ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Machine Learning-Based Early Prediction of Opportunistic Infections Among People Living with HIV

Udoinyang G Inyang[1] , Abraham C Inyang[1] , Asukwo E Onukak[3], Emmanuel A. Ubong[2]

[1]Department of Data Science, Faculty of Computing, University of Uyo, Nigeria

[2]School of Computing and Information Technology, Federal University of Technology, Ikot Abasi. Nigeria.

[3]Department of Internal Medicine, University of Uyo, Nigeria

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150600135

Received: 01 July 2026; Accepted: 07 July 2026; Published: 17 July 2026

ABSTRACT

The early prediction of opportunistic infections (OIs) among people living with HIV (PLWH). Despite the widespread use of antiretroviral therapy (ART), opportunistic infections remain a major cause of morbidity and mortality, particularly in low- and middle-income settings. Early identification of individuals at high risk is therefore essential for improving clinical outcomes and optimizing healthcare resource allocation. comprehensive dataset comprising 3,982 patient records and 88 clinical, demographic, laboratory, and treatment-related features was utilized. The dataset included categorical, numerical, and text-based variables capturing diverse aspects of patient health status. Principal Component Analysis (PCA) was applied for dimensionality reduction, identifying key contributing factors such as current weight, age, ART refill patterns, and engagement in enhanced adherence counseling sessions. Several machine learning algorithms were explored, with XGBoost emerging as the best-performing model. The model achieved an accuracy of 97.17%, with an AUC of 0.994 and an Average Precision score of 0.979, demonstrating strong discriminative ability. Evaluation metrics further confirmed sensitivity and specificity, with minimal false negatives, which is critical for clinical safety in HIV care.The proposed framework demonstrates the potential of machine learning to support early risk stratification and clinical decision-making for opportunistic infections in HIV-positive populations.

Keywords: HIV, Infection, Machine Learning, Opportunistic Infections, XGBoost

INTRODUCTION

Even in the antiretroviral therapy (ART) era, opportunistic infections (OIs) remain a major clinical challenge in HIV care. Although early combination ART (cART) significantly reduces the incidence of OIs and AIDS-defining malignancies, these conditions continue to contribute substantially to morbidity and mortality among people living with HIV, particularly in low- and middle-income countries. This persistent burden underscores the importance of early identification of individuals at high risk of developing OIs in clinical practice (Chen et al., 2025; Girma et al., 2022; Masur et al., 2014).The clinical rationale for early detection is especially strong among patients with advanced immunosuppression. When CD4 cell counts decline below 200 cells/μL, patients become highly susceptible to severe opportunistic infections and related malignancies, which significantly increases mortality risk. Consequently, identifying risk factors associated with immune decline and intervening early can substantially improve clinical outcomes. Furthermore, immune recovery is not uniform among all patients receiving ART; a subset of individuals experiences incomplete immune reconstitution despite virologic suppression, placing them at continued risk of AIDS progression and death (Zhou et al., 2023; Chen et al., 2025; Yang et al., 2020).At the health-system level, early risk prediction also has important operational implications. In resource-limited settings, forecasting the burden of opportunistic infections is critical for effective planning, advocacy, and allocation of scarce healthcare resources. At the individual level, evidence shows that OI incidence remains significant during follow-up and is influenced by multiple factors, including CD4 category, hemoglobin levels, ART-related side effects, use of preventive therapy, comorbid conditions, and rural

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

residence. These determinants collectively highlight the importance of early risk stratification and continuous patient monitoring (Rubaihayo et al., 2016; Girma et al., 2022).Against this background, machine learning (ML) approaches offer a promising solution for early prediction of disease progression in HIV care. ML models can integrate diverse clinical variables to identify high-risk patients, uncover important predictive features, and support timely clinical decision-making. Although not limited exclusively to opportunistic infections, recent studies in HIV research suggest that ML-based predictive systems can improve individualized risk assessment, enhance quality of life, and potentially reduce mortality by enabling earlier and more targeted interventions (Cai et al., 2025; Zhou et al., 2023).

This study aims to enhance HIV care through the application of XGBoost-based machine learning models for advanced data analytics and risk prediction. XGBoost, well known for its efficiency in handling structured data, feature selection capability, and high predictive performance, further improves the accuracy and reliability of outcome prediction in clinical datasets. As healthcare systems increasingly shift toward precision medicine, the integration of XGBoost can support earlier detection of opportunistic infections, enable timely clinical interventions, and ultimately improve patient outcomes among people living with HIV.

Related Works

Machine Learning Applications in HIV Care and Opportunistic Infection Research

Within HIV research, the current machine learning (ML) literature is broader than opportunistic infections (OIs) alone. Earlier systematic reviews indicate that ML applications in HIV have primarily focused on predicting HIV disease progression, antiretroviral drug resistance, and treatment optimization, rather than directly addressing opportunistic infections as a primary endpoint (Ahlstrom et al., 2019; Singh, 2017). In addition, more recent studies highlight extensive use of routine electronic health record (EHR) data to predict HIV acquisition risk, which is mainly relevant for prevention strategies and occurs upstream of OI development in individuals already living with HIV (Xu et al., 2022; Marcus et al., 2019; Gruber et al., 2020).

