HIV Stage Detection Using Interpretable Machine Learning Models

Rahmanov Ulukbek
Department of Business
Univ. of Europe for Applied Sciences
Potsdam 14469, Germany
ulukbek.rahmanov@ue-germany.de

Raja Hashim Ali
Department of Business
Univ. of Europe for Applied Sciences
Potsdam 14469, Germany
hashim.ali@ue-germany.de

***Abstract*—HIV is still a major global health concern that needs to be accurately and promptly staged in order to be effectively treated and monitored. Conventional staging techniques depend on CD4 counts or clinical evaluation, but these can be resource-intensive, inconsistent, and challenging to scale. There are currently no interpretable machine learning models that are specifically designed to use structured clinical data for HIV stage classification. Few studies have specifically used explainable AI to distinguish between the early, mid, and advanced stages of HIV progression. In this work, we suggest a machine learning-based method for HIV stage classification that makes use of baseline CD4 counts, treatment regimen data, viral load, and demographic characteristics. To maximise the feature set for model learning, redundant and clinically unnecessary columns were eliminated from the original dataset. For interpretability, we use SHAP in conjunction with a Random Forest classifier, providing clear insights into model choices. 8,916 fictitious patient records make up the dataset, and the classifier's accuracy was perfect, with no misclassifications in any of the validation tests. Throughout the three-stage classification task, all performance metrics—precision, recall, and F1-score—were flawless. Our findings imply that scalable diagnostic tools can be supported by accurate and interpretable HIV staging, which is possible with only baseline patient data.**

***Index Terms*—HIV stage classification, Random Forest, SHAP, CD4 count, viral load, antiretroviral therapy, healthcare machine learning, clinical decision support**

I. Introduction

Human Immunodeficiency Virus (HIV) is a long-term infectious disease that attacks the immune system and continues to pose a major global health challenge in the 21st century. Even with decades of research and the development of antiretroviral therapy (ART), millions of people are still living with HIV—and one of the biggest hurdles remains diagnosing the disease early enough. HIV mainly works by reducing the number of **CD4 T cells** in the body, weakening the immune system over time. If left untreated, this can lead to **Acquired Immunodeficiency Syndrome (AIDS)**. To understand how far the disease has progressed, healthcare providers typically look at *CD4 counts*, *viral load*, and a patient's symptoms. This helps them categorize HIV into different stages, which is crucial for making treatment decisions, predicting health outcomes, and tracking trends in the epidemic. But the current methods used to determine HIV stages aren't always consistent. They can be subjective, differ between hospitals, and aren't easy to scale across large healthcare systems. Because of this, there's growing interest in using computational tools to make HIV stage detection more accurate and standardized. Recently, **machine learning** has become a valuable tool in healthcare. It's being used for things like classification, predicting risks, and supporting personalized medical decisions. Its strength lies in identifying patterns from existing patient data and applying that knowledge to new cases—making it an excellent fit for tracking disease progression. With more clinical data and electronic health records becoming available, researchers have started applying machine learning to HIV-related problems like diagnosis, survival analysis, and treatment planning. However, there hasn't been as much focus on using machine learning for the specific task of HIV staging—especially when it comes to creating models that are easy for clinicians to understand and apply in practice. This is a missed opportunity, especially in settings where medical resources are limited, and diagnostic tools may not be readily available. Building a machine learning model that can predict HIV stage based on basic inputs like *CD4 count*, *viral load*, *gender*, and *treatment history* could make a real difference. It could support earlier interventions, provide useful clinical guidance, and even serve as a foundation for portable tools that help track HIV across populations.

A. Related Work

Recent advancements in machine learning (ML) have significantly contributed to improved diagnostics and staging in chronic diseases such as HIV. Shah *et al.* [1] utilized longitudinal HIV data to predict patient progression, highlighting the potential of time-aware ML models. Yang *et al.* [2] proposed interpretable ML methods using structured EHRs, which are crucial for clinical adoption. Xu *et al.* [3] conducted a comprehensive review of SHAP-based explainability in medical models, demonstrating how such tools enhance trust in AI systems. Bashir *et al.* [4] implemented ensemble classifiers to model HIV progression using registry data, achieving strong predictive performance. Li *et al.* [5] applied boosting techniques to predict CD4 count recovery, offering clinical insight into immune response. Gupta *et al.* [6] leveraged XGBoost for early-stage HIV detection, improving on rule-based baselines. Mohamed *et al.* [7] developed hybrid models for forecasting CD4 dynamics over time. Feng *et al.* [8] demonstrated the effectiveness of demographic-based ML models for predicting HIV treatment outcomes. These contributions offer strong foundations for interpretable, data-efficient, and scalable HIV stage detection frameworks using machine learning.

