metric	Value	95% CI
Primary Leakage-Safe Model		
ROC-AUC	0.5914	0.4618–0.7231
PR-AUC	0.4968	0.3490–0.6792
Sensitivity	0.4516	0.2812–0.6333
Specificity	0.6875	0.5556–0.8158
Negative Predictive Value	0.6600	0.5208–0.7959
Balanced Accuracy	0.5696	0.4628–0.6783
Brier Score	0.2938	0.2228–0.3681
Exploratory Sensitivity Model		
ROC-AUC	0.9546	0.8816–1.0000
PR-AUC	0.9360	0.8374–1.0000
Sensitivity	0.9677	0.8889–1.0000
Specificity	0.8542	0.7499–0.9474
Negative Predictive Value	0.9762	0.9149–1.0000
Balanced Accuracy	0.9110	0.8461–0.9649
Brier Score	0.0818	0.0412–0.1321

0.6333), specificity was 0.6875 (95% CI: 0.5556–0.8158), and the negative predictive value was 0.6600 (95% CI: 0.5208–0.7959). The balanced accuracy was 0.5696 (95% CI: 0.4628–0.6783), while the Brier score was 0.2938 (95% CI: 0.2228–0.3681). Collectively, these results indicate that the leakage-safe model provided only modest discrimination and limited mortality sensitivity when restricted to information available at the intended Day-0 prediction point. In contrast, the Exploratory Sensitivity Model incorporating Follow-up Duration achieved substantially stronger performance. The ROC-AUC increased to 0.9546 (95% CI: 0.8816–1.0000) and the PR-AUC to 0.9360 (95% CI: 0.8374–1.0000). Sensitivity reached 0.9677 (95% CI: 0.8889–1.0000), specificity 0.8542 (95% CI: 0.7499–0.9474), and the negative predictive value 0.9762 (95% CI: 0.9149–1.0000). Balanced accuracy increased to 0.9110 (95% CI: 0.8461–0.9649), whereas the Brier score decreased to 0.0818 (95% CI: 0.0412–0.1321). These marked differences demonstrate the substantial influence of Follow-up Duration on retrospective predictive performance. Because Follow-up Duration is not available at ART initiation, however, we interpret the higher-performing model as an exploratory sensitivity analysis rather than the primary Day-0 prediction model. The 95% confidence intervals were estimated using bootstrap resampling to quantify uncertainty in the reported performance measures. An exploratory threshold analysis was subsequently performed for the Exploratory Sensitivity Model to examine the sensitivity–specificity trade-off across probability thresholds ranging from 0.30 to 0.70. Sensitivity remained constant at 0.9677 across all evaluated thresholds, whereas specificity increased from 0.7708 at a threshold of 0.30 to 0.9167 at thresholds of 0.60 and 0.70. This pattern indicates that increasing the classification threshold reduced false-positive predictions without decreasing sensitivity within the evaluated threshold range. However, because no clinically validated decision threshold was prespecified and the Exploratory Sensitivity Model incorporated post-baseline Follow-up Duration, these findings are presented strictly as exploratory and should not be interpreted as establishing a recommended clinical cutoff.

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

Table 6. Repeated Stratified Cross-Validation Performance of the Primary Leakage-Safe and Exploratory Sensitivity XGBoost Models

Metric	Mean	SD	95% CI
Primary Leakage-Safe Model			
Accuracy	0.6016	0.0477	0.5880–0.6151
Precision	0.5945	0.0489	0.5806–0.6084
Recall	0.6016	0.0477	0.5880–0.6151
F1-score	0.5942	0.0478	0.5806–0.6078
ROC-AUC	0.5911	0.0686	0.5716–0.6106
Exploratory Sensitivity Model			
Accuracy	0.9181	0.0358	0.9080–0.9283
Precision	0.9210	0.0347	0.9111–0.9308
Recall	0.9181	0.0358	0.9080–0.9283
F1-score	0.9170	0.0368	0.9066–0.9275
ROC-AUC	0.9353	0.0391	0.9242–0.9465

Fig. 5 compares the calibration performance of the Primary Leakage-Safe and Exploratory Sensitivity XGBoost models on the held-out testing dataset. The dashed diagonal line represents perfect calibration, where predicted mortality probabilities correspond directly to the observed mortality proportions. In Fig. 5(a), the Exploratory Sensitivity Model achieved a Brier score of 0.0818, indicating a lower overall mean squared error of the predicted probabilities. Nevertheless, the calibration curve still deviated from the ideal diagonal line across several probability ranges, indicating that strong discrimination and a relatively low Brier score did not necessarily correspond to perfect calibration. In Fig. 5(b), the Primary Leakage-Safe Model showed substantial deviations from the reference line across most probability ranges, indicating limited agreement between predicted probabilities and observed outcomes. This pattern is consistent with its Brier score of 0.2938, indicating relatively large probabilistic prediction error when the model was restricted to predictors available at ART initiation. The observed deviations should also be interpreted cautiously because the held-out test set contained only 79 observations, resulting in few cases within individual probability bins. Overall, the Exploratory Sensitivity Model demonstrated lower probabilistic prediction error than the Primary Leakage-Safe Model. However, because this improvement was obtained after incorporating Follow-up Duration, which is unavailable at the intended Day-0 prediction point, these results should be interpreted only within the exploratory retrospective sensitivity analysis. Further probability recalibration and external validation would be required before the predicted probabilities could be interpreted as reliable absolute mortality-risk estimates for prospective clinical use.

C. Repeated Stratified Cross-Validation

To further assess model stability beyond a single train-test split, repeated stratified cross-validation was performed on the training dataset for both the primary leakage-safe and exploratory sensitivity XGBoost models. This procedure was used to examine whether the observed performance patterns remained consistent across multiple training-validation partitions while preserving the proportions of mortality and survival cases within each fold. A 5-fold stratified cross-validation scheme was repeated ten times, resulting in 50 validation folds for each model. Accuracy, precision, recall, F1-score, and ROC-AUC were calculated for each validation fold and subsequently summarized using the mean, standard deviation (SD), and 95% confidence interval. The primary model excluded Follow-up Duration to maintain consistency with the Day-0 prediction setting, whereas the exploratory sensitivity model included Follow-up Duration to quantify its influence on retrospective predictive performance. The resulting cross-validation performance is presented in Table 6.

Across the repeated validation folds, the primary leakage-safe model achieved a mean accuracy of 0.6016 ± 0.0477 (95% CI: 0.5880–0.6151), mean precision of 0.5945 ± 0.0489 (95% CI: 0.5806–0.6084), mean recall of 0.6016 ± 0.0477 (95% CI: 0.5880–0.6151), and mean F1-score of 0.5942 ± 0.0478 (95% CI: 0.5806–0.6078). Its mean ROC-AUC was 0.5911 ± 0.0686 (95% CI: 0.5716–0.6106). These results indicate modest predictive performance when the model was restricted to baseline predictors available at ART initiation. In contrast, the exploratory sensitivity model demonstrated substantially higher performance across the repeated validation folds. Mean accuracy increased to 0.9181 ± 0.0358 (95% CI: 0.9080–0.9283), while mean precision, recall, and F1-score were 0.9210 ± 0.0347 (95% CI: 0.9111–0.9308), 0.9181 ± 0.0358 (95% CI: 0.9080–0.9283), and 0.9170 ± 0.0368 (95% CI: 0.9066–0.9275), respectively. The mean ROC-AUC reached 0.9353 ± 0.0391 (95% CI: 0.9242–0.9465). The relatively small standard deviations and consistently high performance of the exploratory model indicate stable performance across the repeated validation partitions. However, because this model incorporated Follow-up Duration, its higher performance should be interpreted as retrospective and exploratory rather than as evidence of prospective Day-0 predictive performance.

The substantial differences between the two models across repeated cross-validation further demonstrate the influence of Follow-up Duration on model performance. Importantly, repeated stratified cross-validation was conducted only within the training data, while the held-out test set comprising 79 patient records remained separate and was reserved for independent evaluation. Accordingly, the cross-validation results provide an assessment of model-development stability across repeated partitions, whereas the held-out test results provide an independent assessment on previously unseen observations. The lower performance of the primary leakage-safe model across both evaluation strategies also highlights the limited predictive information available from the baseline variables included in the present dataset and supports the need for richer baseline clinical predictors in future prospective model

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

development.

D. Ablation and Comparative Model Analysis

An ablation and comparative analysis was conducted to examine the individual contributions of Optuna hyperparameter optimization and SMOTE and to compare the performance of the proposed XGBoost framework with conventional machine-learning models. Baseline XGBoost, XGBoost with Optuna, XGBoost with SMOTE, and XGBoost combining Optuna and SMOTE were evaluated using the same held-out test set. Logistic regression, random forest, and gradient boosting were additionally evaluated as benchmark models. The analysis was performed for both the primary leakage-safe setting, which excluded follow-up duration, and the exploratory sensitivity setting, which incorporated follow-up duration. Accuracy, F1-score, and ROC-AUC were used to facilitate direct comparison across model configurations, as summarized in Table 7.