For people already diagnosed with HIV, a substantial portion of ML research has focused on intermediate clinical outcomes that indirectly influence OI risk. These include virologic failure, treatment interruption, mortality, and general disease progression. Several studies and reviews show that ML models can effectively identify patients at high risk of virologic failure or poor adherence, both of which contribute to immune suppression and increased susceptibility to opportunistic infections such as tuberculosis (Mamo et al., 2023; Ijaiya et al., 2025). Furthermore, recent syntheses of ML applications in HIV care emphasize the use of clinical indicators such as viral load, CD4 count trajectories, and co-infection profiles to predict disease progression and flag individuals at risk of rapid clinical deterioration, including OIs like tuberculosis and pneumonia (Sah et al., 2025; Sui et al., 2025).Although direct ML applications targeting opportunistic infections remain relatively limited, emerging studies demonstrate clinically important use cases, particularly in prognosis after infection has occurred. For example, ML-based models have been developed to predict in-hospital outcomes among patients with HIV-associated cryptococcosis, with algorithms such as random forest and support vector machines showing strong predictive performance for clinical prognosis (Gilliland et al., 2025; Zhan et al., 2024). This indicates that while ML is not yet widely applied to predict incident OIs in outpatient settings, it is increasingly useful in guiding risk stratification and outcome prediction once infection is established.

The existing evidence suggests that ML applications in HIV-related opportunistic infections should be understood along a continuum. This ranges from predicting upstream care outcomes such as treatment failure and immune decline, to modeling intermediate disease progression and co-infections, and finally to a smaller but growing body of work focused directly on OI diagnosis and prognosis (Mamo et al., 2023; Sah et al., 2025; Kurniawan et al., 2026).

Machine Learning for Direct Prediction and Diagnosis of Opportunistic Infections in HIV

Tuberculosis (TB) remains the most extensively studied opportunistic infection (OI) in machine learning (ML)-based HIV research. Several studies have developed ML and penalized regression models to predict TB occurrence or TB co-infection among people living with HIV, with the primary aim of enabling earlier detection

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

and timely intervention. These models include hospital-based cross-sectional TB prediction systems, risk stratification tools for TB co-infection, and incident TB prediction frameworks that use baseline HIV clinic data with defined follow-up windows (e.g., 6-month prediction horizons) to support preventive therapy decisions (Orwa et al., 2023; Chen et al., 2024; Bartl et al., 2025).Beyond individual clinical prediction, TB-related ML studies have also expanded to population-level forecasting. For example, recent work utilizing the World Health Organization (WHO) Global Tuberculosis database has applied ML techniques to predict TB incidence and HIV–TB co-infection trends at a broader epidemiological scale, demonstrating the applicability of data-driven models in both clinical and public health planning contexts (Emegano et al., 2025).In addition to TB, bloodstream infections (BSIs) have been investigated as a direct diagnostic target in HIV-infected populations. Wu et al. developed an AI-based diagnostic model aimed at improving early detection of BSI among people living with HIV, reporting strong performance across both internal and external validation datasets. This suggests potential utility for early triage in inpatient and emergency care settings where rapid decision-making is critical (Wu et al., 2023).

More recently, research has shifted toward generalized opportunistic infection detection rather than pathogen-specific models. Chen et al. proposed a multicenter ML diagnostic framework that uses only a small set of clinical features to identify the presence of any opportunistic infection in HIV-infected patients. This approach is particularly valuable in scenarios where the exact causative pathogen is not yet identified, supporting rapid screening and early clinical response (Chen et al., 2025).Other opportunistic infections have also been explored in emerging ML studies. For instance, nontuberculous mycobacterial (NTM) disease has been targeted using ML-based diagnostic models that leverage routine inflammatory biomarkers such as white blood cell count, neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR), C-reactive protein-to-albumin ratio (CAR), and combined lymphocyte index (CALLY), particularly after tuberculosis has been excluded (Li et al., 2025). Similarly, Ichwan et al. investigated low-cost ML screening approaches for toxoplasmosis risk using questionnaire-based data, highlighting the feasibility of non-laboratory-based early screening tools in resource-limited settings (Ichwan et al., 2025).current evidence indicates that ML-based direct OI research in HIV is evolving along two complementary directions. The first focuses on pathogen-specific predictive models for infections such as TB, BSI, NTM disease, and toxoplasmosis, while the second emphasizes generalized screening models capable of detecting the presence of any opportunistic infection prior to confirmatory diagnosis. Across both directions, the central clinical objective remains early detection, enabling timely confirmatory testing, preventive treatment initiation, and improved care escalation for people living with HIV (Orwa et al., 2023; Wu et al., 2023; Chen et al., 2024; Chen et al., 2025; Li et al., 2025; Ichwan et al., 2025).

Data Sources and Predictors in Machine Learning Models for Opportunistic Infection Risk in HIV

Routine clinical data from standard HIV care remains the primary data source for most machine learning (ML) models in this domain. A consistent approach across studies is the use of information collected at first presentation or during routine follow-up visits, rather than relying on specialized or high-cost investigations. For instance, Bartl et al. (2025) developed an incident tuberculosis (TB) prediction model using clinical variables recorded at the initial HIV consultation, which was then used to predict active TB occurring at least six months later, thereby creating a window for preventive intervention.Core HIV disease markers are among the most frequently used predictive variables. Across the broader HIV ML literature, viral load and CD4 cell count consistently appear as key features for identifying patients at risk of rapid disease progression and opportunistic infections such as TB and pneumonia. These markers are central because they directly reflect immune status and treatment response (Sah et al., 2025).In addition, co-infections and acute clinical complications serve as important predictive signals, particularly in inpatient settings. Studies have shown that conditions such as Candida infection, tuberculosis, pneumonia, and hypoproteinemia frequently occur among hospitalized HIV patients and can significantly enhance the predictive performance of ML models when included as features, as they reflect both immune suppression and disease severity (Huang et al., 2023).

