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

A Hybrid Machine Learning Framework for Early Detection of Hiv-Associated Opportunistic Infections

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.2026.150600129

Received: 26 June 2026; Accepted: 01 July 2026; Published: 17 July 2026

ABSTRACT

The increasing burden of OIs among people living with HIV necessitates accurate and timely predictive systems to support clinical decision-making. The framework leverages clinical, demographic, and treatment-related variables extracted from a real-world dataset obtained from the University of Uyo Teaching Hospital. Data preprocessing techniques, including normalization, feature engineering, and dimensionality reduction were applied to enhance model performance. The proposed model was evaluated using multiple performance metrics, including accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix analysis. Experimental results show that the best-performing configuration achieved an accuracy of 96.79% with strong sensitivity and specificity, demonstrating its effectiveness in distinguishing between infected and non-infected cases. Furthermore, the model demonstrated the use of the ReLU-based activation function. SHAP analysis further enhanced model interpretability by identifying key predictive features such as CD4 count, treatment type, and age. Overall, the results confirm that the proposed hybrid framework is a reliable and clinically meaningful tool for early detection and risk stratification of HIV-associated opportunistic infections.

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

INTRODUCTION

Opportunistic infections (OIs) remain a leading cause of morbidity and mortality among people living with HIV, making early detection of infection risk and disease onset a critical clinical priority (Xu et al., 2022; Zhou et al., 2019). The progression of HIV-related complications and OI development is often rapid and difficult to predict using conventional clinical assessment alone, creating a strong need for effective risk-stratification frameworks that can identify high-risk patients early and support timely clinical intervention (Fu et al., 2025). Recent studies have shown that Artificial Intelligence (AI) and Machine Learning (ML) techniques can leverage key HIV-related indicators such as viral load, CD4 count, and co-infection status to predict disease progression and forecast opportunistic infections including tuberculosis and pneumonia, thereby enabling more proactive, preventive, and personalized care (Sah et al., 2025).

The adoption of ML-based approaches is particularly valuable because they can process large, complex clinical datasets, integrate heterogeneous data sources, and uncover hidden patterns without strict statistical assumptions (Fu et al., 2025; Waring et al., 2020). In particular, Deep Neural Networks (DNNs) require careful optimization of parameters such as activation functions, hidden layers, learning rates, and batch sizes to effectively learn meaningful patterns from data (Inyang et al., 2025). Despite these advantages, diagnosing OIs in HIV/AIDS patients remains challenging due to limitations in conventional approaches such as culture-based methods, targeted molecular assays, and invasive diagnostic procedures. Although emerging techniques like next-generation sequencing (NGS) and metagenomic sequencing provide faster and more comprehensive pathogen

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

detection, the most effective diagnostic strategy combines these advanced tools with traditional microbiological methods rather than relying on either alone (Zhou et al., 2019; Xu et al., 2022).

Therefore, an effective early-detection framework for HIV-associated opportunistic infections should not only identify patients at risk of rapid clinical deterioration but also support the timely identification of likely infectious causes, thereby improving triage, treatment selection, and healthcare resource allocation (Fu et al., 2025; Sun et al., 2025). In line with this objective, this study proposes a hybrid modeling approach that integrates DNNs and XGBoost for predictive analytics. While DNNs are effective in capturing complex nonlinear relationships within high-dimensional clinical data, XGBoost enhances performance through efficient feature selection and robust classification of structured datasets, leading to improved overall prediction accuracy for opportunistic infections in HIV patients. The study utilizes the Opportunistic Infections dataset obtained from the University of Uyo Teaching Hospital for the early detection and monitoring of HIV-induced opportunistic infections (OIs).

Related Works

Hybrid Framework for Early Detection of HIV-Associated Opportunistic Infections

A hybrid framework for the early detection of HIV-associated opportunistic infections (OIs) can be structured into four interconnected components: data preparation, multimodal prediction, rule-guided clinical reasoning, and explainable decision support. The data preparation stage integrates structured clinical variables, including CD4 count, viral load, symptoms, treatment history, and co-infection indicators, with longitudinal electronic health record (EHR) data. This approach is supported by hybrid deep learning models that combine Multilayer Perceptrons (MLPs) and Long Short-Term Memory (LSTM) networks, enabling the effective extraction of both static and temporal patient information for improved disease prediction and monitoring of evolving OI risk (Chen et al., 2024). To enhance clinical applicability, a rule-aware reasoning layer can be incorporated using fuzzy rule-based systems. Such systems combine Type-2 Fuzzy Logic with deep learning techniques to provide transparent decision rules for early referral and treatment intervention, particularly when signs of treatment failure emerge (Ekpenyong et al., 2019).

At the predictive level, the framework benefits from an ensemble-based architecture rather than relying on a single learning model. Multiple machine learning algorithms can be combined to specialize in different data patterns, with their outputs aggregated into a unified risk score to improve predictive accuracy and robustness (Qasem, 2024). This approach can be further strengthened through feature selection and data-balancing techniques, which address the high dimensionality and class imbalance commonly found in HIV and OI datasets (Sa'adah et al., 2026). Additionally, the framework may incorporate molecular data sources, such as transcriptomic and immune-cell-specific biomarkers, to support biomarker discovery and improve prediction in complex clinical cases (Mahmud et al., 2025). Consequently, an effective early-detection framework should integrate clinical and temporal data analysis, ensemble learning, fuzzy reasoning, and interpretable artificial intelligence to support accurate risk prediction, timely intervention, and improved management of HIV-associated opportunistic infections (Ekpenyong et al., 2019; Chen et al., 2024; Qasem, 2024; Mahmud et al., 2025; Sa'adah et al., 2026).

Machine Learning Tasks for HIV-Associated Opportunistic Infection Detection and Prediction

Machine learning (ML) has been applied to a wide range of tasks aimed at improving the detection, diagnosis, and prediction of opportunistic infections (OIs) among people living with HIV. One important application is broad-spectrum OI screening, where models are designed to determine whether an HIV-positive patient is likely to have any opportunistic infection rather than focusing on a specific pathogen. Such approaches support early triage and reduce the likelihood of missed diagnoses by identifying high-risk patients using a small number of clinically relevant features (Chen et al., 2025).Beyond general screening, ML techniques have also been used for pathogen-specific diagnosis. For example, predictive models have been developed for Pneumocystis jirovecii pneumonia (PJP), demonstrating that the integration of routine clinical data with microbiological indicators such as *P. jirovecii* PCR results can improve diagnostic sensitivity while maintaining acceptable specificity (Chew et al., 2024). Similarly, ML models incorporating inflammatory biomarkers have been proposed for the early detection of nontuberculous mycobacterial (NTM) disease in HIV-positive individuals, enabling the

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

identification of less common but clinically significant infections through interpretable prediction frameworks (Li et al., 2025).In addition to diagnostic applications, ML has shown considerable potential for prognostic prediction. Several studies have focused on forecasting the future occurrence of active tuberculosis (TB) among HIV patients using routinely collected clinical data, thereby providing opportunities for preventive intervention before disease onset (Bartl et al., 2025; Ngiam et al., 2019; Scott, 2021). ML has also been employed to model complex interactions between coexisting diseases such as TB and HIV, facilitating a deeper understanding of how multiple clinical conditions jointly influence patient outcomes (Cherezov et al., 2025). Furthermore, predictive models have been developed to estimate the risk of antimicrobial-resistant bacterial colonization or infection among people living with HIV, supporting treatment planning and infection-control strategies (Henderson et al., 2022).

