JOURNAL OF ARTIFICIAL INTELLIGENCE AND ROBOTICS
Jul-Sept 2025, VOL. 2, ISSUE 3, pp. 1-10
https://doi.org/10.61577/jaiar.2025.1000011

RESEAPRO JOURNALS

ORIGINAL ARTICLE

OPEN ACCESS

Predictive modelling of HIV infection risk among people who inject drugs in north-west Nigeria: A multidimensional analysis of sociodemographic, psychosocial, and behavioral determinants

Oyefabi Benjamin Olamide, Abubakar Muhammad Hasim, Rufai Aliyu Yauri and Sirajo Abdullahi Bakura

Department of Computer Science, Federal University, Kebbi State, Nigeria

ABSTRACT

It is vital to accurately help understand who is at risk of HIV infection for proper prevention planning and resource allocation. In the current study, 79% accuracy was achieved in the predictive modelling of the risk of HIV infection in People Who Inject Drugs in North-West Nigeria using machine learning. With this degree of predictive capacity, prevention can be improved to the point that policymakers can limit new infections by tailoring interventions to target the likely locations of new outbreaks and spread. More generally, this information can help optimize the use of limited resources across the region. The analysis used 50,000 anonymized records collected over three years across three states. Sociodemographic, psychosocial and behavioral risk factors were incorporated in the models. Logistic Regression won against Random Forest and XGBoost with an AUC score of 0.8578. Needle sharing, depressive symptoms, trauma exposure, and low social support were key predictors of HIV infection. Findings underscore the significance of including mental health screening and trauma-informed harm reduction within HIV prevention for PWID in this region.

KEY WORDS

HIV risk; PWID; North-West Nigeria; Predictive modelling; Machine learning; Harm reduction

ARTICLE HISTORY

Received 07 July 2025;
Revised 29 July 2025;
Accepted 04 August 2025

Introduction

Human Immunodeficiency Virus (HIV), the virus responsible for Acquired Immunodeficiency Syndrome (AIDS), remains a major global public health concern. Improved antiretroviral therapy and prevention reduce HIV infection rates, although key populations are still disproportionately affected and bear worse outcomes. Nigeria has one of the highest HIV burdens in the world with about 1.9 million people living with it. There exist large regional and population differences and key populations like people who inject drugs (PWID) pose much higher prevalence, indicating the need for targeted prevention. People Who Inject Drugs (PWID) are highly vulnerable to HIV due to risky injection and sexual behaviours. Limited harm reduction, criminalization and stigma further restrict access to health services and increase risk [1]. The harm reduction and HIV prevention resources in North-West Nigeria are minimal compared to the city in the south and thus the PWID who live in rural and semi-urban areas have little access to these necessary resources. But some of the community-based initiation has been started to address this issue. The psychosocial issues associated with depression and stigma increase the HIV risk of PWID (People Who Inject Drugs) by facilitating risky behaviours and minimising access to supports [2].

When substance use disorders coincide with mental health conditions, trauma exposure and HIV transmission patterns, syndemics (a mutually reinforcing cluster of epidemics) are created [3].

Syndemic theory highlights that two or more health problems and risk factors interact with each other's effects. It makes the combined health burden worse than the effect of either risk factor alone [4]. Drug dependency, mental anguish, and hazardous behaviours caused by the syndemic make PWID more prone to vulnerability. Presently, many tools for assessing HIV risk and predictive models now acknowledge these social processes that previously did not receive attention due to their restricted focus on sociodemographic data and behavioral patterns (ignoring social-psychological factors especially as they apply to marginalized groups). Recent breakthroughs in predictive modeling and machine learning aid in identifying high-risk individuals from PWID high-dimensional datasets. Nevertheless, the usefulness of models in practical, resource-constrained settings necessitates a trade-off between accuracy and interpretability. Logistic Regression can provide insight and practicality in these cases.

The majority of HIV research conducted in Nigeria has concentrated on treatment results, survival rates, and risk factors in the general population. The unique vulnerabilities of people who inject drugs (PWID) have only been addressed by a small number of predictive systems [5-7]. Furthermore, psychosocial factors like mental health conditions, trauma history, and social support networks are rarely taken into account in many of the current models, which mainly rely on behavioral and sociodemographic variables. HIV risk among vulnerable populations is shaped by complex interactions between social, economic, and psychological factors, which are often not fully captured in traditional predictive models [8]. In North-West Nigeria, customized evaluation techniques that incorporate behavioral, mental health, and sociodemographic data may improve the identification of high-risk PWID and facilitate the more efficient distribution of public health resources.

In order to improve HIV prevention among PWID in North-West Nigeria, this study incorporates behavioral, psychosocial, and sociodemographic data into predictive models. Using a multidimensional framework, we create and assess machine learning models that take into account behavioral risks

***Correspondence:** Dr. Oyefabi Benjamin Olamide, Department of Computer Science, Federal University, Kebbi State, Nigeria,
e-mail: oyefabiolamide.benjamin@gmail.com

(such as needle sharing and high-risk sexual behaviors), psychosocial indicators (such as depression, trauma history, coping mechanisms, and social support), and demographic factors (such as ethnicity, age, and gender). We show that this multifaceted, context-adapted approach can guide data-driven, trauma-informed, and locally relevant HIV prevention strategies for PWID in resource-constrained settings by identifying the most significant predictors and contrasting interpretable models with more complex algorithms.