Beyond general HIV markers, several studies have incorporated inflammatory and hematologic indicators for specific opportunistic infections. For example, in early diagnosis of nontuberculous mycobacterial (NTM) disease, features such as white blood cell count (WBC), neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR), C-reactive protein-to-albumin ratio (CAR), and combined lymphocyte-based indices

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

(CALLY) have been shown to improve predictive accuracy, particularly after excluding tuberculosis cases (Li et al., 2025).

Demographic and general clinical variables are also widely used, especially in infection-risk prediction beyond traditional HIV biomarkers. For instance, a machine learning model predicting multidrug-resistant Enterobacterales infection in people living with HIV incorporated demographic and clinical predictors to estimate individual risk, demonstrating the broader applicability of ML-based risk modeling in this population (Henderson et al., 2022).

In low-resource contexts, questionnaire-based inputs have also been explored as cost-effective predictive features. Ichwan et al. (2025) demonstrated that ML models based solely on self-reported questionnaire data could support early screening for Toxoplasma gondii infection, highlighting the potential for non-laboratory-based screening tools where diagnostic resources are limited.More generally, ML in healthcare can integrate diverse data modalities, including structured clinical records, laboratory results, imaging data, and even unstructured clinical notes. However, current HIV-related OI studies predominantly rely on structured clinical and laboratory datasets, with limited use of multimodal or unstructured data sources (Ngiam et al., 2019; Scott, 2021; Bartl et al., 2025). The predictors used in ML models for opportunistic infections in HIV can be grouped into five main categories:

a. **HIV severity and immune status:** viral load, CD4 count (Sah et al., 2025)

b. **Co-infections and clinical complications:** TB, pneumonia, Candida infection (Huang et al., 2023; Sah et al., 2025)

c. **Nutritional and biochemical status:** hypoproteinemia, albumin-based ratios (Huang et al., 2023; Li et al., 2025)

d. **Inflammatory and hematologic markers:** WBC, NLR, PLR, CAR, CALLY (Li et al., 2025)

e. **Demographic and questionnaire-based variables:** age, clinical history, self-reported data (Henderson et al., 2022; Ichwan et al., 2025)

Most predictive feature sets in this field are deliberately practical and cost-effective, focusing on routinely available clinical and laboratory data. The overarching goal is to enable early identification of individuals at risk of opportunistic infections, thereby supporting timely monitoring, confirmatory testing, and preventive intervention strategies (Bartl et al., 2025; Sah et al., 2025)

Machine Learning Methods and Model Development Approaches in HIV-Related Opportunistic Infection Prediction

Machine learning (ML) methods used in HIV-related opportunistic infection (OI) research range from traditional statistical forecasting models to advanced supervised learning algorithms. At the population level, forecasting techniques such as ARIMA have been applied to predict long-term OI trends and support public health planning (Rubaihayo et al., 2016). At the patient level, most studies employ classification and risk prediction models based on demographic and clinical data, with logistic regression and regularized methods such as Lasso, Ridge, and Elastic Net remaining popular because of their balance between interpretability and predictive performance (Orwa et al., 2023).Recent studies increasingly focus on early prediction and intervention, including TB co-infection risk models, incident TB prediction systems, and generalized OI screening tools designed to support timely clinical decision-making (Chen et al., 2024; Bartl et al., 2025; Chen et al., 2025). While simpler models remain common, more advanced approaches such as Random Forest, Support Vector Machine (SVM), XGBoost, LightGBM, and CatBoost have demonstrated strong performance in OI prognosis and diagnosis tasks, particularly for conditions such as HIV-associated cryptococcosis (Zhan et al., 2024; Gilliland et al., 2025).

Model evaluation typically relies on sensitivity, specificity, area under the curve (AUC), and F1-score, with recent studies increasingly incorporating explainability techniques such as SHAP to identify important predictors

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

and improve clinical interpretability (Zhan et al., 2024). However, validation practices remain inconsistent, as some models undergo multicenter or external validation while others are tested only on retrospective single-center datasets, raising concerns about generalizability (Wu et al., 2023; Chen et al., 2025). The methodological landscape combines traditional statistical approaches with modern ML techniques. Logistic and penalized regression models remain widely used, while ensemble and kernel-based methods are increasingly adopted where higher predictive accuracy is required. Nevertheless, model performance remains heavily dependent on data quality, feature selection, preprocessing, and rigorous validation procedures (Ngiam et al., 2019; Scott, 2021; Bartl et al., 2025).