Table 7. Ablation and Comparative Performance of Model Configurations in the Primary Leakage-Safe and Exploratory Sensitivity Settings

Model	Accuracy	F1-score	ROC-AUC
Primary Leakage-Safe Model			
Baseline XGBoost	0.6329	0.6355	0.5954
XGBoost + Optuna	0.6329	0.6355	0.5867
XGBoost + SMOTE	0.6076	0.6086	0.5860
XGBoost + Optuna + SMOTE	0.5949	0.5923	0.5914
Logistic Regression	0.5570	0.5438	0.5309
Random Forest	0.5823	0.5864	0.5665
Gradient Boosting	0.5443	0.5443	0.6008
Exploratory Sensitivity Model			
Baseline XGBoost	0.9367	0.9371	0.9442
XGBoost + Optuna	0.9367	0.9371	0.9577
XGBoost + SMOTE	0.9494	0.9496	0.9452
XGBoost + Optuna + SMOTE	0.8987	0.8998	0.9546
Logistic Regression	0.9367	0.9371	0.9691
Random Forest	0.9367	0.9371	0.9610
Gradient Boosting	0.9241	0.9244	0.9483

In the primary leakage-safe analysis, predictive performance was generally modest across all evaluated models. Baseline XGBoost achieved an accuracy of 0.6329, an F1-score of 0.6355, and a ROC-AUC of 0.5954. XGBoost with Optuna maintained the same accuracy and F1-score at 0.6329 and 0.6355, respectively, while its ROC-AUC was 0.5867. XGBoost with SMOTE achieved an accuracy of 0.6076, an F1-score of 0.6086, and a ROC-AUC of 0.5860. The complete XGBoost framework combining Optuna and SMOTE achieved an accuracy of 0.5949, an F1-score of 0.5923, and a ROC-AUC of 0.5914. Among the benchmark algorithms, Gradient Boosting achieved the highest ROC-AUC of 0.6008, although its accuracy and F1-score were both 0.5443. Logistic Regression achieved an accuracy of 0.5570, an F1-score of 0.5438, and a ROC-AUC of 0.5309, whereas Random Forest achieved an accuracy of 0.5823, an F1-score of 0.5864, and a ROC-AUC of 0.5665. These findings indicate that neither hyperparameter optimization nor class rebalancing consistently improved performance across all evaluation metrics when prediction was restricted to the available baseline variables.

In contrast, the exploratory sensitivity analysis incorporating Follow-up Duration showed substantially higher performance. Baseline XGBoost achieved an accuracy of 0.9367 and a ROC-AUC of 0.9442, while Optuna optimization maintained the same accuracy and increased the ROC-AUC to 0.9577. XGBoost with SMOTE achieved the highest accuracy and F1-score among the evaluated XGBoost configurations, at 0.9494 and 0.9496, respectively. The complete Optuna-SMOTE XGBoost framework achieved an accuracy of 0.8987, an F1-score of 0.8998, and a ROC-AUC of 0.9546. Logistic Regression and Random Forest achieved ROC-AUC values of 0.9691 and 0.9610, respectively, slightly exceeding that of the complete XGBoost framework. Overall, the pronounced performance difference between the primary and exploratory settings across multiple algorithms indicates that the improvement was not attributable solely to a particular model architecture but was strongly associated with the inclusion of Follow-up Duration. Although standard benchmark models, particularly Logistic Regression (ROC-AUC = 0.9691) and Random Forest (ROC-AUC = 0.9610), achieved slightly higher ROC-AUC values than the complete XGBoost framework (ROC-AUC = 0.9546) in the exploratory sensitivity analysis, XGBoost was retained as the principal machine-learning architecture because of its methodological characteristics and alignment with the study objectives. First, gradient-boosted decision trees can model complex nonlinear relationships and interactions among clinical and demographic predictors without requiring these relationships to be specified a priori. Second, XGBoost incorporates regularization mechanisms within its objective function to control model complexity. Third, its tree-based architecture is compatible with TreeSHAP, enabling consistent local and

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

global attribution of model predictions. Nevertheless, the comparative results demonstrate that XGBoost did not uniformly outperform the benchmark algorithms; therefore, its selection should be interpreted in the context of the study's emphasis on explainable nonlinear modeling rather than superiority based solely on predictive performance.

E. Sensitivity Analysis of Follow-up Duration

To assess the influence of Follow-up Duration on predictive performance, a sensitivity analysis was conducted by comparing models with and without this variable. The model including Follow-up Duration achieved an accuracy of 89.87%, weighted precision of 0.91, weighted recall of 0.90, weighted F1-score of 0.90, and a ROC-AUC of 0.9546. In contrast, the model excluding Follow-up Duration achieved an accuracy of 59.49%, weighted precision of 0.59, weighted recall of 0.59, weighted F1-score of 0.59, and a ROC-AUC of 0.5914. The substantial reduction in performance after excluding Follow-up Duration indicates that this variable contributed strongly to mortality classification in the analyzed cohort. Accordingly, the model excluding Follow-up Duration was designated as the Primary Leakage-Safe Model, whereas the model including Follow-up Duration was retained only as an Exploratory Sensitivity Model to quantify the influence of this post-baseline variable on retrospective predictive performance. This finding reflects a strong predictive association within the retrospective dataset and should not be interpreted as evidence of a causal relationship between Follow-up Duration and mortality.

F. Feature Importance and Explainability Analysis

Feature importance and SHAP analyses were performed to determine which predictors contributed most strongly to mortality predictions in both the Primary Leakage-Safe and Exploratory Sensitivity XGBoost models. Although classification metrics such as accuracy, precision, recall, F1-score, and ROC-AUC provide an overall assessment of predictive performance, they do not indicate how individual predictors contribute to model decisions. Therefore, an interpretation stage was included to improve model transparency and provide a clearer understanding of the factors influencing mortality and survival predictions among PLHIV receiving ART. XGBoost feature importance was first applied to rank predictors according to their relative contributions within each model [38]. However, conventional feature importance provides only the magnitude of relative contribution and does not indicate the direction in which a predictor influences the predicted outcome. Accordingly, SHAP was subsequently applied to quantify both the magnitude and direction of individual predictor contributions [33]. The feature importance obtained from the optimized XGBoost model is presented in Fig. 6.

Fig. 6 compares the feature-importance profiles of the Primary Leakage-Safe and Exploratory Sensitivity XGBoost models. In the Primary Leakage-Safe Model, Functional Status was the most influential predictor, followed by Marital Status and Sex. ART Initiation Year also showed a notable contribution, followed by Clinical Stage, Education Level, Age, Employment Status,

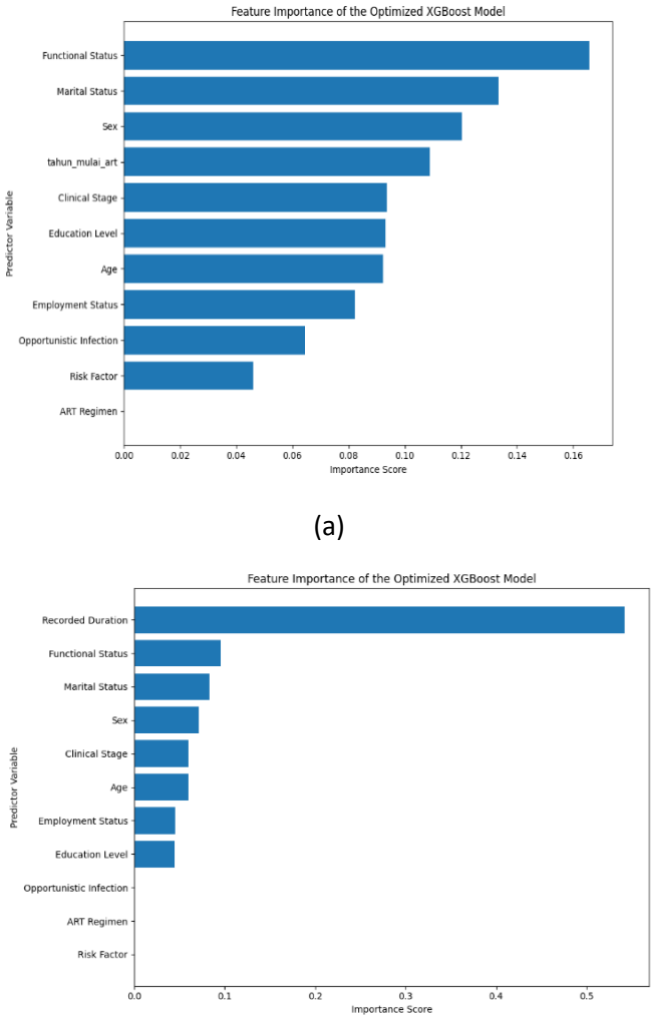


Fig. 6. XGBoost Feature Importance for the (a) Primary Leakage-Safe and (b) Exploratory Sensitivity Models.