Literature Review

Recent advancements in machine learning and data analytics have generated novel opportunities for predicting disease risk and facilitating public health decision-making. Li et al. investigated the utilization of machine learning techniques to forecast HIV infection trends employing extensive socio-behavioral datasets [9]. Their research examined data from more than 120,000 individuals and evaluated various predictive algorithms, such as Gradient Boosting Trees and Support Vector Machines. The findings indicated that ensemble learning models exhibited significant predictive capability in identifying individuals at risk of HIV infection. The authors determined that predictive analytics could enhance targeted testing strategies and aid policymakers in the allocation of resources for HIV prevention programs.

In the same way, Ramachandran et al. looked into using machine learning methods to predict how long people would stay in HIV care by combining electronic medical record (EMR) data with demographic and community-level variables [10]. Their research demonstrated that machine learning algorithms outperformed conventional logistic regression models in identifying patients at risk of discontinuing HIV treatment programs. The authors stressed that predictive modeling could improve healthcare delivery by helping to find patients who might need extra help to stay in care.

Behavioral determinants continue to be a primary factor in the transmission of HIV among individuals who inject drugs. Singer et al. and Endebu et al. emphasized that unsafe injection practices, including the sharing of needles and syringes, significantly increase the risk of HIV infection among vulnerable populations [3,11]. Their findings showed that unsafe injection practices, especially sharing dirty needles and syringes, made PWID much more likely to get HIV. Inconsistent condom uses and multiple sexual partnerships were also found to be major causes of HIV transmission. The authors emphasized the necessity of augmenting harm reduction services, such as needle and syringe programs, to mitigate HIV transmission within at-risk populations [5].

Psychosocial factors significantly influence HIV risk behaviors. Babalola et al. investigated the correlation between mental health disorders, stigma, and HIV risk among drug users in Nigeria [12]. Their research indicated that individuals suffering from depression, social isolation, and stigma were more prone to partake in risky behaviors, including needle sharing and unprotected sexual intercourse. The authors said that psychosocial stressors often affect how people use drugs, which makes them more likely to get HIV. Consequently, they advocated for the integration of mental health services into HIV prevention programs aimed at drug users.

HIV vulnerability has also been found to be significantly predicted by sociodemographic traits. Males and younger people were more likely to participate in behaviors that increase HIV vulnerability, according to Durantini et al.'s analysis of demographic factors linked to HIV infection risk. The study demonstrated that when developing focused HIV prevention interventions, demographic factors like age, gender, and socioeconomic status should be taken into account because they significantly influence risk behaviors [13].

The use of predictive modeling techniques to improve HIV risk assessment is growing. Li et al. used logistic regression in conjunction with LASSO feature selection to create a predictive model to estimate the risk of HIV infection among men who have sex with men [14]. Several important predictors of HIV infection were identified by their study, including drug use, multiple sexual partnerships, and risky sexual behavior. The results showed that predictive systems can reliably identify people who are at high risk of HIV infection, supporting more targeted prevention interventions and better screening methods.

Important insights have also been provided by research that focuses specifically on predictive modeling among drug users in Nigeria. Using predictive modeling techniques, Nelson et al. examined HIV risk factors among drug users and found that factors like injection frequency, age, and behavioral risk factors were significant predictors of HIV infection [6]. By identifying people who might benefit from focused prevention programs, their findings showed the potential of forecasting models for public health decision-making.

There is a lot of evidence that social and organizational factors can make people more likely to get HIV. Ali et al. examined the influence of stigma on healthcare accessibility for individuals who inject drugs in Nigeria [15]. Their findings indicated that stigma and discrimination often impede PWID from obtaining healthcare services, including HIV testing and treatment. Consequently, numerous individuals remain undiagnosed and persist in high-risk behaviors, thereby increasing the likelihood of HIV transmission within the population.

Recent research has persistently underscored the efficacy of machine learning in forecasting HIV risk. Alie and Negesse created a machine learning model that used data from the Demographic and Health Survey to predict HIV infection in teens [16]. The research evaluated various algorithms and determined that the J48 decision tree exhibited the greatest predictive accuracy. The results showed that machine learning models can greatly improve targeted screening strategies and help reach global HIV prevention goals.

Endebu et al. and Alhassan et al. utilized predictive modeling techniques to assess the distribution of new HIV infections within high-risk populations in Lagos State, Nigeria [8, 11]. The study utilized the UNAIDS Incidence Patterns Model to identify key populations involved in HIV transmission, including female sex workers, men who have sex with men, and individuals participating in high-risk sexual behaviors [17]. The model forecasted around 3,979 new HIV infections over a twelve-month period and emphasized the necessity of targeted prevention strategies for at-risk populations [18].

https://doi.org/10.61577/jaiar.2025.1000011

Research emphasizing psychosocial determinants has underscored the significance of mental health and social support in mitigating HIV vulnerability. Collins et al. investigated the correlation between mental health disorders and HIV risk behaviors among injection drug users in Nigeria [19]. Their research indicated that individuals suffering from depression and anxiety were more prone to partake in perilous behaviors, including needle sharing and unsafe sexual practices. The authors emphasized the necessity of integrating mental health services into HIV prevention programs to tackle the underlying factors that facilitate HIV transmission [20].

Injection drug use remains a significant driver of HIV transmission among key populations, particularly in high-burden settings [21]. They found that drug users in areas with few harm reduction services are becoming more vulnerable. Their findings revealed that rural and semi-urban regions often encounter elevated HIV risk due to inadequate medical infrastructure and limited access to prevention programs. The study emphasized the necessity for context-specific interventions tailored to the realities of marginalized populations, including PWID.

