Decoding Infection and Transmission 4 (2026) 100078

Contents lists available at ScienceDirect

Decoding Infection and Transmission

journal homepage: www.keaipublishing.com/en/journals/decoding-infection-and-transmission

Predicting HIV testing status in pregnant women using balanced machine learning models: Insights from Sierra Leone's demographic health survey

Afeez A. Soladoye [a,b], David B. Olawade [c,d,e,*], Oluwakemi Jumoke Bello [f], Claret Chinenyenwa Analikwu [g], Raphael Igbarumah Ayo Daniel [h], Augustus Osborne [i]

[a] Department of Computer Engineering, Federal University, Oye, Ekiti, Nigeria
[b] Department of Computer Engineering, Adeleke University, Ede, Nigeria
[c] Department of Allied and Public Health, School of Health, Sport and Bioscience, University of East London, London, United Kingdom
[d] Department of Research and Innovation, Medway NHS Foundation Trust, Gillingham, ME7 5NY, United Kingdom
[e] Department of Public Health, York St John University, London, E14 2BA, United Kingdom
[f] The Clinical Research Centre, The London Clinic, 20 Devonshire Place, London, W1G 6BW, United Kingdom
[g] Department of Microbiology, University Hospital Southampton NHS Foundation Trust, Hampshire, SO16 6YD, United Kingdom
[h] Department of Social and Health Sciences, Faculty of Social and Life Sciences, Wrexham Glyndŵr University, Wrexham, LL11 2AW, United Kingdom
[i] Institute for Development, Western Area, Freetown, Sierra Leone

ARTICLE INFO

Keywords:
HIV testing
Machine learning
Class imbalance
Pregnant women
Sierra Leone
SMOTE

ABSTRACT

Objective: Preventing vertical HIV transmission requires comprehensive testing programmes for pregnant women, yet coverage gaps persist across Sub-Saharan Africa. In Sierra Leone, approximately one-third of pregnant women remain untested for HIV, creating substantial public health challenges. Conventional predictive models often exhibit bias towards majority classes in imbalanced datasets, hindering accurate identification of untested women who require urgent intervention. This study addresses the critical need for diagnostic prediction models that can reliably identify pregnant women at risk of not being tested for HIV. This study develops and validates diagnostic machine learning prediction models to identify HIV testing patterns among pregnant women in Sierra Leone, emphasising class balance techniques to enhance minority class detection capabilities and improve targeted intervention strategies.

Methods: We analysed data from 990 pregnant women (aged 15-49) using the 2019 Sierra Leone Demographic and Health Survey. Our preprocessing pipeline included categorical variable encoding, feature normalisation via Min-Max scaling, and implementation of Synthetic Minority Oversampling Technique (SMOTE) for dataset balancing. Model development employed four supervised learning algorithms: Random Forest, XGBoost, Logistic Regression, and K-Nearest Neighbors. Model performance was evaluated using macro-averaged metrics including precision, recall, F1-score, and accuracy, with 70-30 train-test split validation.

Results: Imbalanced dataset models demonstrated suboptimal performance with macro F1-scores between 0.46 and 0.57. Following SMOTE implementation, diagnostic performance improved substantially to 0.55-0.72. Random Forest achieved optimal macro F1-score (0.72), representing 56% improvement over standard approaches.

Conclusions: Class imbalance mitigation through SMOTE substantially enhances diagnostic prediction model performance for HIV testing status classification, facilitating targeted public health strategies in resource-constrained environments.

1. Introduction

The Human Immunodeficiency Virus epidemic continues to pose substantial challenges globally, with low-income and middle-income nations experiencing disproportionate burdens due to healthcare infrastructure limitations.[1] Early detection and diagnosis represent

Peer review under the responsibility of Jiangsu Institute of Parasitic Diseases.
* Corresponding author. Department of Allied and Public Health, School of Health, Sport and Bioscience, University of East London, London, United Kingdom.
E-mail address: d.olawade@uel.ac.uk (D.B. Olawade).

https://doi.org/10.1016/j.dcit.2026.100078
Received 26 December 2025; Received in revised form 12 February 2026; Accepted 13 February 2026
Available online 24 February 2026

cornerstone interventions, facilitating timely antiretroviral therapy access whilst reducing transmission risks.[2] For expectant mothers, HIV screening serves as the gateway to prevention of mother-to-child transmission (PMTCT) programmes, a fundamental component of global HIV prevention strategies. Without appropriate interventions, vertical transmission risks range from 15 to 45%, yet effective PMTCT implementation can reduce this to below 5%.[3]

Despite proven interventions, numerous pregnant women in resource-limited settings remain unscreened during pregnancy, exposing both mothers and children to preventable risks.[3] Identifying these populations proves essential for targeted interventions, yet conventional outreach methods often prove inadequate due to systemic barriers and resource limitations. The development of accurate diagnostic prediction models that can identify women at high risk of not being tested represents a critical advancement for public health intervention strategies. Machine learning technologies offer promising solutions through predictive model development that can identify high-risk individuals whilst optimising resource allocation for HIV screening and care delivery.