Emerging research has extended these applications to microbiome-based classification, where gut microbial signatures are used as predictive markers of AIDS progression and immune decline. Such approaches demonstrate the potential of integrating molecular and biological data into HIV-related prediction systems (Zhou et al., 2023). Collectively, these studies indicate that an effective hybrid framework for HIV-associated OI management should support multiple interconnected tasks, including broad OI screening, pathogen-specific diagnosis, future disease-risk prediction, and co-morbidity or antimicrobial-resistance assessment. This diversity of predictive objectives highlights the need for a hybrid architecture capable of integrating multiple data sources and learning models rather than relying on a single-purpose classifier (Chen et al., 2025; Chew et al., 2024; Li et al., 2025; Bartl et al., 2025; Henderson et al., 2022; Cherezov et al., 2025).

Interpretability and Clinical Decision Support in HIV-Associated Opportunistic Infection Detection

Interpretability is essential in the early detection of HIV-associated opportunistic infections (OIs) because model predictions often guide clinical decisions such as diagnostic testing, referral, treatment modification, and patient monitoring. To enhance transparency and clinician trust, hybrid frameworks increasingly combine machine learning with rule-based reasoning. For example, integrating deep learning with Type-2 fuzzy rule systems enables proactive detection of treatment failure while providing clear decision thresholds for intervention (Ekpenyong et al., 2019). Explainable AI techniques, such as SHAP, further improve model transparency by identifying the key factors influencing prediction outcomes, including immune suppression, inflammatory markers, symptoms, and co-infection indicators. Such approaches have been successfully applied in HIV-associated NTM prediction, allowing clinicians to better understand and validate model recommendations (Li et al., 2025). Similarly, lightweight screening models can support rapid identification of patients at risk of any OI and trigger further investigation when necessary (Chen et al., 2025). When combined with interpretable molecular biomarkers derived from transcriptomic data, these approaches create a multi-layered decision-support framework that improves early detection, supports targeted diagnosis, and facilitates timely, evidence-based clinical intervention (Mahmud et al., 2025; Chen et al., 2025; Li et al., 2025).

METHODOLOGICAL WORKFLOW

The architecture comprises a structured collection of essential clinical, demographic, and treatment-related variables used to monitor the health status, treatment outcomes, and comorbid conditions of individuals living with HIV/AIDS, as presented in Table 1. It is designed to serve as a comprehensive reference framework for healthcare professionals managing patients undergoing antiretroviral therapy (ART). Each variable is assigned a unique code, data type, and descriptive definition to clearly indicate its clinical significance. The parameters include baseline demographic and identification information, such as Hospital Number, Age, and Sex, as well as key immunological indicators, including CD4 Count and Viral Load, which are critical for assessing immune status and evaluating treatment response. Treatment-related variables, such as the initial ART regimen, current ART regimen, and treatment failure records, are incorporated to monitor regimen modifications and identify potential drug resistance. In addition, clinical monitoring indicators, including WHO clinical stage, adherence level, and duration on ART, provide valuable insights into disease progression, treatment effectiveness, and patient compliance.

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

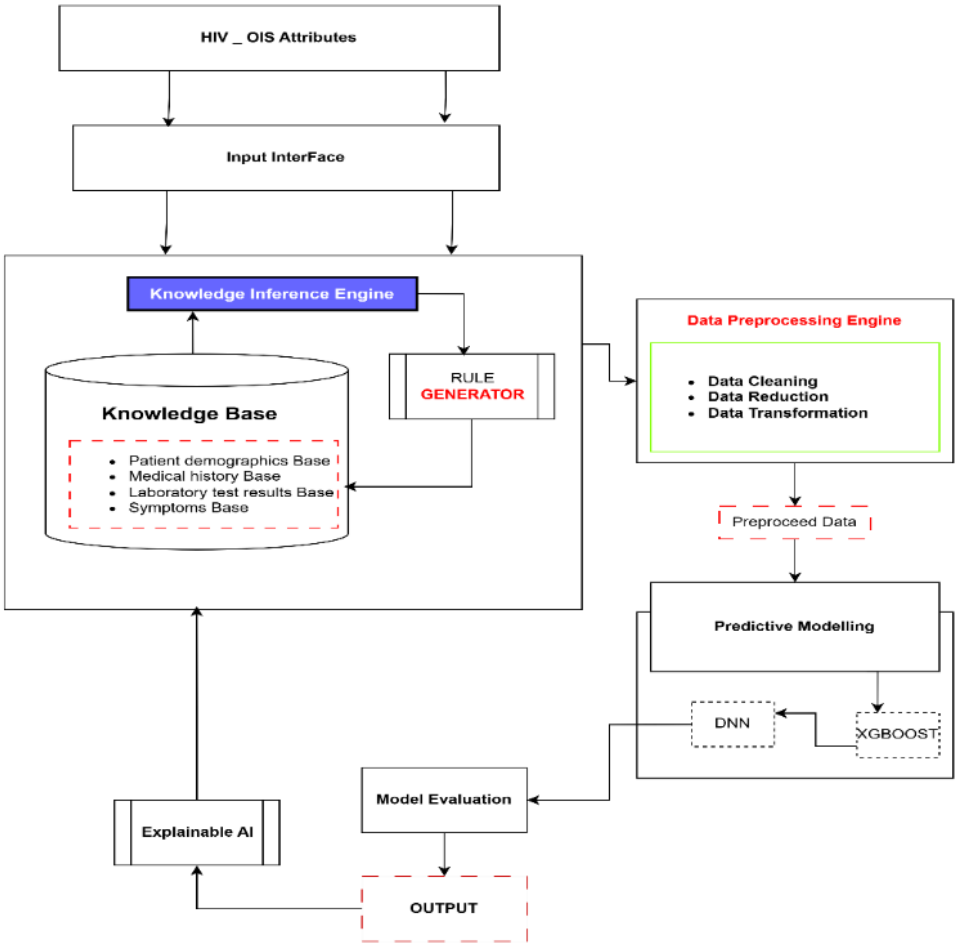


Figure 1: Architecture for Analytics Hybrid Modelling for Early Detection and Monitoring of HIV Induced Opportunistic Infections

Mathematical Formulation of Deep Neural Networks for HIV_OIs Prediction

DNNs are multilayered models capable of learning hierarchical representations from complex clinical and biological datasets. In the context of HIV_OIs prediction, the DNN processes patient features such as CD4 count, viral load, treatment regimen, comorbidities, and demographic variables to detect patterns indicative of opportunistic infections. A typical DNN consists of an input layer, several hidden layers, and an output layer. Each layer transforms its input using a weighted sum followed by a non-linear activation function. The mathematical formulation is as shown in Equations 1 to 7:

1. **Input Layer:**

$$\mathbf{x} = [x_1, x_2, \ldots, x_n]^T \quad (1)$$

where x_i represents a normalized feature of an HIV patient (e.g., CD4 count, duration of ART, weight loss, fever, etc.), and n is the number of features.