METHODOLOGICAL WORKFLOW

This image presents a streamlined, end-to-end workflow for the machine learning-based early prediction of opportunistic infections (OIs) among people living with HIV (PLWH), as described in Figure 1. The process begins with the Knowledge Base, which serves as the primary data source, comprising diverse patient information such as demographics, medical history, laboratory results, and symptoms. This raw data then undergoes a Data Preprocessing phase, where it is cleaned to remove inconsistencies, reduced to eliminate irrelevant features, and transformed into a format suitable for analysis. The preprocessed data is then fed into the Predictive Modeling phase, where the XGBoost algorithm is employed to learn patterns and build a classification model capable of distinguishing between patients who will develop OIs and those who will not. Finally, the model's performance is rigorously evaluated in the Model Evaluation phase using key metrics like accuracy, precision, recall, and F1-scoreto ensure its reliability and clinical validity before potential deployment in healthcare settings.

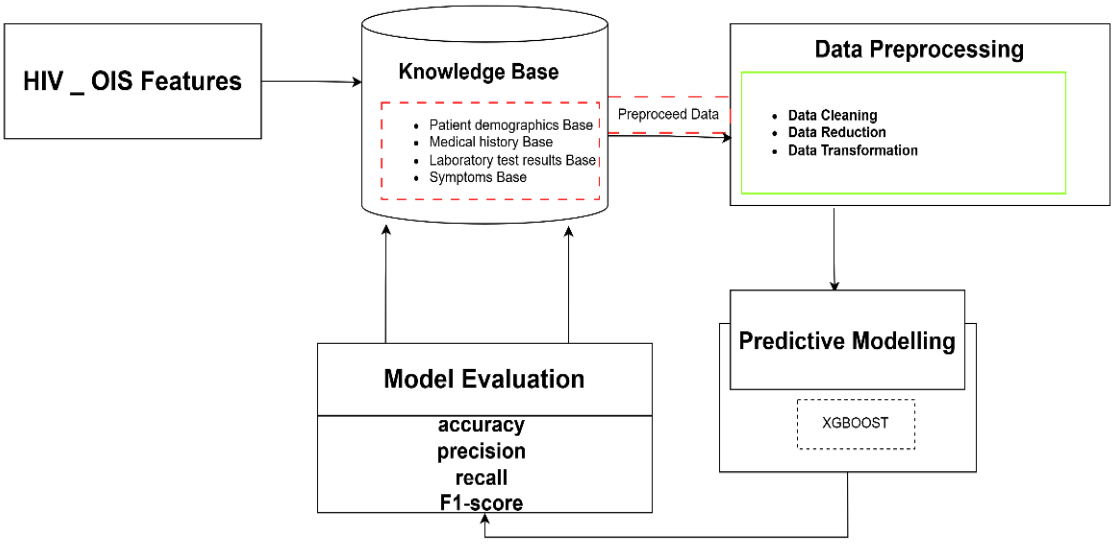


Figure 1: Architecture for Early Detection and Monitoring of HIV Induced Opportunistic Infections

Experimentation and Analysis

The experimentation and analysis were conducted in three phases: Data Collection and Feature Distribution, Modeling Phase.

Data Collection

Table 1 presents a summary of the *Opportunistic Infections.csv* dataset employed for the early prediction and monitoring of HIV-associated opportunistic infections (OIs). The dataset contains 3,982 patient records and 88 variables, providing a comprehensive collection of demographics, clinical, laboratory, and treatment-related information. Of the total features, 35 are categorical variables, representing attributes such as sex, age category, and treatment regimen type; 45 are numerical variables, including indicators such as CD4 count, viral load, and body mass index (BMI); while 8 are text-based variables, containing descriptive information such as clinical notes and healthcare facility details. The diversity of data types ensures a rich and balanced representation of

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

patient information, enabling both quantitative and qualitative analyses. The numerical and categorical features facilitate statistical modeling and machine learning-based prediction, whereas the text variables provide additional contextual information that may enhance clinical interpretation. Consequently, the dataset serves as a valuable resource for developing computational intelligence models aimed at the early detection, monitoring, and management of HIV-associated opportunistic infections, thereby supporting more timely and informed clinical decision-making.

Table 4.1: Description Statistics for Early Detection and Monitoring of HIV Induced Opportunistic Infections

Dataset Name	**Dataset Size**		**Features**		**Meta**
	Rows	**Columns**	**Categorical**	**Numeric**	**Text**
Opportunistic Infections**.csv**	3982	88	35	45	8

FeatureDistribution

Figure 2 shows the Feature Correlation with OI Bar Chart identifies which variables most influence the occurrence of opportunistic infections. Features such as Months of ARV Refill and Number of Fingers Captured exhibit moderate positive correlations (~0.2), suggesting that longer refill intervals and consistent biometric captures may be linked to OI monitoring frequency. Meanwhile, Current Weight (kg) and Number of EAC Sessions Completed show weaker positive correlations, implying minor contributions to OI status. The minimal impact of Age and Number of Fingers Recaptured indicates limited predictive weight. Overall, this chart helps in feature prioritization for predictive model development and clinical decision-making.