Opportunistic Infection, and Risk Factor. ART Regimen showed little to no relative importance. Overall, predictor importance was distributed across several baseline variables, with no single predictor overwhelmingly dominating the model.

In contrast, the Exploratory Sensitivity Model showed a highly uneven distribution of predictor contributions. Follow-up Duration was the dominant predictor, with an importance score exceeding 0.50 and substantially surpassing all other variables. Functional Status ranked second, followed by Marital Status and Sex. Clinical Stage and Age showed smaller but observable contributions, whereas Employment Status and Education Level had relatively low importance scores. Opportunistic Infection, ART Regimen, and Risk Factor showed importance scores close to zero in the exploratory model.

The marked dominance of Follow-up Duration indicates that a substantial proportion of the Exploratory Sensitivity Model's predictive information was derived from this post-baseline variable. This finding is particularly relevant because Follow-up Duration is not available at the intended Day-0 prediction point. Therefore, the feature-importance comparison supports its exclusion from the

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

Primary Leakage-Safe Model and its inclusion only in the Exploratory Sensitivity Model. Importantly, near-zero importance scores should not be interpreted as evidence that Opportunistic Infection, ART Regimen, or Risk Factor lacks clinical relevance; rather, they indicate limited relative contribution within this specific model and dataset. Furthermore, conventional XGBoost feature importance does not indicate the direction in which individual predictors influence mortality predictions. SHAP analysis was therefore performed to examine both the magnitude and direction of predictor contributions, as presented in Fig. 7.

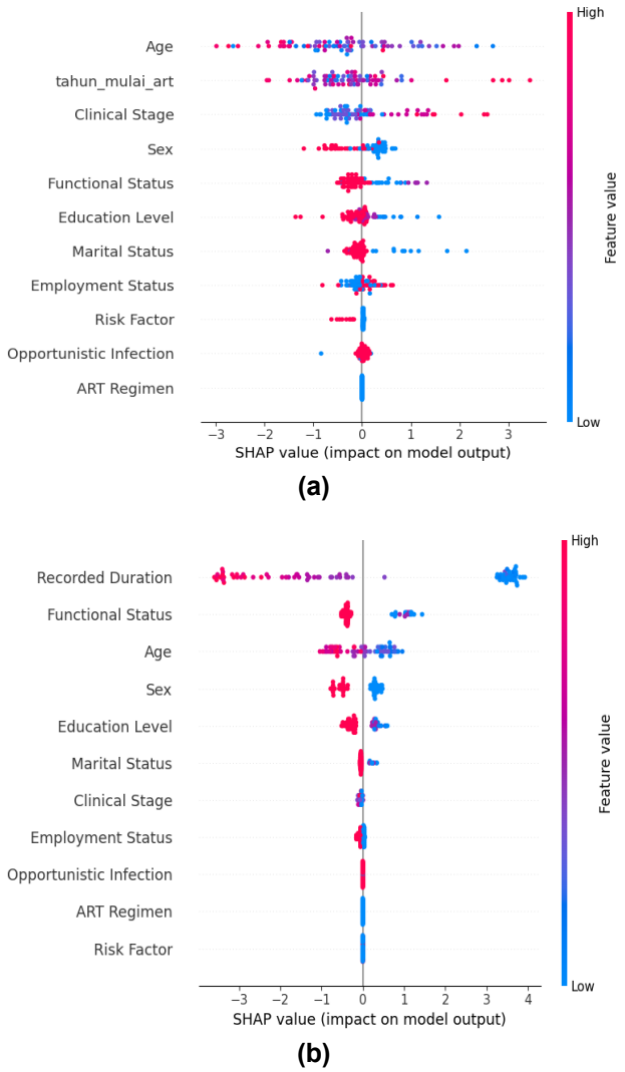


Fig. 7. SHAP Summary Plots of the (a) Primary Leakage-Safe and (b) Exploratory Sensitivity XGBoost Models

Fig. 7. compares the SHAP summary plots of the Primary Leakage-Safe and Exploratory Sensitivity XGBoost models, illustrating both the magnitude and direction of predictor contributions to the model outputs. Predictors are ordered according to their overall SHAP importance, with the most influential variables positioned at the top [39]. Each point represents one patient observation, while its horizontal position indicates the SHAP value. Positive SHAP values shift the model output toward mortality, whereas negative SHAP values shift it toward survival. Feature values are represented by color, with red indicating relatively high values and blue indicating relatively low values.

In the primary leakage-safe model shown in Fig. 7(a), Age exhibited the broadest SHAP distribution, followed by ART Initiation Year, Clinical Stage, Sex, Functional Status, Education Level, and Marital Status. Unlike the Exploratory Sensitivity Model, no single baseline predictor overwhelmingly dominated the SHAP distribution. Instead, contributions were distributed across several demographic and clinical predictors. Employment Status, Risk Factor, Opportunistic Infection, and ART Regimen generally showed more limited contributions, with their SHAP values concentrated closer to zero. This pattern is consistent with the relatively modest predictive performance of the Primary Leakage-Safe Model and suggests that the available baseline predictors provided limited separation between mortality and survival outcomes without post-baseline Follow-up Duration.

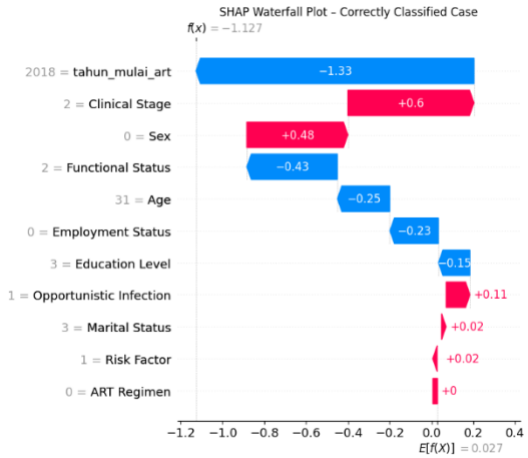
In contrast, the exploratory sensitivity model shown in Fig. 7(b) was strongly dominated by Follow-up Duration. Its SHAP values extended substantially farther from zero than those of the remaining predictors, indicating that Follow-up Duration exerted the largest influence on model predictions. The SHAP distribution indicates that shorter Follow-up Duration tended to shift model predictions toward mortality, whereas longer Follow-up Duration tended to shift predictions toward survival. Functional Status, Age, Sex, and Education Level also contributed to the model output, although their effects were considerably smaller. Marital Status, Clinical Stage, Employment Status, Opportunistic Infection, ART Regimen, and Risk Factor generally showed more limited SHAP contributions.

The substantial difference between the two SHAP profiles reinforces the sensitivity-analysis findings. Once Follow-up Duration was introduced, the model's explanations became strongly concentrated around this post-baseline variable, whereas the Primary Leakage-Safe Model relied on a broader combination of baseline predictors. Because Follow-up Duration is accumulated after ART initiation, its dominant SHAP contribution further supports its exclusion from the Primary Leakage-Safe Model for the intended Day-0 prediction setting and the interpretation of the duration-inclusive model only as an Exploratory Sensitivity Model. For categorical predictors represented through one-hot encoding, SHAP contributions were interpreted with reference to their corresponding encoded categories rather than as ordinal numerical effects.

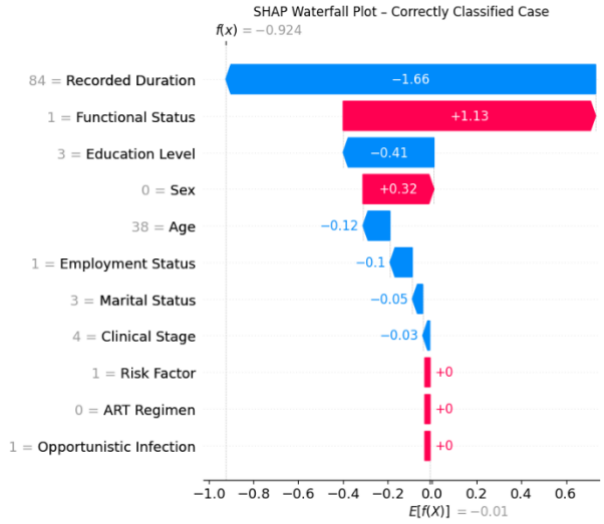
To complement the visual SHAP interpretation of the Exploratory Sensitivity Model, global predictor importance was quantified using the mean absolute SHAP value across the held-out test set. Follow-up Duration had the highest mean absolute SHAP value (2.6861), followed by Functional Status (0.5665), Age (0.5088), Sex (0.4042), and Education Level (0.3092). Marital Status (0.0767), Clinical Stage (0.0517), and Employment Status (0.0504) contributed less, whereas Risk Factor, ART Regimen, and Opportunistic Infection each had a mean absolute

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

SHAP value of 0.0000. These numerical results further demonstrate the dominant contribution of Follow-up Duration within the Exploratory Sensitivity Model. To further examine model behavior at the individual-patient level, local SHAP explanations were evaluated for representative correctly classified, false-positive, and false-negative predictions from the Exploratory Sensitivity Model.