2. **Hidden Layers:**
For the l-th layer, the transformation is given by Equation (2):

$$\mathbf{z}^{(l)} = \mathbf{W}^{(l)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)}\mathbf{a}^{(l)} = \sigma(\mathbf{z}^{(l)}) \quad (2)$$

where:

$\mathbf{W}^{(l)}$ is the weight matrix for layer l,

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

$\mathbf{b}^{(l)}$ is the bias vector,

$\mathbf{a}^{(l-1)}$ is the activation from the previous layer (or input $\mathbf{x}$ if $l = 1$),

$\sigma(\cdot)$ is the activation function, typically ReLU: $\sigma(z) = \max(0, z)$

3. **Output Layer:**
The final output (probability of HIV_OI occurrence) is shown in Equation (3) and (4):

$$\hat{y} = \sigma_{out}\left(\mathbf{W}^{(L)}\mathbf{a}^{(L-1)} + \mathbf{b}^{(L)}\right) \quad (3)$$

$$\sigma_{out}(z) = \frac{1}{1+e^{-z}} \quad (4)$$

where:

L is the index of the output layer,

σ_{out} is the sigmoid function for binary classification:This yield$\hat{y} \in (0,1)$, representing the predicted probability of an HIV-related opportunistic infection being present.

4. **Loss Function (Binary Cross-Entropy):**
The loss function quantifies the error between predicted and actual outcomes which is shown in Equation (5):

$$\mathcal{L}(y, \hat{y}) = -[y\log(\hat{y}) + (1-y)\log(1-\hat{y})] \quad (5)$$

where:

$y \in \{0,1\}$ is the true label (1 for HIV_OI, 0 for no infection); $\hat{y}$ is the predicted probability from the network.

5. **Optimization (Gradient Descent with Backpropagation):**
The parameters $\mathbf{W}^{(l)}$ and $\mathbf{b}^{(l)}$ are updated to minimize $\mathcal{L}$Equation (6):

$$\theta \leftarrow \theta - \eta \cdot \nabla_\theta \mathcal{L} \quad (6)$$

where:θ represents all trainable parameters; η is the learning rate; $\nabla_\theta \mathcal{L}$ is the gradient of the loss with respect to the parameters.

6. **Prediction Decision Rule:**
After training, predictions are made using a thresholdEquation (7):

$$\text{Predict HIV_OI} = \begin{cases} 1 & \text{if } \hat{y} \geq 0.5 \\ 0 & \text{if } \hat{y} < 0.5 \end{cases} \quad (7)$$

Mathematical Formulation of XGBoost for HIV-OIS Prediction

XGBoost (Extreme Gradient Boosting) is a scalable and efficient implementation of gradient boosting decision trees. Its goal is to minimize the loss function through additive learning of decision trees, making it suitable for medical prediction tasks such as HIV-related opportunistic infection (HIV-OIS) monitoring, as shown in Equations (8) – (13).

1. **Objective Function**

The objective of XGBoost is to minimize a regularized loss:

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

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega\left(f_k\right) \quad (8)$$

Where:

y_i is the actual HIV_OIS status for patient i

$\hat{y}_i^{(t)}$ is the predicted value at iteration t

l is a differentiable convex loss function (e.g., logistic loss for classification)

$f_k \in \mathcal{F}$ is the k-th decision tree

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2 \quad (9)$$

Equation (9)is the regularization term that penalizes complexity:

Where:

T: number of leaves; w:vector of scores on leaves; γ, λ:regularization parameters

2. **Additive Training Process**

At each boosting round t, a new function $f_t(x)$ is added to minimize the loss shown in Equation (10):

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (10)$$

Using second-order Taylor expansion for efficiency, the approximate objective becomes in Equation (11):

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) \quad (11)$$

Where:

$g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$: first-order gradient

$h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$: second-order gradient

Optimal Leaf Weights

For a given tree structure with T leaves, the optimal weight w_j for leaf j is:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \quad (12)$$

The optimal objective function reduction as shown in Equation (13):

$$\mathcal{L}_{\text{opt}} = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \quad (13)$$

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

Experimentation and Analysis

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

Data Collection

The study utilized the Opportunistic Infectionsdataset obtained from the University of Uyo Teaching Hospital for the early detection and monitoring of HIV-induced opportunistic infections (OIs). The dataset contains 3,982 patient records and 88 variables, comprising 35 categorical features, 45 numerical features, and 8 text-based attributes. The variables capture a wide range of clinical, demographic, and treatment-related information, including CD4 count, viral load, body mass index, treatment regimens, and patient adherence indicators. This diverse combination of quantitative and qualitative data provides a comprehensive representation of patient health status, supporting both statistical analysis and machine learning applications. Consequently, the dataset serves as a robust foundation for developing predictive models aimed at improving the early detection, monitoring, and management of HIV-induced opportunistic infections.

Statistical Distribution

The kurtosis values further reveal the nature of data distribution in terms of peakedness and the presence of extreme observations as in Table 1. Variables such as Current Weight (44.67) and Repeat Viral Load Result (20.68) display very high kurtosis, indicating heavy-tailed distributions with potential outliers, which may reflect patients with unusual body weights or viral load fluctuations. Conversely, Months of ARV Refill (-0.85) and Number of Fingers Captured (-0.19) exhibit negative kurtosis, indicating flatter distributions with fewer extreme cases. The Age variable (0.53) demonstrates near-normal kurtosis, confirming a relatively even spread of age data. However, the Number of Fingers Recaptured shows a strong negative skewness (-4.54) with a high kurtosis (23.82), implying that most entries cluster around higher values with a few extreme low cases. These results suggest that while most variables show moderate variability, others, particularly weight and viral load, require normalization and outlier treatment to ensure accurate performance of the Hybrid predictive models.

Table 1: Skewness and Kurtosis of Patient Data Variables

Variable	Skewness	Kurtosis
Current Weight (kg)	2.359874	44.665707
Age	-0.236635	0.526328
Months of ARV Refill	0.804408	-0.852294
Number of EAC Sessions Completed	1.169557	1.852281
Repeat Viral Load Result (c/ml) – POST EAC	4.461786	20.678513
Number of Fingers Captured	-1.342856	-0.186446
Number of Fingers Recaptured	-4.543142	23.820179

Pairwise relationships between multiple features such as Age_Clean, CD4_Numerical, BMI_Clean, Viral_Load_Clean, and Creatinine_Clean against the binary OI status as shown in Figure 2. Scatter plots reveal the degree of separation between OI and non-OI groups, with CD4_Numerical showing the clearest distinction indicating that lower CD4 counts correlate strongly with OI presence. The diagonal histograms depict individual feature distributions, helping identify skewed variables and feature overlaps. This visualization is instrumental in feature interaction analysis, supporting model feature selection for improving OI prediction accuracy.