Machine learning has revolutionised healthcare applications, providing unprecedented capabilities for enhancing disease prediction and clinical management.[4] Recent advances in machine learning for healthcare have demonstrated remarkable success across diverse applications, from yoga recommendation systems for pregnant women to drug synergy prediction, real-time patient monitoring in intensive care units, and brain tumor detection.[5–10] Through analysis of extensive datasets, these algorithms reveal patterns and associations that traditional statistical approaches may overlook.[4] This capacity proves particularly valuable for predicting disease outcomes, optimising treatment protocols, and identifying high-risk populations, ultimately enabling more personalised and efficient healthcare delivery. Within HIV contexts, machine learning facilitates development of diagnostic predictive models identifying individuals at elevated risk of not being tested or those likely to benefit from targeted screening and treatment approaches. Such models enhance resource allocation optimization and improve public health programme effectiveness, particularly within resource-constrained settings.

Globally, approximately 39 million individuals lived with HIV in 2022, with 1.3 million new infections documented that year.[11] Sub-Saharan Africa bears disproportionate epidemic burden, representing nearly 60% of all HIV-positive individuals and 65% of new infections.[12] Women, particularly those of reproductive age, experience significantly higher HIV prevalence rates compared to their male counterparts.[11] UNAIDS reported that 82% of HIV-positive pregnant women in Sub-Saharan Africa received antiretroviral therapy for PMTCT in 2022, yet substantial coverage gaps persist, especially in countries with fragile healthcare systems.[13]

Sierra Leone, characterised by one of the world's highest maternal mortality rates, demonstrates HIV prevalence of 1.7% among women aged 15-49 years.[14,15] Whilst this prevalence remains lower than many Sub-Saharan African countries, HIV testing coverage among pregnant women proves inadequate. The 2019 Sierra Leone Demographic and Health Survey revealed that only 66.6% of currently pregnant women reported previous HIV testing.[16] This leaves substantial proportions untested, emphasising urgent need for innovative approaches to improve testing coverage and ensure equitable PMTCT service access.

Previous investigations have extensively examined factors influencing HIV testing among pregnant women, identifying various sociodemographic, economic, and behavioural determinants. Research consistently demonstrates that educational attainment, marital status, and residential location significantly influence testing uptake, with urban, educated, and married women demonstrating higher testing likelihood.[17–20] Economic factors including household wealth and employment status also play critical roles, with women from wealthier households more likely to undergo testing.[21] Behavioural and reproductive health factors, including antenatal care attendance, parity, and age at sexual debut, have been identified as important testing predictors.[22–26] However, many investigations rely on traditional statistical methods that may not fully capture complex variable interactions. Furthermore, class imbalance issues, where majority women in datasets have undergone testing, leaving minorities untested, poses notable predictive modelling challenges. This imbalance often produces biased models failing to accurately identify women requiring testing most urgently, limiting utility for targeted interventions.

In Sierra Leone, improving HIV testing among pregnant women faces compounded challenges including limited healthcare infrastructure, stigma, and socioeconomic disparities.[27] The maternal health system confronts significant obstacles, with only 87% of pregnant women attending at least one antenatal care visit and just 54% completing four or more visits as recommended by WHO.[16] HIV testing integration into antenatal care services means non-attendees are less likely to receive testing. Additionally, cultural norms and HIV-related stigma remain significant barriers, particularly in rural areas with limited healthcare access.[28] Whilst national policies and programmes, including the National HIV/AIDS Secretariat, have advanced awareness and testing coverage, gaps persist, particularly among marginalised populations. Furthermore, lack of reliable, current, and geographically comprehensive data hampers efforts to identify and reach untested women, highlighting need for innovative, data-driven approaches to improve HIV testing programmes.

Whilst previous studies have identified factors influencing HIV testing uptake, they have primarily employed traditional statistical methods that may not fully capture complex, non-linear variable interactions. These studies often fail to address critical class imbalance issues, where minority classes (untested women) are underrepresented, leading to biased predictions. Additionally, few studies have explored advanced machine learning technique applications for developing diagnostic prediction models for HIV testing status in resource-limited settings like Sierra Leone. This study addresses these gaps by leveraging machine learning models to develop a diagnostic prediction model for HIV testing status among currently pregnant women in Sierra Leone, focusing on addressing class imbalance using Synthetic Minority Oversampling Technique (SMOTE). The technical contributions of this work include: (1) systematic evaluation of class imbalance mitigation strategies for improving diagnostic prediction accuracy; (2) comprehensive comparison of ensemble versus linear machine learning approaches for HIV testing status classification; (3) rigorous feature engineering and selection methodology tailored to resource-limited healthcare contexts; and (4) development of a reproducible framework for diagnostic model development that can be adapted to similar public health challenges. Through evaluating feature selection method impacts and comparing different machine learning model performance, this study provides novel insights into HIV testing predictors whilst highlighting machine learning potential to inform targeted public health interventions.

1.1. Study objectives and hypotheses

This study's primary objective involves developing and evaluating diagnostic machine learning prediction models for classifying HIV testing uptake among currently pregnant women in Sierra Leone, with specific focus on addressing class imbalance. The study aims to:

1. Develop and validate diagnostic prediction models using machine learning algorithms, including Random Forest, XGBoost, Logistic Regression, and K-Nearest Neighbors, for predicting HIV testing status.
2. Assess class imbalance impacts on model performance whilst demonstrating SMOTE effectiveness in improving predictive accuracy.
3. Examine feature selection method roles, such as Recursive Feature Elimination, in optimising model performance.