a. Primary Leakage-Safe Model

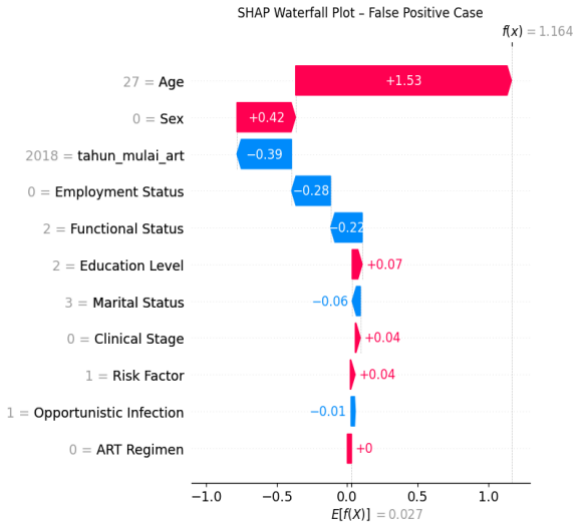


b. Exploratory Sensitivity Model

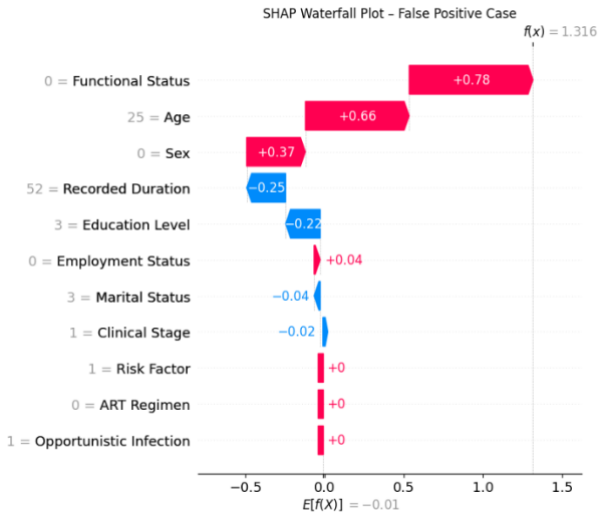
Fig. 8. SHAP Waterfall Plots for Representative Correctly Classified Cases: (a) Primary Leakage-Safe Model and (b) Exploratory Sensitivity Model

Fig. 8. compares local SHAP explanations for representative correctly classified cases from the primary leakage-safe and exploratory sensitivity XGBoost models. In the primary leakage-safe model shown in Fig. 8(a), ART Initiation Year produced the largest negative contribution (SHAP = −1.33), shifting the model output toward survival. Clinical Stage produced the strongest positive contribution (SHAP = +0.60), followed by Sex (+0.48). Functional Status (−0.43), Age (−0.25), Employment Status (−0.23), and Education Level (−0.15) contributed negatively toward survival, whereas Opportunistic Infection (+0.11), Marital Status (+0.02), and Risk Factor (+0.02) produced smaller positive contributions. ART Regimen had a negligible contribution. The combined effects of these baseline predictors resulted in the final correct classification.

In the exploratory sensitivity model shown in Fig. 8(b), Follow-up Duration produced the largest negative contribution (SHAP = −1.66), whereas Functional Status generated the strongest positive contribution (SHAP = +1.13). Education Level (−0.41), Sex (+0.32), Age (−0.12), Employment Status (−0.10), Marital Status (−0.05), and Clinical Stage (−0.03) contributed smaller opposing effects. Despite the positive contribution of Functional Status, the dominant negative contribution of Follow-up Duration shifted the final model output toward the correctly predicted survival class. The comparison illustrates how the Primary Leakage-Safe Model relied on a combination of baseline predictors, whereas the local prediction of the Exploratory Sensitivity Model was strongly influenced by post-baseline Follow-up Duration.



a. Primary Leakage-Safe Model



b. Exploratory Sensitivity Model

Fig. 9. SHAP Waterfall Plots for Representative False-Positive Cases: (a) Primary Leakage-Safe Model and (b) Exploratory Sensitivity Model

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

Fig. 9. compares local SHAP explanations for representative false-positive cases from the primary leakage-safe and exploratory sensitivity XGBoost models. In the primary leakage-safe model shown in Fig. 9(a), Age produced the largest positive contribution (SHAP = +1.53), strongly shifting the model output toward mortality despite the observed survival outcome. Sex also contributed positively (SHAP = +0.42). In contrast, ART Initiation Year (−0.39), Employment Status (−0.28), Functional Status (−0.22), Marital Status (−0.06), and Opportunistic Infection (−0.01) shifted the prediction toward survival. Education Level (+0.07), Clinical Stage (+0.04), and Risk Factor (+0.04) provided smaller positive contributions toward mortality, while ART Regimen had a negligible contribution. The strong positive contribution of Age, together with the contribution of Sex and other smaller positive effects, outweighed the negative contributions and resulted in the false-positive prediction.

In the exploratory sensitivity model shown in Fig. 9(b), Functional Status produced the largest positive contribution (SHAP = +0.78), followed by Age (+0.66) and Sex (+0.37), which together shifted the prediction toward mortality. Follow-up Duration (−0.25) and Education Level (−0.22) contributed toward survival, but their effects were insufficient to counterbalance the stronger positive contributions. Employment Status (+0.04), Marital Status (−0.04), and Clinical Stage (−0.02) had relatively small effects, while Risk Factor, ART Regimen, and Opportunistic Infection had negligible contributions. Consequently, both models produced a false-positive prediction, although the dominant predictors underlying the error differed between the Primary Leakage-Safe and Exploratory Sensitivity settings.

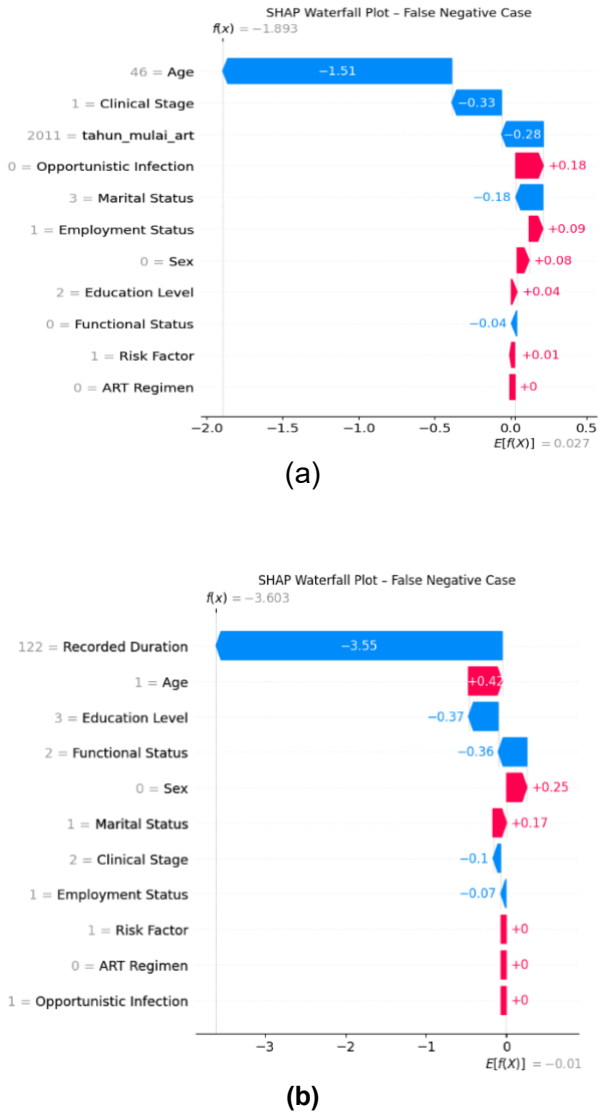


Fig. 10. SHAP Waterfall Plots for Representative False-Negative Cases: (a) Primary Leakage-Safe Model and (b) Exploratory Sensitivity Model

Fig. 10. compares local SHAP explanations for representative false-negative cases from the primary leakage-safe and exploratory sensitivity XGBoost models. In the primary leakage-safe model shown in Fig. 10(a), Age produced the strongest negative contribution (SHAP = −1.51), substantially shifting the model output toward survival despite the observed mortality outcome. Clinical Stage (−0.33) and ART Initiation Year (−0.28) also contributed toward survival. Marital Status (−0.18) and Functional Status (−0.04) provided additional negative contributions. In contrast, Opportunistic Infection (+0.18), Employment Status (+0.09), Sex (+0.08), Education Level (+0.04), and Risk Factor (+0.01) shifted the model output toward mortality, although these positive contributions were insufficient to offset the stronger negative effects. ART Regimen had a negligible contribution. Consequently, the cumulative negative contributions, particularly from Age, caused the model to incorrectly predict the mortality case as survival. In the exploratory sensitivity model shown in Fig. 10(b), Follow-up Duration produced the strongest negative contribution (SHAP = −3.55), substantially shifting the model output toward survival. Age (+0.42), Sex (+0.25), and Marital Status (+0.17) contributed positively toward mortality, whereas Education Level (−0.37), Functional Status (−0.36), Clinical Stage (−0.10), and Employment Status (−0.07) further shifted the prediction toward survival. Risk Factor, ART Regimen, and Opportunistic Infection had negligible contributions. The dominant negative contribution of Follow-up Duration was sufficient to shift the mortality case toward the survival class, resulting in a false-negative prediction.