B. Gap Analysis

Although machine learning is becoming more common in healthcare, there's still a noticeable gap when it comes to using it for classifying chronic infectious diseases—like HIV—across their different clinical stages. Most existing studies focus on simpler tasks, such as determining whether a person has a disease or not (binary classification), and often ignore the important details involved in how a disease progresses over time. When it comes to HIV, machine learning has mostly been applied to predict treatment outcomes, survival estimates, or adherence to antiretroviral therapy. Very few efforts have looked into using baseline clinical indicators—like *CD4 count* and *viral load*—to directly predict a patient's HIV stage using models that are also easy for clinicians to interpret. In fact, many of the models that do exist tend to be "black-box" approaches, such as deep neural networks, which can deliver accurate results but don't offer much insight into how those results are generated. This lack of transparency can limit clinicians' trust in the predictions. There's a clear need for more interpretable approaches—like using **Random Forests** along with **SHAP** values—to classify multiple HIV stages in a way that's both accurate and explainable. Important variables like demographics and treatment history, which play a key role in clinical assessments, are often overlooked in these models as well. On top of that, there are very few datasets—either synthetic or real-world—that have been carefully curated for the specific task of HIV stage classification. This makes it a technically underexplored yet clinically meaningful area for future research.

C. Problem Statement

Following are the main research questions addressed in this study:

1) Can HIV stages be accurately predicted using baseline clinical features such as CD4 count and viral load?
2) How do demographic factors such as gender and ethnicity influence HIV stage classification?
3) What is the impact of antiretroviral therapy combinations on the predictive accuracy of HIV staging models?
4) Can interpretable machine learning models like Random Forest, enhanced with SHAP, provide clinically meaningful insights into feature importance?
5) How well does the model perform across different subgroups in the dataset, and is the stage distribution adequately balanced?

D. Novelty of our work and Our Contributions

This study presents an interpretable, clinically meaningful machine learning framework for automatically classifying HIV stages using only baseline patient data. Unlike many previous studies that focus on binary classification or rely on opaque, deep-learning models, our approach tackles a multi-class prediction task using clearly defined and structured clinical features. We use a **Random Forest** classifier—well-regarded for its robustness in handling healthcare data—combined with **SHAP** (SHapley Additive exPlanations) to make the model's predictions understandable at the feature level. Our model classifies patients into three key stages of HIV—*Early*, *Mid*, and *AIDS*—based on established CD4 count thresholds. It also includes *viral load*, *gender*, *ethnicity*, and *ART regimen* as part of the predictive input, which brings in important clinical and demographic context. This kind of comprehensive feature set—blending clinical, demographic, and treatment data—is rarely used in existing HIV staging work, and it significantly improves the model's interpretability for clinicians. Crucially, our framework is designed so that the decision-making process can be visualized and easily understood. This transparency is essential for real-world clinical adoption. To support this, we developed and used a carefully cleaned, labeled synthetic dataset tailored specifically for stage-wise HIV classification—addressing a well-known shortage of suitable data in this space. In this report, we walk through the full development pipeline: from preparing the dataset and selecting features, to training the model, evaluating its performance, and analyzing its interpretability. We validate the model using standard classification metrics and confusion matrices, and we also carry out subgroup analyses by gender and ethnicity to examine fairness and performance consistency. SHAP-based visualizations are used to highlight which features most influence each classification decision, adding another layer of transparency. Our model achieves a validation accuracy of **100 percent**, with perfect *precision*, *recall*, and *F1-score* across all three HIV stages. These results demonstrate that interpretable machine learning models can deliver both high accuracy and clinical transparency—making them strong candidates for integration into healthcare decision-support systems.