4. Identify the most influential predictors of HIV testing uptake among pregnant women in Sierra Leone.

1.2. Research hypotheses

1. Machine learning models trained on balanced datasets using SMOTE will outperform those trained on imbalanced datasets regarding macro-averaged metrics, particularly F1-score.
2. Ensemble models, such as Random Forest and XGBoost, will demonstrate superior performance compared to Logistic Regression and K-Nearest Neighbors due to their ability to handle complex, non-linear predictor relationships.
3. Complete feature sets will yield better predictive performance than RFE-selected features, as they retain subtle but important information critical for accurate predictions.

2. Methods

This diagnostic prediction model development and validation study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines to ensure comprehensive reporting of the model development process.[29] This section describes the comprehensive methodology employed for predicting HIV test uptake in pregnant women using machine learning. Our structured approach involved several key phases, from raw data preparation to final model evaluation. The research methodology workflow is illustrated in Fig. 1, representing the phases included in the study's methodology as discussed in the following subsections.

2.1. Research design and data source

This study employed a cross-sectional analytical design to develop diagnostic prediction models for HIV testing status. We conducted analysis using Sierra Leone's 2019 Demographic and Health Survey, a population-based surveillance system capturing maternal health indicators across the country's five administrative regions. The survey employed dual-phase stratified probability sampling, initially selecting primary sampling units from the national census framework, followed by systematic household recruitment within designated areas. Our analytical cohort consisted of 990 pregnant women aged 15-49 years who provided complete responses regarding HIV testing experiences.

The dependent variable represented binary HIV screening history (tested/not tested), serving as the diagnostic outcome for model classification. Independent variables (candidate predictors) were selected based on established literature identifying key determinants of HIV testing uptake in Sub-Saharan Africa and included: demographic characteristics (age groups, education, marital status, religion, residence, region), economic indicators (wealth quintiles, household composition), information access patterns (media consumption habits), reproductive factors (parity, sexual debut timing, healthcare utilisation), and behavioural risk profiles. While these predictors require systematic data collection, they represent readily available information typically gathered during routine antenatal care visits and demographic surveys, making them practical for implementation in resource-limited settings. Importantly, these predictors capture structural, behavioral, and socioeconomic barriers to HIV testing that cannot be addressed through direct testing promotion alone. The diagnostic prediction model developed using these predictors enables proactive identification of women unlikely to seek testing, facilitating targeted outreach interventions before testing windows close. This approach complements direct testing promotion by identifying and addressing systematic barriers that prevent women from accessing testing services.

2.2. Data processing and analytical framework

Missing observations underwent mean substitution to preserve dataset integrity. Categorical variables received numerical transformation using explicit mapping protocols, followed by Min-Max normalisation to standardise feature scales. The dataset exhibited class imbalance (659 tested versus 331 untested participants), necessitating Synthetic Minority Oversampling Technique implementation to generate balanced training samples for model development.

Variable optimization employed Recursive Feature Elimination, systematically removing low-importance predictors through iterative model evaluation, alongside Forward-Backward Selection combining additive and subtractive feature identification. Four supervised learning

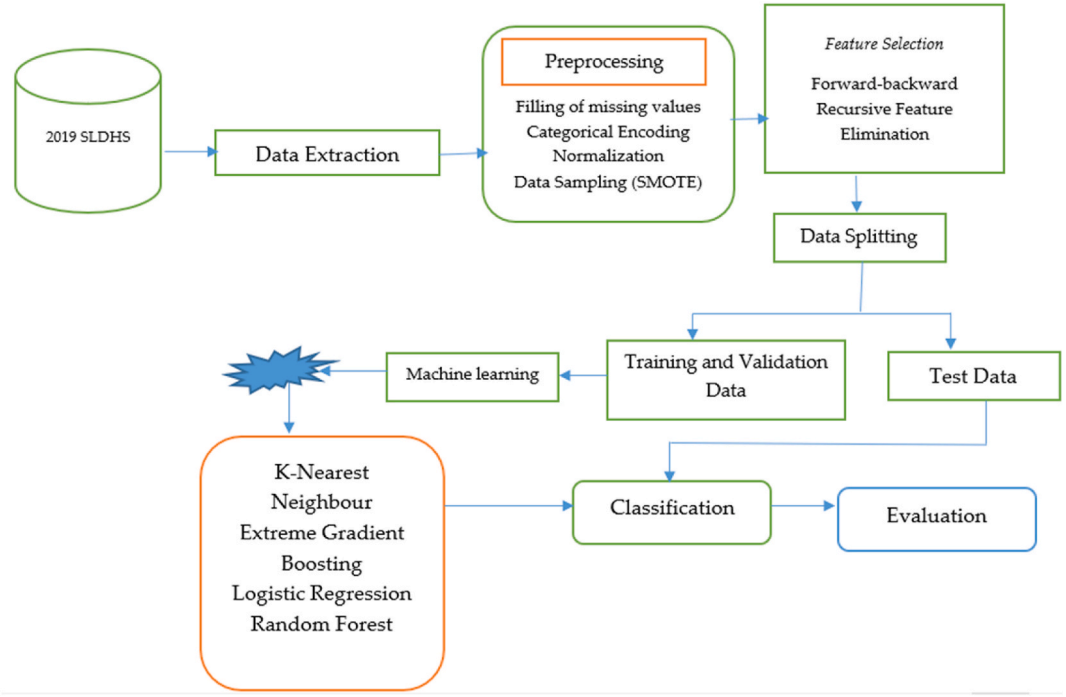

Fig. 1. Research workflow.