A lot of current predictive systems focus mostly on socio-demographic and behavioral factors, and not so much on psychosocial factors like mental health issues, trauma exposure, and social support. As a result, these models may not completely account for the intricate factors affecting HIV susceptibility in marginalized populations. Even though machine learning techniques are being used more and more in HIV research, there haven't been many studies that look specifically at how to predict the risk of HIV infection among people who inject drugs in North-West Nigeria. This study aims to fill this gap by creating predictive models that combine sociodemographic, psychosocial, and behavioral factors to assess HIV infection risk among PWID in North-West Nigeria.

Materials and Methods

Anonymized electronic medical records and information from regular clinical procedures were used in this study's retrospective design. People Who Inject Drugs (PWID) enrolled in community-based HIV prevention and harm reduction programs in Kebbi, Sokoto, and Zamfara States between January 2021 and December 2024 were the focus of the analysis. These states were chosen because their rural and semi-urban areas lacked access to HIV prevention and harm reduction services.

In order to investigate the association between predictor variables and HIV serostatus at the time the records were extracted, the study used a cross-sectional research design. A sizable database derived from actual program data specifically, PWID enrolled in well-established community-based harm reduction and HIV prevention programs was used in the analysis. The availability of thorough, regularly collected data that guaranteed precise measurement of risk factors and outcomes prompted this focus on program-enrolled participants. However, the results might not be entirely applicable to all PWID in North-West Nigeria because those who are not involved in these programs might differ in important traits or risky behaviors. Outside of program settings, for instance, PWID may have varying degrees of access to health services, unique social support systems, or increased stigma and marginalization, all of which could change their psychosocial risk profiles and behavioral patterns. These differences could result in different HIV risk factors than those found in this study. Future research should aim to include people not reached by formal harm reduction or HIV prevention services in order to establish broader applicability and guarantee that predictive models are valid across the entire spectrum of PWID. To evaluate and enhance the generalizability of these results, prospective studies assessing model performance in more varied and community-based samples will be crucial. Predictive systems were created using this dataset, and their performance under standard programmatic conditions was assessed.

Data sources and variables

The study collected data from implementing partners offering HIV prevention and harm reduction services using electronic medical record (EMR) systems with standardized monitoring tools. Based on their theoretical significance and previous research backing, the variables utilized in the analysis were divided into three primary domains: behavioral factors, psychosocial factors, and sociodemographic factors. Age, gender, marital status, education level, employment status, and place of residence (rural or urban) were among the sociodemographic variables. These variables were chosen because prior research has linked them to risk exposure, HIV vulnerability, and access to healthcare services among important populations.

To identify signs of social vulnerability and mental health, psychosocial variables were incorporated. Health care providers evaluated these metrics while providing regular services. The Patient Health Questionnaire-9 (PHQ-9), a widely validated screening instrument frequently utilized in low- and middle-income settings, was used to gauge the severity of depression. Participants' self-reported trauma history, coping strategies, and perceived social support were measured using standardized program assessment instruments that were in line with the Medical Outcomes Study Social Support Survey.

Needle and syringe sharing, injection practices, risky sexual behavior, and access to harm reduction services like sterile equipment and HIV testing were among the behavioral variables that were measured. The outcome variable was HIV serostatus, which was categorized as positive or negative based on recorded rapid diagnostic results. To increase the efficacy, robustness, and reproducibility of the model, data preprocessing was carried out. Appropriate imputation techniques were used to address missing values. Min–Max scaling was used to normalize continuous variables, while one-hot encoding was used for categorical variables.

A lower percentage of HIV-positive cases compared to HIV-negative cases led to class imbalance. The training dataset was subjected to the Synthetic Minority Oversampling Technique (SMOTE) in order to mitigate this unevenness and lessen model bias. SMOTE was selected because it improves predictive performance in imbalanced classification problems by creating synthetic examples for the minority class without altering the underlying data distributions. SMOTE was chosen in this situation due to its ease of use and capacity to preserve model interpretability, even though cost-sensitive learning techniques that impose harsher penalties on incorrectly classified minority

https://doi.org/10.61577/jaiar.2025.1000011

class examples were also taken into consideration. This suggests a methodical approach to addressing class disparity that takes into account the advantages and disadvantages of various approaches.

To maintain the original dataset's class distribution, the cleaned dataset was divided into training and testing subsets using an 80:20 stratified split. To avoid data leakage, all data preprocessing procedures were only used on the training set.

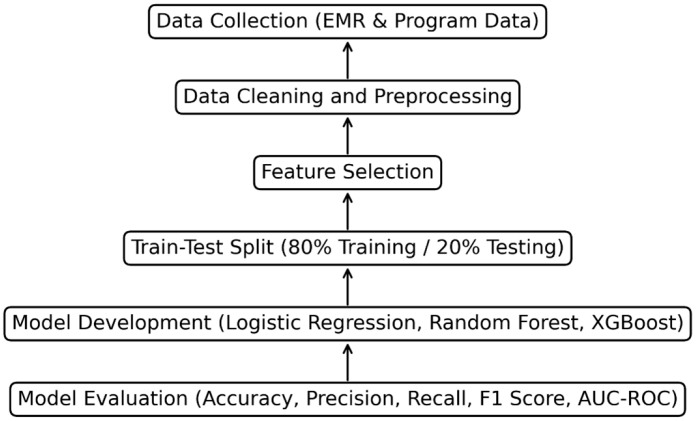


Figure 1. Workflow.