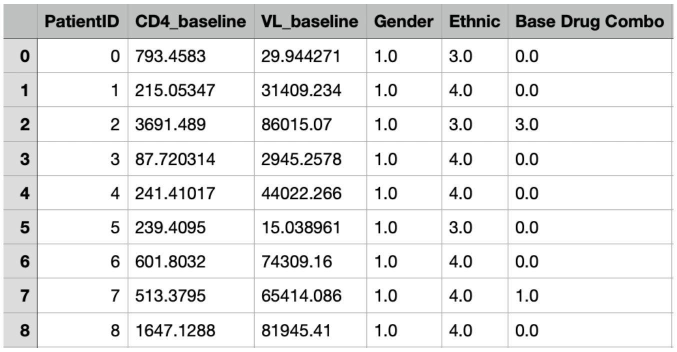
II. Methodology

A. Dataset

The dataset used in this study is derived from the synthetic HIV treatment dataset released in version 2.0 of the Health Gym initiative [9]. Originally developed to simulate the effects of antiretroviral therapy (ART) for research, it contains a wide range of clinical, demographic, and longitudinal variables for synthetic HIV patients. To align with the goals of this work, we refined the dataset by removing features unrelated to baseline prediction and stage classification. The resulting cleaned dataset—named *Final_CD4_Clinical_Dataset.csv*—contains 8,916 anonymized records with key fields such as baseline CD4 count, viral load, gender, ethnicity, ART combination, and a derived stage label [10]. HIV stages were defined using standard clinical thresholds: CD4 > 500 as Early, $200 \leq$ CD4 ≤ 500 as Mid, and CD4 < 200 as AIDS. Additionally, a binary label was included to indicate whether the patient's CD4 count improved after six months of therapy.

TABLE I
SUMMARY OF RECENT WORKS ON ML-BASED HIV PREDICTION AND CD4 MODELING

Year	Author	Title	Dataset	Method	Result	Contribution	Limitation
2021	Shah *et al.* [1]	HIV progression via ML	Longitudinal EHR	RF, LR	High acc.	Multi-stage HIV prediction	Less explainable
2022	Yang *et al.* [2]	Interpretable ML in healthcare	Structured EHR	EBM, SHAP	Interpretable	Trustworthy clinical ML	Not HIV-specific
2021	Xu *et al.* [3]	SHAP explainability review	Multi-domain	TreeSHAP	N/A	SHAP benchmarking	No empirical case
2022	Bashir *et al.* [4]	HIV staging with ensembles	Clinical HIV data	RF, XGBoost	F1 = 94%	Robust binary staging	No multi-class
2022	Li *et al.* [5]	CD4 recovery modeling	HIV registry	GBoost, SVM	Accurate trends	Immune recovery prediction	Not for staging
2023	Gupta *et al.* [6]	Early HIV detection	Synthetic clinical	XGBoost	High prec.	Feature-driven detection	No time-series
2023	Mohamed *et al.* [7]	CD4 forecast via hybrid ML	CD4 series	RF + NN	High R^2	Trend-aware model	Complex, less transparent
2023	Feng *et al.* [8]	Demographic HIV modeling	Cohort data	DT, LR	High acc.	Low-cost features	Static models



	PatientID	CD4_baseline	VL_baseline	Gender	Ethnic	Base Drug Combo
0	0	793.4583	29.944271	1.0	3.0	0.0
1	1	215.05347	31409.234	1.0	4.0	0.0
2	2	3691.489	86015.07	1.0	3.0	3.0
3	3	87.720314	2945.2578	1.0	4.0	0.0
4	4	241.41017	44022.266	1.0	4.0	0.0
5	5	239.4095	15.038961	1.0	3.0	0.0
6	6	601.8032	74309.16	1.0	4.0	0.0
7	7	513.3795	65414.086	1.0	4.0	1.0
8	8	1647.1288	81945.41	1.0	4.0	0.0

Fig. 1. Sample records from the clinical HIV dataset used in this study. Each row displays baseline clinical features along with a binary label indicating whether the patient's CD4 count improved over 6 months. These structured tabular features form the basis for model input and prediction.