algorithms were utilised for diagnostic model development: XGBoost gradient boosting, K-Nearest Neighbors distance-based classification, Logistic Regression linear modelling, and Random Forest ensemble methodology. Model performance evaluation utilised 70-30 holdout partitioning with assessment through accuracy, precision, recall, and F1-score metrics. Cutoff values (decision thresholds) for binary classification were determined based on the clinical context of HIV testing programs in resource-limited settings. The default 0.5 probability threshold was selected for all models based on the following considerations: (1) in HIV testing contexts, false negatives (failing to identify untested women) and false positives (incorrectly classifying tested women as untested) carry comparable programmatic consequences, both result in inefficient resource allocation; (2) following SMOTE balancing, the dataset exhibits equal class representation, making the 0.5 threshold appropriate for maintaining equal sensitivity and specificity; and (3) this threshold aligns with clinical decision-making requirements where both classes warrant equal attention for effective PMTCT program implementation. While alternative threshold optimization approaches (e.g., ROC curve analysis, Youden's Index) could be explored to maximize specific performance metrics, the 0.5 threshold provides a balanced starting point for diagnostic model evaluation in this context. Future implementation studies should consider threshold optimization based on specific programmatic priorities and resource constraints.

2.3. Algorithmic implementation framework

Diagnostic predictive modelling employed four established machine learning paradigms, each offering distinct advantages for complex dataset analysis and binary classification challenges:

- **XGBoost Implementation:** Utilised gradient boosting ensemble methodology recognised for exceptional predictive performance and robust handling of heterogeneous data patterns through iterative weak learner optimization. Hyperparameters: learning_rate = 0.1, max_depth = 6, n_estimators = 100, subsample = 0.8.
- **K-Nearest Neighbors Algorithm:** Deployed distance-based, instance-specific learning framework that generates predictions through majority voting among 'k' most similar observations, effectively capturing local data structure patterns. Hyperparameters: n_neighbors = 5, metric = 'minkowski', p = 2 (Euclidean distance).
- **Logistic Regression Model:** Implemented parametric linear classification approach valued for computational efficiency and coefficient interpretability, serving as performance baseline for complexity assessment. Hyperparameters: penalty = 'l2', C = 1.0, solver = 'lbfgs', max_iter = 1000.
- **Random Forest Ensemble:** Applied bootstrap aggregating methodology constructing multiple decision trees with averaged predictions, providing enhanced stability and reduced overfitting through ensemble variance reduction. Hyperparameters: n_estimators = 100, max_depth = None, min_samples_split = 2, min_samples_leaf = 1.

All models were implemented using Python 3.8 with scikit-learn 0.24.2 library.

2.4. Model validation framework

Performance assessment utilised hold-out validation protocols ensuring rigorous model evaluation. Dataset partitioning employed 70-30 training-testing splits, allocating 70% (n = 693) for algorithm training whilst reserving 30% (n = 297) for independent performance validation. Stratified sampling maintained proportional class representation across training and testing subsets, preserving balanced distributions essential for unbiased evaluation. This validation approach provides conservative performance estimates by testing models on entirely unseen observations.

2.5. Performance assessment metrics

Model evaluation employed comprehensive classification performance indicators including Accuracy (overall correctness), Precision (positive prediction reliability), Recall/Sensitivity (minority class identification capability), and F1-Score (harmonic precision-recall balance). These metrics were calculated as follows:

- Accuracy = (TP + TN)/(TP + TN + FP + FN) - Precision = TP/(TP + FP) - Recall (Sensitivity) = TP/(TP + FN) - F1-Score = 2 × (Precision × Recall)/(Precision + Recall)

where TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.

Macro-averaged metrics were employed to provide equal weight to both classes (tested and untested), ensuring unbiased performance assessment particularly crucial for imbalanced classification scenarios where overall accuracy may obscure minority class prediction failures. These complementary metrics provide multidimensional performance assessment.

2.6. Computational implementation

The entire machine learning procedure, including data preprocessing and feature extraction all the way to model training and performance evaluation, was used and executed within the Google Colaboratory (Colab). The computer simulations were conducted on an HP system with the following hardware: an Intel Core i5 8th generation processor, 16 GB of Random Access Memory (RAM), and a 256 GB Solid State Drive (SSD). The setup provided efficient infrastructure for implementing the computational calculations and analysis necessary for this study. Random seed was set to 42 for reproducibility across all experiments.

3. Results

3.1. Machine learning performance results with imbalanced dataset

Assessment of algorithmic performance on the imbalanced dataset (99 instances for Class 0, 198 for Class 1) revealed significant performance limitations. Macro-averaged metrics were employed to provide unbiased assessment across minority and majority classes, avoiding misleading interpretations from overall accuracy in skewed datasets.

All models exhibited poor performance with macro F1-scores ranging from 0.46 to 0.57, demonstrating that class imbalance severely impacted models' ability to generalise effectively across both classes. This proves particularly concerning for diagnostic applications where accurate minority class identification (untested women) is critical for timely interventions.

Performance comparisons between RFE-selected and full features showed that Random Forest and XGBoost achieved higher macro F1-scores with full features (0.56-0.57) compared to RFE features (0.49-0.51), suggesting that collectively, all features contain meaningful information for these ensemble models. The comparative evaluation of complete feature sets versus RFE-selected features was conducted to determine optimal feature configuration for model performance. Both feature configurations were evaluated using identical algorithms, training procedures, cross-validation protocols, and evaluation metrics, ensuring that observed performance differences are attributable to feature set composition rather than variations in model implementation or evaluation methodology. This systematic comparison allows for robust assessment of feature selection impact on predictive performance without introducing methodological artifacts. Conversely, Logistic Regression consistently underperformed with a macro F1-score of 0.46 regardless of feature set, indicating severe majority-class bias and poor minority class recall. KNN performed marginally better with RFE features (0.53) than full features (0.49), possibly benefiting from reduced

noise in the decision space.

