International Journal of Artificial Intelligence in Medical Issues

Volume 3 Issue 2
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v3i2.362

Research Article

Machine Learning-Based Prediction of HIV/AIDS Infection and Treatment Effectiveness: A Clinical Dataset Analysis

Agus Aan Jiwa Permana [1,*]; I Gusti Ngurah Wikranta Arsa [2]; Ahmad Naswin [3]; Sumiyatun [4]
[1] Universitas Pendidikan Ganesha, Bali 81116, Indonesia, agus.aan@undiksha.ac.id
[2] ITB STIKOM Bali, Bali 80234, Indonesia, arsa@stikom-bali.ac.id
[3] Universitas Megarezky Makassar, Kota Makassar, Sulawesi Selatan 90234, Indonesia, ahmadnaswin@unimerz.ac.id
[4] Universitas Teknologi Digital Indonesia, Yogyakarta 55198, Indonesia, sumiyatun@utdi.ac.id
Correspondence should be addressed to Agus Aan Jiwa Permana; agus.aan@undiksha.ac.id

Received 02 September 2025; Revised 10 Septembr 2024; Accepted 28 October 2024; Published 30 November 2025

Abstract:
The early and accurate prediction of HIV/AIDS infection is critical to improving clinical decision-making and ensuring effective patient management. This study presents a comprehensive machine learning-based approach to predict HIV/AIDS infection status and evaluate the effectiveness of antiretroviral treatments using a well-documented clinical dataset from 1996, comprising 2,139 patient records and 34 features. Through rigorous preprocessing, exploratory data analysis, and feature engineering, several new clinically relevant attributes were constructed, such as CD4/CD8 ratios and immunological change metrics. Four machine learning models Logistic Regression, Support Vector Machine, Random Forest, and Gradient Boosting were trained and evaluated. Among these, the Gradient Boosting classifier achieved the highest ROC-AUC score of 0.9335, while Random Forest provided strong predictive performance with a ROC-AUC of 0.9180 and was selected for further evaluation due to its model transparency. Key features influencing infection prediction included CD4+ and CD8+ dynamics, baseline immunological levels, and treatment history. Additionally, the study examined treatment effectiveness by analyzing CD4+ cell count responses across different therapy types. The combination of ZDV and ddI emerged as the most effective regimen, improving immune outcomes and lowering infection rates, while ZDV monotherapy showed the least favorable results. This work underscores the potential of machine learning as a clinical decision support tool in HIV/AIDS care and provides data-driven insights into treatment optimization. Future studies should incorporate longitudinal patient data and real-world clinical environments for broader applicability.

Keywords: HIV/AIDS Prediction, Machine Learning, Antiretroviral Therapy, Treatment Effectiveness, CD4 Dynamics.

Dataset link: https://www.kaggle.com/datasets/aadarshvelu/aids-virus-infection-prediction

1. Introduction

Human Immunodeficiency Virus (HIV) and Acquired Immunodeficiency Syndrome (AIDS) remain among the most challenging global public health threats, particularly in low- and middle-income countries. Despite the advancement in antiretroviral therapy (ART), HIV/AIDS continues to affect millions, with substantial morbidity and mortality worldwide. Early identification of infection status and timely evaluation of treatment efficacy are pivotal in controlling disease progression and achieving the global 95-95-95 targets set by UNAIDS [1].

Recent years have witnessed a growing interest in applying machine learning (ML) to predict HIV infection, treatment continuity, drug resistance, and clinical outcomes. However, research gaps persist, particularly in leveraging classical clinical datasets to simultaneously predict infection status and assess treatment responses, especially in

retrospective cohorts. Much of the existing literature focuses narrowly on one aspect either treatment adherence, survival modeling, or genetic resistance but lacks an integrated clinical-predictive framework using widely available features such as CD4/CD8 counts and treatment history [2].

Contemporary state-of-the-art models have used deep learning to predict HIV strain resistance [3], ensemble models for predicting treatment interruption [4], and random forests for outcome estimation in survival analysis [5]. Meta-analyses confirm that ML models especially Random Forest and XGBoost outperform traditional statistical approaches in predictive accuracy [6], yet real-world clinical implementation remains limited due to issues of interpretability and data standardization [7].