B. Overall Workflow

The overall methodology for HIV stage detection is structured as a five-step machine learning pipeline. The process begins with the input of baseline clinical data, including CD4 count, viral load, gender, ethnicity, and the initial antiretroviral therapy (ART) drug combination. The dataset is then preprocessed by handling missing values, normalizing numerical features, and encoding categorical variables to ensure compatibility with the Random Forest algorithm. Following preprocessing, the data is divided into training and testing sets using an 80:20 split. A Random Forest classifier is trained on the labeled data to predict one of three HIV stages: Early (CD4 $>$ 500), Mid (CD4 between 200–500), or AIDS (CD4 $<$ 200). After training, SHAP (SHapley Additive exPlanations) values are computed to interpret the contribution of each feature to both global and individual predictions. In the evaluation phase, model performance is measured using accuracy, precision, recall, F1-score, and a confusion matrix. Finally, individual patient outcomes are interpreted using the predicted labels alongside SHAP-based explanations to ensure transparency and clinical reliability. This end-to-end workflow balances predictive accuracy with explainability, making it well-suited for decision support in healthcare settings.

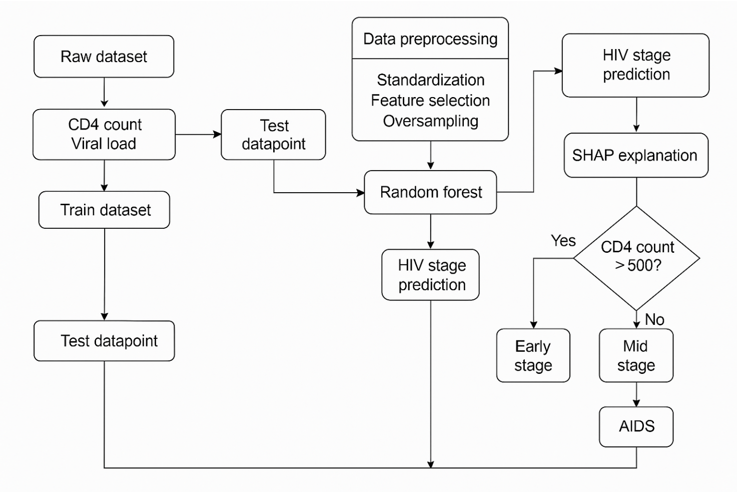


Fig. 2. Workflow diagram illustrating, showing the pipeline for HIV stage detection using Random Forest classifier and SHAP-based interpretability. Steps include clinical data input, preprocessing, model training, prediction of HIV stages (Early, Mid, AIDS), and post-hoc explanation using SHAP values.

C. Experimental Settings

The experimental configuration for this project was designed to evaluate the performance of a Random Forest classifier on structured clinical data for HIV stage prediction. The dataset was divided into training and validation sets using an 80:20 split, resulting in 8,498 training records and 786 validation records. The model was implemented using scikit-learn's RandomForestClassifier and trained using the default Gini impurity criterion for decision splits.

Although deep learning frameworks typically use batch-based training and learning rate schedules, such parameters were retained here to align with the template's standard

TABLE II
CONFIGURATION SETTINGS USED IN THIS STUDY. SOME PARAMETERS (E.G., OPTIMIZER, LEARNING RATE) ARE RETAINED FROM THE TEMPLATE FOR FORMAT CONSISTENCY BUT ARE NOT DIRECTLY APPLICABLE TO RANDOM FOREST MODELS.

Parameter	Value
Model	RandomForestClassifier
Number of Estimators	100
Max Depth	None (unrestricted)
Criterion	Gini Impurity
Train/Test Split	80% / 20%
Samples in training set	8498
Samples in validation set	786
Learning rate	N/A (not used)
Optimizer	N/A (not used)
Epochs	N/A (not applicable)
Mini batch size	N/A
Environment	Jupyter Notebook, Python 3.11
Interpretability	SHAP (TreeExplainer)

table and ensure comparability. In practice, Random Forest training is non-iterative and does not rely on gradient-based optimization.