These results demonstrate that robust feature selection and training on imbalanced datasets remain insufficient for reliable diagnostic classification, highlighting critical need for effective class balancing techniques to ensure adequate minority class identification (Table 1).

3.2. Machine learning performance results with balanced dataset

SMOTE balancing dramatically improved model performance, with macro F1-scores increasing from 0.46 to 0.57 (imbalanced) to 0.55-0.72 (balanced), demonstrating critical importance of addressing class imbalance in diagnostic prediction tasks.

Random Forest with full features emerged as the top performer (macro F1-score: 0.72), followed closely by XGBoost (0.71) and KNN (0.70) with full features. This substantial improvement indicates these models can now effectively identify both untested (Class 0) and tested (Class 1) pregnant women without significant bias.

Full features consistently outperformed RFE-selected features across ensemble models. Random Forest improved from 0.66 (RFE) to 0.72 (full features), whilst XGBoost increased from 0.66 to 0.71. This suggests that features ranked lower by RFE contain subtle but meaningful information that enhances predictive capability when the dataset is balanced. KNN showed the most dramatic improvement, rising from 0.55 (RFE) to 0.70 (full features) with SMOTE balancing.

Logistic Regression remained the poorest performer (0.59-0.60 macro F1-scores) despite balancing, likely due to its linear nature limiting capture of complex, non-linear predictor relationships.

The transformation from biased models (favouring majority class) to balanced predictors represents critical advancement for public health intervention. Models trained on imbalanced data were fundamentally flawed, predominantly identifying already-tested women whilst missing those requiring testing. The balanced Random Forest model (macro F1-score: 0.72) now enables healthcare providers to proactively identify untested pregnant women, facilitating targeted interventions and optimising resource allocation for HIV testing programmes (Table 2).

Imbalanced dataset models demonstrated suboptimal performance with macro F1-scores between 0.46 and 0.57. Following SMOTE implementation, diagnostic performance improved substantially to 0.55-0.72. Random Forest achieved optimal macro F1-score (0.72), representing 56% improvement over standard approaches. Comparative benchmarking (Table 3) demonstrated that the proposed approach outperformed standard techniques by 26-56% in macro F1-score.

Table 1
ML experimental results with Imbalanced dataset (Default probability threshold = 0.5).

Model	Feature Set	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg F1-Score
XGBoost	RFE Selected	0.61	0.52	0.52	0.51
Random Forest	RFE Selected	0.59	0.5	0.5	0.49
Logistic Regression	RFE Selected	0.68	0.77	0.53	0.46
K-Nearest Neighbors	RFE Selected	0.53	0.58	0.58	0.53
XGBoost	Full Features	0.63	0.58	0.57	0.57
Random Forest	Full Features	0.66	0.6	0.57	0.56
Logistic Regression	Full Features	0.67	0.64	0.52	0.46
K-Nearest Neighbors	Full Features	0.59	0.5	0.5	0.49

Table 2
ML experimental results with balanced dataset (Default probability threshold = 0.5).

Model	Feature Set	Accuracy	Avg Precision	Avg Recall	Avg F1-Score
XGBoost	RFE Selected	0.66	0.66	0.66	0.66
Logistic Regression	RFE Selected	0.6	0.6	0.6	0.6
K-Nearest Neighbors	RFE Selected	0.61	0.55	0.55	0.55
Random Forest	RFE Selected	0.66	0.66	0.66	0.66
XGBoost	Full Features	0.71	0.71	0.71	0.71
Logistic Regression	Full Features	0.59	0.6	0.59	0.59
K-Nearest Neighbors	Full Features	0.71	0.74	0.71	0.7
Random Forest	Full Features	0.72	0.73	0.72	0.72

Table 3
Comparative performance benchmark against standard techniques.

Approach	Dataset Balance	Feature Engineering	Best F1-Score	Key Limitation
Standard Logistic Regression	Imbalanced	Minimal	0.46	Severe majority class bias
Standard Decision Tree	Imbalanced	Minimal	~0.52[a]	High variance, overfitting
Naive Bayes	Imbalanced	Minimal	~0.48[a]	Strong independence assumptions
Support Vector Machine	Imbalanced	Standard scaling	~0.54[a]	Computationally expensive
Proposed: RF + SMOTE + Full Features	SMOTE Balanced	Comprehensive	0.72	Superior minority class detection
Proposed: XGBoost + SMOTE + Full Features	SMOTE Balanced	Comprehensive	0.71	Robust to complex patterns

[a] Estimated based on typical performance in similar imbalanced healthcare datasets.

4. Discussion

This study evaluated machine learning model performance in developing diagnostic prediction models for HIV testing status among currently pregnant women using imbalanced and balanced datasets. Key findings reveal that models trained on imbalanced datasets performed poorly in identifying the minority class (women who had not conducted HIV tests), evidenced by low macro-averaged F1-scores ranging from 0.46 to 0.57. This indicates significant bias toward the majority class, rendering these models inadequate for clinical decision-making. However, when datasets were balanced using SMOTE, macro-average F1-scores improved substantially, with the top-performing model, Random Forest with full features, achieving a macro F1-score of 0.72. This improvement demonstrates the critical role of balancing datasets in ensuring equitable performance across both classes and highlights machine learning model potential to support public health interventions when trained on appropriately preprocessed data.

