Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predicting HIV treatment interruption in Ghana: development and evaluation of an individual-level Machine Learning risk model

Williams Kwarah

Kwarah@gmail.com

University of Ghana

Frances Baaba da-Costa Vroom

University of Ghana

Duah Dwomoh

University of Ghana

Samuel Bosomprah

University of Ghana

Research Article

Keywords: HIV/AIDS, Treatment Interruption, Machine Learning, Predictive Modeling, Ghana, eXtreme Gradient Boosting, SHAP analysis

Posted Date: March 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9125259/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Treatment interruption (TI) among people living with HIV (PLHIV) remains a critical barrier to achieving sustained viral suppression and reducing HIV-related mortality in Ghana. Approximately 41% of PLHIVs experience TI within the first six months of treatment. This study developed and validated an individual-level machine learning (ML) predictive model to identify individuals at high risk of TI, enabling precision-based retention strategies in Ghana.

Methods

A retrospective predictive study was conducted using routinely collected data from Ghana's national HIV eTracker database for 33,613 PLHIV (contributing 227,319 clinic visits) from 245 health facilities across six regions between January 2019 and December 2023. Seven ML algorithms - XGBoost, Random Forest, Logistic Regression, Naïve Bayes, AdaBoost, K-Nearest Neighbors, and Support Vector Machines - were trained and evaluated using 10-fold cross-validation. Model performance was primarily assessed using the Area Under the Curve (AUC-ROC), alongside sensitivity, specificity, and calibration analysis.

Results

The XGBoost algorithm demonstrated superior discriminative performance with an AUC-ROC of 0.899 (95% CI: 0.869–0.911). On the hold-out test dataset, the tuned XGBoost model achieved an accuracy of 0.832, a sensitivity of 0.791, and a specificity of 0.851. SHapley Additive exPlanations (SHAP) analysis revealed that behavioural features were the strongest predictors of TI, with the "maximum days between visits" and "proportion of late visits" ranking as the most influential factors. Calibration plots confirmed that the model's predicted risk scores closely aligned with actual observed frequencies, ensuring clinical reliability.

Conclusions

We developed and validated an individual-level ML model that accurately predicts TI among PLHIV in Ghana and offered transparent explanations of key risk drivers. This model could facilitate targeted interventions thereby optimizing resource allocation and improving long-term retention in HIV care.

Introduction

Treatment interruption (TI) among people living with HIV (PLHIV) remains a critical challenge in the global effort to achieve sustained viral suppression and reduce HIV-related morbidity and mortality. Globally, 39 million people live with HIV, and Africa bears approximately 88% of the cases [1]. Yet even

with effective antiretroviral therapy (ART) available, adherence gaps persist with only 46–85% of clients remaining on ART two years after initiation as reported in a previous study [2, 3]. In Ghana, HIV prevalence is relatively low (~ 1.5% in 2024), but retention on ART remain a challenge [4]. A recent study in Ghana reported about 68% of PLHIV missed at least one scheduled visit over a five-year cohort follow-up, with TI at 41% within the first six months [5]. Conventional statistical analyses such as multivariable logistic regression or Cox regression models have shown that social factors significantly reduce the odds of lost to follow-up (LTFU) [4]. These regression-based models typically identify a few baseline predictors of LTFU, but treat large client subgroups as homogeneous and assume risk factors remain static over time [6]. However, the reality is that adherence is driven by complex, time-varying behavioural, social, clinical, and structural factors that traditional models may miss [7, 8].

Machine learning (ML) models offer an alternative to static models by harnessing large, complex clinical datasets to uncover nonlinear risk patterns. Unlike conventional regression, ML algorithms can integrate longitudinal health and behavioural data to better predict individuals at risk of dropping out of care [9]. In Ghana, there is no documented ML-based retention tool currently in use; but published studies have identified cost, clinic access, and stigma as risk factors for treatment interruptions [10]. Predictive analytics could enable “precision” care; for example, by stratifying clients according to risk, scarce resources (e.g. adherence counseling or reminder systems), and can focus on those most likely to disengage [11]. Indeed, ML-based risk scores have been shown to quantify a client’s risk of future disengagement and prioritise support for those most in need elsewhere - an approach valuable in Ghana’s resource-constrained setting [12].

In this study, we developed and validated a ML-based risk prediction model for HIV TI in Ghana using client behaviour, clinical and system-level factors to inform how precise HIV retention interventions are designed and delivered.

Methods

Study design and participants

We conducted a retrospective predictive study with data routinely collected from Ghana’s national HIV programme eTracker database. The eTracker is an electronic medical records platform used in all public-sector health facilities that provide HIV services. It captures transactional data including client’s demographics, clinical assessments, pharmacy refills, and visit outcomes. For this study, we extracted all ART clinics visit records from the HIV eTracker database between 1 January 2019 and 31 December 2023 from 245 health facilities in Western, Western North, Ahafo, Volta, Upper east and Bono Regions of Ghana. The criteria for selecting these regions have been explained in a previous study [5]. The data extraction was done on August 14, 2024. The resulting dataset had 33,613 PLHIV who contributed 227,319 clinic visits during the study period.

The study population included PLHIV enrolled into HIV care at the selected facilities with at least one documented clinic visit within the study period. Individuals were eligible if they had a confirmed HIV diagnosis, a unique client identifier, and at least one recorded visit within the study period. For the analysis, we focused on clients receiving ART and excluded records with missing or implausible variables such as participant on family planning, herbal medication, prophylaxis for opportunistic infections, tuberculosis (TB) treatment, adverse drug reaction, and TB prevention therapy. Each participant was followed from their first qualifying visit within the study period until their last recorded clinic encounter, documented death, transfer out of the health facility, or administrative censoring on 31 December 2023, whichever occurred first. The final dataset included 54 variables - demographic, clinical, facility, and engineered behaviour variables.

Data cleaning and preprocessing

First, all date variables were standardized into unambiguous standard formats to ensure data consistency before analysis. This preprocessing step ensured accurate calculation of the time intervals. To reconstruct the trajectory of each study participant's engagement with care, the entire dataset was sequentially sorted first by the unique participant's identifier and then by the follow-up visit date to obtain the correct sequence of visits, which was fundamental for generating the behaviour features. The final dataset was divided into two; 70% train set and 30% test set. The train set was used to build and train the model; while the 30% test set was used to test the developed model. Splitting the data into train and test sets improved the model selection process, and avoided overfitting which enabled an evaluation of the model's performance on the unseen test set [13–18]. Data cleaning, preprocessing and analysis was done using R version 4.5.1.

Missing data handling and management

Missing data, which is an inherent problem in routine health records such as the national HIV eTracker, were managed through a multi-stage preprocessing strategy. An initial assessment of missingness patterns was made which led to the exclusion of variables with more than 35% missing values to minimize noise. For the retained variables with less than 35% missingness, we applied K-Nearest Neighbors (KNN) imputation for the continuous variables to capture non-linear dependencies, and mode imputation for the categorical variables. Instances of structural missingness such as pending viral loads tests were retained as distinct categories. To ensure rigorous model evaluation, the imputation pipeline was fitted on the training dataset and applied unchanged to the test set, preventing information leakage. The derived behavioural features were computed after the imputation. The data integrity was confirmed through validation checks on value ranges and post-imputation distributions.