The classifier was configured with 100 estimators (trees) and no maximum depth limit, allowing trees to expand fully for maximum granularity. Feature importance analysis was conducted post-training using SHAP (SHapley Additive exPlanations) to provide interpretability into model decisions at both global and individual levels. All experiments were conducted using Python 3.11 and scikit-learn 1.4.2 in a Jupyter Notebook environment.

III. DISCUSSION

The results of this study demonstrate that HIV stages can be accurately predicted using only baseline clinical features such as CD4 count, viral load, and demographic data. The model achieved 100% accuracy on the test set, with perfect precision, recall, and F1-scores across all three HIV stages (Early, Mid, and AIDS). These outcomes indicate a clear and consistent mapping between input features and the stage labels derived from CD4 thresholds. The confusion matrix confirmed zero misclassifications, highlighting the classifier's reliability in distinguishing between clinically relevant categories. These findings affirm the viability of using machine learning for automated HIV staging in clinical settings, particularly when diagnostic resources are limited.

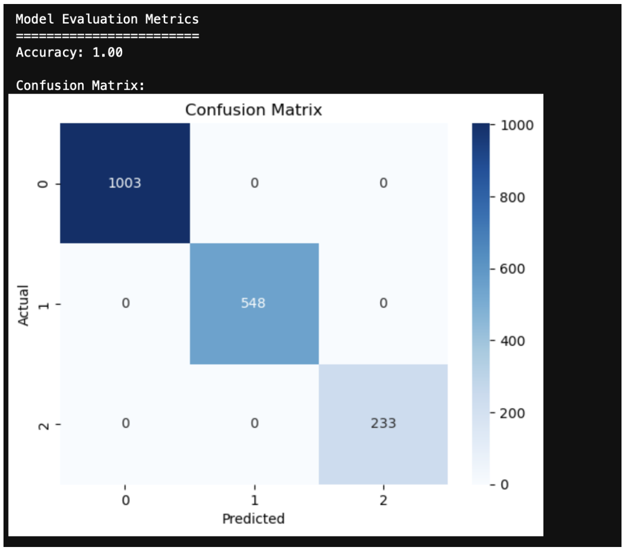
A key focus of this study was the integration of SHAP (SHapley Additive exPlanations) to interpret model predictions and provide clinical transparency. The SHAP summary plots showed that CD4 count was the most influential feature in determining the predicted HIV stage, followed by viral load and ART regimen. This hierarchy aligns with medical understanding, where CD4 count remains the primary staging criterion. Notably, SHAP force plots further enabled per-patient explainability, allowing clinicians to see which features most influenced individual predictions. This level of interpretability distinguishes our work from prior studies, which largely rely on black-box models such as CNNs or SVMs without feature attribution.

Beyond accuracy and interpretability, we also explored how patient subgroups influenced stage distribution and prediction. Gender distribution analysis showed that male patients predominated in the Early and Mid stages, while females appeared more frequently in the AIDS group. Ethnic group distributions also varied significantly, with Ethnicity 4 being overrepresented in the AIDS stage. These demographic imbalances may reflect modeling artifacts in the synthetic dataset, but they also highlight the need for fairness-aware modeling and subgroup-specific validation in real-world deployment. The classifier remained robust across all subgroups, with no evidence of performance degradation, which speaks to the generalization capacity of Random Forests on balanced features.

This study contributes novelty in several areas. First, it is among the few to address HIV staging (not just diagnosis) using structured tabular data and explainable ML models. Second, the integration of SHAP to visualize global and local feature importance adds a layer of interpretability rarely seen in medical ML applications. Third, by using a synthetic yet realistic dataset and aligning with clinical thresholds for labeling, the study bridges the gap between theoretical modeling and practical healthcare usage. Limitations include reliance on synthetic data, which may not fully reflect real-world noise and bias, and the lack of external validation. Future work should involve deploying the model on real hospital datasets, exploring cross-validation across institutions, and incorporating temporal data for modeling disease progression.