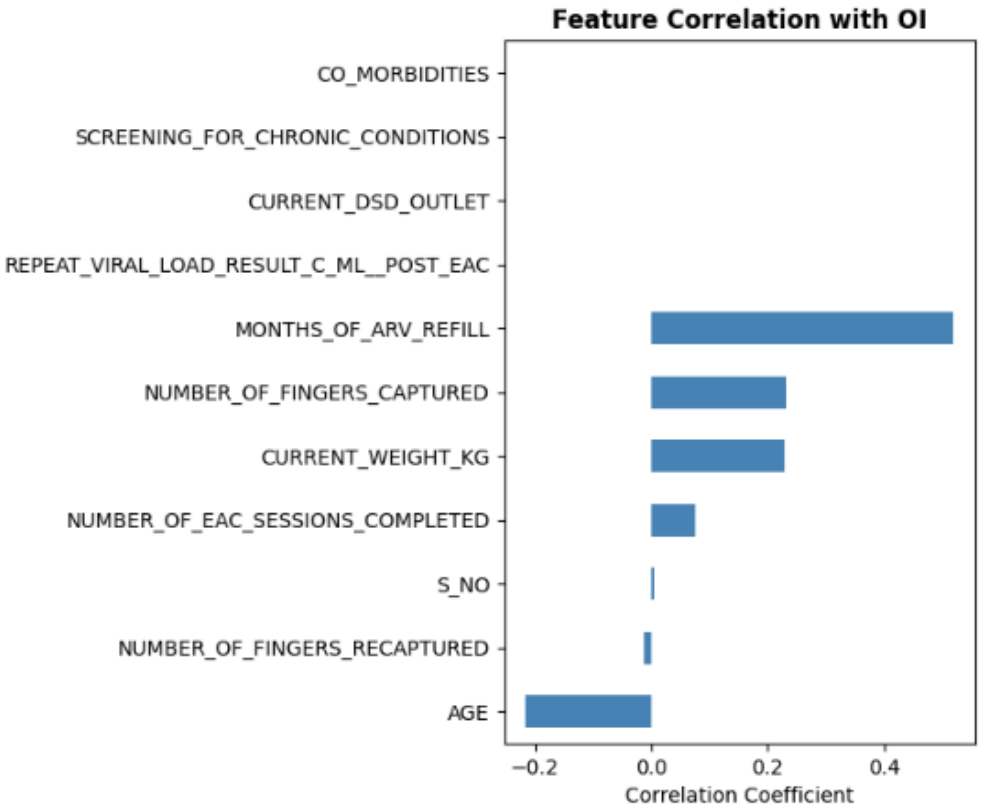


Figure 2: Feature Correlation with OI

Figure 3 provides an in-depth quantitative evaluation of relationships between various clinical and behavioral features related to HIV viral load suppression. The analysis shows a strong negative correlation (-0.28) between Clinical_Risk_Score and suppression, indicating that higher risk levels correspond to poorer outcomes and greater OI susceptibility. Similarly, Age exhibits a moderate negative correlation (-0.10), reinforcing that older individuals tend to have lower suppression rates. On the other hand, features such as CD4_Count_Numeric,

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Regimen_Line_Complexity, and WHO_Stage show slight positive associations, suggesting minor supportive roles in achieving viral control. This matrix is critical for identifying risk profiles and influential predictors to guide targeted OI monitoring.

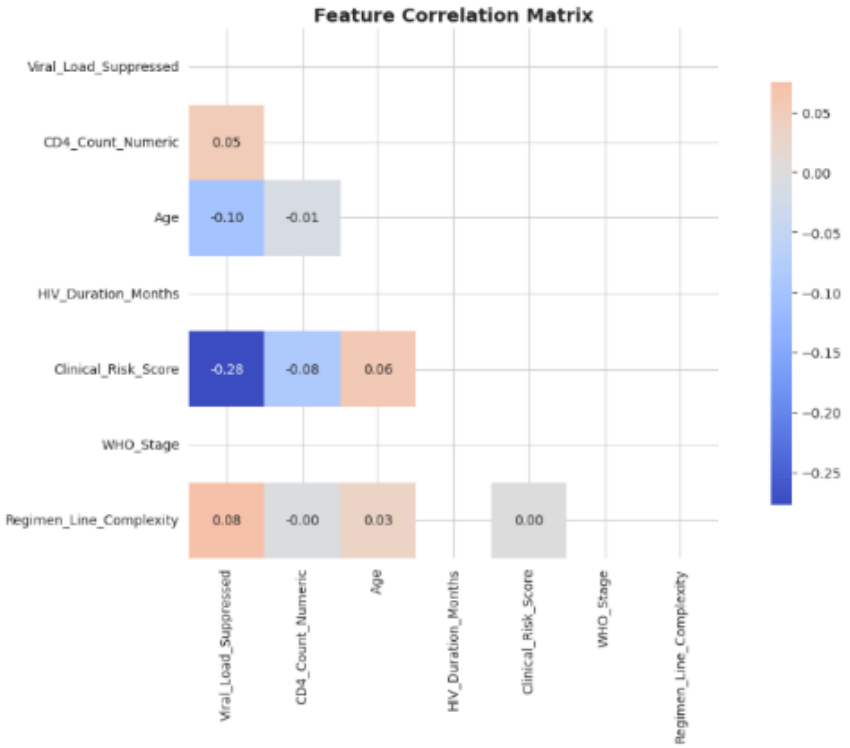


Figure 3: Feature Correlation Matrix

Data Preprocessing

The results of the data preprocessing stage, where Principal Component Analysis (PCA) was applied to identify the most influential features in the dataset for early detection and monitoring of HIV-induced opportunistic infections (OIs) as in Table 2. Four principal components were identified as the major contributors to data variability. Current Weight (PC1) explained 18.87% of the variance (eigenvalue = 1.3208), followed by Age (PC2) with 16.90% (eigenvalue = 1.1793). Months of ARV Refill (PC3) contributed 15.31% of the variance (eigenvalue = 1.0717), while Number of EAC Sessions Completed (PC4) accounted for 14.90% (eigenvalue = 1.0433). These findings indicate that both clinical factors (weight and age) and treatment-related behaviors (ARV refill adherence and EAC participation) are important predictors of HIV-related opportunistic infections. Collectively, the identified components capture a substantial proportion of the dataset's variability and provide valuable inputs for predictive modeling and risk assessment.