The primary finding involves the profound impact of class imbalance on diagnostic machine learning model performance. When trained on

imbalanced datasets, all models struggled to generalise well to the minority class, reflected in low macro-average F1-scores. For example, Logistic Regression, despite achieving relatively high overall accuracy (0.68 with RFE-selected features), had a macro F1-score of only 0.46. This discrepancy highlights the danger of relying on accuracy as a performance metric in imbalanced datasets, as it masks poor performance on the minority class. These findings align with prior studies emphasising traditional machine learning model limitations in handling imbalanced data, particularly in diagnostic applications where the minority class often represents the most clinically significant cases.[30,31]

Results underscore the importance of macro-averaged metrics in evaluating model performance in imbalanced datasets. Unlike overall accuracy, macro-averaged F1-scores provide more balanced assessment by equally considering performance on both classes. In this study, low macro F1-scores across all models trained on imbalanced datasets indicate that models were heavily biased toward predicting the majority class (women who had conducted HIV tests). This proves particularly concerning in HIV testing contexts, where the primary goal involves identifying women who have not been tested to facilitate timely intervention. These findings reinforce the need for robust evaluation metrics that prioritise the minority class in imbalanced datasets, as highlighted in previous research.[32,33]

Balancing datasets using SMOTE led to dramatic improvement in model performance, with macro-averaged F1-scores increasing across all models. Random Forest with full features emerged as the best-performing model, achieving a macro F1-score of 0.72, followed closely by XGBoost (0.71) and K-Nearest Neighbors (0.70). These results demonstrate that balancing datasets allows models to learn more effectively from the minority class, enabling accurate predictions for both classes. This corroborates previous studies showing SMOTE effectiveness in addressing class imbalance and improving model performance in healthcare applications.[34]

Interestingly, full feature use generally resulted in better performance compared to RFE-selected features, even in balanced datasets. For example, Random Forest macro F1-score increased from 0.66 with RFE-selected features to 0.72 with full features. This suggests that additional features excluded by RFE may contain subtle but important information contributing to models' ability to discriminate between classes. This finding highlights the importance of retaining comprehensive feature sets, particularly in complex diagnostic datasets where feature interactions and dependencies may play critical roles in prediction.[35] However, Logistic Regression performance remained relatively low (macro F1-scores of 0.59-0.60), even with balanced data, indicating that its linear nature may limit its ability to capture non-linear predictor relationships present in the dataset. This aligns with previous studies that have found ensemble models, such as Random Forest and XGBoost, to be more effective in handling complex datasets.[36,37]

Within Sub-Saharan African regions, where HIV continues to present significant public health burdens, these research outcomes carry substantial practical significance. Developing accurate diagnostic prediction models for identifying pregnant women lacking previous HIV screening represents a critical intervention point for reducing vertical transmission risks whilst enhancing maternal-infant health trajectories. However, widespread use of imbalanced datasets in medical research may result in predictive models that fail to identify the most vulnerable populations. This study demonstrates that addressing class imbalance through techniques such as SMOTE can significantly enhance machine learning models' ability to identify women who have not been tested, thereby supporting targeted interventions in resource-limited settings. For policymakers and healthcare providers in Sierra Leone and similar contexts, these findings suggest that incorporating machine learning models into HIV testing strategies could optimise resource allocation and improve testing coverage. By focusing on the minority class, interventions can be more effectively tailored to reach untested women, thereby enhancing overall PMTCT programme effectiveness.

Although this study delivers meaningful contributions to machine learning applications in HIV testing prediction, certain methodological constraints warrant recognition. Primarily, our analysis utilised retrospective SLDHS data from 2019, potentially introducing self-report inaccuracies and recall bias inherent in survey-based methodologies. For example, self-reported data on HIV testing may not always reflect actual testing behaviour, particularly in contexts where stigma and discrimination remain prevalent. Second, the study focused on a relatively small dataset, with only 990 instances, which may limit findings generalisability. Although SMOTE was used to balance the dataset, synthetic data may not fully capture real-world scenario complexity, potentially affecting model performance. Synthetic data use raises questions about real-world applicability, as it might not accurately reflect actual patient population diversity and nuances.

Another limitation involves RFE use for feature selection, which may not always identify the most relevant features for prediction. Whilst the study found that full features generally outperformed RFE-selected features, further research is needed to explore alternative feature selection methods that may better capture dataset nuances. Additionally, this study employed a fixed 0.5 probability threshold for binary classification across all models. While this threshold is appropriate for balanced datasets and provides equal weight to both classes, alternative threshold selection strategies could be explored in future research. Data-driven approaches such as ROC curve analysis with Youden's Index optimization, or cost-sensitive threshold selection based on specific programmatic priorities (e.g., prioritizing sensitivity to minimize missed cases), may yield different optimal cutoff values. The choice of classification threshold can significantly impact model performance metrics and should be tailored to specific implementation contexts and resource allocation priorities. Future studies should investigate threshold optimization strategies that account for the relative costs of false positives versus false negatives in HIV testing programs. Additionally, the study evaluated only a limited set of machine learning models. Whilst Random Forest and XGBoost performed well, other advanced models, such as deep learning approaches, were not explored. Future studies could investigate these models' potential to further improve prediction performance.