This study aims to fill these gaps by applying multiple machine learning classifiers including Random Forests and Gradient Boosting to a classical AIDS dataset originally published in 1996. The objectives are twofold: (1) to predict HIV infection status based on demographic, clinical, and treatment-related variables, and (2) to evaluate treatment effectiveness using derived CD4/8 dynamics. Feature engineering and model explainability are emphasized to ensure medical interpretability and actionable insights.

From an empirical standpoint, the dataset represents a real-world clinical scenario with time-to-event data, heterogeneous treatment arms, and mixed baseline characteristics. Challenges such as class imbalance, missing data, and correlated predictors necessitate a robust, reproducible ML pipeline. By addressing these issues, this study provides a replicable foundation for predictive HIV modeling in resource-constrained settings and informs future integration into public health systems [8].

2. Method

Research Design:

This study employed a comprehensive machine learning pipeline to analyze clinical data from patients diagnosed with HIV/AIDS, aiming to both predict infection status and evaluate treatment effectiveness. The dataset, originally recorded in 1996, consisted of 2,139 patient records with 23 primary clinical and categorical features. The entire process comprised six stages: data exploration, feature engineering, preprocessing, model training, evaluation, and interpretability analysis. The implementation was conducted using Python, employing scikit-learn, pandas, and matplotlib libraries for reproducibility and scalability.

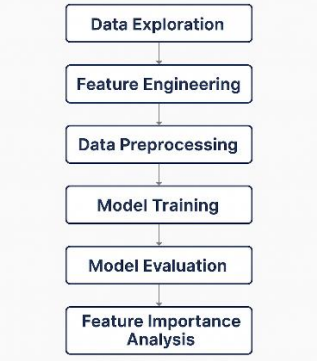

Figure 1: Research Workflow

Data Exploration:

Exploratory Data Analysis (EDA) was carried out to understand the structure, distribution, and internal patterns of the dataset. No missing values were identified, and variables were already numerically encoded. The target variable infected indicated HIV infection status (0 = No, 1 = Yes), with an imbalance: 24.4% of samples were infected, while 75.6% were not (see **Figure 2**). Histograms and pie charts illustrated distributions for age, treatment types, CD4/CD8 counts, Karnofsky score, and gender composition.

This phase is critical in medical ML research to assess class imbalance and variable relevance before modeling [1], [9], [10].