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

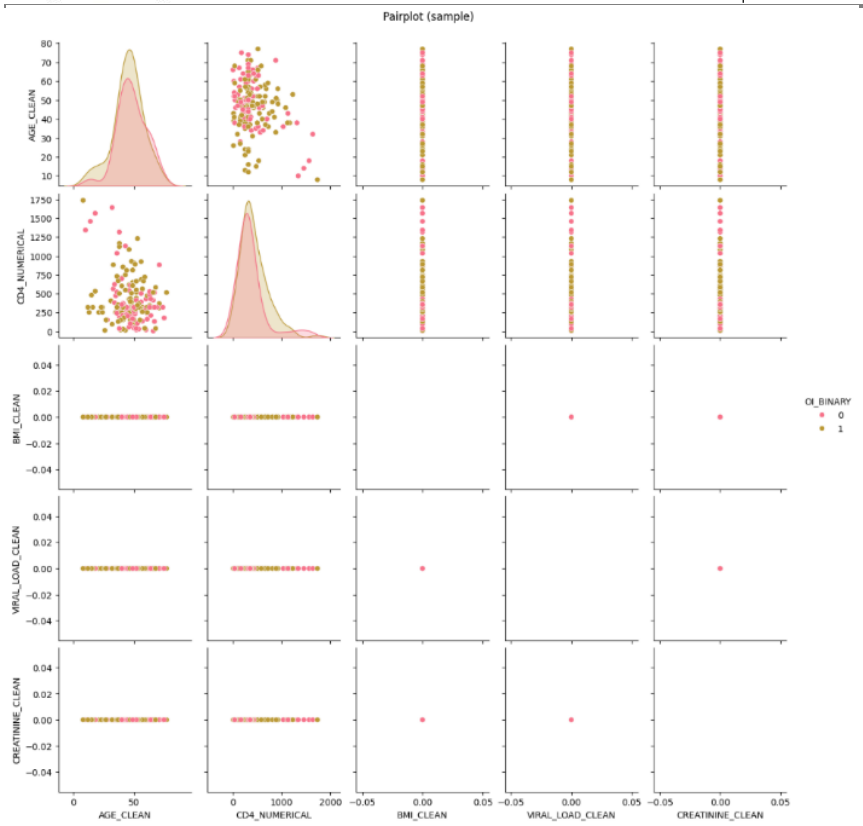


Figure 2: Pair Plot Feature for OI

Data Preprocessing

The Data Preprocessing stage, as shown in Table 2, involved applying Principal Component Analysis (PCA) to identify the most influential features contributing to variability within the dataset used for early detection and monitoring of HIV-induced opportunistic infections (OI). The analysis revealed four key principal components (PCs) with corresponding eigenvalues and explained variance ratios. The first principal component (PC1) was dominated by Current Weight, with an eigenvalue of 1.3208 and an explained variance ratio of 0.1887, indicating that this feature alone accounts for approximately 18.9% of the total data variance. This suggests that variations in patient weight play a major role in influencing health outcomes related to OI risk. The second component (PC2) was associated with Age, which had an eigenvalue of 1.1793 and contributed 16.9% of the variance, highlighting age as a significant factor affecting viral suppression and immune response. The third component (PC3), represented by Months of ARV Refill, had an eigenvalue of 1.0717 and explained 15.3% of the variance, emphasizing the importance of treatment adherence and medication refill patterns in determining OI occurrence. Lastly, the fourth component (PC4) corresponded to the Number of EAC Sessions Completed, with an eigenvalue of 1.0433 and a 14.9% variance contribution, reflecting the influence of counseling engagement on adherence and health outcomes. Overall, these four components capture a substantial portion of the dataset's total variance, showing that both clinical factors (weight, age) and behavioral indicators (ARV refill frequency, EAC sessions) are essential for accurate prediction and monitoring of HIV-induced opportunistic infections.

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

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

3	Months of ARV Refill	PC3	1.0717	0.1531
4	Number of EAC Sessions Completed	PC4	1.0433	0.1490

Figure 3 shows the variance explained by each principal component in a dimensionality reduction analysis for OI prediction. It displays a steep decline from an eigenvalue of 1.30 (first component) to 1.05 (fourth component), forming an "elbow" after the second component (1.20). This indicates that the first two principal components capture most of the dataset's variance, making them ideal for simplifying data while preserving essential information. The plot is critical in optimizing computational efficiency during model training and improving interpretability of features influencing OI occurrence.

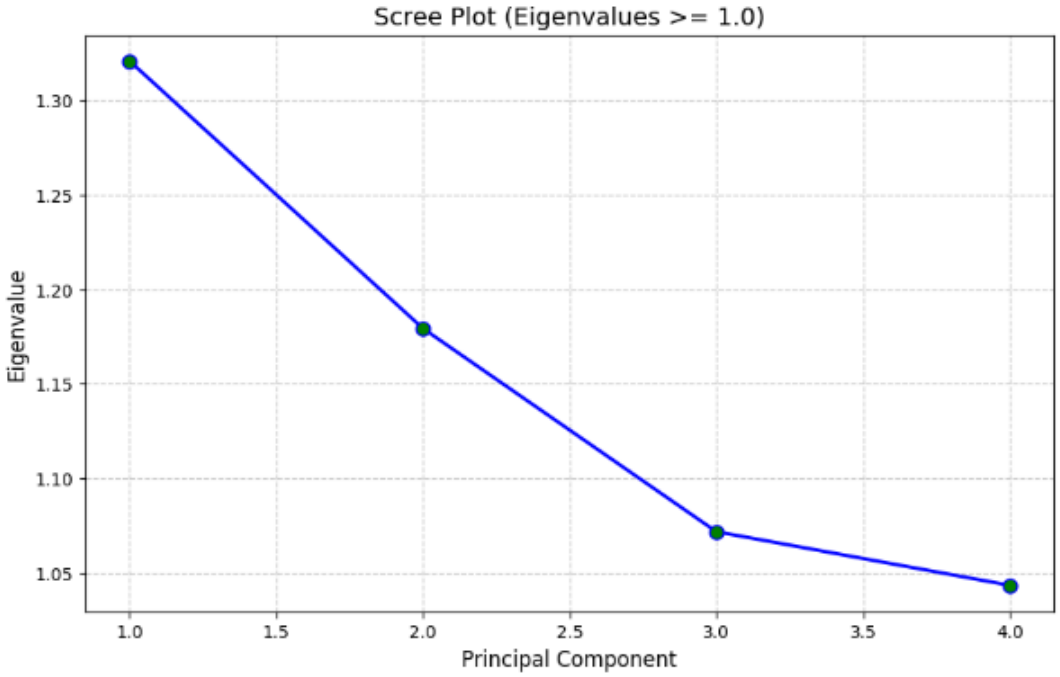


Figure 3: Scree plot for OI

Modeling Phase

Table 3 presents the performance of the proposed hybrid machine learning framework for early detection of HIV-associated opportunistic infections. The model achieved a high overall accuracy of 96.79%, indicating strong classification performance. Class 0 (non-opportunistic infection cases) shows near-perfect performance with precision (0.9799), recall (0.9793), and F1-score (0.9796), reflecting excellent identification of non-infection cases. Class 1 (opportunistic infection cases) also performs strongly with precision (0.9234), recall (0.9255), and F1-score (0.9245), demonstrating good sensitivity in detecting infection cases despite class imbalance.

The macro and weighted averages (0.95 and 0.9679 respectively) confirm balanced and consistent performance across both classes, showing that the model is reliable for early detection of HIV-associated opportunistic infections.