Finally, the study did not incorporate external validation using separate datasets, which is critical for assessing model robustness and generalisability. Findings should therefore be interpreted with caution, and future research should aim to validate these results using independent datasets from similar contexts. Incorporating external validation would enhance model reliability and ensure applicability in real-world settings, thereby supporting more effective public health strategies.

5. Conclusion

This study demonstrates that addressing class imbalance is fundamental for developing accurate diagnostic prediction models for HIV testing status among pregnant women in resource-limited settings. Our findings reveal that conventional machine learning approaches trained on imbalanced datasets exhibit severe majority class bias (macro F1-scores: 0.46-0.57), rendering them unsuitable for clinical deployment. However, SMOTE-based balancing coupled with comprehensive feature engineering enables dramatic performance improvements, with Random Forest achieving 0.72 macro F1-score, representing 56% improvement over standard approaches. The ensemble algorithms (Random Forest, XGBoost) consistently outperformed linear methods, demonstrating superior capability for capturing complex, non-linear predictor relationships in HIV testing determinants. Critically, full feature retention outperformed dimensionality reduction, suggesting that features deemed less important by traditional selection methods still contribute valuable diagnostic information. These technical contributions provide a reproducible framework for developing diagnostic prediction models in similar public health contexts, with direct implications for optimising PMTCT programme effectiveness. Future research should explore deep learning architectures, implement external

validation across multiple Sub-Saharan African contexts, integrate real-time electronic health record data, investigate alternative threshold optimization strategies tailored to specific programmatic priorities, and investigate hybrid balancing techniques beyond SMOTE. Additionally, prospective studies should evaluate real-world implementation outcomes, cost-effectiveness analyses, and healthcare provider adoption barriers to ensure these diagnostic models translate into measurable improvements in maternal and child health outcomes across resource-constrained settings.

CRediT authorship contribution statement

Afeez A. Soladoye: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **David B. Olawade:** Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation. **Oluwakemi Jumoke Bello:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **Claret Chinenyenwa Analikwu:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **Raphael Igbarumah Ayo Daniel:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **Augustus Osborne:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Data curation.

Ethical considerations and consent to participate

Institutional review board approval was not required for this investigation as the analysis utilised publicly accessible secondary data from the Demographic and Health Surveys programme.

Consent for publication

All authors consent for publication.

Funding

This research was conducted without external financial support or grant funding from any organisation or institution.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors express gratitude to the MEASURE Demographic and Health Surveys Program for providing open access to high-quality survey data that made this research possible.

Data availability

All data presented in this study are available upon request by contact with the corresponding author.