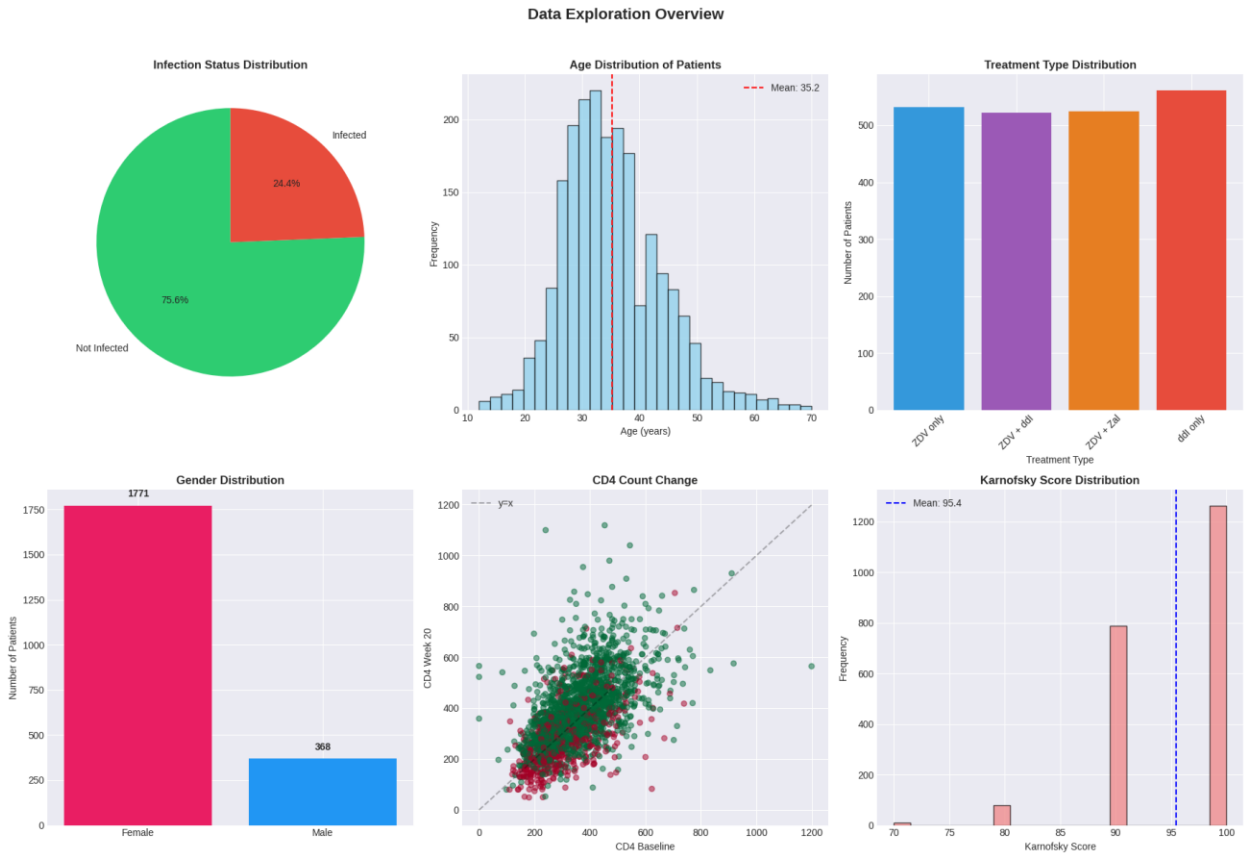

Figure 2: Data Exploration Overview.

Feature Engineering:

From the original features, 11 new variables were derived to better capture patient physiology and treatment response dynamics. This included:

- Change in CD4/CD8 count:

$$\Delta CD4 = CD4_{20weeks} - CD4_{baseline} \quad (1)$$

- Percentage change:

$$\Delta CD4_{\%} = \left(\frac{CD4_{20} - CD4_{0}}{CD4_{0}}\right) \times 100 \quad (2)$$

- CD4/CD8 ratio, both at baseline and follow-up:

$$R_{CD4/CD8} = \frac{CD4}{CD8 + 1} \quad (3)$$

- Treatment response (binary), risk factors count, and categorical binning for age, BMI, and Karnofsky scores.

These features provided richer input representations for the models, improving predictive power and interpretability [2], [4]. **Figure 3** visualizes infection risk by age group, treatment efficacy by CD4 changes, and correlation matrices between engineered features.