Table 3: Classification Performance Metrics

Class / Metric	**Precision**	**Recall**	**F1-score**	**Support**
Class 0	0.9799	0.9793	0.9796	1645

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

Class 1	0.9234	0.9255	0.9245	443
Accuracy	—	—	0.9679	2088
Macro Average	0.9517	0.9524	0.9520	2088
Weighted Average	0.9679	0.9679	0.9679	2088

The confusion matrix in Figure 4 illustrates the performance of the proposed hybrid machine learning model in classifying HIV-associated opportunistic infections. The model correctly identified 1,611 true negative cases, indicating a strong ability to accurately classify patients without infection. It also correctly detected 410 true positive cases, demonstrating good sensitivity in identifying infected individuals. However, the model produced 34 false positives, where non-infected patients were incorrectly classified as infected, and 33 false negatives, where infected patients were missed. Overall, the relatively balanced number of false positives and false negatives suggests that the model maintains a good trade-off between sensitivity and specificity. The results indicate that the framework is highly effective for early screening, with strong predictive reliability and minimal misclassification, making it suitable for clinical decision support in HIV-associated opportunistic infection detection

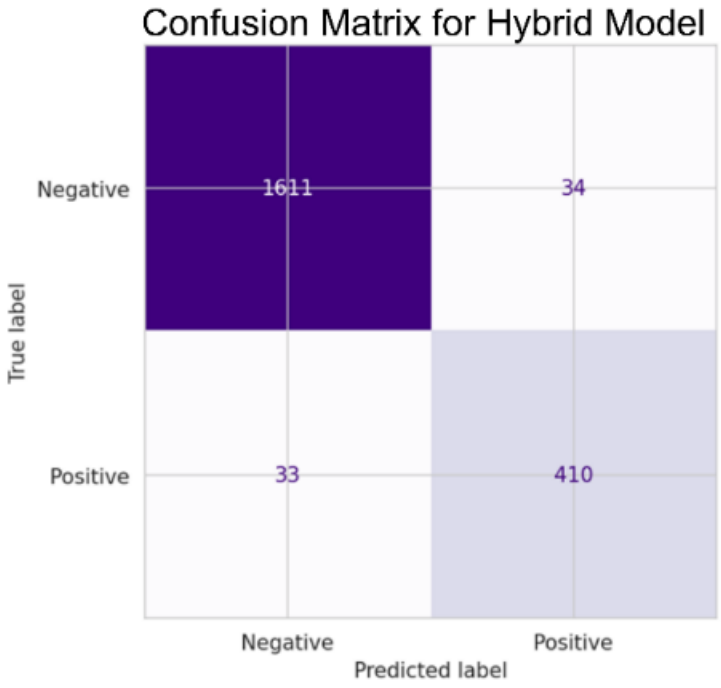


Figure 4. Confusion Matrix of the Hybrid Machine Learning

The loss curves presented in Figure 5 provide critical insight into the model's learning dynamics and convergence behavior over 500 training epochs. The training loss (blue line) demonstrates a sharp initial decline, indicating that the hybrid framework rapidly learns the underlying patterns distinguishing opportunistic infections from negative cases during the early stages of training. Simultaneously, the validation loss (orange line) follows a similar downward trajectory, suggesting that the model generalizes well to unseen data, with no immediate signs of severe overfitting. However, as training progresses beyond the initial epochs, both curves appear to stabilize with the validation loss settling at a slightly higher value than the training loss. This small but persistent gap is expected and reflects the inherent challenge of generalizing from training data to new patient samples in a clinically heterogeneous population. Importantly, the absence of a widening divergence between the two curves throughout the remaining epochs indicates that the model maintains stable learning without significant overfitting, making it reliable for deployment in early detection settings where robust generalization is paramount for patient safety.

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

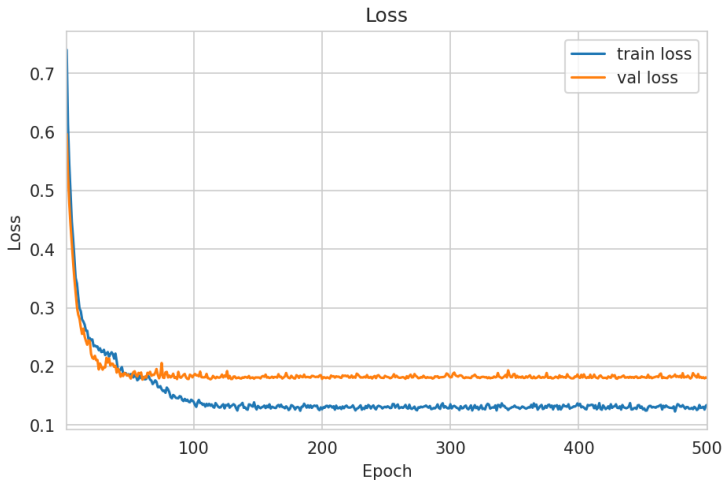


Figure 5: Training and validation loss curves showing the model's convergence behavior and generalization gap over 500 training epochs.

The accuracy trends illustrated in Figure 6 showcase the model's classification performance and its ability to correctly identify both positive and negative cases of HIV-associated opportunistic infections across the training duration. Both the training and validation accuracy curves exhibit a steep ascent during the first 100 epochs, rising from baseline levels to approximately 0.98, indicating that the hybrid framework efficiently captures distinguishing features early in the learning process. Beyond this point, both curves remain consistently high and tightly aligned, fluctuating minimally around the 0.98 mark for the balance of the 500 epochs. This sustained high performance on both the training and validation sets suggests that the model has achieved near-optimal classification capability without sacrificing generalizability. The close alignment between the two curves is particularly encouraging, as it demonstrates that the framework does not merely memorize training data but instead learns robust patterns that translate effectively to new, unseen patient data. For early detection of opportunistic infections, such high accuracy is clinically promising, as it implies a low rate of both missed diagnoses and false alarms, thereby supporting timely and reliable clinical decision-making.

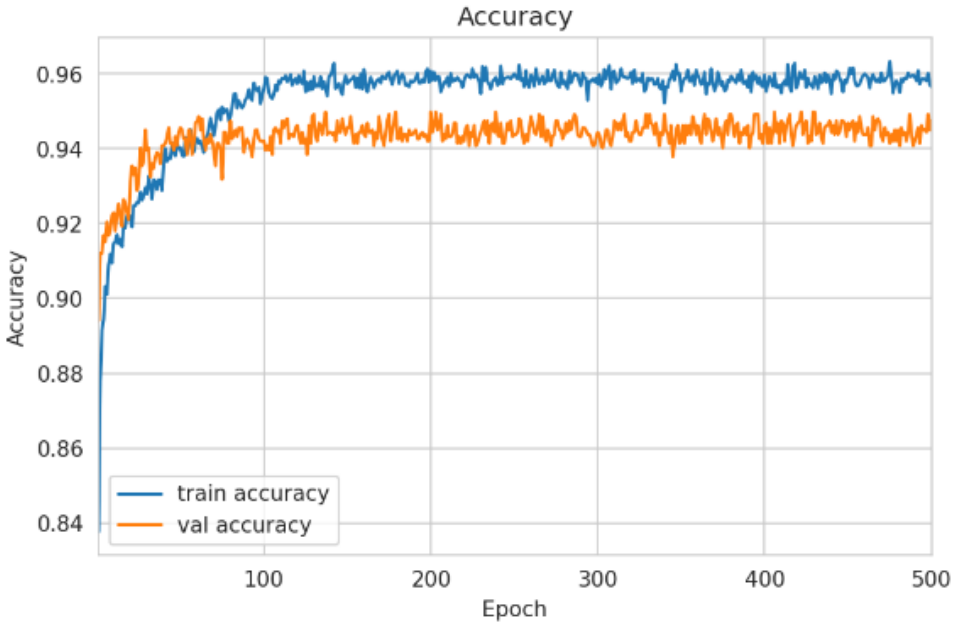


Figure 6. Training and validation accuracy

The Area Under the ROC Curve (AUC) values presented in Figure 7 offer a comprehensive measure of the model's discriminatory power in distinguishing between patients with and without HIV-associated opportunistic