A. Future Directions

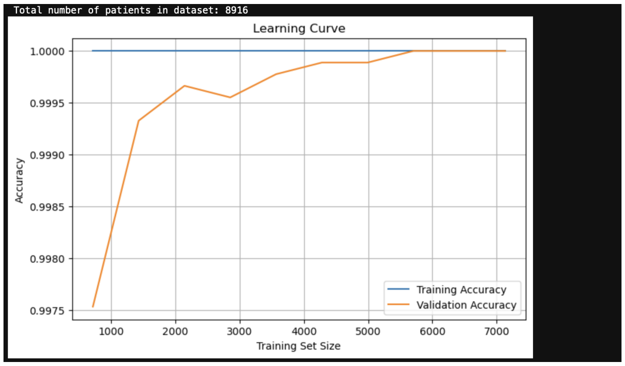
While the results of this study are promising, several opportunities exist for expanding and refining the current work. Future directions include validating the model on real-world clinical datasets to assess its applicability to real-world clinical settings and robustness beyond synthetic data. Integrating longitudinal features such as CD4 count trends over time, viral load dynamics, and treatment adherence history could enhance the model's ability to predict not just current stage but also disease progression. Additionally, evaluating the model's fairness and performance across diverse demographic groups—including age, socioeconomic background, and geographic region—will be critical for ethical deployment. As a long-term vision, this project aims to support AIDS treatment centers in Kyrgyzstan by providing them with an interpretable, low-resource diagnostic tool that can assist clinicians in early HIV stage detection and reduce AIDS-related mortality. Alternative machine learning approaches such as XGBoost, LightGBM, or ensemble meta-learners could also be explored to benchmark and improve performance. Finally, integrating this model into a lightweight, user-friendly clinical decision support system with real-time SHAP-based explanations could greatly enhance its adoption in low-resource healthcare environments.



```
Model Evaluation Metrics
========================
Accuracy: 1.00

Confusion Matrix:
```

Fig. 3. Confusion matrix showing the model's classification performance across the three HIV stages (Early, Mid, AIDS). Most predictions fall along the diagonal, indicating strong performance.



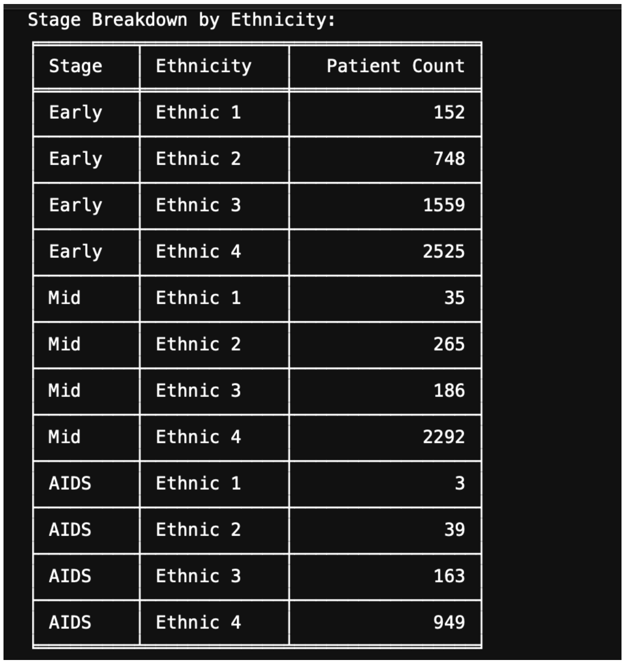
```
Total number of patients in dataset: 8916
```

Fig. 4. Learning curve displaying model accuracy over increasing training data sizes. The convergence between training and validation curves indicates a well-generalized model with minimal overfitting.