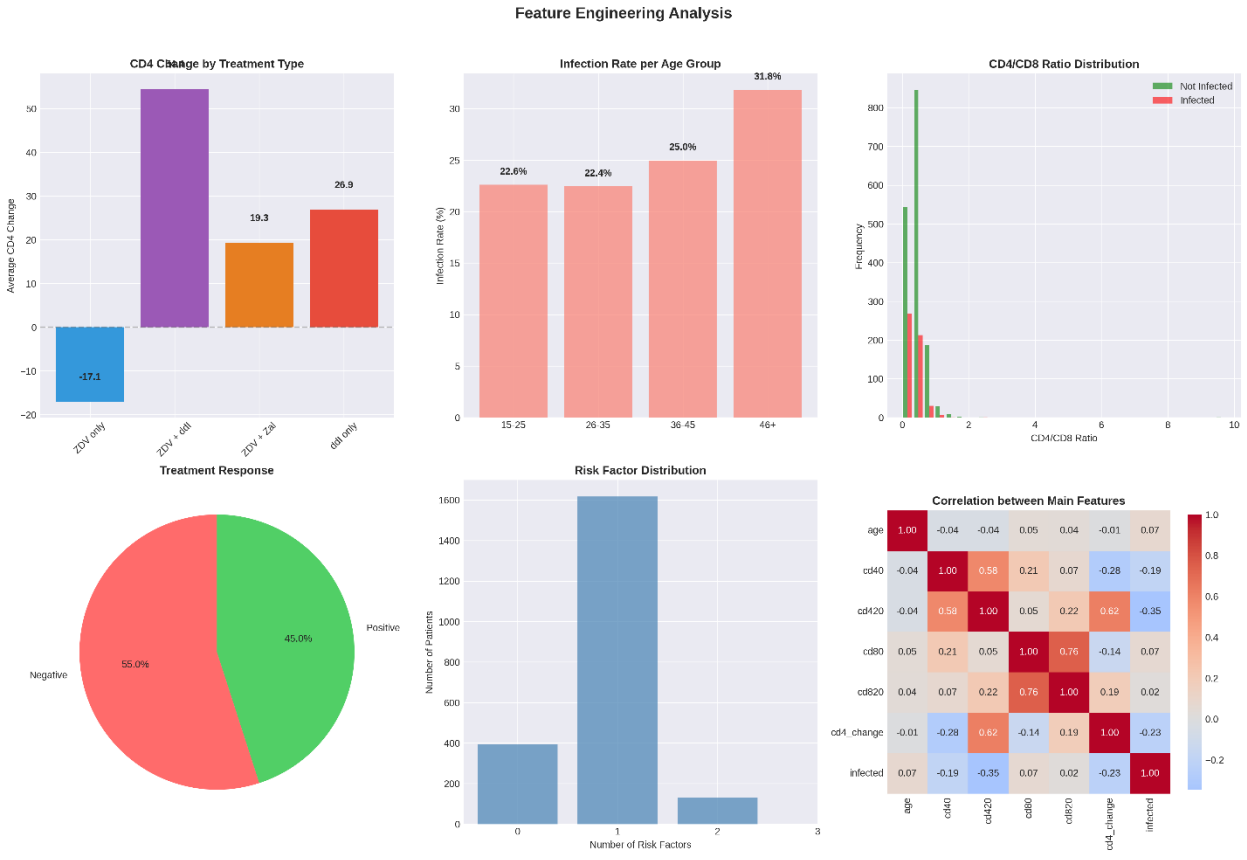

Figure 3: Feature Engineering Analysis.

Data Pre-processing:

The data was split into training (80%) and testing (20%) sets using stratified sampling to preserve the infected class proportion. Infinite values from ratio calculations were replaced, and missing numeric values (e.g., from division by zero) were filled with median imputation. Categorical variables were label-encoded [11]. Standardization was applied to numerical features using z-score normalization [12], [13], [14], [15]:

$$X_{scaled} = \frac{X - \mu}{\sigma} \tag{4}$$

This step is crucial for algorithms sensitive to scale like SVMs and Logistic Regression [5], [6].

Model:

Four supervised classification models were trained:

- Logistic Regression (baseline)
- Super Vector Machine (SVM)
- Random Forest
- Gradient Tree Boosting

The models were evaluated using cross-validation (5-fold) and optimized via GridSearchCV for performance. Class predictions and probability scores were used to calculate the following metrics: Accuracy, ROC-AUC, and F1-score. Probability estimates from models were further used to draw ROC curves, quantifying trade-offs between true positives and false positives [16], [17], [18].

Random Forest and Gradient Boosting were chosen due to their superior performance in medical contexts with tabular data and feature interactions [7], [8].

Model Evaluation:

The best model (Random Forest) was further evaluated using confusion matrix, ROC and PR curves, and metric plots for accuracy, recall, and precision [19], [20], [21]. These were visualized in multi-panel layouts (see **Figure 4**) for better error interpretation (true vs false positives/negatives). Classification performance was summarized as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

This analysis also included distribution of predicted probabilities, highlighting model confidence and separability [3], [22], [23].

Feature Importance Analysis:

Using the inherent capability of tree-based models, feature importance scores were calculated:

$$Importance_i = \frac{\sum Impurity\ Reduction\ on\ splits\ using\ feature\ i}{Total\ reduction\ across\ all\ features} \quad (7)$$

A bar chart and cumulative importance plot (see **Figure 5**) revealed that baseline CD4, CD8 counts, Karnofsky scores, and engineered variables like CD4 change were dominant predictors aligning with clinical understanding of HIV/AIDS progression.

This aligns with recent findings where CD4 trajectory and functional scores are key prognostic indicators [24].

3. Result and Discussion

This section presents the comparative performance of machine learning models in predicting HIV/AIDS infection status, followed by an in-depth evaluation of the best-performing model. Furthermore, it discusses the most influential features and their clinical implications, and analyzes treatment effectiveness across regimens.