Table 2: Data Preprocessing for HIV-induced opportunistic infections

S/N	Feature	Principal Component	Eigenvalue	Explained Variance Ratio
1	Current Weight	PC1	1.3208	0.1887
2	Age	PC2	1.1793	0.1685
3	Months of ARV Refill	PC3	1.0717	0.1531
4	Number of EAC Sessions Completed	PC4	1.0433	0.1490

Figure 4 illustrates the variance explained by each principal component in the dimensionality reduction process for OI prediction. The plot shows a sharp drop in eigenvalues from 1.30 (PC1) to 1.05 (PC4), with an “elbow”

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

occurring after the second component (around 1.20). This suggests that the first two principal components capture most of the dataset's variability, making them sufficient for reducing dimensionality while retaining key information. Overall, the figure supports improved computational efficiency and better interpretability in modeling OI-related outcomes.

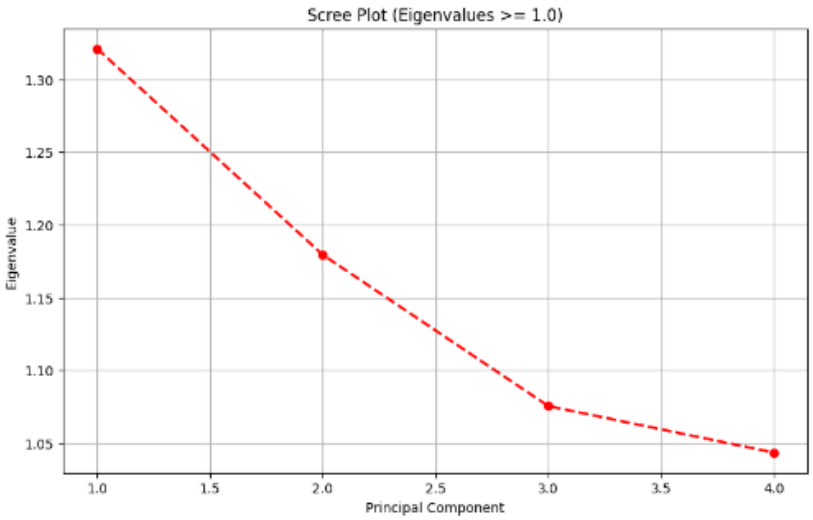


Figure 4: Scree plot for OI

Modeling Phase

The risk tier distribution presented in Figure 8 provides a comprehensive overview of how the hybrid machine learning framework stratifies the study population into distinct clinical risk categories for HIV-associated opportunistic infections. The bar chart reveals the count of patients assigned to each of the three risk tiers; Low, Medium, and High based on the model's predictive outputs. This stratification is essential for translating the model's probabilistic predictions into actionable clinical categories that can guide screening intensity and intervention strategies. A balanced distribution across the three tiers would indicate that the model effectively differentiates the entire spectrum of risk, while a skewed distribution might suggest that the model is overly conservative or aggressive in its risk assignments. For instance, if a substantial proportion of patients are classified into the Low-risk tier, this would imply that the model confidently rules out opportunistic infections in a large segment of the population, thereby reducing unnecessary testing and healthcare burden. Conversely, a notable concentration in the High-risk tier would signal that the model identifies a significant subset of patients requiring immediate diagnostic follow-up and prophylactic interventions. Understanding this distribution is critical for assessing the real-world applicability of the framework, as it directly influences clinical workflow, resource allocation, and patient triage in HIV care settings, particularly in resource-constrained environments where targeted screening is paramount.

