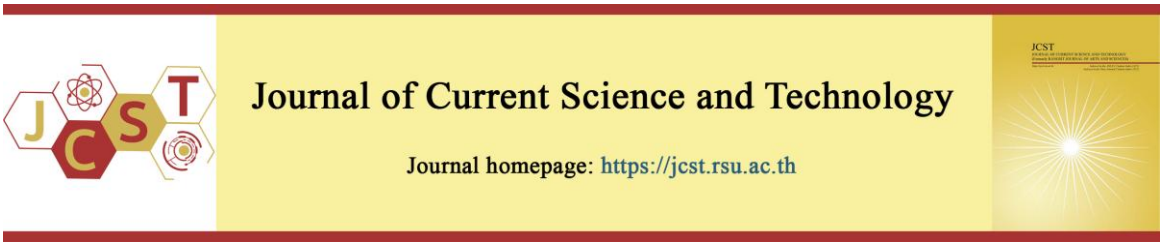
Journal of Current Science and Technology, January-March 2026 Vol. 16 No. 1, Article 158
ISSN 2630-0656 (Online)

Cite this article: Pechprasarn, S., Santichanyaphon, P., Triyaprasertporn, R., & Asawarachan, A. (2026). A machine learning approach to predicting survival of clinical AIDS patients via feature selection algorithms. *Journal of Current Science and Technology*, *16*(1), Article 158. https://doi.org/10.59796/jcst.V16N1.2026.158



A Machine Learning Approach to Predicting Survival of Clinical AIDS Patients via Feature Selection Algorithms

Suejit Pechprasarn[1, 2, *], Punn Santichanyaphon[3], Rawisara Triyaprasertporn[3], and Achiraya Asawarachan[3]

[1]College of Biomedical Engineering, Rangsit University, Pathum Thani 12000, Thailand
[2]Center of Excellence in AI and Supercomputing, Rangsit University, Pathum Thani 12000, Thailand
[3]Satriwithaya School, Wat Bowon Niwet, Phra Nakhon, Bangkok 10200, Thailand

*Corresponding author; E-mail: suejit.p@rsu.ac.th

Received 30 April 2025; Revised 2 September 2025; Accepted 15 October 2025; Published online 25 December 2025

Abstract

Numerous individuals deal with incurable illnesses, such as acquired immunodeficiency syndrome (AIDS), daily. One of the objectives of this study is to assist AIDS patients and healthcare professionals by creating selective machine-learning models in MATLAB R2024b, which could reduce the medical costs as well as the time required to examine the patients, to aid in personalizing patient treatment, and optimize healthcare management through feature selection techniques by lessening the number of attributes used in the operational model. The dataset named AIDS Clinical Trials Group Study 175 analyzed, is obtained from open-source Kaggle and consists of 23 predictors, including time, treatment indicator, age, weight, hemophilia, homosexual activity, history of IV drug use, Karnofsky score, History of non-ZDV antiretroviral therapy, ZDV history 30 days before dataset collecting period, ZDV history, antiretroviral therapy history, race, gender, symptom indicator, treatment indicator, treatment of off-treatment before 96 ± 5 weeks, CD4 count, and CD8 count. It was randomly split into a training and testing dataset in an 80:20 ratio to train 34 machine learning models and identify the best-performing model. Feature selection methods include Minimum Relevance and Maximum Relevance, Minimum Redundancy, Chi Square (χ^2), ANOVA, and Kruskal-Wallis to highlight the importance of each clinical feature. Here, the Boosted Tree (Tree) achieved the highest accuracy of 87.86%. Each model was then tested on the test dataset, and the results were compared with those from the previous procedure. The models also underwent feature-selection analysis to determine the significance of each predictor and the minimum number of predictors required to function efficiently. Finally, we conclude that the model with the best performance is the Linear SVM (SVM), with 85.0% accuracy and 5 or 6 predictors, including (1) time, (2) cd420, (3) karnof, (4) cd40, (5) z30, and (6) str2.