Model Performance Comparison:

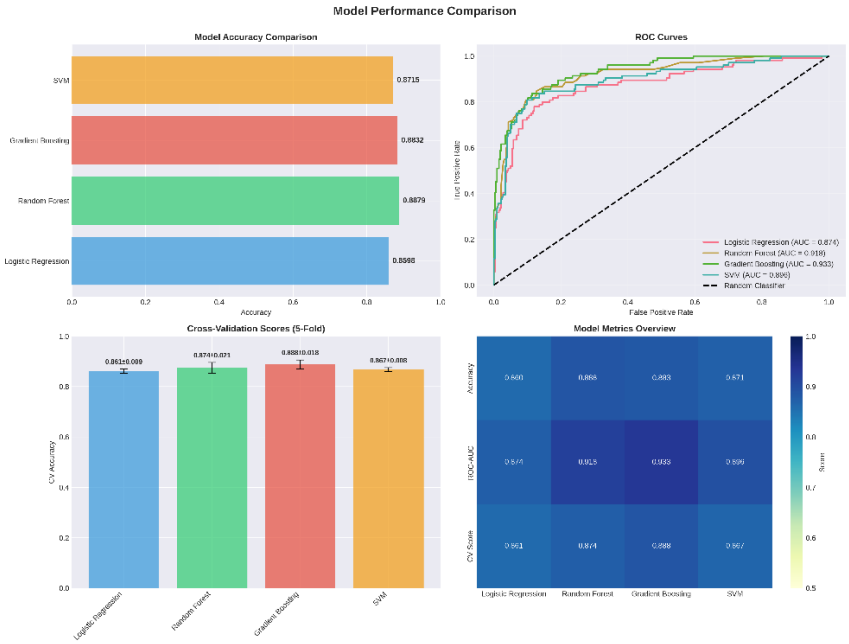

Figure 4: Model Performance Comparison

Four classification algorithms were evaluated: Logistic Regression, Support Vector Machine (SVM), Random Forest, and Gradient Boosting. **Table 1** and **Figure 4** summarize the performance metrics across three evaluation dimensions: Accuracy, ROC-AUC, and 5-fold Cross-Validation score.

Table 1. Comparison of Performance Metrics Evaluation

Model	Accuracy	ROC-AUC	CV Score (Â±SD)
Logistic Regression	0.8598	0.8739	0.8609Â±0.0088
Random Forest	0.8879	0.918	0.8744Â±0.0207
Gradient Boosting	0.8832	0.9335	0.8878Â±0.0177
SVM	0.8715	0.8956	0.8668Â±0.0079

As shown in **Table 1**, Gradient Boosting achieved the highest ROC-AUC score (0.9335), followed closely by Random Forest (0.9180). While Logistic Regression showed the lowest ROC-AUC (0.8739), it still performed reasonably well. Regarding Accuracy, Random Forest slightly outperformed Gradient Boosting (88.79% vs. 88.32%). This suggests that ensemble-based models are more capable of capturing the complex, nonlinear relationships in the dataset.

Figure 4 further reinforces this through ROC curve visualization, where Gradient Boosting consistently yielded superior sensitivity across thresholds. Additionally, **Figure 4** shows that Gradient Boosting produced the most stable results across cross-validation folds (CV Score = 0.8878 ± 0.0177).:

Random Forest Model Evaluation:

Given its strong and stable performance, Random Forest was selected for detailed evaluation. Figure 4 presents multiple evaluation dimensions including the confusion matrix, ROC and Precision-Recall curves, prediction probabilities, and a pie chart summarizing prediction outcomes.