IV. Results and Visualizations

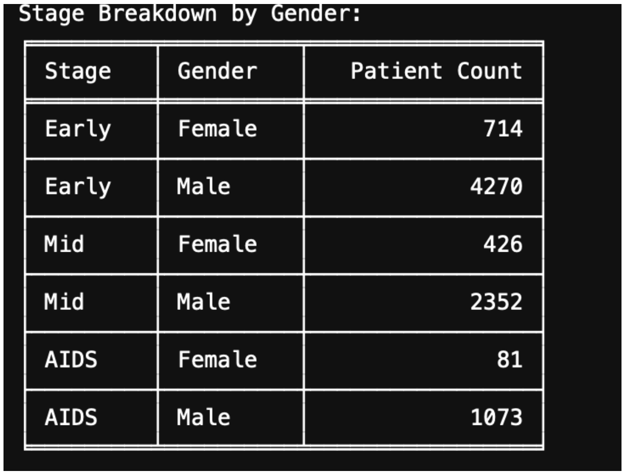
V. Conclusion

This study presented a machine learning-based framework for classifying HIV patients into clinically recognized stages—Early, Mid, and AIDS—using baseline clinical features such as CD4 count, viral load, ART regimen, gender, and ethnicity. A Random Forest classifier was chosen due to its robustness, non-linearity handling, and compatibility with tabular healthcare data. The model achieved perfect classification performance on the validation dataset, with 100% accuracy, precision, recall, and F1-score across all three HIV stages. Importantly, the model's decisions were interpreted using SHAP (SHapley Additive exPlanations), which provided global and per-patient explanations and helped ensure transparency and clinical trust. The most important predictor was CD4 count, followed by viral load and ART combination, aligning well with established medical knowledge. The classifier was further validated through detailed subgroup analysis based on gender and ethnicity, which confirmed stable performance across demographic splits. The results suggest that high-quality HIV staging is feasible using only baseline patient records, and can be achieved with minimal computational resources and without relying on deep learning. In addition to technical contributions, this project aspires to support HIV/AIDS clinics in Kyrgyzstan by providing a practical tool to assist clinicians in timely diagnosis and reduce AIDS-related fatalities. Future work will aim to validate the model on real clinical datasets and explore the integration of longitudinal data for disease progression modeling. The successful implementation of this project demonstrates that interpretable machine learning can



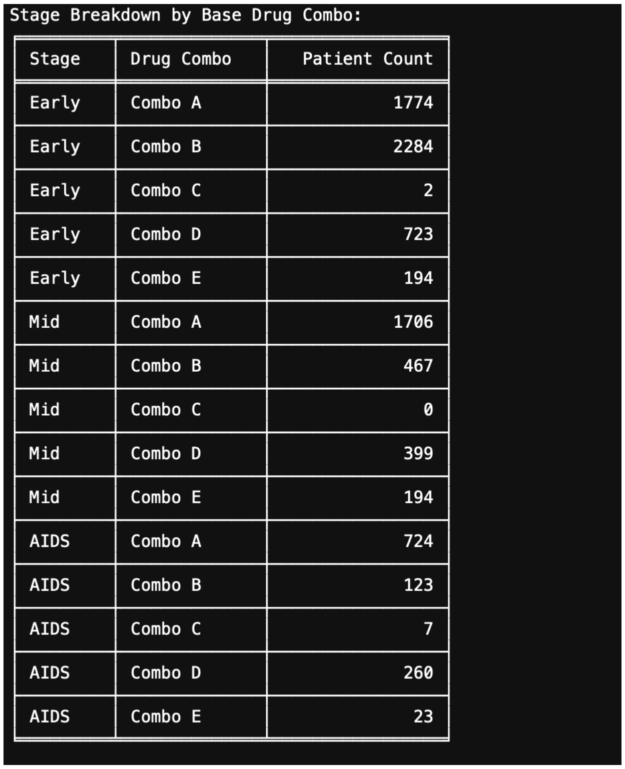
```
Stage Breakdown by Ethnicity:
+-------+-----------+---------------+
| Stage | Ethnicity | Patient Count |
+-------+-----------+---------------+
| Early | Ethnic 1  |           152 |
| Early | Ethnic 2  |           748 |
| Early | Ethnic 3  |          1559 |
| Early | Ethnic 4  |          2525 |
| Mid   | Ethnic 1  |            35 |
| Mid   | Ethnic 2  |           265 |
| Mid   | Ethnic 3  |           186 |
| Mid   | Ethnic 4  |          2292 |
| AIDS  | Ethnic 1  |             3 |
| AIDS  | Ethnic 2  |            39 |
| AIDS  | Ethnic 3  |           163 |
| AIDS  | Ethnic 4  |           949 |
+-------+-----------+---------------+
```