Keywords: *AIDS; algorithm; artificial intelligence; classification learner; feature selection; HIV; machine learning*

1. Introduction

Acquired immunodeficiency syndrome (AIDS) represents the most advanced stage of human immunodeficiency virus (HIV) infection, characterized by severe immunosuppression and increased susceptibility to opportunistic infections (Chanthara et al., 2025; Van Heuvel et al., 2022). HIV transmission occurs through sexual contact, exposure to contaminated blood or body fluids via needle sharing, and vertical transmission during pregnancy, delivery, or breastfeeding (Sahoo et al., 2017). As a member of the family Retroviridae, genus Lentivirus, HIV was first identified in 1983 (Gallo & Montagnier, 1987). Two distinct viral subtypes exist: HIV-1 and HIV-2, which originated from separate zoonotic transmissions of simian immunodeficiency virus (SIV) (Motomura et al., 2008). Despite structural similarities in their genetic material, these subtypes

exhibit substantial molecular divergence, sharing only approximately 60% amino acid sequence identity and 48% nucleotide sequence identity (Motomura et al., 2008).

The global HIV/AIDS epidemic continues to pose a critical public health challenge. As of 2022, approximately 40.4 million people were living with HIV worldwide (Wu et al., 2024), with no definitive cure currently available. In 2023 alone, HIV-related illnesses claimed 630,000 lives, while 1.3 million new infections were reported (UNAIDS, 2024). Although global prevention efforts have achieved substantial progress, new infection rates have declined by 60% since their 1995 peak (Payagala & Pozniak, 2024) the persistent disease burden necessitates improved prognostic tools and treatment optimization strategies.

Clinical management of HIV/AIDS relies heavily on CD4 and T-lymphocyte monitoring, as these immune cells are the primary targets of HIV infection. Normal CD4 counts range from 500 to 1,500 cells/mm³; a decline below 200 cells/mm³ indicates progression to AIDS and signals critical immunodeficiency (Battistini et al., 2023). While antiretroviral therapy (ART) has transformed HIV from a fatal diagnosis to a manageable chronic condition, significant treatment gaps persist. Although more than 26 million people living with HIV receive ART, over 12 million remain without access to life-saving treatment due to inadequate healthcare infrastructure and limited harm reduction programs in many regions (Kumah et al., 2023). These disparities in access to care, combined with the complexity of individualized treatment planning, underscore the need for advanced predictive tools to optimize resource allocation and personalize patient management. Machine learning approaches have shown promise in HIV prevention research, supporting diagnostic accuracy and informing clinical decision-making (Marcus et al., 2020).

Recent advances in artificial intelligence have demonstrated the utility of machine learning in HIV/AIDS care. For instance, predictive models have been deployed to identify high-risk populations for targeted testing interventions and to optimize prevention strategies through geospatial analysis (Bukachi et al., 2024). Similarly, AI-assisted spatial mapping has enabled identification of vulnerable communities requiring enhanced public health interventions (Tiribelli et al., 2024). Machine learning-based risk assessment tools have also been developed to predict HIV infection risk using large-scale clinical data. One notable example analyzed over one million patient visits to sexual health centers between 2008 and 2022, demonstrating the feasibility of personalized risk stratification to guide prevention efforts (Latt et al., 2024).

Despite these advances, machine learning applications in HIV/AIDS research have primarily focused on infection risk prediction and diagnostic support rather than prognostic modeling for patients already diagnosed with AIDS. Several studies have explored mortality prediction for specific HIV-associated opportunistic infections. For example, Shi et al. (2022) developed a machine learning model to predict mortality in HIV patients with Talaromycosis, a severe fungal infection caused by Talaromyces *marneffei*. While valuable, such disease-specific models have limited generalizability to the broader AIDS patient population and typically require extensive clinical parameters, potentially limiting their practical utility in resource-constrained settings.