IV. Discussion

A. Performance Evaluation of the Optimized XGBoost Model

The Primary Leakage-Safe XGBoost Model, which excluded Follow-up Duration and used only predictors available at ART initiation, demonstrated modest predictive performance on the independent testing dataset. The complete XGBoost configuration combining Optuna and SMOTE achieved an accuracy of 59.49%, an F1-score of 0.5923, and a ROC-AUC of 0.5914. These findings indicate that the available baseline predictors provided limited discrimination between mortality and survival at the intended Day-0 prediction point. This result is important because the primary model represents the

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

Table 8. Comparison of the Proposed Framework with Existing HIV/AIDS Mortality Prediction Studies.

Study	Target Population	Model Architecture	Sample Size (N)	Accuracy (%)	ROC-AUC	Sensitivity	Specificity
Shi et al. (2022) [43]	In-hospital mortality in HIV/AIDS patients with *Talaromyces marneffei* infection	XGBoost	1.927	0.9344	9.210	0.6912	0.9667
Chen et al. (2025) [13]	Mortality risk in AIDS patients with AIDS-related diseases or symptoms	XGBoost	478	0.675	0.729	0.717	0.636
Proposed Framework (Primary Baseline)	ART Initiation Mortality	XGBoost (Leakage-Safe)	393	0.5949	0.5914	0.4516	0.6875
Proposed Framework (Sensitivity Model)	Retrospective Cohort	XGBoost + Follow-up	393	89.87	0.9546	0.9677	0.8542

clinically appropriate prediction setting in which only information available at ART initiation is used. In contrast, the Exploratory Sensitivity XGBoost Model incorporating Follow-up Duration achieved substantially higher performance, with an accuracy of 89.87% and an ROC-AUC of 0.9546. The weighted precision, recall, and F1-score were 0.91, 0.90, and 0.90, respectively. Notably, the mortality-class recall reached 0.97, with 30 of 31 mortality cases correctly identified. In mortality-risk prediction, high recall is clinically important because false-negative predictions may result in high-risk patients being incorrectly classified as survivors and potentially receiving insufficient clinical attention [40]. Only one mortality case was misclassified as survival, whereas the mortality-class precision of 0.81 indicates that seven surviving patients were incorrectly classified as mortality cases. Thus, the exploratory model demonstrated high sensitivity for mortality detection, although some false-positive predictions remained.

The ROC-AUC of 0.9546 indicates that the exploratory sensitivity model could distinguish mortality from survival across a range of decision thresholds. Unlike accuracy, which reflects classification performance at a specific threshold, ROC-AUC summarizes discrimination across multiple thresholds [37]. However, this strong performance should not be interpreted as prospective Day-0 predictive capability because follow-up duration is accumulated after ART initiation and is therefore unavailable at the intended prediction point. The substantial reduction in performance after excluding follow-up duration further indicates that this post-baseline variable contributed strongly to the apparent retrospective predictive performance.

The observed performance was also examined in relation to the modeling procedures used in the proposed framework. Class imbalance was addressed by applying SMOTE exclusively to the training dataset [41], while Optuna was used to select suitable hyperparameter values for the XGBoost classifier [42]. The ablation analysis showed that the effects of these procedures were metric-dependent. In the primary leakage-safe setting, neither SMOTE nor Optuna produced strong discrimination, whereas substantially higher performance was observed when follow-up duration was included in the exploratory sensitivity analysis. Therefore, the performance improvement cannot be attributed solely to SMOTE or hyperparameter optimization and appears to be strongly influenced by the inclusion of follow-up duration.

B. Comparison with Existing Studies

To contextualize the predictive performance and clinical utility of the proposed explainable XGBoost framework, the results were compared with two previously published machine learning studies addressing mortality prediction among people living with HIV/AIDS. Although the studies differed in clinical populations, prediction settings, outcome definitions, and available predictors, both employed XGBoost and reported complete classification performance metrics, thereby enabling a more transparent methodological comparison, as presented in Table 8.

As shown in Table 8, substantial variation in predictive performance was observed across studies, reflecting differences in patient populations, prediction timing, and predictor availability. Shi et al. [43] developed an XGBoost model for predicting in-hospital mortality among 1,927 HIV/AIDS patients with *Talaromyces marneffei* infection. On the testing dataset, their model achieved an accuracy of 0.9344, ROC-AUC of 0.9008, sensitivity of 0.6912, and specificity of 0.9667. Chen et al. [13] developed an XGBoost-based mortality prediction model among 478 AIDS patients presenting with AIDS-related diseases or symptoms. Their test-set model achieved an accuracy of 0.6750, an ROC-AUC of 0.7290, a sensitivity of 0.7170, and a specificity of 0.6360. These results demonstrate that XGBoost can provide useful mortality discrimination in HIV/AIDS populations when informative clinical and laboratory predictors are available. However, direct

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

numerical comparison should be interpreted cautiously because the studies differed in outcome definitions, patient characteristics, predictor sets, and prediction settings. In the present study, the Primary Leakage-Safe Model was intentionally restricted to predictors available at ART initiation (Day 0). Under this prediction constraint, the model achieved an accuracy of 0.5949, an ROC-AUC of 0.5914, a sensitivity of 0.4516, and a specificity of 0.6875. The modest discrimination of the primary model suggests that the available baseline demographic and clinical information provided limited predictive information for mortality classification in this cohort. This finding should therefore be interpreted in relation to predictor availability at the intended clinical decision point rather than as evidence of an inherent limitation of the XGBoost algorithm.

In contrast, the Exploratory Sensitivity Model incorporating Follow-up Duration achieved substantially higher predictive performance, with an accuracy of 0.8987, ROC-AUC of 0.9546, sensitivity of 0.9677, and specificity of 0.8542. However, Follow-up Duration represents post-baseline information that would not be available at the intended ART-initiation prediction point. Therefore, the higher performance of the Exploratory Sensitivity Model should not be interpreted as evidence of superior prospective clinical utility. Instead, this sensitivity analysis demonstrates the substantial predictive information contained in Follow-up Duration and reinforces the importance of excluding post-baseline information from the Primary Leakage-Safe Model to minimize potential target leakage. From an algorithmic perspective, XGBoost offers the ability to capture nonlinear relationships and higher-order interactions that may not be adequately represented by conventional linear models without explicitly specifying nonlinear or interaction terms. Its regularized boosting framework can also control model complexity. Furthermore, integration with SHAP provides both global and patient-level feature attribution, enabling the contribution of individual predictors to model predictions to be examined transparently. These characteristics support the use of an explainable XGBoost framework for investigating mortality risk patterns in clinical data.

C. Interpretation of the Most Influential Predictors

The feature importance and SHAP analyses identified follow-up duration as the most influential predictor in the developed model. The SHAP summary plot demonstrated that reduced follow-up duration shifted SHAP values toward the positive range, thereby increasing the likelihood of mortality predictions. This pattern indicates that shorter recorded treatment or follow-up durations were more frequently associated with mortality outcomes in the analyzed dataset. The dominant influence of follow-up duration may be partly explained by the timing of mortality and changes in ART initiation policies during the 2006–2023 study period. HIV treatment policies evolved during the study period, particularly with the transition toward universal test-and-treat, which expanded ART initiation regardless of clinical or immunological status [44]. Patients who began ART at an advanced stage of HIV had a greater risk of dying during the first months of treatment. This finding may partially explain why deceased patients had shorter recorded follow-up durations [45]. In later periods, HIV treatment shifted toward the Treat All approach, under which ART is recommended for all PLHIV irrespective of the WHO clinical stage or CD4 cell count [46]. This policy enabled patients to initiate ART earlier after diagnosis, potentially before substantial clinical deterioration occurred. Patients who initiated treatment earlier and survived longer consequently accumulated higher follow-up durations. Therefore, differences in ART initiation timing, clinical condition at treatment initiation, and early mortality may partly explain why totalbulan emerged as the most influential predictor.