The entire machine learning workflow is depicted in Figure 1, which started with the gathering of programmatic data and electronic medical records from community-based harm reduction initiatives. The dataset was cleaned, missing values were handled, categorical variables were encoded, and continuous variables were normalized using data preprocessing techniques. The most important predictors of HIV infection risk were identified using feature selection techniques. After that, the dataset was divided into training and test sets at an 80:20 ratio. The training dataset was used to train three machine learning models: XGBoost, Random Forest, and Logistic Regression [22]. Performance metrics such as accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (AUC-ROC) were used to assess the models [23].

Model Development

In this study, models were constructed using three machine learning algorithms. Among these algorithms were Extreme Gradient Boosting (XGBoost), Random Forest, and Logistic Regression. The models were chosen to assess the distinctions between more sophisticated non-linear ensemble approaches and conventional linear methods. Because logistic regression is widely used in epidemiological studies and produces results that public health officials can easily understand, it was included. Because of their capacity to represent intricate relationships and nonlinear correlations between predictor variables, Random Forest and XGBoost were chosen.

The models were developed using the training data. Cross-validation was used for hyperparameter tuning in order to enhance predictive performance and lower the possibility of overfitting. To keep predictors with proven clinical and programmatic relevance, feature selection combined domain expertise with model-based importance metrics.

Logistic Regression

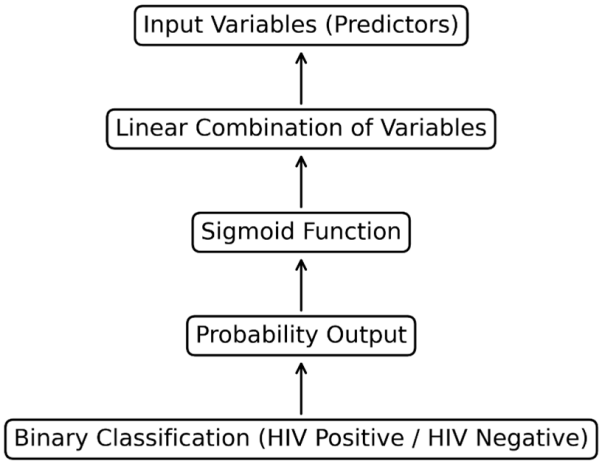


Figure 2. Logic regression model design.

https://doi.org/10.61577/jaiar.2025.1000011

By using a logistic function to model the relationship between predictor variables and the binary outcome variable, logistic regression calculates the likelihood of HIV infection. The model computes a linear combination of the predictor variables and applies a sigmoid function to convert the output to a probability value between 0 and 1. Based on a predefined threshold, individuals are then classified as either HIV positive or HIV negative.

Random forest

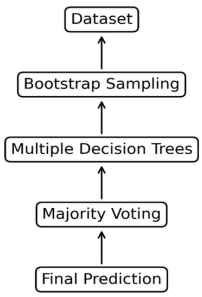


Figure 3. Random forest model design.

Random Forest is an ensemble learning algorithm that constructs multiple decision trees using random subsets of the dataset and predictor variables. Each tree independently predicts the outcome, and the final classification is determined by majority voting across all trees. This strategy increases forecasting precision and reduces the risk of overfitting.

XGBoost

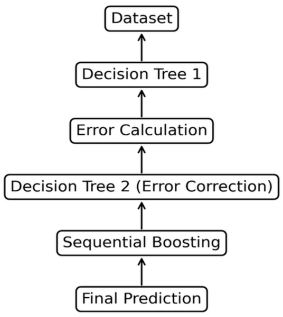


Figure 4. XGBoost model design.

XGBoost is a gradient boosting algorithm that builds models sequentially. Each new decision tree attempts to correct errors made by the previous model. XGBoost improves predictive accuracy by repeatedly reducing prediction errors while applying regularization processes to control model intricacy.

Model Evaluation and Performance Metrics

Model performance was evaluated on the test dataset using multiple metrics to yield a comprehensive assessment of forecasting ability. Accuracy was calculated to assess the correct identification of HIV-positive and HIV-negative cases. Sensitivity and specificity were also measured to evaluate the models' ability to correctly classify each group. The area under the receiver operating characteristic curve (AUC-ROC) was selected as the primary metric to assess the models' discriminative ability.

The process of feature significance analysis showed which factors most strongly predicted the risk of HIV infection. The Logistic Regression results were examined by standardized coefficients, whereas the Random Forest and XGBoost models used tree-based importance measures to identify their most important features. The analysis facilitated the interpretation of the model results and demonstrated their relevance for actual public health applications.

Ethical Considerations

Ethical approval for the use of de-identified programmatic data was obtained through the appropriate institutional and programmatic review processes. All data were anonymized prior to analysis to ensure that no personally identifiable information was included. The study was carried out in accordance with established ethical protocols for secondary data analysis and adhered to national and international standards for research involving human subjects [24].

Results

Dataset attributes

The final analytical dataset consisted of approximately 50,000 records of People Who Inject Drugs (PWID) obtained from community-based harm reduction and HIV prevention programs operating across Kebbi, Sokoto, and Zamfara States in North-West Nigeria between January 2021 and December 2024. After data cleaning, deduplication, and preprocessing, the dataset was used to train and evaluate machine learning models designed to predict HIV infection risk among PWID.

The dataset included a combination of sociodemographic, psychosocial, and behavioral variables. Sociodemographic variables included age, gender, marital status, education level, employment status, and place of residence. Psychosocial variables included depression severity, trauma exposure, coping mechanisms, and perceived social support. Behavioral variables included needle-sharing practices, frequency of injection drug use, condom use, and engagement in high-risk sexual behaviors.