Feature selection and engineering

Of a total of 54 potential features, 33 features were used in the final model development. These features included participant demographics (age, sex, highest educational level, occupation, marital status, has a mobile phone, has physical house address, has health insurance, and client type), facility characteristics (region, settings, number of staff, years ART clinic operated, ART clinic status, facility type, and facility ownership), clinical features (HIV type, ART initiation timing, TB screening at initiation, TB screening at follow-up, antiretroviral (ARV) class combination, multi-month dispensing, family planning use, and first viral load test time), and behavioural features (maximum days between visits, proportion of late visits, number of late visits, total visits, and average number of days between visits). We systematically transformed the original visit-level dataset into a client-level analytic set, which facilitated temporal patterns to be quantified. The behavioural features were created to numerically represent each participant's history and pattern of engagement with the health facilities. This approach did not only improve predictive accuracy but also enhanced the model interpretability.

To construct these behavioural features, the transformed client-level dataset had each row representing a unique participant (n = 33,613). For each participant, the number of ARV pills dispensed at the previous visit was used to calculate the expected return date. A visit was then flagged as "late" if it occurred more than seven days after the expected return date. This 7-day threshold reflected the degree of flexibility commonly found in clinical practice while aligning with the concept of ‘at-risk of TI’ as has been reported in prior studies [19]. To quantify the behavioural patterns, we created the following new features:

1. *proportion_late_visits*: This measured the consistency of lateness by calculating the percentage of a participant’s total visits that were flagged as late. It distinguished a long-term participant who was late twice in thirty visits versus a new participant who was late twice in three visits.
2. *count_late_visits*: This provided a simple frequency count of the total number of times a participant was late for an appointment.
3. *max_days_between_visits*: This captured historical periods of interruption by identifying the single longest time gap a participant ever had between two consecutive visits.
4. *avg_days_between_visits* & *std_dev_days_between_visits*: These metrics measured the regularity of a client's visit schedule. A high standard deviation *(std_dev_days_between_visits)* indicated an erratic and unpredictable pattern of attendance.

Model development and evaluation strategy

Data partitioning

The final dataset was divided into training and test set. A stratified sampling method was used for this split, ensuring that the proportion of participants who experienced TI was the same in both the training and test sets (approximately 68.3%). This sampling method was crucial for preventing bias and ensuring that both datasets are representative of the overall participant population. The training set (70%) comprised of 23,530 participants was used for training the algorithms, model development and comparing the model performance. The test set (30%) contained 10,083 participants and was used only

at the end to assess the best model's performance on data it had never encountered before, thereby simulating a real-world application.

Candidate ML algorithms

To identify the most effective predictive approach, seven machine learning algorithms - eXtreme Gradient Boosting (XGBoost), Random Forest, logistic regression, Naïve Bayes, AdaBoost, K-Nearest Neighbors, and Support Vector Machines - were trained and evaluated using a 10-fold cross validation. These algorithms were chosen because of their robustness, effectiveness, and utility in clinical predictive model development as demonstrated in a prior studies [20]. These diverse sets of algorithms reflected the complementary strengths of each model for binary outcome prediction. Logistic regression provides a strong linear baseline with stable estimates and clear coefficient interpretation, which supports transparency and clinical plausibility [21, 22]. Random Forest captures nonlinear relationships and higher-order interactions through ensemble decision trees, offering robustness to noise and strong performance with mixed data types [23–25]. XGBoost extends the tree-based learning algorithms through gradient boosting, sequentially correcting errors and improving discrimination in complex, high-dimensional data, which suits longitudinal behavioural patterns [26–29]. Naive Bayes offers a probabilistic framework with fast training and strong performance under a conditional independence assumptions, serving as a useful benchmark [30, 31]. K-nearest neighbors represents a distance-based approach that detects local structure in the feature space without parametric assumptions [32]. AdaBoost using the AdaBag R package trains classifiers sequentially, focusing on correcting previous errors by increasing weights of misclassified data [33–35]. Support Vector Machines provides a margin-based classifier designed to separate classes in high-dimensional spaces, offering a contrast to tree-based and probabilistic methods [36, 37]. Together, these algorithms span linear, nonlinear, probabilistic, distance-based, and ensemble learning paradigms, supporting a rigorous and comparative evaluation for the most appropriate algorithms to predict treatment interruption using the study dataset.

Model validation

To ensure that the performance estimates for each model were stable and reliable, the dataset was split into 70% training set used for model development and 30% test set for validating the final model. A 10-fold cross-validation technique was implemented on the training set. This process involved randomly dividing the training set into 10 equal-sized folds or subsets and iterating model training 10 times. In each iteration, one-fold was held out as a temporary validation set, while the model was trained on the remaining nine folds. The trained model was then tested on the held-out fold, and its performance was recorded. The performance metrics from the 10 iterations were then averaged to produce a single, robust estimate of the model's performance. This technique was crucial for mitigating the risk of overfitting, where a model learns the training data too well (including its noise) and fails to generalize to new data.

Predictive model evaluation metric

The primary metric used to compare the performance of the candidate models was the AUC-ROC. The AUC-ROC is a standard metric for binary classification tasks that measures a model's overall ability to discriminate between two classes (in this case, participants who have interrupted treatment versus those who did not interrupt treatment). An AUC-ROC value of 1.0 represents a perfect model discrimination, while a value of 0.5 represents a model with no better-than-random-chance predictive ability [38]. This metric was chosen as the primary metric because it provided a comprehensive performance summary that is independent of any specific classification threshold. In addition to this performance measure, model accuracy, sensitivity, and specificity were reported.

Model optimisation and evaluation

To maximize the predictive power of the best performing model, hyperparameter tuning was conducted. Hyperparameters are the internal settings of the algorithm that were not learned from the data but were set before the training process begun. A grid search methodology was used to systematically explore a predefined range of hyperparameter settings to obtain the optimal combination that provided the best model performance. These key hyperparameter settings were *nrounds* which is the total number of decision trees to be built sequentially and was set at 100, 200,300; *max_depth* which controlled the complexity of the model by ensuring maximum depth of each decision tree was attained and was set at 4, 6, 8; *learning rate (eta)* which controlled the step size at each iteration, preventing the model from making overly aggressive corrections and was set at 0.05, 0.1; *gamma* which is the minimum loss reduction required to split a node and was set at 0; *colsample_bytree* which is the fraction of predictors randomly sampled for each tree and was set at 0.80; *min_child_weight* which is the minimum sum of instance weights required in a leaf node and was set at 1; and *subsample* which is the fraction of training observations used to grow each tree and was set at 0.80.

Using the 10-fold cross-validation with 3 values for the *nrounds*, 3 values for *max_depth*, and 2 values for *eta* gave 18 combinations for the models' evaluation. The combination that yielded the highest mean AUC-ROC values was reported for the final model, ensuring it was well-optimized for the dataset. The final tuned model's performance was evaluated on the hold-out test data set, providing an unbiased assessment of the model's ability to generalize new, unseen data. A comprehensive set of performance metrics, including accuracy, sensitivity, specificity, and the AUC-ROC values with 95% confidence interval, were calculated and reported for the final tuned and tested model.