Nevertheless, the association between shorter follow-up durations and mortality must not be construed as evidence of a direct causal relationship between abbreviated ART duration and patient death. Instead, lower values may indicate that mortality occurred relatively early after ART initiation, particularly among patients who began treatment with advanced clinical conditions. Furthermore, interpreting follow-up duration depends on how the variable was defined and calculated. If it represents the duration from ART initiation until death, the last recorded follow-up, or the end of observation, its strong predictive contribution may partly reflect information closely associated with the outcome. Inclusion of variables encoding information unavailable at the operational point of prediction can introduce target leakage, thereby yielding spuriously optimistic performance estimates [47]. The sensitivity analysis showed a substantial decline in predictive performance after follow-up duration was excluded, indicating that this variable contributed strongly to mortality classification in the analyzed cohort. Accordingly, follow-up duration was retained in the developed model, while the model excluding this variable was used as a comparative sensitivity analysis. However, because follow-up duration reflects information accumulated after ART initiation, its strong association with mortality should be interpreted as a predictive relationship within the retrospective dataset rather than as evidence of a direct causal effect. Functional status also emerged as an influential predictor because it captures the extent to which patients can perform everyday tasks and may reflect the severity of their general clinical condition [48]. Earlier research has consistently linked impaired functional status with advanced disease and unfavorable treatment outcomes among PLHIV receiving ART [3], [49]. The SHAP analysis also showed contributions from age, sex, and education level. Demographic and socioeconomic characteristics have previously been associated with variations in access to healthcare, treatment adherence, disease progression, and clinical outcomes among PLHIV [50], [51]. Categorical predictors were represented using one-hot encoding to avoid imposing artificial ordinal relationships among nominal categories. Accordingly, their SHAP contributions were interpreted with reference to the corresponding original categories rather than as ordered numerical values. For nominal variables, numerical codes do not necessarily represent a clinically meaningful ordinal relationship between categories [52].

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

D. Explainability of the Optimized XGBoost Model

The feature rankings obtained from XGBoost feature importance and SHAP were not completely identical. This difference is expected because XGBoost feature importance is derived from statistics related to the tree-building process [53]. In contrast, the global SHAP analysis reflects the overall influence of each predictor by averaging its contribution across all predictions [54]. By complementing the feature importance analysis, SHAP provided a clearer interpretation of the optimized XGBoost model. Rather than presenting the classifier solely as a black-box system, SHAP provided an interpretable representation of how individual predictors influenced the model outputs. Explainability is particularly important in healthcare applications because clinicians need to understand and critically evaluate the factors underlying model predictions before considering them in patient management [55].

E. Clinical and Practical Implications

Although the high recall observed for the mortality class underscores the framework's preliminary capacity to identify high-risk individuals, such predictive metrics merely represent an initial algorithmic benchmark rather than proof of readiness for routine clinical screening [56]. Automated risk stratification ought to serve as a supportive tool rather than a replacement for clinical intuition; hence, model predictions must always be interpreted alongside a patient's broader clinical context, past medical background, and physician expertise [57]. Apart from the occurrence of false-positive outputs, several methodological vulnerabilities restrict the direct translational value of these results. Specifically, the framework relies on single-center retrospective data collected between 2006 and 2023—a prolonged period characterized by major transitions in ART eligibility, care protocols, and treatment guidelines—alongside the potential threat of feature leakage. Given that prognostic models often suffer from reduced transportability when applied to evolving temporal cohorts or distinct patient populations [58], any assertions regarding clinical utility must be conservatively framed. Consequently, rigorous model calibration, temporal validation across distinct policy eras, external replication in multicenter settings, and prospective real-world trials remain indispensable prerequisites prior to endorsing this model for routine clinical use.

F. Study Limitations and Future Work

Several inherent limitations warrant careful consideration when interpreting the conclusions of this study. Primarily, sourcing data exclusively from a single medical center potentially restricts the external generalizability of these findings to alternative clinical environments. Furthermore, owing to the retrospective design, reliance on routine medical records necessitated missing value imputation, which inevitably introduces an element of statistical uncertainty. Sample size represents another notable constraint; the finalized dataset comprised 393 patient records, with a subset of 79 set aside for independent model testing, highlighting the necessity of larger and more heterogeneous cohorts to confirm predictive stability. Additionally, while one-hot encoding was applied to categorical features to prevent imposing artificial ordinal hierarchies, interpreting these variables remains tied to the distribution of the original clinical data. Finally, although sensitivity analysis showed a substantial drop in predictive power when excluding follow-up duration, this feature accumulates progressively over time and should be framed cautiously as a retrospective explanatory factor rather than a baseline predictor. Selection bias is another critical methodological vulnerability in this work. Of the 1,579 initially evaluated patient records, 1,186 (75.1%) were excluded due to facility transfers (n = 557), loss to follow-up (n = 359), or unrecorded treatment status (n = 264). Individuals lost to care or referred elsewhere may have socioeconomic, geographic, or baseline clinical risk profiles that systematically differ from those retained in continuous care. Furthermore, omitting individuals lost to follow-up (LTFU) introduces a distinct risk of selection bias, given that patients who disengage from routine clinical care in real-world settings frequently suffer undocumented fatal outcomes. Omission of these cases may have attenuated the observed predictive impact of key clinical mortality determinants, thereby constraining the direct applicability of the trained classifiers to broader patient populations operating under different care retention protocols. To overcome these constraints, subsequent investigations should leverage prospective, multicenter datasets, execute external and temporal validations across evolving treatment guideline eras, apply formal probability recalibration techniques, and evaluate model performance directly within real-world clinical workflows.

V. Conclusion

The principal objective of this research was to construct and validate an interpretable machine learning framework for predicting mortality outcomes among people living with HIV (PLHIV) receiving antiretroviral therapy (ART). Using a single-center retrospective dataset, the optimized Primary Leakage-Safe XGBoost Model, which was restricted to predictors available at ART initiation, achieved an accuracy of 59.49% and an ROC-AUC of 0.5914. These findings indicate modest discriminative performance and highlight the limited predictive information provided by the available baseline variables at the intended Day-0 prediction point. In contrast, the Exploratory Sensitivity Model incorporating Follow-up Duration achieved substantially higher performance, with 89.87% accuracy and a ROC-AUC of 0.9546. Feature-importance and SHAP analyses further showed that Follow-up Duration was the dominant predictor in the exploratory model, demonstrating its substantial influence on retrospective mortality classification. A key contribution of this study is the empirical demonstration of how the inclusion of post-baseline information can substantially influence apparent predictive performance in retrospective HIV mortality modeling. The strong performance of the Exploratory Sensitivity Model should therefore not be interpreted as prospective clinical predictive capability because Follow-up Duration is unavailable at the intended ART-initiation prediction point. Future research should prioritize the inclusion of richer

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

baseline clinical and laboratory predictors, such as baseline CD4 count and viral load; prospective validation in larger multicenter cohorts; probability recalibration; and evaluation of model transportability across different populations and evolving ART treatment guidelines.

Acknowledgment

The authors thank Universitas Udayana for facilitating access to the retrospective dataset used in this study. The authors also acknowledge RSUD Sanjiwani, Gianyar, Bali, Indonesia, as the healthcare institution that provided the clinical records. Furthermore, the authors appreciate the support of the Faculty of Engineering and Informatics, Universitas Pendidikan Nasional, during the conduct of this research.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Author Contributions

Ni Putu Likayuni Viona conceived and designed the study, performed data preprocessing, developed and evaluated the machine learning model, conducted the SHAP analysis, interpreted the results, and prepared the initial draft of the manuscript. Ngakan Nyoman Kutha Krisnawijaya contributed to the study design, provided methodological supervision, assisted with result interpretation, and critically revised the manuscript for important intellectual content. Desak Nyoman Widyanthini facilitated data acquisition, provided clinical context, validated the study findings, and reviewed the manuscript. All authors read and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Data Availability

The clinical dataset used in this study is not publicly available because it contains confidential patient information and is subject to institutional data-sharing restrictions.

Code Availability

The source code for data preprocessing, model development, hyperparameter optimization, performance evaluation, and SHAP analysis is not publicly available. However, it may be obtained from the corresponding author upon reasonable request.

Ethical Approval

Ethical approval was obtained from the Research Ethics Committee of the Faculty of Medicine, Udayana University (Ethical Clearance No. 1567/UN14.2.2.VII.14/LT/2024). Institutional permission to access the retrospective medical records was obtained from RSUD Sanjiwani, Gianyar, Bali, Indonesia. The study was conducted in accordance with the principles of the Declaration of Helsinki. All patient records were anonymized before data processing and analysis.

Consent to Participate

The requirement for written informed consent was waived by the Research Ethics Committee, Faculty of Medicine, Udayana University, because of the retrospective study design and the use of fully anonymized patient records

Consent for Publication

Not applicable. This manuscript contains no personally identifiable patient information, and only anonymized and aggregated data are reported.

Competing Interests

The authors declare no competing financial or non-financial interests related to this study.

Reference

[1] U. D. 2023, “UNAIDS DATA 2023. Geneva: Joint United Nations Programme on HIV/AIDS; 2023. Licence: CC BY-NC-SA 3.0 IGO.,” *Medizinische Klin. - Intensivmed. und Notfallmedizin*, vol. 119, no. 8, pp. 614–623, 2024.

[2] X. Wu *et al.*, “Associations of modern initial antiretroviral therapy regimens with all-cause mortality in people living with HIV in resource-limited settings: a retrospective multicenter cohort study in China,” *Nat. Commun.* , vol. 14, no. 1, 2023, doi: 10.1038/s41467-023-41051-w.