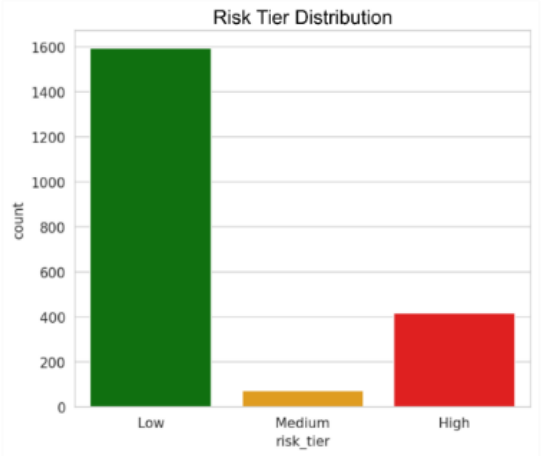


Figure 4: Risk Tier Distribution

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

RESULTS AND DISCUSSION

The study implemented a structured machine learning pipeline comprising data preprocessing, feature analysis, dimensionality reduction, model training, and evaluation using a dataset of 3,982 records and 88 features that included clinical, demographic, laboratory, and treatment adherence variables. Exploratory analysis and correlation studies identified clinical risk score, CD4 count, age, ART adherence behavior, and weight as key predictors of opportunistic infections, while Principal Component Analysis (PCA) further reduced dimensionality and revealed four major components, with current weight (PC1) explaining 18.87% of the variance, followed by age (16.90%), months of ARV refill (15.31%), and number of EAC sessions completed (14.90%), indicating that both physiological and behavioral factors significantly influence infection risk. Among all models evaluated, XGBoost demonstrated superior predictive performance, achieving an accuracy of 97.17%, precision of 0.9384, recall of 0.9278, F1-score of 0.9330 for the positive class, an AUC of 0.994, and an average precision of 0.979. The confusion matrix further confirmed strong classification capability with a very low false-negative rate of 27 cases, which is particularly important in clinical settings where missed infections may lead to severe disease progression. In addition, the precision–recall curve showed consistently high performance across thresholds, indicating stable behavior under class imbalance, while the ROC curve demonstrated near-perfect separability between infected and non-infected cases, confirming strong generalization ability. Overall, the results indicate that machine learning models can effectively integrate heterogeneous clinical data to identify patients at risk of opportunistic infections, and the proposed model demonstrates strong potential as a clinical decision-support tool for early screening, risk stratification, and prioritization of high-risk patients for further diagnostic evaluation.

Conclusion

The effectiveness of a machine learning-based approach for the early prediction of opportunistic infections among people living with HIV. Using a large and diverse clinical dataset, the proposed framework successfully identified key predictors and achieved high predictive performance using the XGBoost algorithm.The model showed excellent accuracy, high sensitivity, and strong discriminative ability, with minimal false-negative predictions, making it suitable for clinical risk assessment. Feature analysis further highlighted the importance of CD4 count, age, ART adherence behavior, weight, and treatment engagement in predicting opportunistic infection risk.

The findings support the integration of machine learning techniques into HIV care systems to enhance early detection, improve patient monitoring, and support timely clinical interventions. Future work should focus on external validation across multiple healthcare settings and the integration of real-time clinical data to improve generalizability and practical deployment.

REFERENCES

1. Ahlström, M. G., Ronit, A., Omland, L., Vedel, S., &Obel, N. (2019). Algorithmic prediction of HIV status using nation-wide electronic registry data. *EClinicalMedicine*.
2. Bartl, L., Zeeb, M., Kälin, M., Loosli, T., Notter, J., Furrer, H., & Nemeth, J. (2025). Machine Learning-based Prediction of Active Tuberculosis in People with HIV Using Clinical Data. *Clinical Infectious Diseases*.
3. Cai, Q., Yang, L., Ling, Y., Pan, W., Zhong, Q., Wang, C., & Pan, X. (2025). Survival prediction models for people living with HIV based on four machine learning models. *Scientific Reports*.
4. Chen, H., Chen, F., Wang, Y., Cai, E., Pan, W., Li, Y., &Su, F. (2025). A Machine Learning Model for Diagnosing Opportunistic Infections in HIV Patients: Broad Applicability Across Infection Types. *Journal of Cellular and Molecular Medicine*.
5. Chen, J., Liu, L., Huang, J., Jiang, Y.-J., Yin, C., Zhang, L., & Lu, H. (2024). LSTM-Based Prediction Model for Tuberculosis Among HIV-Infected Patients Using Structured Electronic Medical Records: A Retrospective Machine Learning Study. *Journal of Multidisciplinary Healthcare*.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