Model interpretability for clinical use

A central requirement for this study was to ensure the final model was not a "black box". To achieve this, the SHapley Additive exPlanations (SHAP) analysis technique was used to assess the final model's decision-making process. SHAP analysis is based on principles from cooperative game theory that works by calculating the contribution of each feature in a model to each individual prediction [39]. The output of this process was a SHAP value for every feature for every participant. This value quantified the

contribution of each feature and the direction of the prediction (in this case towards a higher or lower risk of TI). Mean Absolute SHAP values were calculated and represented the average magnitude of a feature's impact on the model's output, regardless of the direction, allowing for a clear ranking of the most influential predictors.

Model calibration

The reliability of the model's predicted probabilities was assessed to ensure reliability for application within clinical settings. A calibration plot was generated to investigate the model's reliability. The plot compared the predicted probabilities from the model against the actual observed proportion of TI cases within different probability bins. For example, the calibration plot answered the question 'for all participants that the model assigned an 80% risk of TI, what percentage of participants actually had a TI?'. A model is considered well-calibrated if its predictions align closely with the observed outcomes, following a diagonal line on the calibration plot [40].

Results

Study participants characteristics

A total of 33,613 participants contributed 227,319 follow-up visits in the analysis. The mean age of the participants was 35.6 (± 14) years, ranging from 0 years (less than 1 year) to 89 years. Females were the majority (71.6%) while 28.4% were males. More than half, (64.9%), had at least secondary education, 17.2% had primary education, while 18.9% had no education. Unemployment or no paid work was among 35.1% participants, while 64.9% participants were employed or engaged in paid work. About 53.6% of the participants were married or cohabiting, while 15.8% were divorced/separated/widowed, and 30.9% of participants were single. A small proportion (2.1%) of the participants ever smoked, while 75.0% of participants provided mobile phone numbers at initiation. More than a quarter (28.4%) of participants were in the Western region, 23.0% in the Bono region, and 20.3% in the Volta region. More than half (55.4%) of the participants lived in rural areas. The participant characteristics have been reported in a previous study [5].

Predictive model development and selection

The XGBoost model had the best AUC-ROC of 0.899 (95% CI: 0.869–0.911) compared to the other six candidate algorithms. The full model performance for each algorithm with a 10-fold cross validation is summarized in Table 1. The AUC-ROC curves for each model are shown in Fig. 1. The Support Vector Machine (SVM) algorithm failed to converge in 90% of validation folds. This computational instability suggests that SVM is ill-suited for this dataset, likely due to the high dimensionality introduced by categorical facility and regional variables.

Table 1
Model performance assessment across the 10 validation folds

Algorithm	ROC-AUC	Sensitivity	Specificity
XGBoost	0.899	0.773	0.849
Random Forest	0.896	0.726	0.865
Logistic Regression	0.887	0.783	0.837
Naive Bayes	0.867	0.639	0.878
AdaBoost	0.841	0.782	0.833
KNN	0.815	0.573	0.841
SVM	*0.881*	*Unstable*	*Unstable*

For the final model selection, the XGBoost algorithm demonstrated the highest mean AUC-ROC, suggesting superior discriminative capability and performance. The final tuned XGBoost model evaluated on the hold-out test dataset had an accuracy of 0.832 (95% CI: 0.825–0.840), sensitivity of 0.791, specificity of 0.851 and Kappa score of 0.624, suggesting stable and consistent high discriminative capability.

Reliability of the final predictive model

Beyond discriminative accuracy, a crucial characteristic of the predictive model was the reliability of its probability estimates. The calibration analysis conducted on the final tuned XGBoost model determined that the predicted risk scores corresponded to the actual observed frequencies of the outcome as shown in the calibration plot in Fig. 2.

The calibration plot (left panel) compared the mean predicted probabilities against the observed actual proportion of TI cases within deciles of risk. The dashed diagonal line represented a state of perfect calibration, where the predicted probabilities perfectly matched the observed frequencies. From the results, the plot demonstrated that the XGBoost model was well-calibrated, as the plotted points were very close to the diagonal line across the full spectrum of risk. Further, the narrow confidence intervals for each point underscored the statistical reliability of the model. The distribution plot on the right panel is a histogram of the predicted probabilities for all study participants in the test set. A bimodal distribution is observed with a large concentration of study participants assigned very high-risk scores with risk probabilities near 1.0 and a wider distribution of participants at the lower end of the risk spectrum. This indicated that the model is effective at separating the participant population into distinct risk strata.

Interpretability of the tuned XGBoost model

To ensure the final model was not an uninterpretable "black box", the SHAP analysis provided a transparent and quantitative assessment of the model's logic process by explaining how each feature contributed to the risk prediction for individual participants. The engineered behaviour features were predominantly prioritized in the model prediction as shown in Fig. 3.

The SHAP summary plot in Fig. 4 visualized the distribution, direction, and magnitude of each feature's impact across all participants in the test set. The results indicated that the behavioural features were the dominant drivers of the model's predictions, ranking the top four behaviour features higher than any other static demographic or clinical feature.

The most influential feature was *max_days_between_visits* with a Mean Absolute SHAP value of 1.89. Figure 4 emphasises the finding that *max_days_between_visits* was primarily a protective factor where yellow dots represented a history of a long gap between visits and is associated with large negative SHAP values. This quantitatively demonstrated that the model learned to interpret re-engagement after a significant lapse in care as a strong signal of a participant's resilience thereby leading to a substantially reduced calculated risk for future interruption.

The proportion of late visits (*proportion_late_visits*) was the second significant factor with a Mean Absolute SHAP value of 0.890. From Fig. 4, the trend showed that as the proportion of participants' with late visits increased (moving from yellow to purple dots), the corresponding SHAP values became strongly positive, indicating that a consistent pattern of late attendance significantly contributed a positive term to a participant's risk score. This finding is supported by the third-ranked feature which is the number of late visits (*count_late_visits*) with a Mean Absolute SHAP value of 0.3438, demonstrating that the absolute frequency of lateness is a factor the model learned to associate with increased risk. In other words, a higher value of late visits was associated with higher risk of interrupting treatment.

Discussions

Our study successfully developed and validated an individual level machine learning (ML) model using the XGBoost algorithm to predict treatment interruption among PLHIV in Ghana. By leveraging routinely collected longitudinal data from the national HIV eTracker, the model achieved high predictive performance with an AUC-ROC of 0.899. These results extend the emerging evidence on the utility of ML techniques for HIV retention interventions in resource-limited settings. Our model performance was in the upper range of predictive performance compared to similar studies in Africa and Asia which reported AUC-ROC values were between 0.75 and 0.90.

The superior performance of the XGBoost algorithm is consistent with growing evidence that gradient boosting methods outperform traditional regression and simpler classifiers when modeling complex clinical and behavioural patterns. Generally, tree-based boosting methods can iteratively correct prediction errors and capture nonlinear relationships and feature interactions, which are common characteristics inherent in adherence behaviour and health service utilization [41]. Further, our model demonstrated strong calibration, meaning the predicted probabilities closely matched observed

outcomes. Calibration is critical for clinical decision-making and in this instance, the risk scores guide clinical intervention thresholds. For example, a poorly calibrated model may inform the misallocation of limited resources in settings where a significant portion of the population may miss scheduled visits. Unlike prior studies that failed to report calibration models [20, 42], our calibration measures demonstrated the robustness of the analytical approach and potential for adoption in clinical workflows.