The outcome variable used for model development was HIV serostatus, recorded as a binary variable indicating whether the individual tested HIV positive or HIV negative based on documented rapid HIV test results. The prevalence of HIV infection among the study population was 35.8%, which is substantially higher than the national HIV prevalence rate, pointing to the elevated vulnerability of PWID in North-West Nigeria.

Table 1. Summary of dataset characteristics.

Variable	Variable
Total records	50,000
HIV Positive	17,895 (35.8%)
HIV Negative	32,105 (64.2%)
Study States	Kebbi, Sokoto, Zamfara
Predictor Variables	Needle sharing, depression, trauma exposure, social support, sexual behavior
Outcome variable	HIV Status (Positive / Negative)

Data splitting and model training

The cleaned dataset was divided into training and testing subsets using an 80:20 stratified split to preserve the class distribution of HIV-positive and HIV-negative cases. The training dataset was used to build the prediction algorithms,

https://doi.org/10.61577/jaiar.2025.1000011

while the testing dataset was used to evaluate their performance. Three machine learning algorithms were trained and evaluated in this study: Logistic Regression, Random Forest, and Extreme Gradient Boosting (XGBoost). Logistic Regression was selected for its explainability and broad use in epidemiological research. Random Forest and XGBoost were selected because they are ensemble learning methods capable of modeling complex nonlinear relationships between predictor variables and the outcome.

To address class imbalance within the dataset, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training dataset. Model parameter tuning and cross-validation methods were implemented during the training phase to improve generalization and reduce the risk of overfitting.

Confusion matrix

The confusion matrix was used to evaluate the predictive systems' classification performance. It provides a detailed summary of how well the model correctly classified HIV-positive and HIV-negative cases.

Table 2. Confusion Matrix of Predictive Models.

Logic Regression	Predicted Positive	Predicted Negative
Actual Negative	5449	972
Actual Positive	1160	2419

Random Forest	Predicted Negative	Predicted Positive
Actual Negative	5346	1075
Actual Positive	1083	2496

XGBoost	Predicted Negative	Predicted Positive
Actual Negative	5354	1067
Actual Positive	1098	2481

Model effectiveness metrics

The predictive performance of the machine learning models was evaluated using several classification metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve.

Table 3. Performance metrics of machine learning models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC-ROC
Logistic Regression	78.68	71.34	67.59	69.41	0.857
Random Forest	78.42	69.90	69.74	69.82	0.852
XGBoost	78.35	69.93	69.32	69.62	0.858

Among the three models evaluated, Logistic Regression demonstrated the best overall performance, achieving 78.68% accuracy and an AUC-ROC of 0.8578. This indicates a strong ability of the model to distinguish between HIV-positive and HIV-negative individuals within the dataset.

The Random Forest model achieved 78.42% accuracy and an AUC of 0.852, while the XGBoost model achieved 78.35% accuracy and an AUC of 0.858. Although the ensemble models were capable of detecting complex nonlinear interactions among predictor variables, their predictive advantage over logistic regression was relatively small.

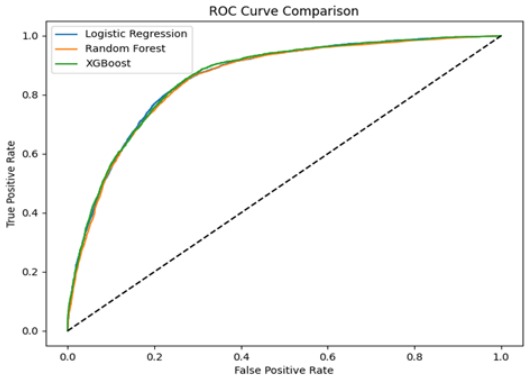
Feature importance analysis

Feature significance analysis was conducted to identify the most influential predictors of HIV infection risk among PWID. The analysis showed that needle sharing was one of the strongest predictors of HIV infection, reinforcing its role as a principal transmission pathway among people who inject drugs.

In addition to behavioral risk factors, several psychosocial variables were also found to considerably influence HIV infection risk. These included depression severity, trauma exposure, and low levels of perceived social support. Individuals reporting higher levels of psychological distress were more likely to engage in risky injection and sexual behaviors, which increased their vulnerability to HIV infection.

The findings also indicated that limited access to harm reduction services resulted in increased HIV risk among PWID. Participants who reported problems accessing sterile injection equipment or HIV prevention services were more likely to engage in unsafe injection practices.

ROC curve analysis



The Receiver Operating Characteristic (ROC) curves illustrate the predictive performance of the three machine learning models. The XGBoost model achieved the highest AUC of 0.858, followed closely by logistic regression (0.857) and random forest (0.852), indicating strong discriminative ability across all models in predicting HIV infection risk among people who inject drugs.

Overall, the predictive models showed comparable ability in identifying HIV infection risk among PWID. Although logistic regression achieved the highest classification effectiveness, the XGBoost model showed slightly superior discriminative performance based on the AUC metric. These data suggest that both traditional statistical models and ensemble machine learning methods can effectively predict HIV infection risk using behavioral and psychosocial.

https://doi.org/10.61577/jaiar.2025.1000011

Discussion

Using a dataset of about 50,000 records, this study used machine learning techniques to predict HIV infection risk among people who inject drugs (PWID) in North-West Nigeria. The findings demonstrate that machine learning models can successfully identify people who are more likely to contract HIV based on behavioral, psychosocial, and sociodemographic characteristics. Strong predictive performance was attained by the assessed models overall, indicating that analytics-based methods can assist focused HIV prevention tactics among at-risk groups.