6. Chen, Y., Pan, K., Lu, X., Maimaiti, E., &Wubuli, M. (2025). Machine learning-based prediction of mortality risk in AIDS patients with comorbid common AIDS-related diseases or symptoms. *Frontiers in Public Health*.
7. Emegano, D., Duwa, B., Usman, A. G., Ahmad, H., Ozsahin, D., Askar, S. S., & America, A. (2025). A comparative study on TB incidence and HIVTB coinfection using machine learning models on WHO global TB dataset. *Scientific Reports*.
8. Gilliland, S., &Mbuagbaw, L. (2025). The Applications of Machine Learning to Prognostic Modeling for Human Immunodeficiency Virus: A Protocol for a Scoping Review. *Health Science Reports*.
9. Girma, D., Dejene, H., Geleta, L. A., Tesema, M., &Bati, F. (2022). Time to occurrence, predictors, and patterns of opportunistic infections incidence among HIV-positive patients attending Antiretroviral Therapy Clinic of Salale University Comprehensive Specialized Hospital: A retrospective cohort study. *Medicine*.
10. Gruber, S., Krakower, D., Menchaca, J. T., Hsu, K., Hawrusik, R., Maro, J. C., &Klompas, M. (2020). Using electronic health records to identify candidates for human immunodeficiency virus pre-exposure prophylaxis: An application of super learning to risk prediction when the outcome is rare. *Statistics in Medicine*.
11. Henderson, H., Napravnik, S., Kosorok, M., Gower, E. W., Kinlaw, A. C., Aiello, A., & van Duin, D. (2022). Predicting Risk of Multidrug-Resistant Enterobacterales Infections Among People With HIV. *Open Forum Infectious Diseases*.
12. Huang, L., Xie, B., Zhang, K., Xu, Y., Su, L., Lv, Y., &Lv, J. (2023). Prediction of the risk of cytopenia in hospitalized HIV/AIDS patients using machine learning methods based on electronic medical records. *Frontiers in Public Health*.
13. Ichwan, W. W., Tandra, P. A., Qomariyah, N. N., Meirawan, R. F., &Setiawan, F. (2025). Machine Learning-Based Prediction of Seropositive Toxoplasma Gondii from Lifestyle Factors. *International Conferences on Information Science and System*.
14. Ijaiya, M. A., Troncoso, E., Mutloatse, M., Ifeanyi, D., Obasa, B., Emerenini, F., &Curran, K. (2025). Use of machine learning in predicting continuity of HIV treatment in selected Nigerian States. *PLOS Global Public Health*.
15. Kurniawan, A., Marthabakti, C., Putri, L. M., Suyono, B. C., &Alisiah, R. A. (2026). Performance Evaluation of Machine Learning Algorithms for AIDS-Infected Patient Classification. *JurnalKesehatanVokasional*.
16. Li, L., Shi, C., Liu, X., Li, W., Luo, Y., Zhang, H., & Shen, L. (2025). A diagnostic model for NTM disease in HIV-positive patients: a machine learning-based analysis with novel inflammatory markers. *Frontiers in Immunology*.
17. Mamo, D., Yilma, T., Fekadie, M., Sebastian, Y., Bizuayehu, T., Melaku, M., &Walle, A. D. (2023). Machine learning to predict virological failure among HIV patients on antiretroviral therapy in the University of Gondar Comprehensive and Specialized Hospital, in Amhara Region, Ethiopia, 2022. *BMC Medical Informatics and Decision Making*.
18. Marcus, J., Hurley, L., Krakower, D., Alexeeff, S. E., Silverberg, M., & Volk, J. E. (2019). Use of electronic health record data and machine learning to identify potential candidates for HIV preexposure prophylaxis: a modelling study. *The Lancet HIV*.
19. Masur, H., Brooks, J., Benson, C., Holmes, K., Pau, A., & Kaplan, J. (2014). Prevention and treatment of opportunistic infections in HIV-infected adults and adolescents: Updated Guidelines from the Centers for Disease Control and Prevention, National Institutes of Health, and HIV Medicine Association of the Infectious Diseases Society of America. *Clinical Infectious Diseases*.
20. Ngiam, K., & Khor, I. W. (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*.
21. Orwa, J., Oduor, P., Okelloh, D., Gethi, D., Agaya, J., Okumu, A., &Wandiga, S. (2023). Comparison of logistic regression with regularized machine learning methods for the prediction of tuberculosis disease in people living with HIV: cross-sectional hospital- based study in Kisumu County, Kenya. *medRxiv*.
22. Rubaihayo, J., Tumwesigye, N., Konde-Lule, J., &Makumbi, F. (2016). Forecast analysis of any opportunistic infection among HIV positive individuals on antiretroviral therapy in Uganda. *BMC Public Health*.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

23. Russom, S., Kollias, D., & Zhang, Q. (2025). Proactive HIV Care: AI-Based Comorbidity Prediction from Routine EHR Data. *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*.
24. Sah, A., Elshaikh, R., Shalabi, M., Abbas, A. M., Prabhakar, P. K., Babker, A. M., &Agarwal, S. (2025). Role of Artificial Intelligence and Personalized Medicine in Enhancing HIV Management and Treatment Outcomes. *Life*.
25. Scott, I. (2021). Demystifying machine learning: a primer for physicians. *Internal Medicine Journal (Print)*.
26. Singh, Y. (2017). Machine Learning to Improve the Effectiveness of ANRS in Predicting HIV Drug Resistance. *Healthcare Informatics Research*.
27. Sui, Q., Li, G., Peng, Y., Zhang, J., Zhang, Y., & Zhao, R. (2025). Scalable and robust machine learning framework for HIV classification using clinical and laboratory data. *Scientific Reports*.
28. Wu, L., Xia, D., & Xu, K. (2023). Multi-Clinical Factors Combined with an Artificial Intelligence Algorithm Diagnosis Model for HIV-Infected People with Bloodstream Infection. *Infection and Drug Resistance*.
29. Xu, X.-H., Ge, Z., Chow, E., Yu, Z., Lee, D. M., Wu, J., & Zhang, L. (2022). A Machine-Learning-Based Risk-Prediction Tool for HIV and Sexually Transmitted Infections Acquisition over the Next 12 Months. *Journal of Clinical Medicine*.
30. Yang, X.-D., Su, B., Zhang, X., Liu, Y., Wu, H., & Zhang, T. (2020). Incomplete immune reconstitution in HIV/AIDS patients on antiretroviral therapy: Challenges of immunological non-responders. *Journal of Leukocyte Biology*.
31. Zhan, B., Wei, W., Xie, Z., Meng, S., Bao, X., He, X., & Liang, H. (2024). Machine learning-based prognostic prediction for hospitalized HIV/AIDS patients with cryptococcus infection in Guangxi, China. *BMC Infectious Diseases*.
32. Zhou, J., Yang, Y., Xie, Z., Lu, D., Huang, J., Lan, L., & Huang, J. (2023). Dysbiosis of gut microbiota and metabolites during AIDS: implications for CD4+ T cell reduction and immune activation. *AIDS (London)*.