A primary contribution to this work is the inclusion of behaviour features, which emerged as the strongest predictors of TI, consistently ranking above the static demographic and clinical variables. These features were the proportion of late visits, count of late visits and visit interval variability for a study participant. These findings suggest that TI is not a sudden event but rather the culmination of a "waning" engagement process by the recipient of care. The inclusion of behaviour features is consistent with findings from prior studies that have shown that past behaviours can potentially predict future disengagement from HIV care [43]. Interestingly, the SHAP analysis revealed that a history of long gaps between visits followed by successful re-engagement appeared protective, suggesting that a certain level of resilience among individuals who re-enter care either due to effective tracing mechanism or improved personal circumstances. This inclusion of the SHAP analysis illustrated the added value of explainable AI (XAI), which allowed clinicians to distinguish between chronic disengagement and temporary TI followed by a renewed commitment to care, emphasising the need to align predictive tools with clinical workflows and utility [44]. Furthermore, the use of SHAP analysis enhanced interpretability, addressing concerns regarding “black box” decision-making in clinical applications and fostering clinician trust [45, 46].

The inclusion of facility characteristics such as workload in terms of the number of staff at the ART clinics and regional infrastructure enabled the model to reflect structural barriers rather than attributing disengagement solely to individual behaviour. In Ghana, regional variations in health infrastructure and service delivery models influence retention patterns. Prior studies have highlighted that structural barriers played critical roles in PLHIVs disengaging in care [47]. Though we did not find these factors to strongly predict a participant’s risk of interrupting care, their inclusion in the model was important to account for any confounding.

The model enabled risk stratification to guide precision retention interventions in resource-limited systems where universal intensive follow-up is not feasible. Automated risk scoring within the eTracker platform could prioritise high-risk individuals for targeted adherence counseling, active tracing, and digital adherence tools like SMS reminders. Additionally, these scores can inform differentiated service delivery (DSD), prioritizing intensive contact for high-risk clients while safely extending multi-month dispensing of ARVs for low-risk clients. Such precision approaches improve efficiency and may increase viral suppression by focusing on resources where risk is highest as suggested in other studies [48, 49].

Despite these strengths, some limitations must be acknowledged. First, our inability to externally validate the model due to data access challenges limits the generalizability of the model application. Second, routine health data may be subject to recording errors and incomplete documentation leading to

missing and incomplete data. Third, although imputation strategies were applied, residual bias may remain. Finally, important psychosocial and structural determinants such as stigma, mental health, and exact travel costs were not captured in the dataset and could only be inferred through proxy features.

Future research should focus on prospective implementation studies to evaluate the model's impact on long-term clinical outcomes and cost-effectiveness within clinical workflows. Research should also formally evaluate algorithmic bias to ensure equitable risk stratification across key demographic groups. Also, external validation studies are critical to ensure model generalizability. Finally, exploring the integration of community-level and socioeconomic data may further improve predictive accuracy and model robustness.

Conclusion

Our model offered a robust and actionable tool for identifying individuals at risk of TI, thereby supporting efforts to improve retention and improve HIV treatment outcomes. This study demonstrated that machine learning can be applied to routine HIV program data to provide a robust tool for precise public health planning, optimizing resource allocation and strengthening HIV treatment outcomes in resource poor settings such as Ghana.

Abbreviations

Abbreviation	Meaning
HIV	Human Immunodeficiency Virus
AIDS	Acquired Immunodeficiency syndrome
PLHIV	People living with HIV
ART	Antiretroviral therapy
LFTU	Loss to follow up
ML	Machine Learning
AI	Artificial Intelligence
TI	Treatment interruption in treatment
SD	Standard Deviation
CI	confidence interval
NACP	National HIV/AIDS & STI Control Program
DSD	Differentiated Service Delivery
SVM	Support Vector Machines
AUC ROC	Area Under Curve Receiver Operating Characteristic
SHAP	Shapley Additive exPlanation
XAI	eXplainable AI
ARV	Anti-Retroviral drugs
XGBoost	eXtreme Gradient Boosting
KNN	K-Nearest Nerghbours

Declarations

Ethics approval and consent to participate

The study protocol was reviewed and approved by the Ghana Health Service Ethical Review Committee with approval number GHS-ERC:003/08/24. Adequate permissions were received from the National HIV/AIDS & Sexually Transmitted Infection (STI) Control Program (NACP) prior to the extraction of the datasets from the HIV eTracker. This secondary data analysis used de-identified records, and informed consent was waived. All ethical principles were followed in this study. Consent to participate is not applicable.

Consent for publication

Not applicable

Availability of data and materials

The datasets analysed are not publicly available due to the sensitivity but are available from the corresponding author on reasonable request.

Competing interests

All authors declare that they have no competing interests.

Funding

Not applicable

Authors' contributions

WK and SB conceived the study, WK led data management, model analysis and wrote the first draft of the manuscript. FBV, DD, and SB contributed to the methods, analysis, and reviewed the entire manuscript. All authors read and approved of the final manuscript.

Acknowledgements

We would like to express our sincere gratitude to the National HIV/AIDS & STI Control Programme, Ghana Health Service, for granting permission and access for the data to be used in this study. Also, our appreciation goes to Mr. Ekow Wiah and Mr. Philip Boakye at the NACP for extracting the data from the HIV eTracker for this work and being available to respond to follow-up queries and provide feedback.

References

1. 2023 - The path that ends AIDS UNAIDS Global AIDS Update.pdf. 2023. https://www.unaids.org/sites/default/files/media_asset/2023-unaids-global-aids-update_en.pdf. Accessed July 19.
2. Brinkhof, Martin W. G., François Dabis, Landon Myer, David R. Bangsberg, Andrew Boulle, Denis Nash, Mauro Schechter, et al. 2008. Early loss of HIV-infected patients on potent antiretroviral therapy programmes in lower-income countries. *Bulletin of the World Health Organization* 86: 559–567. https://doi.org/10.2471/blt.07.044248.
3. Dabis, Francois. 2008. Eearly loss of HIV-infected patients on potent antiretroviral therapy programmes in lower-income countries. *Bulletin of the World Health Organization* 86: 559–567. https://doi.org/10.2471/BLT.07.044248.
4. Mensah, Benedicta Ayiedu, Henrietta E. Mensah-Brown, Frederica D. Partey, Christabel Addo, Gertrude Buah, Gifty A. Afudego, Daniel Okyere, et al. 2025. Identifying risk factors for loss to follow-

up in adults living with HIV in a high-burden district in Ghana. *BMC Public Health* 25: 1042. https://doi.org/10.1186/s12889-025-22254-w.