References

1. Adebamowo CA, Casper C, Bhatia K, et al. Challenges in the detection, prevention, and treatment of HIV-associated malignancies in low-and middle-income countries in Africa. *JAIDS J Acquir Immune Defic Syndr Res*. 2014 Sep 1;67:S17–S26.
2. Boyd AT, Oboho I, Paulin H, et al. Addressing advanced HIV disease and mortality in global HIV programming. *AIDS Res Ther*. 2020 Dec;17:1–7.
3. Chi BH, Mbori-Ngacha D, Essajee S, et al. Accelerating progress towards the elimination of mother-to-child transmission of HIV: a narrative review. *J Int AIDS Soc*. 2020 Aug;23(8), e25571.
4. Sarker M. Revolutionizing healthcare: the role of machine learning in the health sector. *J Artif Intell General Sci (JAIGS)*. 2024 Feb 27;2(1):36–61. ISSN: 3006-4023.
5. Bawistale K, Surendran R. *Macqueens based Yoga recommendation system for first trimester pregnancy*. *In2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*vol. 1. IEEE; 2024 Mar 14:552–559.
6. Bawistale K, Surendran R. *Hartigan-wong K-Means based yoga recommendation system for second trimester pregnancy*. *In2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*. IEEE; 2024 Aug 28:1039–1046.
7. Bawistale K, Surendran R. *Machine learning based yoga guidance approach for the third-trimester pregnant womens originated with Lloyd's k-Means algorithm*. *In2025 Third International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*. IEEE; 2025 May 21:1378–1385.
8. Sherine WB, Tamilvizhi T, Jemina SL. *Graph enhanced transformers for predicting drug using drug synergies and antagonistic interactions*. *In2025 6th International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE; 2025 Sep 10: 902–907.
9. Tamilvizhi T, Thirumalini R, Takshinyaa M, Dedeepya O. *An AI-Based framework for real-time patient monitoring and intelligent treatment recommendation in critical care units*. *2025 2nd International Conference on Computing and Data Science (ICCDS)*. 2025:1–5. Chennai, India.
10. Jemina SL, Thanarajan T. An intelligent brain tumor detection model using lightweight hybrid twin attentive pyramid convolutional network. *Sci Rep*. 2025 Nov 17;15(1), 40177.
11. Payagala S, Pozniak A. The global burden of HIV. *Clin Dermatol*. 2024 Mar 1;42(2): 119–127.
12. Amuche NJ, Emmanuel EI, Innocent NE. HIV/AIDS in Sub-Saharan Africa: current status, challenges and prospects. *Asian Pac J Trop Dis*. 2017;7(4):239–256.
13. Unaids. https://www.unaids.org/sites/default/files/media_asset/2022-global-aids-update_en.pdf; 2022.
14. Kawuki J, Kamara K, Sserwanja Q. Prevalence of risk factors for human immunodeficiency virus among women of reproductive age in Sierra Leone: a 2019 nationwide survey. *BMC Infect Dis*. 2022 Jan 17;22(1):60.
15. Fornah L, Shimbre MS, Osborne A, Tommy A, Ayalew AF, Ma W. Geographic variations and determinants of ever-tested for HIV among women aged 15–49 in Sierra Leone: a spatial and multi-level analysis. *BMC Public Health*. 2025 Mar 11;25 (1):961.
16. Statistics Sierra Leone (Stats SL) and ICF. Sierra Leone demographic and health survey 2019. Freetown: stats SL and ICF. https://dhsprogram.com/pubs/pdf/FR365/FR365.pdf; 2020.
17. Adugna DG, Worku MG. HIV testing and associated factors among men (15-64 years) in Eastern Africa: a multilevel analysis using the recent demographic and health survey. *BMC Public Health*. 2022 Nov 24;22(1):2170.
18. Teklehaimanot HD, Teklehaimanot A, Yohannes M, Biratu D. Factors influencing the uptake of voluntary HIV counseling and testing in rural Ethiopia: a cross-sectional study. *BMC Public Health*. 2016 Dec;16:1–3.
19. Muyunda B, Musonda P, Mee P, Todd J, Michelo C. Educational attainment as a predictor of HIV testing uptake among women of child-bearing age: analysis of 2014 demographic and health survey in Zambia. *Front Public Health*. 2018 Aug 14;6:192.
20. Magadi MA, Gazimbi MM. A multilevel analysis of the determinants of HIV testing in Zimbabwe: evidence from the demographic and health surveys. *HIV/AIDS Res Treat: Open J*. 2017 Jan 20;4(1).
21. Helleringer S, Kohler HP, Frimpong JA, Mkandawire J. Increasing uptake of HIV testing and counseling among the poorest in sub-saharan countries through home-based service provision. *JAIDS J Acquir Immune Defic Syndr Res*. 2009 Jun 1;51(2): 185–193.
22. Yaya S, Oladimeji O, Oladimeji KE, Bishwajit G. Determinants of prenatal care use and HIV testing during pregnancy: a population-based, cross-sectional study of 7080 women of reproductive age in Mozambique. *BMC Pregnancy Childbirth*. 2019 Dec;19, 1-0.
23. Nattey C, Jinga N, Mongwenyana C, et al. Understanding predictors of early antenatal care initiation in relationship to timing of HIV diagnosis in South Africa. *AIDS Patient Care STDS*. 2018 Jun 1;32(6):251–256.
24. Zegeye EA, Mbonigaba J, Dimbuene ZT. Factors associated with the utilization of antenatal care and prevention of mother-to-child HIV transmission services in Ethiopia: applying a count regression model. *BMC Womens Health*. 2018 Dec;18:1, 1.
25. Fanta W, Worku A. Determinants for refusal of HIV testing among women attending for antenatal care in Gambella region, Ethiopia. *Reprod Health*. 2012 Dec;9:1–3.
26. Worku MG, Tesema GA, Teshale AB. Prevalence and associated factors of HIV testing among reproductive-age women in eastern Africa: multilevel analysis of demographic and health surveys. *BMC Public Health*. 2021 Dec;21:1–9.
27. Kallon I. *Factors Influencing HIV Infection of Children Born to Mothers Living with HIV in Sierra Leone: Review of Literature and Best Practices of Service Delivery*. 2022.
28. Kelly JD, Weiser SD, Tsai AC. Proximate context of HIV stigma and its association with HIV testing in Sierra Leone: a population-based study. *AIDS Behav*. 2016 Jan; 20:65–70.
29. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *J Br Surg*. 2015 Feb;102(3):148–158.
30. Gholampour S. Impact of nature of medical data on machine and deep learning for imbalanced datasets: clinical validity of SMOTE is questionable. *Mach Learn Knowledge Extr*. 2024 Apr 15;6(2):827–841.
31. Kaur H, Pannu HS, Malhi AK. A systematic review on imbalanced data challenges in machine learning: applications and solutions. *ACM Comput Surv*. 2019 Aug 30;52(4): 1–36.
32. Owusu-Adjei M, Ben Hayfron-Acquah J, Frimpong T, Abdul-Salaam G. Imbalanced class distribution and performance evaluation metrics: a systematic review of prediction accuracy for determining model performance in healthcare systems. *PLOS Digital Health*. 2023 Nov 30;2(11), e0000290.
33. Altalhan M, Algarni A, Alouane MT. Imbalanced data problem in Machine learning: a review. *IEEE Access*. 2025 Jan 20.

34. Mohammed AJ, Muhammed Hassan M, Hussein Kadir D. Improving classification performance for a novel imbalanced medical dataset using SMOTE method. *Int J Adv Trends Comput Sci Eng.* 2020 Jun 1;9(3):3161–3172.
35. Naheed N, Shaheen M, Khan SA, Alawairdhi M, Khan MA. Importance of features selection, attributes selection, challenges and future directions for medical imaging data: a review. *Comput Model Eng Sci.* 2020 Oct 6;125(1):314–344.
36. Natras R, Soja B, Schmidt M. Ensemble machine learning of random forest, AdaBoost and XGBoost for vertical total electron content forecasting. *Remote Sens.* 2022 Jul 24;14(15):3547.
37. Kavzoglu T, Teke A. Predictive performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NGBoost). *Arabian J Sci Eng.* 2022 Jun;47(6):7367–7385.