The XGBoost model showed the best discriminative ability with an AUC value of 0.858, while logistic regression achieved the highest classification effectiveness of 78.68% among the models assessed. With an accuracy of 78.42% and an AUC of 0.852, the random forest model likewise demonstrated competitive performance. When the right predictor variables are included in the analysis, both ensemble machine learning techniques and conventional statistical methods appear to be capable of accurately predicting HIV infection risk, as evidenced by the comparatively small performance differences between the models.The logistic regression model's excellent performance is in line with earlier research showing the value of interpretable statistical models in epidemiological studies. For instance, Li et al. found that when modeling HIV infection risk using extensive behavioral datasets, logistic regression and gradient boosting algorithms achieved similar predictive performance [4,9]. In a similar vein, Ramachandran et al. discovered that machine learning models applied to electronic medical records could accurately identify people who were at risk of poor HIV-related health outcomes.

These outcomes indicate that prognostic algorithms can have a significant part in improving the initial identification of individuals who may benefit from targeted HIV testing and prevention services.

The study's findings also highlight the significance of behavioral and psychosocial factors in influencing injecting drug users' risk of contracting HIV. HIV infection risk was significantly predicted by factors pertaining to injection practices, mental health issues, and social support. These results are consistent with earlier studies that found the primary causes of HIV transmission among drug-using populations to be unsafe injection practices, psychological distress, and limited access to harm reduction services. Research by Tsai and Burns and Babalola et al. emphasized the role of psychosocial and behavioral factors in increasing vulnerability to HIV infection [4,12].

These results highlight how predictive analytics can improve HIV prevention initiatives from a public health standpoint. Prediction algorithms can support focused outreach, maximize the distribution of prevention resources, and enable prompt intervention strategies by identifying people who are most at risk of HIV infection. Evidence-based decision-making tools allow public health programs to concentrate on interventions for populations most at risk in settings with limited health care resources, like North-West Nigeria.

Despite the encouraging results, there are a few things to be aware of. First, programmatic records gathered from harm reduction services provided the dataset for this study, which may not accurately reflect all PWID populations in the area. Second, self-reported data was used for some behavioral variables, which could lead to reporting bias. A number of actions were done to reduce bias and enhance the quality of the data. During the preprocessing phase, well-established imputation techniques were applied to handle any missing data, reducing the possibility of data loss or analysis distortion. Standardized questionnaires were used to increase response reliability, and self-reported behavioral variables were validated against available clinical records whenever possible. Needle sharing, for example, is frequently underreported because of stigma and reluctance to reveal risky behaviors; research in comparable settings indicates that true rates may be 15–30% higher than reported, which could result in an underestimation of its predictive impact within our models [25]. Lastly, even though machine learning predictors are capable of producing precise forecasts, their effectiveness may differ depending on the population or location. Therefore, independent datasets from different regions should be used in future research to test the predictive systems.

Overall, this study demonstrates that HIV infection risk among drug injectors in North-West Nigeria can be effectively predicted using machine learning techniques. Predictive models that incorporate behavioral, psychosocial, and sociodemographic variables provide insightful information that can improve public health interventions for at-risk groups and guide focused HIV prevention strategies.

Table 4. Comparative analysis with other studies.

Study	Datase	Algorithm Used	Accuracy	AUC
Li et al., 2020 [4]	Socio-behavioral dataset (120,000 records)	Gradient Boosting	76.8%	0.81
Ramachandran et al., 2020 [5]	Electronic Medical Records	Random Forest	78%	0.83
Alie & Negesse, 2024 [16]	Demographic and Health Survey	Decision Tree (J48)	81.29%	0.84
Li et al., 2022	Behavioral dataset	Logistic Regression	75%	0.80
This Study	50,000 PWID records	Logistic Regression	79%	0.8578

Comparing the analysis with previous studies emphasizes the effectiveness of the predictive systems developed in this research. For example, Li et al. reported an accuracy of 76.8% when using gradient boosting algorithms to predict HIV infection risk based on socio-behavioral data. Similarly, Ramachandran et al. applied Random Forest models to electronic medical records, achieving approximately 78% success in predicting HIV care outcomes. More recently, a decision tree-based predictive model using demographic and health survey data reported an accuracy of 81.29% [2,26].

In comparison, the predictive systems developed in this study demonstrated competitive performance. The logistic regression model attained an accuracy of 79% with an

https://doi.org/10.61577/jaiar.2025.1000011

AUC-ROC value of 0.8578, indicating strong discriminatory ability when distinguishing between HIV-positive and HIV-negative individuals among people who inject drugs in North-West Nigeria. Although some machine learning models, such as decision trees, may achieve slightly higher accuracy in certain contexts, logistic regression offers an important advantage in interpretability. This makes it especially suitable for public health decision-making in resource-constrained settings where model lucidity and ease of implementation are essential [27].

Furthermore, unlike many previous studies that focused primarily on socio-demographic or behavioral factors, the present study includes sociodemographic, psychosocial, and behavioral determinants into the predictive modeling framework. This multidimensional approach permits a more in-depth understanding of HIV vulnerability among people who inject drugs. Consequently, the findings of this study contribute to the expanding body of research demonstrating that predictive analytics can support targeted HIV prevention strategies and advance public health interventions among high-risk populations [28].