5. Kwarah, Williams, Jasmin Kwarah, Yakubu Alhassan, Frances Baaba da-Costa Vroom, Duah Dwomoh, and Samuel Bosomprah. 2026. Predictors of HIV treatment interruption among people living with HIV in 6 regions in Ghana: a retrospective cohort study. *BMC Public Health* 26: 503. https://doi.org/10.1186/s12889-026-26243-5.
6. Maskew, Mhairi, Kieran Sharpey-Schafer, Lucien De Voux, Thomas Crompton, Jacob Bor, Marcus Rennick, Admire Chirowodza, et al. 2022. Applying machine learning and predictive modeling to retention and viral suppression in South African HIV treatment cohorts. *Scientific Reports* 12. Nature Publishing Group: 12715. https://doi.org/10.1038/s41598-022-16062-0.
7. Fox, Matthew P., and Sydney Rosen. 2010. Patient retention in antiretroviral therapy programs up to three years on treatment in sub-Saharan Africa, 2007–2009: systematic review. *Tropical Medicine & International Health* 15: 1–15. https://doi.org/10.1111/j.1365-3156.2010.02508.x.
8. Kim, Gwang Suk, Youngin Kim, Min Kyung Park, Sang A. Lee, Youngjin Lee, and Mi-So Shim. 2022. Barriers and facilitators on the HIV care continuum among adults living with HIV in high-income countries: a scoping review protocol. *Systematic Reviews* 11: 222. https://doi.org/10.1186/s13643-022-02097-x.
9. Shumba, C., L. Atuhaire, R. Imakit, R. Atukunda, and P. Memiah. 2013. Missed Doses and Missed Appointments: Adherence to ART among Adult Patients in Uganda. *International Scholarly Research Notices* 2013: 270914. https://doi.org/10.1155/2013/270914.
10. Kwarah, Williams, Jasmin Kwarah, Yakubu Alhassan, Frances Baaba da-Costa Vroom, Duah Dwomoh, and Samuel Bosomprah. 2025. Predictors of HIV interruption in treatment among people living with HIV in 6 regions in Ghana: a retrospective cohort study. Research Square. https://doi.org/10.21203/rs.3.rs-7871647/v1.
11. Fahey, Carolyn A., Linqing Wei, Prosper F. Njau, Siraji Shabani, Sylvester Kwilasa, Werner Maokola, Laura Packel, Zeyu Zheng, Jingshen Wang, and Sandra I. McCoy. 2022. Machine learning with routine electronic medical record data to identify people at high risk of disengagement from HIV care in Tanzania. *PLOS Global Public Health* 2. Public Library of Science: e0000720. https://doi.org/10.1371/journal.pgph.0000720.
12. Ridgway, Jessica P., Aswathy Ajith, Eleanor E. Friedman, Michael J. Mugavero, Mari M. Kitahata, Heidi M. Crane, Richard D. Moore, et al. 2022. Multicenter Development and Validation of a Model for Predicting Retention in Care Among People with HIV. *AIDS and Behavior* 26: 3279–3288. https://doi.org/10.1007/s10461-022-03672-y.
13. ÇETİN, Kardel Rüveyda. 2024. The Importance of Splitting Datasets into Training, Validation, and Test Sets. *Medium*.
14. Dobbin, Kevin K., and Richard M. Simon. 2011. Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics* 4: 31. https://doi.org/10.1186/1755-8794-4-31.

15. Joseph, V. Roshan, and Akhil Vakayil. 2022. SPlit: An Optimal Method for Data Splitting. *Technometrics* 64. ASA Website: 166–176. https://doi.org/10.1080/00401706.2021.1921037.
16. Lee, Loong Chuen, Choong-Yeun Liong, and Abdul Aziz Jemain. 2018. Validity of the best practice in splitting data for hold-out validation strategy as performed on the ink strokes in the context of forensic science. *Microchemical Journal* 139: 125–133. https://doi.org/10.1016/j.microc.2018.02.009.
17. Puzyn, Tomasz, Aleksandra Mostrag-Szlichtyng, Agnieszka Gajewicz, Michał Skrzyński, and Andrew P. Worth. 2011. Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models. *Structural Chemistry* 22: 795–804. https://doi.org/10.1007/s11224-011-9757-4.
18. Xu, Yun, and Royston Goodacre. 2018. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing* 2: 249–262. https://doi.org/10.1007/s41664-018-0068-2.
19. PEPFAR COP/ROP 2021 Guidance. 2025. https://www.state.gov/wp-content/uploads/2021/02/PEPFAR-COP21-Guidance-Final.pdf. Accessed October 25.
20. Kwarah, Williams, Frances Baaba da-Costa Vroom, Duah Dwomoh, and Samuel Bosomprah. 2025. Evaluating predictive performance, validity, and applicability of machine learning models for predicting HIV treatment interruption: a systematic review. *BMC global and public health* 3: 64. https://doi.org/10.1186/s44263-025-00184-4.
21. Bisong, Ekaba. 2019. Logistic Regression. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, ed. Ekaba Bisong, 243–250. Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4842-4470-8_20.
22. Sufriyana, Herdiantri, Atina Husnayain, Ya-Lin Chen, Chao-Yang Kuo, Onkar Singh, Tso-Yang Yeh, Yu-Wei Wu, and Emily Chia-Yu Su. 2020. Comparison of Multivariable Logistic Regression and Other Machine Learning Algorithms for Prognostic Prediction Studies in Pregnancy Care: Systematic Review and Meta-Analysis. *JMIR Medical Informatics* 8. JMIR Publications Inc., Toronto, Canada: e16503. https://doi.org/10.2196/16503.
23. Aslan, Manar, and Ergin Toros. 2025. Machine Learning in Optimising Nursing Care Delivery Models: An Empirical Analysis of Hospital Wards. *Journal of Evaluation in Clinical Practice* 31: e70001. https://doi.org/10.1111/jep.70001.
24. Random Forest - an overview | ScienceDirect Topics. 2025. https://www.sciencedirect.com/topics/engineering/random-forest. Accessed October 25.
25. Taye, Eliyas Addisu, Eden Yitbarek Woubet, Gabrela Yimer Hailie, Fetlework Gubena Arage, Tigabu Eskeziya Zerihun, Adem Tsegaw Zegeye, Tarekegn Cheklie Zeleke, and Abel Temeche Kassaw. 2025. Application of the random forest algorithm to predict skilled birth attendance and identify determinants among reproductive-age women in 27 Sub-Saharan African countries; machine learning analysis. *BMC Public Health* 25: 901. https://doi.org/10.1186/s12889-025-22007-9.