[3] G. S. Argaw *et al.*, “Survival and predictors of mortality among HIV-infected adults after initiation of antiretroviral therapy in Eastern Ethiopia Governmental hospitals, from January 2015 to December 2021 (multi-center retrospective follow-up study),” *BMC Infect. Dis.*, vol. 24, no. 1, 2024, doi: 10.1186/s12879-024-10225-2.

[4] B. Yadecha, G. Fetensa, M. Abebe, and M. Mekuria, “Risk factors for early mortality among HIV-positive adults on antiretroviral therapy in Ethiopian health facilities,” *Sci. Rep.*, vol. 15, no. 1, pp. 1–10, 2025, doi: 10.1038/s41598-025-23129-1.

[5] E. Sackeya *et al.*, “Survival trends among people living with human immunodeficiency virus on antiretroviral treatment in two rural districts in Ghana,” *PLoS One*, vol. 19, no. 3 March, pp. 1–13, 2024, doi: 10.1371/journal.pone.0290810.

[6] World Health Organization, “Identifying common opportunistic infections among people with advanced HIV disease: Policy Brief,” *World Heal. Organ.*, pp. 1–28, 2023, [Online].

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

Available: https://www.who.int/publications/i/item/9789240084957

[7] B. Z. Woldegeorgis, Y. sisay Asgedom, A. Y. Gebrekidan, G. A. Kassie, U. D. Borko, and M. S. Obsa, "Mortality and its predictors among human immunodeficiency virus-infected children younger than 15 years receiving antiretroviral therapy in Ethiopia: a systematic review and meta-analysis," *BMC Infect. Dis.*, vol. 24, no. 1, pp. 1–14, 2024, doi: 10.1186/s12879-024-09366-1.

[8] M. K. Gabre, T. B. Tafesse, L. A. Geleta, C. K. Asfaw, and H. A. Delelegn, "The effect of late presentation on HIV related mortality among adolescents in public hospitals of north showa zone Oromiya, Ethiopia; 2022: a retrospective cohort study," *BMC Infect. Dis.*, vol. 24, no. 1, pp. 1–9, 2024, doi: 10.1186/s12879-024-09550-3.

[9] J. Ma *et al.*, "Antiretroviral treatment interruption and resumption within 16 weeks among HIV-positive adults in Jinan, China: a retrospective cohort study," *Front. Public Heal.*, vol. 11, 2023, doi: 10.3389/fpubh.2023.1137132.

[10] Q. Cai *et al.*, "Survival prediction models for people living with HIV based on four machine learning models," *Sci. Rep.*, vol. 15, no. 1, pp. 1–11, 2025, doi: 10.1038/s41598-025-16479-3.

[11] N. N. K. Krisnawijaya, B. Tekinerdogan, C. Catal, and R. van der Tol, "Multi-Criteria decision analysis approach for selecting feasible data analytics platforms for precision farming," *Comput. Electron. Agric.*, vol. 209, no. March, p. 107869, 2023, doi: 10.1016/j.compag.2023.107869.

[12] Q. Sui, G. Li, Y. Peng, J. Zhang, Y. Zhang, and R. Zhao, "Scalable and robust machine learning framework for HIV classification using clinical and laboratory data," *Sci. Rep.*, vol. 15, no. 1, pp. 1–27, 2025, doi: 10.1038/s41598-025-00085-4.

[13] Y. Chen, K. Pan, X. Lu, E. Maimaiti, and M. Wubuli, "Machine learning-based prediction of mortality risk in AIDS patients with comorbid common AIDS-related diseases or symptoms," *Front. Public Heal.*, vol. 13, no. March, pp. 1–18, 2025, doi: 10.3389/fpubh.2025.1544351.

[14] K. A. Yeneakal, G. H. Teferi, T. T. Mihret, A. K. Mengistu, S. B. Tizie, and M. M. Tadele, "Predicting antiretroviral therapy adherence status of adult HIV-positive patients using machine-learning Northwest, Ethiopia, 2025," *BMC Med. Inform. Decis. Mak.*, vol. 25, no. 1, 2025, doi: 10.1186/s12911-025-03106-4.

[15] D. Yehadji *et al.*, "Development of machine learning algorithms to predict viral load suppression among HIV patients in Conakry (Guinea)," *Front. Artif. Intell.*, vol. 8, no. March, pp. 1–10, 2025, doi: 10.3389/frai.2025.1446876.

[16] Y. Hu *et al.*, "Beyond Comparing Machine Learning and Logistic Regression in Clinical Prediction Modelling: Shifting from Model Debate to Data Quality," *J. Med. Internet Res.*, vol. 27, 2025, doi: 10.2196/77721.

[17] W. Srivilaithon and P. Thanasarnpaiboon, "Performance of machine learning models in predicting difficult laryngoscopy in the emergency department: a single-centre retrospective study comparing with conventional regression method," *BMC Emerg. Med.*, vol. 25, no. 1, 2025, doi: 10.1186/s12873-025-01185-0.

[18] Q. Mao *et al.*, "Predicting disease progression in people living with HIV using machine learning and a nomogram: a 10-year cohort study based in Xinjiang, China," *BMJ Open* , vol. 15, no. 11, 2025, doi: 10.1136/bmjopen-2025-105026.

[19] Z. Sadeghi *et al.*, "A review of Explainable Artificial Intelligence in healthcare," *Comput. Electr. Eng.*, vol. 118, no. PA, p. 109370, 2024, doi: 10.1016/j.compeleceng.2024.109370.

[20] L. H. Lai *et al.*, "The Use of Machine Learning Models with Optuna in Disease Prediction," *Electron.*, vol. 13, no. 23, pp. 1–20, 2024, doi: 10.3390/electronics13234775.

[21] L. Aissaoui Ferhi, M. Ben Amar, F. Choubani, and R. Bouallegue, "Enhancing diagnostic accuracy in symptom-based health checkers: a comprehensive machine learning approach with clinical vignettes and benchmarking," *Front. Artif. Intell.*, vol. 7, no. October, 2024, doi: 10.3389/frai.2024.1397388.

[22] P. Cerda, G. Varoquaux, and B. Kégl, "Similarity encoding for learning with dirty categorical variables," *Mach. Learn.*, vol. 107, no. 8–10, pp. 1477–1494, 2018, doi: 10.1007/s10994-018-5724-2.

[23] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach. Learn.*, vol. 113, no. 7, pp. 4903–4923, 2024, doi: 10.1007/s10994-022-06296-4.

[24] C. Meaney, X. Wang, J. Guan, and T. A. Stukel, "Comparison of methods for tuning machine learning model hyper-parameters: with application to predicting high-need high-cost health care users," *BMC Med. Res. Methodol.*, vol. 25, no. 1, 2025, doi: 10.1186/s12874-025-02561-x.

[25] B. Bischl *et al.*, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 13, no. 2, pp. 1–43, 2023, doi: 10.1002/widm.1484.

[26] K. Nam and J. Kim, "Determination and application of hyperparameters for landslide susceptibility assessment using Optuna: the

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

Eunsan shallow landslides occurred on August 14, 2022, in Buyeo City, Chungcheong Province, Korea," *Geosci. J.*, vol. 29, no. 6, pp. 978–994, 2025, doi: 10.1007/s12303-025-00054-z.

[27] K. Shen, H. Qin, J. Zhou, and G. Liu, "Runoff Probability Prediction Model Based on Natural Gradient Boosting with Tree-Structured Parzen Estimator Optimization," *Water (Switzerland)*, vol. 14, no. 4, 2022, doi: 10.3390/w14040545.

[28] M. Wiens, A. Verone-Boyle, N. Henscheid, J. T. Podichetty, and J. Burton, "A Tutorial and Use Case Example of the eXtreme Gradient Boosting (XGBoost) Artificial Intelligence Algorithm for Drug Development Applications," *Clin. Transl. Sci.*, vol. 18, no. 3, 2025, doi: 10.1111/cts.70172.

[29] N. H. Alhumaidi, D. Dermawan, H. F. Kamaruzaman, and N. Alotaiq, "The Use of Machine Learning for Analyzing Real-World Data in Disease Prediction and Management: Systematic Review," *JMIR Med. Informatics*, vol. 13, 2025, doi: 10.2196/68898.

[30] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-Augu, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[31] Z. Lu *et al.*, "Interpretable machine-learning strategy for soft-magnetic property and thermal stability in Fe-based metallic glasses," *npj Comput. Mater.*, vol. 6, no. 1, pp. 1–9, 2020, doi: 10.1038/s41524-020-00460-x.

[32] M. Skliarov, R. El Shawi, C. Dhaoui, and N. Ahmed, "A comparative evaluation of explainability techniques for image data," *Sci. Rep.*, vol. 15, no. 1, pp. 1–18, 2025, doi: 10.1038/s41598-025-25839-y.