A critical gap exists in the literature regarding machine learning models specifically designed to predict survival outcomes among AIDS patients using streamlined clinical features. This research addresses this gap by developing and validating machine learning models that predict survival duration in AIDS patients using a substantially reduced set of clinical predictors. Through systematic feature selection analysis, we identify the minimum number of clinical parameters required to maintain high predictive accuracy, thereby reducing diagnostic burden, lowering healthcare costs, and facilitating implementation in resource-limited settings. Furthermore, this study provides comprehensive insights into the relative importance of various HIV/AIDS-related clinical factors in survival prediction, establishing a methodological framework for future predictive analytics research in HIV/AIDS care. By demonstrating that accurate survival prediction can be achieved with fewer clinical inputs, our findings have the potential to improve treatment planning, optimize healthcare resource allocation, and enhance patient care outcomes, particularly in settings where comprehensive diagnostic testing may be unavailable or prohibitively expensive.

2. Objectives

1. To develop and validate a machine-learning model capable of predicting survival probability and personalized survival time for AIDS patients, as well as facilitating early disease detection.
2. To explore and gain in-depth knowledge about AIDS and HIV through algorithmic analysis of

patient data, which may assist in future treatment strategies.

3. To provide fundamental tools and knowledge related to machine learning for future clinical research on AIDS and HIV.

4. To develop an interpretable AI-based diagnostic tool with high accuracy for AIDS patients, based on everyone's profile.

3. Materials and Methods

This section describes the procedures and materials used in this study, including the predictors and labels in the dataset, the data source, the data handling process, and the software used. Furthermore, the procedures will be demonstrated as a flowchart in Figure 1 below.

3.1 Dataset and Study Population

This study utilized data from the AIDS Clinical Trials Group Study 175 (ACTG 175), a randomized, double-blind clinical trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults (Hammer et al., 1996). The dataset, accessed through the Kaggle open-source platform in February 2025 under the Open Database License (ODbL), contains clinical and demographic information from 2,139 patients. All participants had CD4 cell counts between 200 and 500 cells/mm³ at enrollment. The dataset comprises 23 predictor variables and one binary outcome variable (survival status at study conclusion).

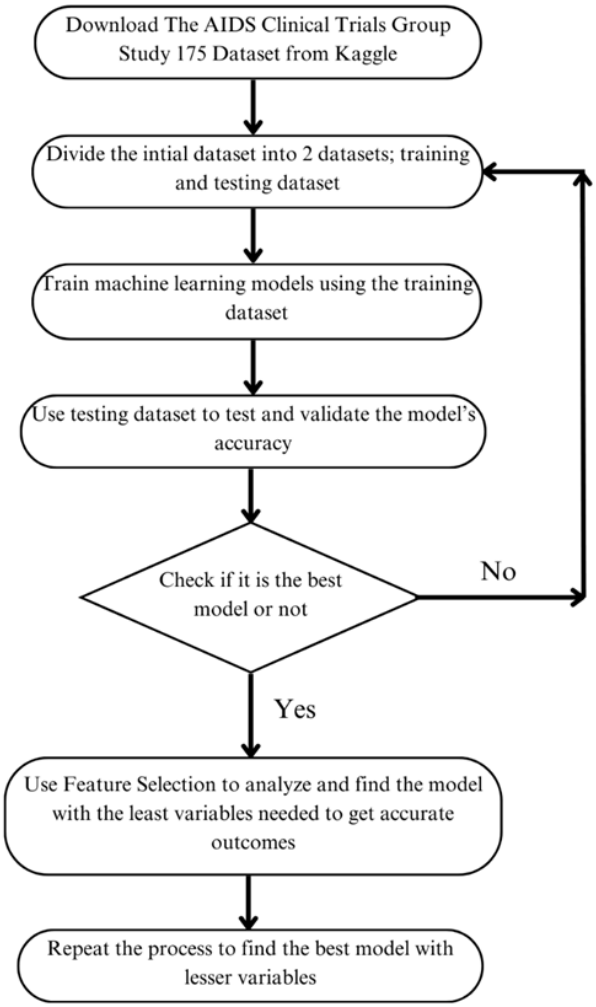


Figure 1 Flowchart representing the flow of data and procedures in the proposed system. The figure is modified from Pechprasarn et al. (2025)