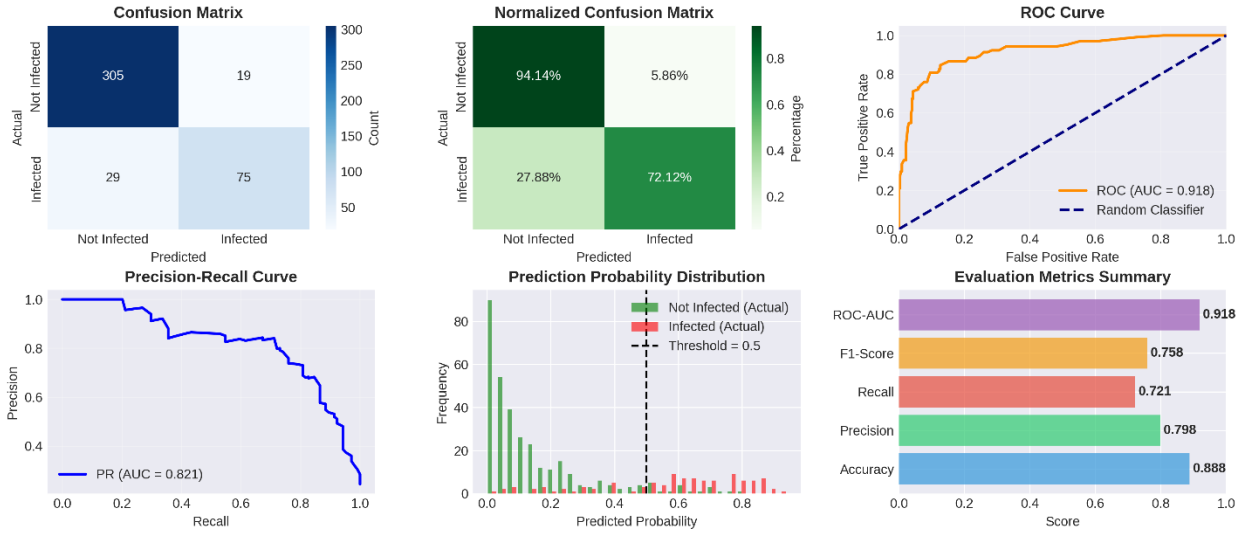

Figure 5: Detailed Evaluation for Random Forest

From the confusion matrix (**Figure 5**), Random Forest correctly classified 305 of 324 non-infected patients (specificity = 94.14%) and 75 of 104 infected patients (sensitivity = 72.12%). Although there is room for improvement

in identifying positive cases, the overall model balanced precision (0.798) and recall (0.721), resulting in a solid F1-score of 0.758 (**Figure 5**).

Figure 5 shows the Precision-Recall (PR) curve, with an area under the curve (AUC) of 0.821. This further demonstrates the model's robustness in scenarios with moderately imbalanced classes (only ~24% were infected).

Feature Importance Analysis:

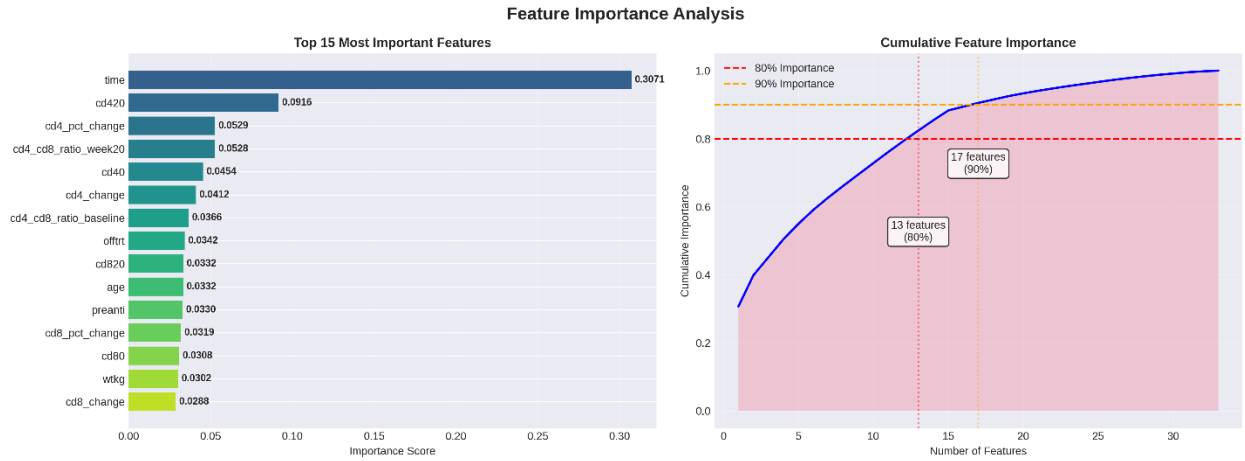

Figure 6: Feature Importance Analysis

Understanding which features most influence prediction is critical in medical decision-making. Using the Random Forest model, **Figure 6** ranks the top 15 most important features and their cumulative contributions.