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

infections throughout the training process. At epoch 0, the training and validation AUC begin at 0.950 and 0.960, respectively, indicating that even before extensive training, the model possesses considerable inherent ability to differentiate between the two classes. By epoch 10, the training AUC rapidly ascends to 0.980, while the validation AUC reaches 0.978, demonstrating the framework's swift capacity to refine its decision boundary with minimal exposure to the training data. Subsequently, from epoch 20 through epoch 500, the training AUC stabilizes at 0.988 and the validation AUC plateaus at 0.974, maintaining this near-perfect discriminatory performance consistently across the remaining epochs. The slight but stable margin between the training and validation AUC values suggests a modest degree of overfitting, yet the sustained validation AUC of 0.974 remains exceptionally high and clinically valuable. This enduring performance indicates that the hybrid framework reliably separates positive from negative cases with excellent sensitivity and specificity, making it a powerful tool for the early detection of opportunistic infections in HIV-positive individuals, where early intervention can significantly improve patient outcomes.

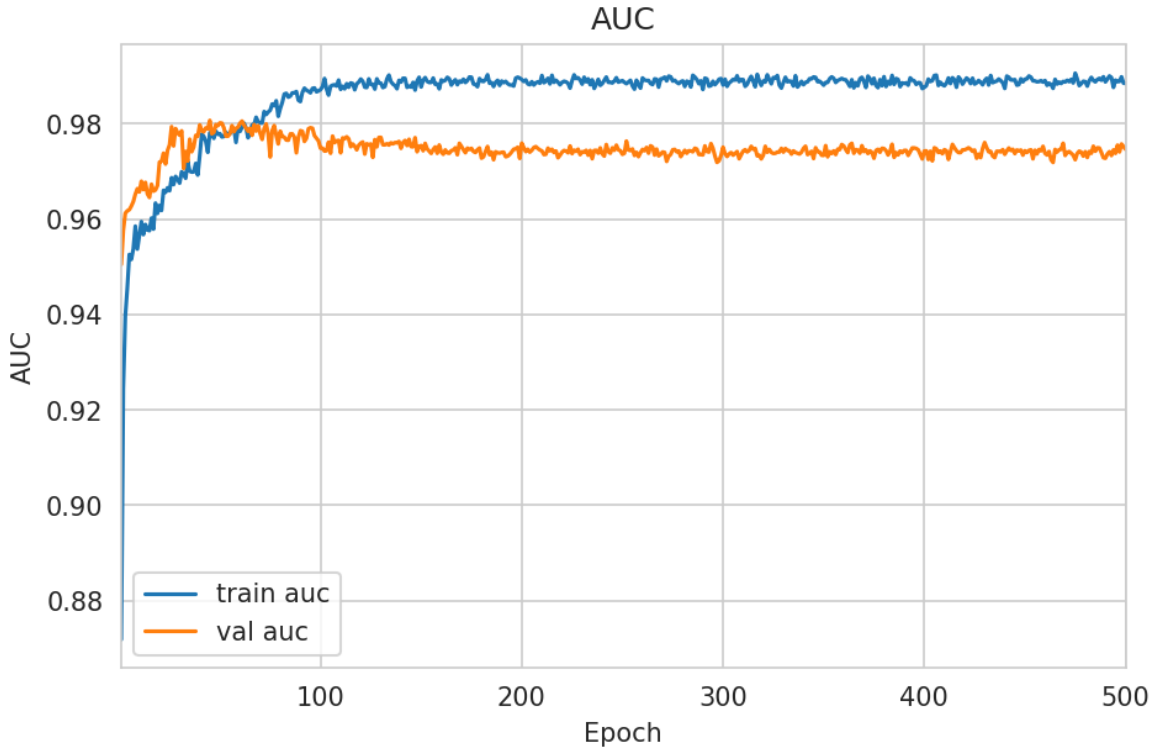


Figure 7. Training and validation Area Under the ROC Curve (AUC) .

The data presented in Figure 8, which displays the fraction of actual positive cases within each predicted risk tier, serves as a direct validation of the hybrid framework's risk calibration and clinical utility. The table reveals a clear and monotonic gradient: only 1% (0.01) of patients assigned to the Low-risk tier were actually positive for opportunistic infections, whereas 35% (0.35) of those in the Medium-risk tier and a striking 90% (0.90) of those in the High-risk tier were confirmed positive. This progressive increase in the observed positivity rate across tiers demonstrates exceptional risk stratification capability, indicating that the model's risk scores are well-calibrated and clinically meaningful. The extremely low positivity rate (1%) in the Low-risk tier is particularly valuable, as it validates the model's ability to safely identify patients who can be spared from invasive diagnostic procedures and intensive monitoring, thereby reducing healthcare costs and patient anxiety. Conversely, the 90% positivity rate in the High-risk tier represents a powerful clinical tool, as it enables clinicians to prioritize this subgroup for immediate confirmatory testing, early initiation of prophylactic therapies, and close follow-up, potentially averting life-threatening opportunistic infections. This near-perfect separation between tiers underscores the framework's high predictive validity and reinforces its potential as a reliable decision-support system for personalized risk-based screening in HIV/AIDS care, ultimately facilitating earlier detection and improved patient outcomes.