The 23 predictor variables include: (1) time to study event or censoring (days), (2) age at enrollment (years), (3) weight (kg), (4) hemophilia diagnosis, (5) homosexual activity, (6) history of intravenous drug use, (7) Karnofsky performance score, (8) race, (9) gender, (10) antiretroviral therapy history prior to study, (11) symptomatic HIV indicator, (12) CD4 count at 96 ± 5 weeks, (13) CD4 count at baseline, (14) CD8 count at 96 ± 5 weeks, (15) CD8 count at baseline, (16) treatment indicator for non-ZDV antiretroviral therapy, (17) ZDV use in the 30 days before study enrollment, (18) days of prior ZDV use, (19) ZDV use indicator, (20) stratification indicator, (21) treatment indicator, (22) treatment of off-treatment status before 96 ± 5 weeks, and (23) days of prior non-ZDV antiretroviral use. Complete definitions are provided in Table 1.

The outcome variable indicates patient survival status at study conclusion: 0 = alive (n = 1,618; 75.6%) and 1 = deceased (n = 521; 24.4%). Although this dataset is publicly available and exempt from additional ethical review, all analyses were conducted in accordance with ethical principles for medical research involving human data.

3.2 Data Preprocessing and Class Imbalance Management

The substantial class imbalance in the original dataset (1,618 survivors vs. 521 deceased) posed a risk of model bias toward the majority class. Two primary approaches exist for addressing class imbalance: oversampling minority-class instances or undersampling majority-class instances. Oversampling techniques, such as Synthetic Minority Over-sampling Technique (SMOTE), generate synthetic minority-class samples but may introduce noise, increase the risk of overfitting, and perform poorly on high-dimensional data (Wongvorachan et al., 2023; Matharaarachchi et al., 2024).

Table 1 Predictors and their definition from the Kaggle AIDS Clinical Trials Group Study 175 Dataset

Predictors	**Details**
time	The number of days to failure or censoring.
trt	Treatment indicator. 0: ZDV only, 1= ZDV + ddl, 2=ZDV + zal, and 3= ddl only
age	Age in years.
wtkg	Weight in kilograms.
hemo	Hemophilia. 0 = No and 1 = Yes
homo	Homosexual activity. 0 = No and 1 = Yes
drugs	History of IV drug use. 0 = No and 1 = Yes
karnof	Karnofsky score on the scale of 1 to 100
oprior	Non-ZDV antiretroviral therapy was performed before the dataset collection period. 0 = NO and 1 = Yes
z30	ZDV in the 30 days before the dataset collection period. 0 = NO and 1 = Yes
zprior	ZDV before the dataset collection period. 0 = NO and 1 = Yes
preanti	The number of days before the dataset collecting period, the patient was on antiretroviral therapy.
race	Race. 0 = White and 1 = Non-white
gender	Gender. 0 = Female and 1 = Male
str2	Antiretroviral history. 0 = Naive and 1 = Exposed
strat	Antiretroviral stratification, 1=Antiretroviral naive, 2= >1 but ≤ 52 weeks of prior and 3= > 52 weeks of prior
symptom	Symptomatic indicator. 0: Asymptomatic and 1: Symptomatic
treat	Treatment indicator. 0 = ZDV only and 1 = Others
offtrt	An indicator of an off-treatment period before 96 ± 5 weeks. 0 = No and 1 = Yes
cd40	CD4 counts.
cd420	CD4 counts at 20 ± 5 weeks.
cd80	CD8 counts.
cd820	CD8 counts at 20 ± 5 weeks.
Label	**Details**
label	Patient's status. 0: Censoring and 1: Failure

We employed random undersampling of the majority class for three reasons: (1) the majority class contained sufficient cases for robust model training after reduction, (2) undersampling preserves the original data distribution without introducing synthetic cases, and (3) reduced dataset size improved computational efficiency while maintaining statistical power. While undersampling discards potentially informative majority class instances, the remaining sample size (832 patients for training) was adequate for the modeling objectives.