[33] A. V. Ponce-Bobadilla, V. Schmitt, C. S. Maier, S. Mensing, and S. Stodtmann, "Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development," *Clin. Transl. Sci.*, vol. 17, no. 11, pp. 1–15, 2024, doi: 10.1111/cts.70056.

[34] M. M. Ahamad *et al.*, "Early-Stage Detection of Ovarian Cancer Based on Clinical Data Using Machine Learning Approaches," *J. Pers. Med.*, vol. 12, no. 8, pp. 1–16, 2022, doi: 10.3390/jpm12081211.

[35] T. F. Monaghan *et al.*, "Foundational statistical principles in medical research: Sensitivity, specificity, positive predictive value, and negative predictive value," *Med.*, vol. 57, no. 5, pp. 0–6, 2021, doi: 10.3390/medicina57050503.

[36] L. Hoessly, "On misconceptions about the Brier score in binary prediction models," *Glob. Epidemiol.*, vol. 11, no. December 2025, p. 100242, 2026, doi: 10.1016/j.gloepi.2025.100242.

[37] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, no. 1, pp. 1–14, 2024, doi: 10.1038/s41598-024-56706-x.

[38] S. Chang, X. Wang, Y. Luo, and L. Jia, "Data augmentation alters feature importance in XGBoost for CVD prediction," *Sci. Rep.*, vol. 15, no. 1, pp. 1–11, 2025, doi: 10.1038/s41598-025-26228-1.

[39] Y. Jin *et al.*, "SHAP-based interpretable machine learning for Parkinson's disease severity prediction: integrated analysis of clinical and environmental features," *Front. Neurol.*, vol. 16, no. September, pp. 1–20, 2025, doi: 10.3389/fneur.2025.1678463.

[40] T. Miguel-Medina *et al.*, "Development of a Machine Learning-Based Predictive Model and Clinically Oriented Web Application for 30-Day Mortality Following Cardiac Surgery," *Sensors*, vol. 26, no. 5, pp. 1–18, 2026, doi: 10.3390/s26051656.

[41] Y. Jang, "Feature-based ensemble modeling for addressing diabetes data imbalance using the SMOTE, RUS, and random forest methods: a prediction study," *Ewha Med. J.*, vol. 48, no. 2, p. e32, 2025, doi: 10.12771/emj.2025.00353.

[42] R. M. Munshi *et al.*, "Optimising hyperparameters with a tree structured Parzen estimator to improve diabetes prediction," *Sci. Rep.*, vol. 15, no. 1, pp. 1–10, 2025, doi: 10.1038/s41598-025-19295-x.

[43] M. Shi *et al.*, "Machine learning-based in-hospital mortality prediction of HIV/AIDS patients with Talaromyces marneffei infection in Guangxi, China," *PLoS Negl. Trop. Dis.*, vol. 16, no. 5, pp. 1–18, 2022, doi: 10.1371/journal.pntd.0010388.

[44] L. Mugenyi *et al.*, "Effect of the 'universal test and treat' policy on the characteristics of persons registering for HIV care and initiating antiretroviral therapy in Uganda," *Front. Public Heal.*, vol. 11, no. June, 2023, doi: 10.3389/fpubh.2023.1187274.

[45] M. K. Kitenge, G. Fatti, I. Eshun-Wilson, O. Aluko, and P. Nyasulu, "Prevalence and trends of advanced HIV disease among antiretroviral therapy-naïve and antiretroviral therapy-experienced patients in South Africa between 2010-2021: a systematic review and meta-analysis," *BMC Infect. Dis.*, vol. 23, no. 1, pp. 1–16, 2023, doi: 10.1186/s12879-023-08521-4.

[46] P. H. A. C. Leite *et al.*, "Early mortality in a cohort of people living with HIV in Rio de Janeiro, Brazil, 2004–2015: a persisting problem," *BMC Infect. Dis.*, vol. 22, no. 1, pp. 1–12, 2022, doi: 10.1186/s12879-022-07451-x.

[47] M. Rosenblatt, L. Tejavibulya, R. Jiang, S. Noble, and D. Scheinost, "Data leakage inflates prediction performance in connectome-based

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

machine learning models," *Nat. Commun.* , vol. 15, no. 1, pp. 1–15, 2024, doi: 10.1038/s41467-024-46150-w.

[48] N. P. H. Lugito, M. I. Simanjuntak, S. Sumantri, Y. Kho, S. Kosayuz, and R. Liauw, "Functional status as a predictor of loss to follow-up among people living with HIV on antiretroviral therapy in Tangerang District, Indonesia," *BMC Infect. Dis.*, vol. 26, no. 1, 2026, doi: 10.1186/s12879-025-12470-5.

[49] B. Z. Woldegeorgis, Y. S. Asgedom, A. Habte, G. A. Kassie, and A. S. Badacho, "Highly active antiretroviral therapy is necessary but not sufficient. A systematic review and meta-analysis of mortality incidence rates and predictors among HIV-infected adults receiving treatment in Ethiopia, a surrogate study for resource-poor settings," *BMC Public Health*, vol. 24, no. 1, pp. 1–19, 2024, doi: 10.1186/s12889-024-19268-1.

[50] T. W. Menza, L. K. Hixson, L. Lipira, and L. Drach, "Social Determinants of Health and Care Outcomes among People with HIV in the United States," *Open Forum Infect. Dis.*, vol. 8, no. 7, pp. 1–10, 2021, doi: 10.1093/ofid/ofab330.

[51] M. Sohail *et al.*, "Assessing the Impact of Social Determinants of Health on HIV Care Engagement in the Southern United States: A Cross-Sectional Study," *J. Int. Assoc. Provid. AIDS Care*, vol. 23, pp. 1–10, 2024, doi: 10.1177/23259582241251728.

[52] F. Pargent, F. Pfisterer, J. Thomas, and B. Bischl, "Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features," *Comput. Stat.*, vol. 37, no. 5, pp. 2671–2692, 2022, doi: 10.1007/s00180-022-01207-6.

[53] M. J. Raihan, M. A. M. Khan, S. H. Kee, and A. Al Nahid, "Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP," *Sci. Rep.*, vol. 13, no. 1, pp. 1–15, 2023, doi: 10.1038/s41598-023-33525-0.

[54] J. B. Guimbaud *et al.*, "Machine learning-based health environmental-clinical risk scores in European children," *Commun. Med.*, vol. 4, no. 1, 2024, doi: 10.1038/s43856-024-00513-y.

[55] O. Wysocki *et al.*, "Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making," *Artif. Intell.*, vol. 316, 2023, doi: 10.1016/j.artint.2022.103839.

[56] L. Coombs *et al.*, "A machine learning framework supporting prospective clinical decisions applied to risk prediction in oncology," *npj Digit. Med.*, vol. 5, no. 1, pp. 1–9, 2022, doi: 10.1038/s41746-022-00660-3.

[57] E. Sezgin, "Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers," *Digit. Heal.*, vol. 9, pp. 1–5, 2023, doi: 10.1177/20552076231186520.

[58] B. H. Chew and K. Y. Ngiam, "Artificial intelligence tool development: what clinicians need to know?," *BMC Med.*, vol. 23, no. 1, 2025, doi: 10.1186/s12916-025-04076-0.

Author Biography

Ni Putu Likayuni Viona is an undergraduate student in the Department of Information Technology at Universitas Pendidikan Nasional. Her academic and research interests focus on machine learning, data analytics, artificial intelligence, and health information systems. Her current research explores the application of explainable machine learning for clinical outcome prediction, particularly mortality prediction among people living with HIV receiving antiretroviral therapy. She is interested in the development of data-driven approaches that combine predictive performance with model interpretability to support healthcare decision-making. In addition to her academic research, she is also interested in the broader application of artificial intelligence and data analytics in healthcare and information systems. She is open to further research and collaboration in machine learning, medical informatics, health information systems, and related data analytics domains.

Ngakan Nyoman Kutha Krisnawijaya is an assistant professor at the Department of Information Technology. His research focuses on data management and analytics, with particular emphasis on AI development and parallel computing. He has published journal articles in several highly reputable journals across computer science, information technology, and data science domains. He is open to further collaboration in the Health Information Systems domain and in other data analytics-related domains. In addition to research, he is actively involved in teaching activities and professional organizations. He currently leads the AI development research center at the Universitas Pendidikan Nasional and conducts several research projects to provide AI-based solutions.

Desak Nym Widyanthini is an Assistant Professor in the Public Health Study Program under the Faculty of Medicine at Udayana University, Bali, Indonesia. She holds a Master of Public Health (M.Kes) degree from Udayana University, which she completed in 2014.. Her primary research focus addresses critical issues in

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383

public health, maternal and child health (MCH), and adolescent reproductive health, with specialized attention directed toward infectious disease prevention, particularly HIV/AIDS and sexually transmitted infections (STIs). A recipient of two Australia Awards Fellowships in 2016, she holds advanced training in public health program evaluation and infectious disease epidemiology.

Manuscript received 20 June 2026; Revised 25 July 2026; Accepted 15 Agusut 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.383