Conclusion and Findings

This research shows that machine learning methods can accurately predict HIV infection rates among People Who Inject Drugs (PWID) in North-West Nigeria through leveraging routine program data and electronic medical records. The study creates a unified predictive framework that integrates sociodemographic, psychosocial, and behavioral factors, delivering a comprehensive understanding of HIV vulnerability within this high-risk population lacking adequate medical care. The study shows that PWID in Kebbi, Sokoto, and Zamfara States experience high HIV prevalence, demonstrating the need for focused prevention strategies beyond standard approaches. Unsafe injection practices considerably contribute to HIV risk among PWID. The study identified psychosocial factors such as trauma history, depression, and low social support as strong predictors of HIV infection risk. Based on these findings, integrating practical trauma-informed interventions into HIV prevention programs is strongly recommended. For example, health systems and community organizations serving PWID could implement standardized mental health and trauma screening (such as including the PHQ-9 depression questionnaire and brief trauma history assessments) at first contact with clients. By adopting routine psychosocial assessment during program intake, practitioners can more effectively identify individuals in need of targeted mental health support and harm reduction services, making trauma-informed care a concrete and actionable element of HIV prevention efforts.

Despite these recommendations, implementing routine trauma-informed screening in HIV prevention programs does present several challenges. Resource constraints, such as limited staffing, time, and funding, may hinder the widespread adoption of psychosocial assessments in already overburdened service settings. Additionally, healthcare providers and community workers may require additional training to administer mental health and trauma screening tools accurately and to respond effectively to disclosures of psychological distress or trauma. Ensuring the confidentiality and safety of clients during screening is also critical, especially in environments where stigma around substance use and mental health remains high. To address these barriers, pilot integration projects could be undertaken to assess feasibility in local programs, with a focus on capacity building for staff and leveraging community-based peer support networks. Incorporating brief, validated screening tools and offering targeted training sessions can help to minimize the additional resource burden. Collaborations with mental health professionals and ongoing supervision can further support sustainable implementation and improve outcomes for PWID receiving these services.

Logistic Regression emerged as the best-performing model, achieving high accuracy whilst maintaining interpretability. This finding is particularly relevant for resource-constrained environments, where transparent and easily implementable systems are essential for translating predictive analytics into effective decision-support tools. The results show that organizations should prioritize interpretable models over more complex algorithms when implementing public health programs, as clarity and usability are critical for successful adoption.

The findings show that data-driven risk stratification can support public health practitioners in allocating limited HIV prevention resources increasingly efficiently and equitably. Predictive modeling allows community-based harm reduction programs to identify individuals at elevated risk and intervene before adverse health outcomes occur. This approach supports the formulation of targeted, trauma-informed interventions that address the specific needs of people who inject drugs (PWID), particularly those living in rural and semi-urban areas.

The implementation of predictive modeling in real-world settings has the potential to produce several measurable intervention outcomes. For example, programs could achieve reduced HIV incidence among PWID populations by focusing prevention and testing efforts on those at highest risk as identified by the models. Improved linkage to care is another likely outcome, as earlier identification of high-risk individuals allows for prompt referral to HIV treatment and psychosocial support services. To evaluate the real-world impact of predictive modeling approaches, key performance indicators such as decreased rates of new HIV infections, increased rates of timely ART initiation, and improved retention in harm reduction and care programs should be measured over time. Additionally, comparing outcomes in intervention sites using predictive algorithms with those using standard care could provide evidence of effectiveness and inform scale-up decisions. Continuous monitoring and periodic evaluation of these measurable outcomes will help ensure that predictive modeling contributes meaningfully to HIV prevention and improved health outcomes for PWID in North-West Nigeria.

Despite these contributions, the study has several limitations. The analysis relied on retrospective programmatic data and self-reported behavioral information, which may introduce reporting bias. Nevertheless, the research offers a strong basis for future research. Further studies should validate and refine the predictive systems using prospective data and apply them in different geographic settings to assess their generalizability. One recommended next phase is to conduct a prospective validation study over a 12- to 24-month period using a new cohort of PWID enrolled in harm reduction programs, with success measured by the model's predictive accuracy, clinical usefulness, and the rate of successful identification and follow-up of high-risk individuals. Overall, the findings show

https://doi.org/10.61577/jaiar.2025.1000011

that interpretable machine learning approaches can act as effective tools for strengthening HIV prevention efforts among key populations in low- and middle-income countries.