3.3 Training and Testing Dataset Partitioning

The original 2,139 patient records were partitioned into training and testing subsets using an 80:20 ratio, a standard practice in machine learning that balances model learning capacity with adequate validation sample size. The partition was implemented in two stages:

First, to create a balanced training set, we randomly selected 416 cases from each outcome class (832 total: 416 survivors, 416 deceased) using random undersampling of the majority class. This balanced distribution prevents model bias toward the majority class during training. A fixed random seed (not specified in the original implementation) was used to ensure reproducibility.

Second, the remaining 1,307 patient records (1,202 survivors, 105 deceased) constituted the independent testing dataset. Critically, the testing set intentionally preserved the original class distribution (approximately 92% survivors, 8% deceased) to reflect real-world population characteristics and provide a realistic evaluation of model performance in clinical settings.

All preprocessing and partitioning were conducted before model training to prevent data leakage, ensuring that no information from the testing set influenced model development.

3.4 Machine Learning Model Development and Training

We trained and evaluated 34 classification models representing diverse algorithmic approaches to identify the optimal model for predicting AIDS patient survival. Models were implemented using the Classification Learner Toolbox in MATLAB R2024b, which provides built-in hyperparameter optimization capabilities.

Model Selection and Categories: Models were selected from seven prominent algorithm families:

- Decision Trees: Fine Tree, Medium Tree, Coarse Tree
- Support Vector Machines (SVM): Linear, Quadratic, Cubic, Fine Gaussian, Medium Gaussian, Coarse Gaussian, Efficient Linear
- Logistic Regression: Binary logistic regression and Efficient
- Naive Bayes: Gaussian and Kernel variants
- K-Nearest Neighbors (KNN): Fine, Medium, Coarse, Cosine, Cubic, Weighted
- Ensemble Methods: Boosted Trees, Bagged Trees, RUSBoosted Trees
- Subspace Discriminant, Subspace KNN
- Neural Networks: Narrow, Medium, Wide, Bilayered, Trilayered configurations
- Discriminant: Linear, Quadratic
- Kernel: SVM, Logistic Regression

Hyperparameter Optimization: Automated hyperparameter optimization was performed for each model using Bayesian optimization with the following settings:

- Optimization objective: Minimize classification error
- Maximum iterations: 30 per model
- Cross-validation: 5-fold (described below)

The hyperparameters optimized varied by algorithm but typically included regularization parameters, kernel functions, number of neighbors, tree depth, learning rates, and ensemble size, where applicable.

Cross-Validation Strategy: All models were trained on the balanced training dataset (832 patients) using 5-fold cross-validation. This approach partitions the training data into five equal subsets, training each model five times on four subsets while validating on the remaining subset. The 5-fold configuration balances computational efficiency with reliable performance estimation and is standard practice for datasets of this size. Cross-validation performance metrics (Table 2) represent the average across all five folds.

Model Performance Evaluation: Following training and cross-validation, all 34 models were evaluated on the independent testing dataset (1,307 patients) to assess generalization performance. Model performance was quantified using five standard classification metrics, as shown in Equations (1) to (5).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (4)$$

$$\text{F1 score} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

TP and TN are regarded as true positive cases and true negative cases, respectively, and FP and FN are regarded as false positive and false negative cases, respectively.

Where TP = true positives (correctly predicted deceased), TN = true negatives (correctly predicted survivors), FP = false positives (survivors incorrectly predicted as deceased), and FN = false negatives (deceased incorrectly predicted as survivors).