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

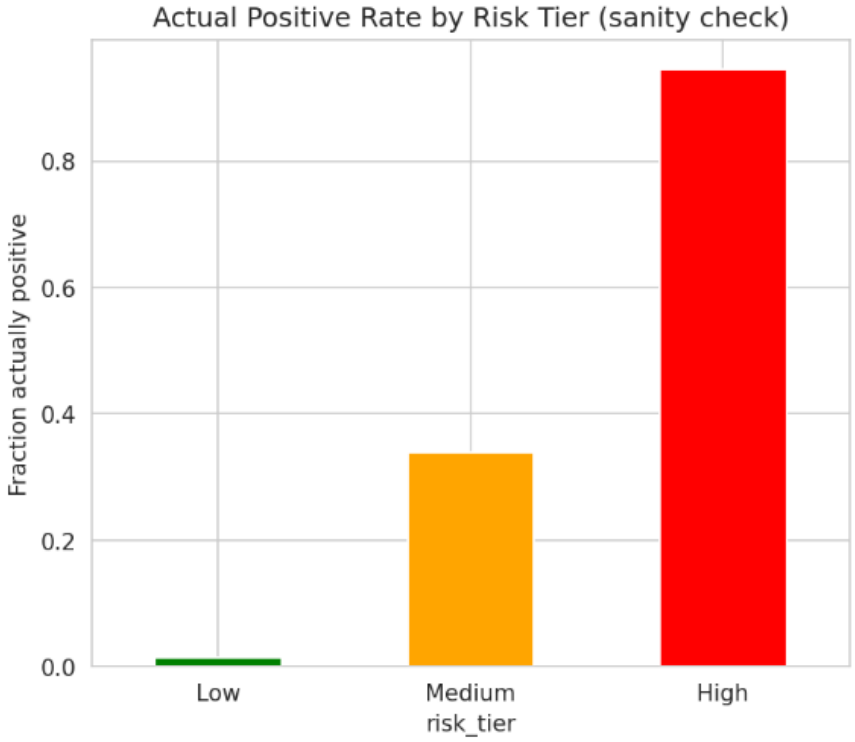


Figure 8: Fraction of Actual Positive Cases Within Each Predicted Risk Tier

The SHAP analysis provides valuable insights into the relative importance of features used in predicting HIV-induced opportunistic infections. The results reveal that TPT_TYPE_ENCODED, CD4_NUMERICAL, and AGE_CLEAN are the most influential predictors, with higher SHAP values indicating a stronger contribution to the model's output. Among these variables, TPT_TYPE_ENCODED demonstrates the greatest predictive influence, followed by CD4 count and age, underscoring the importance of treatment type, immune status, and demographic characteristics in assessing infection risk. As illustrated in Figure 9, the distribution of SHAP values enhances model interpretability by showing how individual features affect prediction outcomes, thereby providing actionable, data-driven insights that can support clinical decision-making and improve the early detection and management of opportunistic infections.

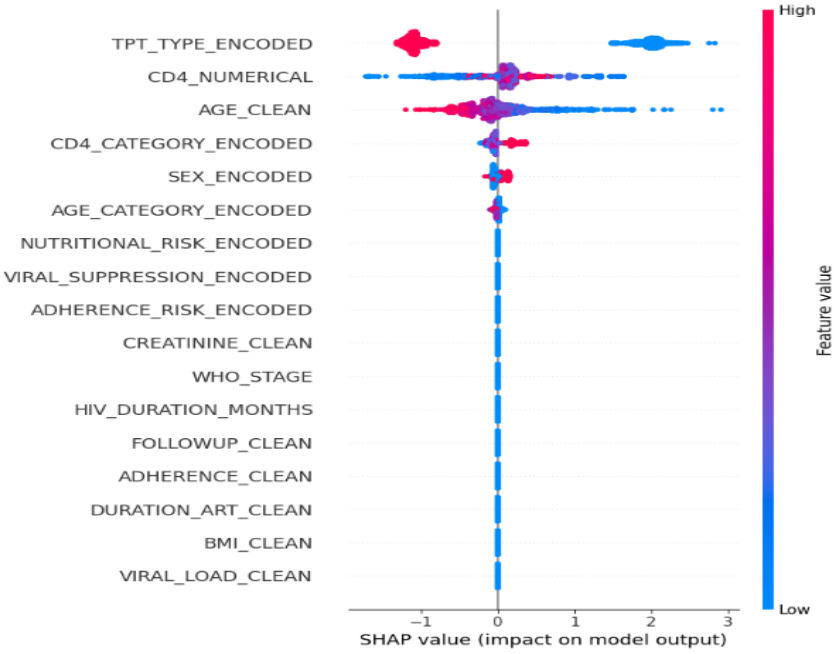


Figure 9: SHAP value for OI

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

Figure 10 presents a comprehensive analysis of the linear relationships between clinical and demographic variables associated with viral load suppression among HIV patients. The results indicate that "Clinical_Risk_Score_28" has a strong correlation coefficient of -0.28, establishing it as a valuable indicator in understanding suppression outcomes. "Age" shows a smaller correlation of -0.10, while "CD4_Count_Numeric" (0.05), "Regimen_Line_Complexity" (0.08), and "WHO_Stage" (0.12) exhibit positive correlations, signifying that these factors positively support effective viral suppression. These relationships enhance the interpretability of the models, offering valuable insights for identifying patient patterns and improving monitoring precision in HIV-induced OI prediction

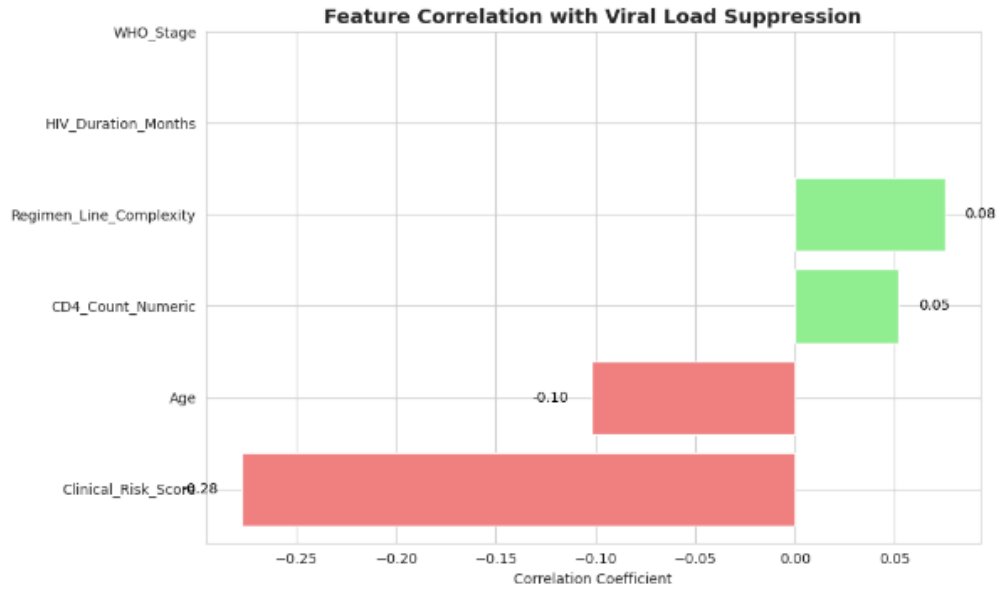


Figure 10: Accuracy by Activation and layers

5.1 Results and Discussion

The results of this study demonstrate the effectiveness of the proposed hybrid machine learning framework for the early detection of HIV-associated opportunistic infections. The model achieved a high classification performance, with an overall accuracy of 96.79%, indicating strong capability in distinguishing between infected and non-infected cases. The classification report further shows that the model performs exceptionally well on Class 0 (non-infection cases), with precision (0.9799), recall (0.9793), and F1-score (0.9796), reflecting its reliability in correctly identifying healthy patients.For Class 1 (opportunistic infection cases), the model also achieved strong performance with precision (0.9234), recall (0.9255), and F1-score (0.9245), indicating good sensitivity in detecting infected individuals even under class imbalance conditions. The macro-average (0.9520) and weighted-average (0.9679) scores further confirm that the model maintains balanced performance across both classes without significant bias toward the majority class.