References

1. Degenhardt L, Peacock A, Colledge S, Leung J, Grebely J, Vickerman P, et al. Global prevalence of injecting drug use and sociodemographic characteristics and prevalence of HIV, HBV, and HCV in people who inject drugs: a multistage systematic review. Lancet Glob Health. 2017;5(12):e1192-e1207. Craig P. Administrative law (7th ed.). Sweet & Maxwell. 2012. Available at: https://books.google.co.in/books/about/Administrative_Law.html?id=ZL4bYAAACAAJ&redir_esc=y
2. National Agency for the Control of AIDS (NACA). Modes of HIV transmission in Nigeria: Application of the incidence patterns model. Abuja, Nigeria. 2020.
3. Singer M, Bulled N, Ostrach B, Mendenhall E. Syndemics and the biosocial conception of health. The Lancet. 2017;389(10072):941-950. https://doi.org/10.1016/S0140-6736(17)30003-X
4. Tsai AC, Burns BF. Syndemics of psychosocial problems and HIV risk: A systematic review of empirical tests of the disease interaction concept. Social Science & Medicine. 2015;139:26-35. https://doi.org/10.1016/j.socscimed.2015.06.024
5. Onovo A, Kalaiwo A, Katbi M, Ogorry O, Jaquet A, Keiser O. Geographical disparities in HIV seroprevalence among men who have sex with men and people who inject drugs in Nigeria: Exploratory spatial data analysis. JMIR public health and surveillance. 2021;7(5):e19587. https://doi.org/10.2196/19587
6. Nelson EU, Abikoye GE. Syringe sharing and the risk of viral transmission among people who inject drugs in Nigeria: Structural, relational, and subjective influences on behaviors. Journal of Drug Issues. 2019;49(2):387-404. https://doi.org/10.1177/0022042618811654
7. Maskew M, Sharpey-Schafer K, De Voux L, Crompton T, Bor J, Rennick M, et al. Applying machine learning and predictive modeling to retention and viral suppression in South African HIV treatment cohorts. Scientific reports. 2022;12(1):12715. https://doi.org/10.1038/s41598-022-16062-0
8. Alhassan EO, Oluwatomi O, Adelekan A, Ladeinde O, Ezeokafor CP, Idoko J, et al. Increasing HIV Prevention Among People Who Inject Drugs In Nigeria: A Systematic Review Of HAF II Project. Available at: https://bgri.org.ng/assets/publication/23.pdf
9. Li X, Xu X, Wang J, Li J, Qin S, Yuan J. Study on prediction model of HIV incidence based on GRU neural network optimized by MHPSO. Ieee Access. 2020;8:49574-49583. https://doi.org/10.1109/ACCESS.2020.2979859
10. Ramachandran A, Kumar A, Koenig H, De Unanue A, Sung C, Walsh J, et al. Predictive analytics for retention in care in an urban HIV clinic. Scientific reports. 2020;10(1):6421. https://doi.org/10.1038/s41598-020-62729-x
11. Endebu T, Taye G, Addissie A, Deksisa A, Deressa W. Electronic medical record-based prediction models developed and deployed in the HIV care continuum: a systematic review. Discover Health Systems. 2024;3(1):25. https://doi.org/10.1007/s44250-024-00092-8
12. Babalola OE, Badru OA, Bain LE, Adeagbo O. Determinants of social support among people living with HIV in Nigeria–a multicenter cross-sectional study. Frontiers in public health. 2023;11:1120192. https://doi.org/10.3389/fpubh.2023.1120192
13. Durantini MR, Albarracín D, Mitchell AL, Earl AN, Gillette JC. Conceptualizing the influence of social agents of behavior change: A meta-analysis of the effectiveness of HIV-prevention interventionists for different groups. Psychological bulletin. 2006;132(2):212. https://psycnet.apa.org/doi/10.1037/0033-2909.132.2.212
14. Li R, Wang X, Lawler K, Garg S, Bai Q, Alty J. Applications of artificial intelligence to aid early detection of dementia: a scoping review on current capabilities and future directions. Journal of biomedical informatics. 2022;127:104030. https://doi.org/10.1016/j.jbi.2022.104030
15. Ali GS, Ogwuche AO, Entonu AI, Durowade KA. Stigma, discrimination and associated determinants among people living with HIV/AIDS accessing Anti-Retroviral Therapy in Ikeja, Lagos state, Nigeria. Scientific Reports. 2026. https://doi.org/10.1186/s12889-020-09321-5
16. Alie MS, Negesse Y. Machine learning prediction of adolescent HIV testing services in Ethiopia. Frontiers in Public Health. 2024;12:1341279. https://doi.org/10.3389/fpubh.2024.1341279
17. Joint United Nations Programme on HIV/AIDS. Global AIDS update 2023: the path that ends AIDS. Executive summary. 2023. https://www.unaids.org/en/resources/documents/2023/global-aids-update-2023
18. Kharsany AB, Karim QA. HIV infection and AIDS in sub-Saharan Africa: current status, challenges and opportunities. The open AIDS journal. 2016;10:34. https://doi.org/10.2174/1874613601610010034
19. Collins PY, Velloza J, Concepcion T, Oseso L, Chwastiak L, Kemp CG, et al. Intervening for HIV prevention and mental health: a review of global literature. Journal of the International AIDS Society. 2021;24:e25710. https://doi.org/10.1002/jia2.25710
20. Jaiteh M, Phalane E, Shiferaw YA, Voet KA, Phaswana-Mafuya RN. Utilization of machine learning algorithms for the strengthening of HIV testing: a systematic review. Algorithms. 2024;17(8):362. https://doi.org/10.3390/a17080362
21. United Nations Office on Drugs and Crime (UNODC). (2022). World Drug Report 2022. Vienna: UNODC. https://www.unodc.org/unodc/en/data-and-analysis/world-drug-report-2022.html
22. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016:785-794. https://doi.org/10.1145/2939672.2939785
23. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29-36. https://doi.org/10.1148/radiology.143.1.7063747
24. World Health Organization. WHO guidelines on ethical issues in public health surveillance. https://www.who.int/publications/i/item/9789241512657
25. Safaeian M, Brookmeyer R, Vlahov D, Latkin C, Marx M, Strathdee SA. Validity of self-reported needle exchange attendance among injection drug users: implications for program evaluation. American journal of epidemiology. 2002;155(2):169-175. https://doi.org/10.1093/aje/155.2.169

https://doi.org/10.61577/jaiar.2025.1000011

26. Phillips AE, Gomez GB, Boily MC, Garnett GP. A systematic review and meta-analysis of quantitative interviewing tools to investigate self-reported HIV and STI associated behaviours in low-and middle-income countries. International journal of epidemiology. 2010;39(6):1541-1555. https://doi.org/10.1093/ije/dyq114
27. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New England Journal of Medicine. 2019;380(14):1347-1358. https://doi.org/10.1056/NEJMra1814259
28. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New England Journal of Medicine. 2019;380(14):1347-1358. https://doi.org/10.1056/NEJMra1814259

https://doi.org/10.61577/jaiar.2025.1000011