26. Dong, Tim, Iyabosola Busola Oronti, Shubhra Sinha, Alberto Freitas, Bing Zhai, Jeremy Chan, Daniel P. Fudulu, Massimo Caputo, and Gianni D. Angelini. 2024. Enhancing Cardiovascular Risk Prediction: Development of an Advanced Xgboost Model with Hospital-Level Random Effects. *Bioengineering* 11. Multidisciplinary Digital Publishing Institute: 1039. https://doi.org/10.3390/bioengineering11101039.
27. Inoue, Tomoo, Daisuke Ichikawa, Taro Ueno, Maxwell Cheong, Takashi Inoue, William D. Whetstone, Toshiki Endo, Kuniyasu Nizuma, and Teiji Tominaga. 2020. XGBoost, a Machine Learning Method, Predicts Neurological Recovery in Patients with Cervical Spinal Cord Injury. *Neurotrauma Reports* 1: 8–16. https://doi.org/10.1089/neur.2020.0009.
28. Supervised learning. 2025. https://doi.org/10.1016/B978-0-12-821929-4.00004-4.
29. Zheng, Jing, Jianjun Li, Zhengyu Zhang, Yue Yu, Juntao Tan, Yunyu Liu, Jun Gong, Tingting Wang, Xiaoxin Wu, and Zihao Guo. 2023. Clinical Data based XGBoost Algorithm for infection risk prediction of patients with decompensated cirrhosis: a 10-year (2012–2021) Multicenter Retrospective Case-control study. *BMC Gastroenterology* 23: 310. https://doi.org/10.1186/s12876-023-02949-3.
30. Kazmierska, Joanna, and Julian Malicki. 2008. Application of the Naïve Bayesian Classifier to optimize treatment decisions. *Radiotherapy and Oncology* 86: 211–216. https://doi.org/10.1016/j.radonc.2007.10.019.
31. Naïve Bayes Model - an overview | ScienceDirect Topics. 2026. https://www.sciencedirect.com/topics/computer-science/naive-bayes-model. Accessed February 1.
32. Lakshmi, Dr G. Vijaya. 2025. A Review on K nearest Neighbour Classification Technique in Machine Learning. *International Journal of Scientific Research in Science, Engineering and Technology* 12: 257–260. https://doi.org/10.32628/IJSRSET25121168.
33. Chengsheng, Tu, Liu Huacheng, and Xu Bing. 2017. AdaBoost typical Algorithm and its application research. *MATEC Web of Conferences* 139. EDP Sciences: 00222. https://doi.org/10.1051/matecconf/201713900222.
34. Beja-Battais, Perceval. 2023. AdaBoost: A theoritical review.
35. Cao, Ying, Qi-Guang Miao, Jia-Chen Liu, and Lin Gao. 2013. Advance and Prospects of AdaBoost Algorithm. *Acta Automatica Sinica* 39: 745–758. https://doi.org/10.1016/S1874-1029(13)60052-X.
36. Zhang, Yongli. 2012. Support Vector Machine Classification Algorithm and Its Application. In *Information Computing and Applications*, ed. Chunfeng Liu, Leizhen Wang, and Aimin Yang, 179–186. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-34041-3_27.
37. Guido, Rosita, Stefania Ferrisi, Danilo Lofaro, and Domenico Conforti. 2024. An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review. *Information* 15. publisher. https://doi.org/10.3390/info15040235.
38. Çorbacıoğlu, Şeref Kerem, and Gökhan Aksel. 2023. Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turkish Journal of Emergency Medicine* 23: 195–198. https://doi.org/10.4103/tjem.tjem_182_23.

39. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. 2022. *Computer Methods and Programs in Biomedicine* 214. Elsevier: 106584. https://doi.org/10.1016/j.cmpb.2021.106584.
40. A discussion of calibration techniques for evaluating binary and categorical predictive models. 2018. *Preventive Veterinary Medicine* 149. Elsevier: 107–114. https://doi.org/10.1016/j.prevetmed.2017.11.018.
41. Hu, Liangyuan, and Lihua Li. 2022. Using Tree-Based Machine Learning for Health Studies: Literature Review and Case Series. *International Journal of Environmental Research and Public Health* 19: 16080. https://doi.org/10.3390/ijerph192316080.
42. Van Calster, Ben, David J. McLernon, Maarten van Smeden, Laure Wynants, Ewout W. Steyerberg, Patrick Bossuyt, Gary S. Collins, et al. 2019. Calibration: the Achilles heel of predictive analytics. *BMC Medicine* 17: 230. https://doi.org/10.1186/s12916-019-1466-7.
43. Okeke, Nwora Lance, Meredith E Clement, Mehri S McKellar, and Jason E Stout. 2018. Health Care Utilization Behaviors Predict Disengagement From HIV Care: A Latent Class Analysis. *Open Forum Infectious Diseases* 5: ofy088. https://doi.org/10.1093/ofid/ofy088.
44. Salimparsa, Mozhgan, Kamran Sedig, Daniel J. Lizotte, Sheikh S. Abdullah, Niaz Chalabianloo, and Flory T. Muanda. 2025. Explainable AI for Clinical Decision Support Systems: Literature Review, Key Gaps, and Research Synthesis. *Informatics* 12. Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/informatics12040119.
45. Ennab, Mohammad, and Hamid Mcheick. 2024. Enhancing interpretability and accuracy of AI models in healthcare: a comprehensive review on challenges and future directions. *Frontiers in Robotics and AI* 11: 1444763. https://doi.org/10.3389/frobt.2024.1444763.
46. Hur, Sujeong, Yura Lee, Joongheum Park, Yeong Jeong Jeon, Jong Ho Cho, Duck Cho, Dobin Lim, Wonil Hwang, Won Chul Cha, and Junsang Yoo. 2025. Comparison of SHAP and clinician friendly explanations reveals effects on clinical decision behaviour. *npj Digital Medicine* 8. Nature Publishing Group: 578. https://doi.org/10.1038/s41746-025-01958-8.
47. Magura, Judie, Sibongile R. Nhari, and Thokozani I. Nzimakwe. 2025. Barriers to ART adherence in sub-Saharan Africa: a scoping review toward achieving UNAIDS 95-95-95 targets. *Frontiers in Public Health* 13. Frontiers. https://doi.org/10.3389/fpubh.2025.1609743.
48. HIGA, Darrel H., Nicole CREPAZ, Mary M. MULLINS, Adebukola ADEGBITE-JOHNSON, Jayleen K L GUNN, Christa DENARD, and Yuko MIZUNO. 2022. Strategies to improve HIV care outcomes for people with HIV who are out of care. *AIDS (London, England)* 36: 853–862. https://doi.org/10.1097/QAD.0000000000003172.
49. Mbewe, Alinafe, Hammond N Sarkwah, Olugbenga Ijaodola, Jirair Ratevosian, Solomon Simba, Tyler R Smith, Wisdom K Atiwoto, et al. 2026. The role of digital health and artificial intelligence in improving the reach and effectiveness of HIV prevention in Africa. *The Lancet Global Health* 14: e131–e142. https://doi.org/10.1016/S2214-109X(25)00447-4.