These metrics provide complementary perspectives on model performance: accuracy indicates overall correctness, precision reflects positive prediction reliability, recall measures sensitivity to the minority class (deceased patients), specificity indicates majority class performance, and F1-score balances precision and recall. For clinical applications, recall (identifying patients at risk of death) and specificity (avoiding false alarms) are particularly important for resource allocation and treatment planning.

Model Selection Criteria: The optimal model was selected based on: (1) high testing accuracy, (2) minimal performance degradation between cross-validation and testing (indicating good generalization), (3) balanced performance across all metrics, and (4) stability across the cross-validation folds.

3.5 Feature Selection and Importance Analysis

To identify the minimum set of clinical predictors necessary for accurate survival prediction, we conducted systematic feature selection analysis using four complementary statistical methods: Minimum Redundancy Maximum Relevance (MRMR), Chi-Square (χ^2), Analysis of Variance (ANOVA), and Kruskal-Wallis test. These methods were selected to capture different aspects of feature importance: MRMR identifies features with maximum relevance to the outcome and minimum redundancy with other features, χ^2 and ANOVA evaluate association strength using parametric approaches, and Kruskal-Wallis provides a non-parametric alternative suitable for non-normal distributions.

Feature Importance Ranking: Each method independently ranked all 23 predictors according to their contribution to survival prediction, generating four separate importance rankings (Table 4). By comparing rankings across methods, we identified predictors consistently ranked as important, providing robust evidence of their prognostic value.

Sequential Feature Addition Analysis: Following feature importance ranking, we employed a sequential forward selection approach to determine the minimum predictor set. Starting with the highest-ranked feature, we iteratively added predictors one at a time according to their importance ranking. At each step, we retrained the best-performing model (Linear SVM, identified in Section 4.2) using only the selected features and evaluated performance on both training (via 5-fold cross-validation) and testing datasets.

This sequential analysis generated a series of models using 1, 2, 3, ..., up to 10 predictors, allowing us to identify the point at which additional features provided minimal performance improvement. The optimal feature set was defined as the minimum number of predictors that achieved performance comparable to the complete 23-predictor model (within 2–3% accuracy).

Consensus Feature Selection: To ensure robustness, we prioritized features consistently ranked highly across multiple selection methods. The final reduced predictor set was validated by: (1) achieving comparable or superior accuracy to the full model, (2) demonstrating consistent performance across cross-validation folds, and (3) showing stable rankings across different feature selection algorithms.

This comprehensive feature selection procedure serves two purposes: identifying the most prognostically important clinical factors and establishing a streamlined model suitable for resource-constrained clinical settings where comprehensive diagnostic testing may be impractical.

3.6 Statistical Analysis and Software

All data preprocessing, model training, cross-validation, and feature selection were performed using MATLAB R2024b (MathWorks, Natick, MA, USA) with the Statistics and Machine Learning Toolbox and Classification Learner App. Performance metrics were calculated using standard confusion matrix-based formulas. Model comparisons were based on point estimates of accuracy and other performance metrics, with full confusion matrices reported to enable comprehensive performance assessment.

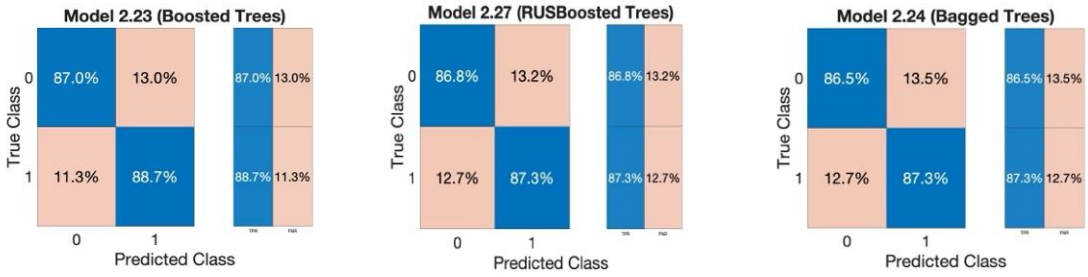
4. Results

4.1 Model Performance Using Complete Predictor Set

We trained and evaluated 34 machine learning models representing seven algorithm families using all 23 clinical and demographic predictors (Table 1). Training was conducted on the balanced dataset (832 patients) using 5-fold cross-validation. Complete training performance results for all models are presented in Table 2.