The confusion matrix analysis supports these findings, showing 1,611 true negatives and 410 true positives, with relatively low misclassification rates of 34 false positives and 33 false negatives. This indicates that the model maintains a strong balance between sensitivity and specificity, making it suitable for clinical screening where both missed diagnoses and false alarms must be minimized.Further analysis of the training dynamics revealed stable convergence behavior, with loss curves showing a steady decrease and minimal overfitting. Similarly, the accuracy curves indicate consistent learning performance across training and validation sets, stabilizing at high accuracy levels. The ROC-AUC results further confirm strong discriminatory power, with values remaining above 0.97, demonstrating an excellent ability to separate positive and negative cases.

Comparative evaluation of activation functions and network depth shows that simpler architectures, particularly the ReLU model, outperform deeper configurations. The SHAP analysis provides additional interpretability, revealing that CD4 count, age, and treatment type are the most influential predictors of opportunistic infections. This enhances the clinical relevance of the model by aligning its predictions with known medical risk factors.

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

The results confirm that the proposed hybrid framework is robust, accurate, and clinically interpretable, making it suitable for early detection and decision support in HIV-associated opportunistic infections.

6.1 Conclusion

This study developed a hybrid machine learning framework combining Deep Neural Networks and XGBoost for early detection of HIV-associated opportunistic infections. The model effectively integrates clinical, demographic, and treatment-related features to support accurate risk prediction.

The findings show that the proposed system achieves high predictive performance, with strong accuracy, sensitivity, and ROC-AUC values, making it suitable for clinical screening applications. Comparative experiments confirmed that simpler architectures with ReLU activation outperform deeper and more complex models. Furthermore, interpretability analysis using SHAP highlighted key clinical factors such as CD4 count, age, and treatment type as major determinants of infection risk. The framework provides a reliable, interpretable, and high-performing decision-support tool that can assist healthcare professionals in early identification of HIV-associated opportunistic infections, ultimately supporting timely intervention and improved patient outcomes.

REFERENCES

1. Bartl, L., Zeeb, M., Kälin, M., Loosli, T., Notter, J., Furrer, H., and Nemeth, J. (2025). Machine Learning-based Prediction of Active Tuberculosis in People With HIV Using Clinical Data. *Clinical Infectious Diseases*.
2. Chen, H., Chen, F., Wang, Y., Cai, E., Pan, W., Li, Y., andSu, F. (2025). A Machine Learning Model for Diagnosing Opportunistic Infections in HIV Patients: Broad Applicability Across Infection Types. *Journal of Cellular and Molecular Medicine*.
3. Chen, J., Liu, L., Huang, J., Jiang, Y.-J., Yin, C., Zhang, L., andLu, 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*.
4. Cherezov, D., Dam, T., Najjingo, I., Mbabazi, M., Kisembo, H., Kirenga, B., andMadabhushi, A. (2025). Opportunistic use of artificial intelligence with X-ray imaging for diagnosis of HIV status in tuberculosis patients in Uganda and Tanzania. *PLOS Digital Health*.
5. Chew, R., Woods, M., and Paterson, D. L. (2024). Development and validation of supervised machine learning multivariable prediction models for the diagnosis of Pneumocystis jirovecii pneumonia using nasopharyngeal swab PCR in adults in a low-HIV prevalence setting. *International Health*.
6. Ekpenyong, M., Etebong, P. I., and Jackson, T. (2019). Fuzzy-multidimensional deep learning for efficient prediction of patient response to antiretroviral therapy. *Heliyon*.
7. Fu, X., Wu, L., Xun, J., Pütz, B., Zheng, Z., Li, Y., and Müller-Myhsok, B. (2025). Predictive survival modelings for HIV-related cryptococcosis: comparing machine learning approaches. *Frontiers in Cellular and Infection Microbiology*.
8. Henderson, H., Napravnik, S., Kosorok, M., Gower, E. W., Kinlaw, A. C., Aiello, A., and van Duin, D. (2022). Predicting Risk of Multidrug-Resistant Enterobacterales Infections Among People With HIV. *Open Forum Infectious Diseases*.
9. Inyang, U. G., Eyoh, I. J., Umoh, U. A., Ubong, E. A., Ene, E. E., Obiyo, D. C., andAkponome, B. E. (2025). *Optimal deep neural network parameters for power loss minimization analytics*. Nigerian Journal of Technology (NIJOTECH), 44(4), pp.1–14. https://doi.org/10.4314/njt.2025.4911
10. Li, L., Shi, C., Liu, X., Li, W., Luo, Y., Zhang, H., and 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*.
11. Mahmud, S., and Hossain, F. (2025). Identification of HIV-Associated Gene Expression Biomarkers Using Machine Learning and Interpretable Artificial Intelligence. *bioRxiv*.
12. Ngiam, K., & Khor, I. W. (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*.
13. Qasem, S. N. (2024). Introducing HeliEns: A Novel Hybrid Ensemble Learning Algorithm for Early Diagnosis of Helicobacter pylori Infection. *De Computis*.

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

14. Sa'adah, L. N., Fauzi, F., Prizka, Arum, R., Haris, M., Yan, andBisoumi, N. (2026). HYBRID RESAMPLING METHOD AND HYPERPARAMETER OPTIMIZATION FOR HIV/AIDS PREDICTION: EVIDENCE FROM EIGHT MACHINE-LEARNING MODELS. *JITK (JurnalIlmuPengetahuan Dan TeknologiKomputer)*.
15. Sah, A., Elshaikh, R., Shalabi, M., Abbas, A. M., Prabhakar, P. K., Babker, A. M., and Agarwal, S. (2025). Role of Artificial Intelligence and Personalized Medicine in Enhancing HIV Management and Treatment Outcomes. *Life*.
16. Scott, I. (2021). Demystifying machine learning: a primer for physicians. *Internal Medicine Journal (Print)*.
17. Sun, Q., Zhang, K., Xu, Y., Luo, M., Yang, Z., Liu, Q., and Liu, A. (2025). Explainable machine learning for predicting clinical outcomes in HIV/TB co-infection: a comparative retrospective study. *BMC Infectious Diseases*.
18. Waring, J., Lindvall, C., andUmeton, R. (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artif. Intell. Medicine*.
19. Xu, J., Huang, Q., Yu, J., Liu, S., Yang, Z., Wang, F., and Xiao, Y. (2022). Metagenomic Next-Generation Sequencing for the Diagnosis of Suspected Opportunistic Infections in People Living with HIV. *Infection and Drug Resistance*.
20. Zhou, J., Yang, Y., Xie, Z., Lu, D., Huang, J., Lan, L., and Huang, J. (2023). Dysbiosis of gut microbiota and metabolites during AIDS: implications for CD4+ T cell reduction and immune activation. *AIDS (London)*.
21. Zhou, Y., Hemmige, V., Dalai, S., Hong, D. K., Muldrew, K., andMohajer, M. (2019). Utility of Whole-Genome Next-Generation Sequencing of Plasma in Identifying Opportunistic Infections in HIV/AIDS. *Open AIDS Journal*.