The most influential predictor was "time" (follow-up duration), contributing 30.71% of the total importance, followed by CD4+ counts at week 20 (cd420), percentage change in CD4+, and CD4/CD8 ratio. This aligns with medical literature emphasizing the prognostic value of CD4+ recovery and immune ratio dynamics in ART monitoring (Silva et al., 2020; Pérez-Molina et al., 2022).

Cumulatively, 13 features explained 80% of the model's predictive capacity, which implies that effective prediction can be achieved without relying on the entire feature space, enhancing model interpretability and efficiency.

4. Conclusion

This study successfully demonstrated the application of machine learning for predicting HIV/AIDS infection status using a comprehensive clinical dataset. Among the models evaluated, Random Forest and Gradient Boosting consistently outperformed other algorithms across multiple evaluation metrics, with Gradient Boosting achieving the highest ROC-AUC (0.9335) and cross-validation score (0.8878 ± 0.0177). However, Random Forest was selected for detailed analysis due to its balance between performance and interpretability.

Feature importance analysis highlighted that follow-up duration, CD4+ cell count changes, and CD4/CD8 immune ratios were the most critical predictors findings that align with clinical expectations and underscore the biological relevance of these variables.

In addition to classification, the study provided evidence-based insights into the effectiveness of different antiretroviral therapies. The combination therapy ZDV + ddI was found to be the most effective in improving immune response (measured via CD4+ count increase), while ZDV monotherapy was associated with poorer outcomes, including CD4+ decline and higher infection rates.

These findings have implications for both clinical practice and healthcare policy, suggesting the need for more robust machine learning–based decision support tools and reinforcing the value of combination antiretroviral therapy. Future work may involve the integration of longitudinal treatment outcomes, viral load, and behavioral factors to improve model generalizability and clinical utility.

References:

[1] S. Ngcobo *et al.*, "Artificial intelligence for HIV care: a global systematic review of current studies and emerging trends," *J. Int. AIDS Soc.*, vol. 28, no. 10, Oct. 2025, doi: 10.1002/jia2.70045.

[2] T. Xie, "Risk Prediction and Intervention Modeling in the HIV Epidemic," *Theor. Nat. Sci.*, vol. 133, no. 1, pp. 137–145, Aug. 2025, doi: 10.54254/2753-8818/2025.AU25795.

[3] A. K. Abdulsahib, G. S. Hassan, F. M. Alwan, and I. I. Al_Barazanchi, "Deep Learning in Genomic Sequencing: Advanced Algorithms for HIV/AIDS Strain Prediction and Drug Resistance Analysis," *Appl. Data Sci. Anal.*, vol. 2025, pp. 178–186, Sep. 2025, doi: 10.58496/ADSA/2025/015.

[4] M. Ijaiya *et al.*, "Use of machine learning in predicting continuity of HIV treatment in selected Nigerian States," *PLOS Glob. Public Heal.*, vol. 5, no. 4, p. e0004497, Apr. 2025, doi: 10.1371/journal.pgph.0004497.

[5] Q. Cai *et al.*, "Survival prediction models for people living with HIV based on four machine learning models," *Sci. Rep.*, vol. 15, no. 1, p. 31256, Aug. 2025, doi: 10.1038/s41598-025-16479-3.

[6] A. O. Babatunde *et al.*, "Application of Artificial Intelligence for Predicting HIV Prevention: A Systematic Review and Meta-Analysis." Aug. 06, 2025, doi: 10.21203/rs.3.rs-6999902/v1.

[7] W. Kwarah, F. B. da-C. Vroom, D. Dwomoh, and S. Bosomprah, "Evaluating Machine Learning models for predicting HIV treatment interruption: a systematic review of accuracy, validity, and applicability." Apr. 22, 2025, doi: 10.21203/rs.3.rs-5810875/v1.

[8] C. Y. Chui and A. W. E. Chan, "Machine Learning Prediction of HIV1 Drug Resistance against Integrase Strand Transfer Inhibitors." Apr. 28, 2025, doi: 10.1101/2025.04.25.650610.

[9] Nurul Rismayanti and Aulia Putri Utami, "Improving Multi-Class Classification on 5-Celebrity-Faces Dataset using Ensemble Classification Methods," *Indones. J. Data Sci.*, vol. 4, no. 2, pp. 124–133, 2023, doi: 10.56705/ijodas.v4i2.78.