Figures

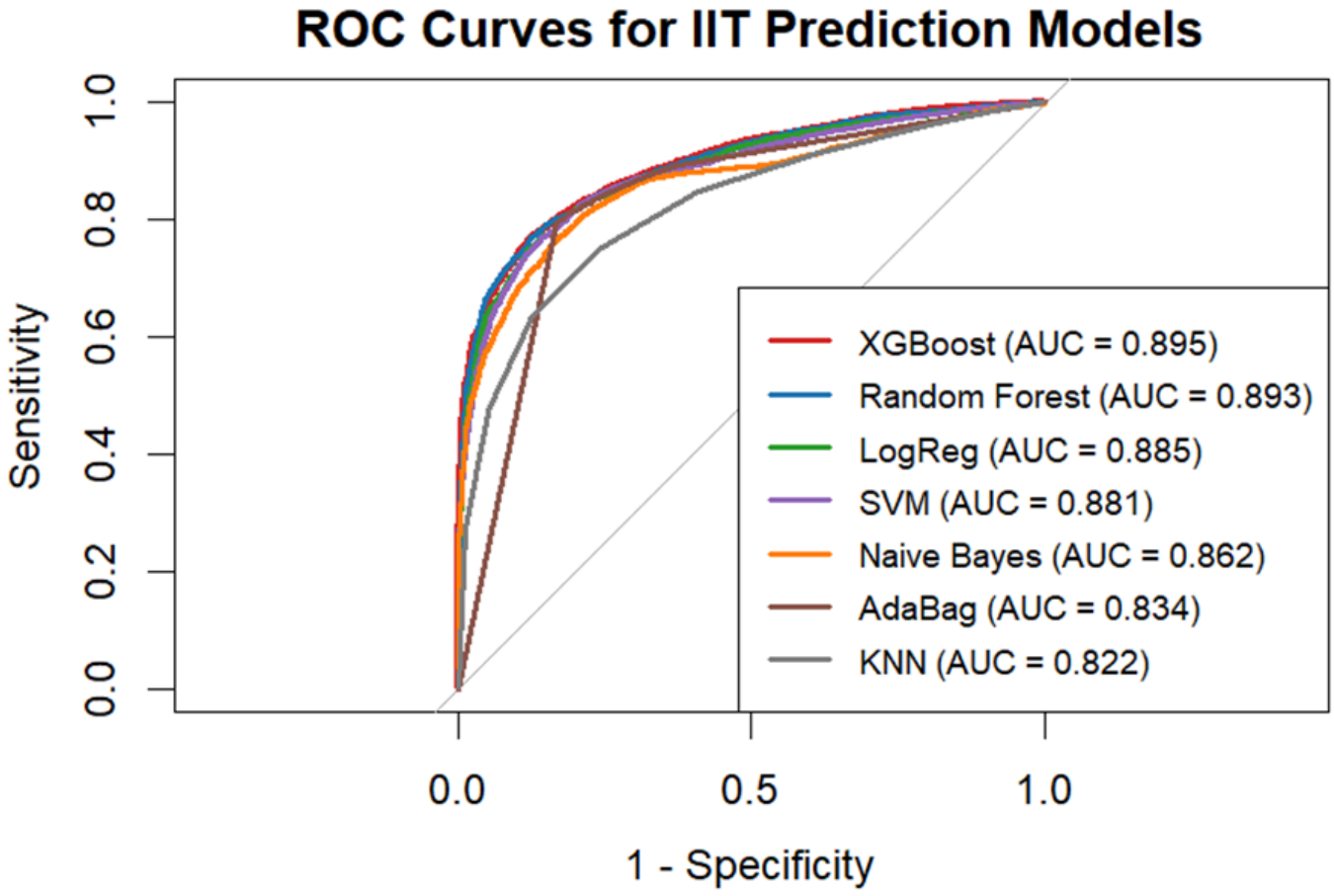


Figure 1

AUC-ROC curves for candidate models on the test set

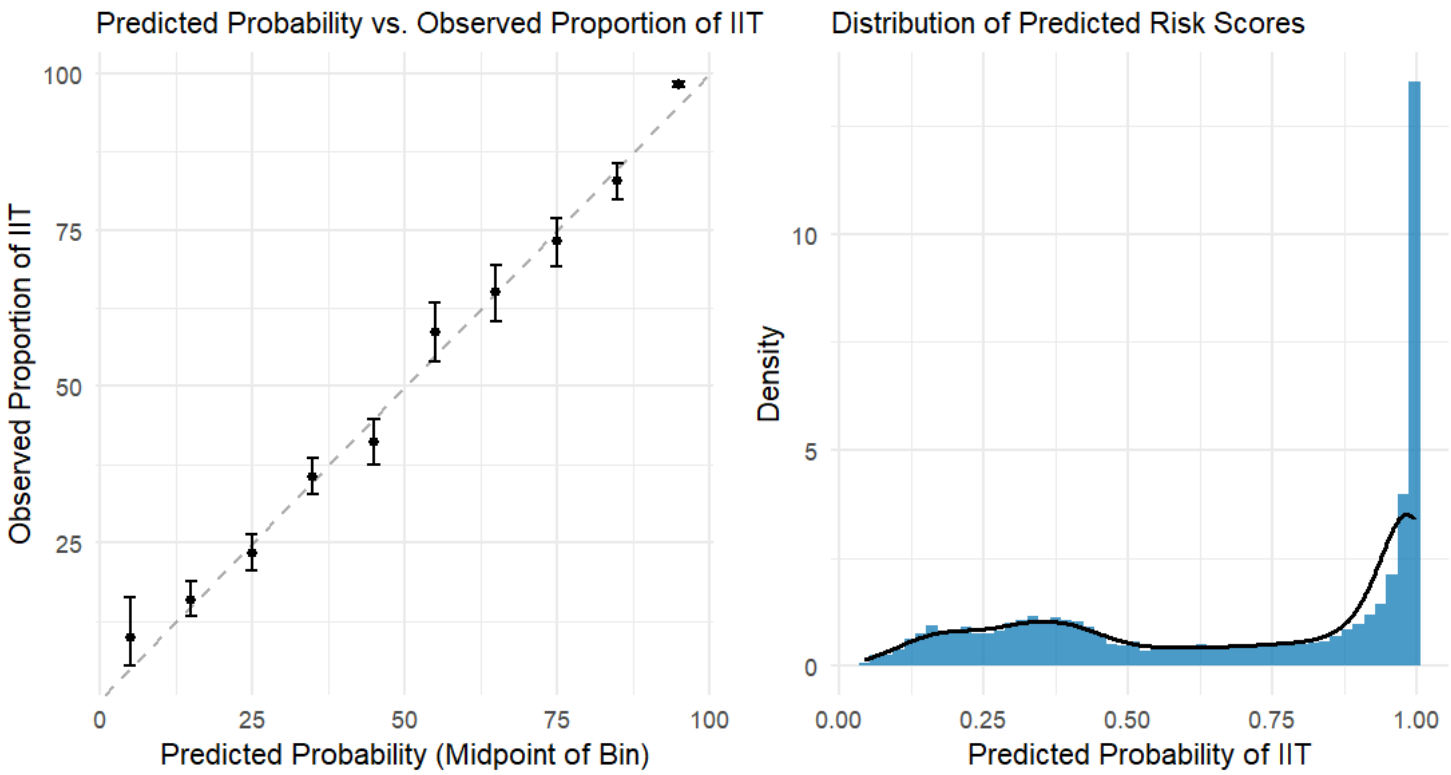


Figure 2

Calibration and distribution plot for the final XGBoost model