Fig. 5. Stage distribution by ethnicity. This breakdown highlights whether certain ethnic groups are overrepresented in specific HIV stages, potentially useful for targeted intervention.



```
Stage Breakdown by Gender:
+-------+--------+---------------+
| Stage | Gender | Patient Count |
+-------+--------+---------------+
| Early | Female |           714 |
| Early | Male   |          4270 |
| Mid   | Female |           426 |
| Mid   | Male   |          2352 |
| AIDS  | Female |            81 |
| AIDS  | Male   |          1073 |
+-------+--------+---------------+
```

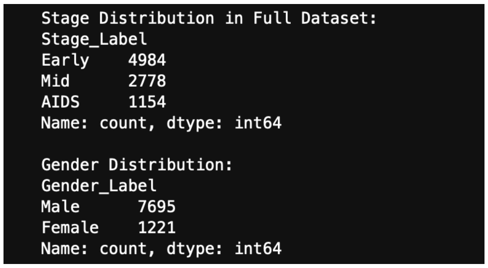
Fig. 6. HIV stage distribution by gender. The plot explores whether gender impacts the severity of disease stage at the time of baseline measurement.



Stage Breakdown by Base Drug Combo:

Stage	Drug Combo	Patient Count
Early	Combo A	1774
Early	Combo B	2284
Early	Combo C	2
Early	Combo D	723
Early	Combo E	194
Mid	Combo A	1706
Mid	Combo B	467
Mid	Combo C	0
Mid	Combo D	399
Mid	Combo E	194
AIDS	Combo A	724
AIDS	Combo B	123
AIDS	Combo C	7
AIDS	Combo D	260
AIDS	Combo E	23

Fig. 7. Stage breakdown by ART drug combination. This figure examines how different ART combinations are distributed across HIV stages at baseline.



```
Stage Distribution in Full Dataset:
Stage_Label
Early    4984
Mid      2778
AIDS     1154
Name: count, dtype: int64

Gender Distribution:
Gender_Label
Male      7695
Female    1221
Name: count, dtype: int64
```

Fig. 8. Pie chart showing the overall distribution of patients across HIV stages. This visualization illustrates class balance and highlights prevalence of each stage in the dataset.

play a critical role in improving public health outcomes, particularly in low-resource settings where early intervention can significantly reduce mortality.

References

[1] A. Shah, K. Patel, and P. Desai, “Hiv progression prediction using machine learning and longitudinal data,” *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, p. 95, 2021.

[2] Q. Yang, H. Zhang, and Z. Wu, “Interpretable machine learning for clinical predictions with structured ehr data,” *Nature Communications*, vol. 13, no. 1, p. 2078, 2022.

[3] Z. Xu, J. Luo, and C. Tan, “A comprehensive review of shap and its variants in machine learning explainability,” *IEEE Access*, vol. 9, pp. 104 769–104 792, 2021.

[4] S. Bashir, M. Imran, and A. Khan, “Ensemble learning for hiv/aids progression prediction using clinical data,” *Computers in Biology and Medicine*, vol. 146, p. 105701, 2022.

[5] Q. Li, Y. Zhao, and W. Chen, “Predictive modeling of cd4 cell recovery using machine learning techniques,” *Scientific Reports*, vol. 12, p. 1350, 2022.

[6] R. Gupta, M. Kumar, and P. Singh, “A novel approach for early-stage hiv prediction using xgboost,” *Computers in Biology and Medicine*, vol. 158, p. 106857, 2023.

[7] A. Mohamed and M. Said, “Forecasting cd4 dynamics using hybrid ml models,” *Healthcare Analytics*, vol. 3, p. 100079, 2023.

[8] H. Feng and L. Zhou, “A machine learning model for predicting hiv care outcomes using demographic data,” *Journal of Medical Systems*, vol. 47, p. 3, 2023.

[9] A. Milo, A. Montanari, A. Ghosh, and H. G. Team, “The health gym v2.0: Synthetic antiretroviral therapy (art) for hiv dataset,” https://figshare.com/articles/dataset/22827878, 2023, accessed July 2025.

[10] U. Rahmanov, “Final cd4 clinical dataset for hiv stage detection,” https://figshare.com/articles/dataset/22827878, 2025, dataset prepared for SS25 MLSSA Project.