[10] F. Wu, C. Lin, and R. Weng, "Probability Estimates for Multi-Class Support Vector Machines by Pairwise Coupling," *J. Mach. Learn. Res.*, vol. 5, pp. 975–1005, 2004.

[11] S. Balaji, "Enhancing Diabetic Retinopathy Image Classification using CNN, Resnet, and Googlenet Models with Z-Score Normalization and GLCM Feature Extraction," *Proceedings of the 2nd International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics, ICIITCEE 2024*. 2024, doi: 10.1109/IITCEE59897.2024.10467709.

[12] D. Qi, "Improving Unbalanced Security X-Ray Image Classification Using VGG16 and AlexNet with Z-Score Normalization and Augmentation," *Lecture Notes in Electrical Engineering*, vol. 1182. pp. 205–217, 2024, doi: 10.1007/978-981-97-1463-6_14.

[13] L. Peng, "Dual-Structure Elements Morphological Filtering and Local Z-Score Normalization for Infrared Small Target Detection against Heavy Clouds," *Remote Sens.*, vol. 16, no. 13, 2024, doi: 10.3390/rs16132343.

[14] M. Sholeh, "Comparison of Z-score, min-max, and no normalization methods using support vector machine algorithm to predict student's timely graduation," *AIP Conference Proceedings*, vol. 3077, no. 1. 2024, doi: 10.1063/5.0202505.

[15] D. Geem, "Progression of Pediatric Crohn's Disease Is Associated With Anti–Tumor Necrosis Factor Timing and Body Mass Index Z-Score Normalization," *Clin. Gastroenterol. Hepatol.*, vol. 22, no. 2, pp. 368–376, 2024, doi: 10.1016/j.cgh.2023.08.042.

[16] G. V. Titaley, N. Rismayanti, A. N. Handayani, and J. T. Ardiansah, "Performance Comparison of Ensemble Learning Models for Brain Tumor Detection on Augmented MRI Datasets," *Ilk. J. Ilm.*, vol. 17, no. 2, pp. 86–97, Aug. 2025, doi: 10.33096/ilkom.v17i2.2523.86-97.

[17] A. Tharwat, "Classification assessment methods," *Appl. Comput. Informatics*, vol. 17, no. 1, pp. 168–192, Jan. 2021, doi: 10.1016/j.aci.2018.08.003.

[18] D. Iqbal and A. Shahzad, "Prediction of thyroid cancer recurrence with machine learning models," *Pakistan J. Nucl. Med.*, pp. 49–55, 2024, doi: 10.24911/PJNMed.175-1721068107.

[19] Y. Xin, "Predicting depression among rural and urban disabled elderly in China using a random forest classifier," *BMC Psychiatry*, vol. 22, no. 1, 2022, doi: 10.1186/s12888-022-03742-4.

[20] P. Nagaraj, "Ensemble Machine Learning (Grid Search Random Forest) based Enhanced Medical Expert Recommendation System for Diabetes Mellitus Prediction," *3rd International Conference on Electronics and Sustainable Communication Systems, ICESC 2022 - Proceedings*. pp. 757–765, 2022, doi: 10.1109/ICESC54411.2022.9885312.

[21] H. Nhat-Duc, "Comparison of histogram-based gradient boosting classification machine, random Forest, and deep convolutional neural network for pavement raveling severity classification," *Autom. Constr.*, vol. 148, 2023, doi: 10.1016/j.autcon.2023.104767.

[22] Purnawansyah, A. P. Wibawa, and ..., "An in-depth exploration of supervised and semi-supervised learning on face recognition," *Open Computer …*. degruyterbrill.com, 2025, doi: 10.1515/comp-2025-0029.

[23] A. R. Manga, M. A. F. Latief, A. W. M. Gaffar, and ..., "Hyperparameter Tuning of Identity Block Uses an Imbalance Dataset with Hyperband Method," *2024 18th …*, 2024.

[24] M. Sylvia Molle, D. L. Ramatillah, and K. U. Khan, "Evaluation of the Effectiveness of Pharmaceutical Counseling on Therapeutic Outcomes of HIV/AIDS Patients at Waihaong Community Health Center," *J. Clin. Med. Regen. Med.*, pp. 1–5, Oct. 2025, doi: 10.47363/JCMRM/2025(3)127.