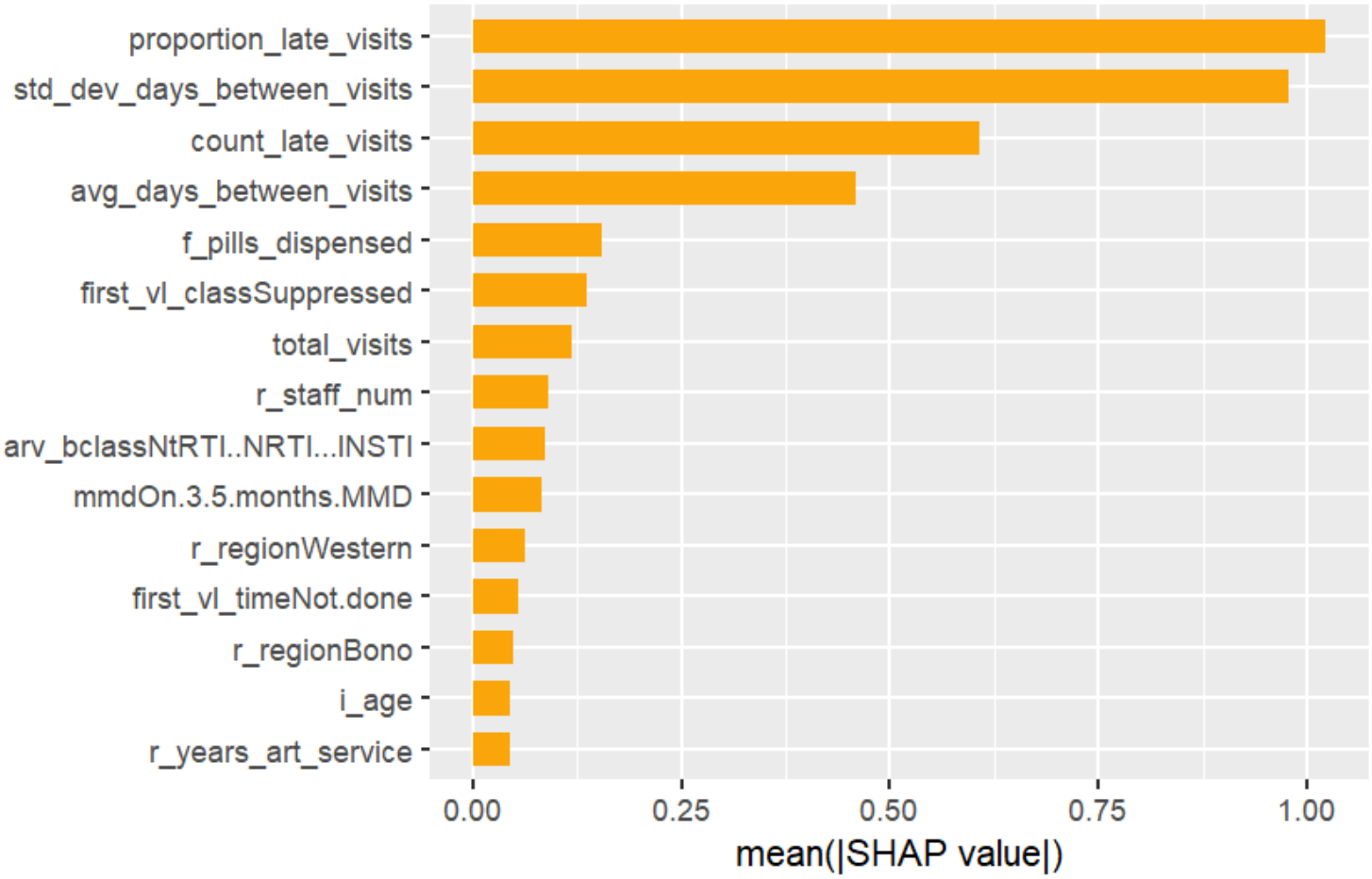


Figure 3

Hierarchy of factors prioritized in the model prediction

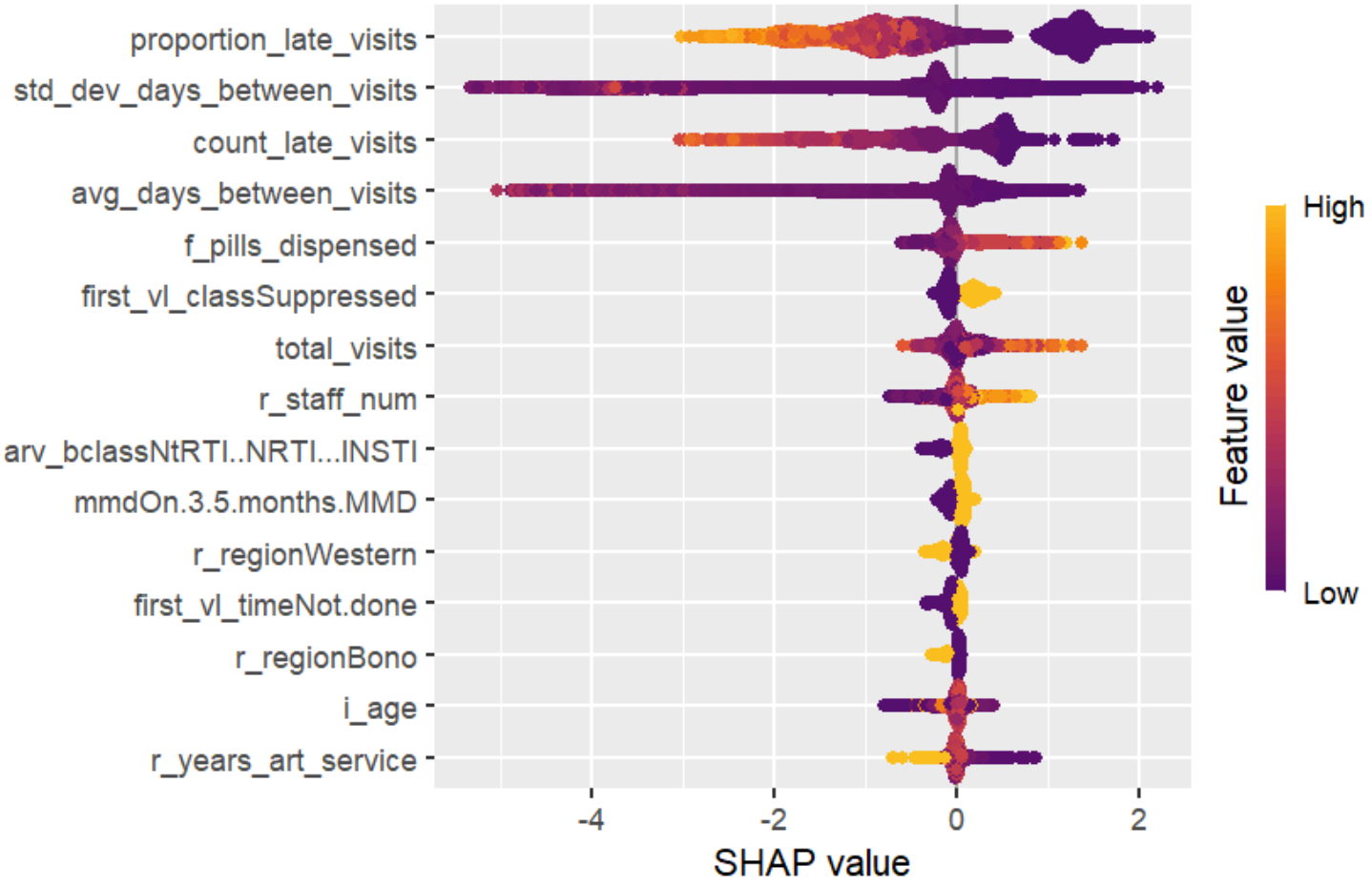


Figure 4

SHAP plot of the most important drivers of TI