Three models achieved the highest cross-validation accuracy: Boosted Trees (87.86%), RUS Boosted Trees (87.02%), and Bagged Trees (86.90%), with their confusion matrices shown in Figures 2a-c. These ensemble methods substantially outperformed single-classifier approaches, with the next-highest accuracy being 85.00% for Linear SVM. Among individual algorithm families, Support Vector Machines demonstrated the strongest and most consistent performance, with Linear, Quadratic, and Cubic SVM variants achieving accuracy between 80–85%.

Three models Linear Discriminant Analysis, Quadratic Discriminant Analysis, and Gaussian Naïve Bayes could not be trained successfully due to mathematical constraints arising from the data structure and distributional assumptions and are marked as failed in Table



(a) (b) (c)

Figure 2 Confusion matrices of the top 3 models with the highest accuracy, trained with 23 predictors (a) Boosted Trees (Ensemble), (b) RUS Boosted Trees (Ensemble), and (c) Bagged Trees (Ensemble)

Table 2 Classification validation accuracy of models trained with 23 predictors using a 5-fold cross-validation method

Model	Details	Accuracy	Specificity	Precision	Recall	F1 Score	AUC	Average of Performance Metrics
Tree	Fine Tree	0.8389	0.8438	0.8422	0.8341	0.8382	0.8818	0.8398
	Medium Tree	0.8690	0.8702	0.8699	0.8678	0.8688	0.9115	0.8692
	Coarse Tree	0.8462	0.8293	0.8349	0.8630	0.8487	0.8542	0.8433*
Discriminant	Linear Discriminant	-	-	-	-	-	-	-
	Quadratic Discriminant	-	-	-	-	-	-	-
Binary GLM Logistic Regression	Binary GLM Logistic Regression	0.8522	0.8967	0.8865	0.8077	0.8453	0.9051	0.8608
Efficient Logistic Regression	Efficient Logistic Regression	0.8281	0.8606	0.8509	0.7957	0.8224	0.8943	0.8338
Efficient Linear SVM	Efficient Linear SVM	0.8233	0.8654	0.8530	0.7813	0.8156	0.8956	0.8307
Naïve Bayes	Gaussian Naïve Bayes	-	-	-	-	-	-	-
	Kernel Naïve Bayes	0.8005	0.8510	0.8342	0.7500	0.7899	0.8658	0.8089

Table 2 Cont.

Model	Details	Accuracy	Specificity	Precision	Recall	F1 Score	AUC	Average of Performance Metrics
SVM	Linear SVM	0.8377	0.8870	0.8747	0.7885	0.8293	0.9120	0.8470*
	Quadratic SVM	0.8341	0.8510	0.8458	0.8173	0.8313	0.9119	0.8370
	Cubic SVM	0.8065	0.8221	0.8164	0.7909	0.8034	0.8839	0.8090
	Fine Gaussian SVM	0.6058	0.3221	0.5675	0.8894	0.6929	0.7138	0.5962
	Medium Gaussian SVM	0.8486	0.8702	0.8643	0.8269	0.8452	0.9159	0.8525
	Coarse Gaussian SVM	0.8257	0.8966	0.8796	0.7548	0.8124	0.9035	0.8392*
KNN	Fine KNN	0.6851	0.6875	0.6860	0.6827	0.6843	0.6851	0.6853
	Medium KNN	0.7548	0.8389	0.8064	0.6707	0.7323	0.8143	0.7677
	Coarse KNN	0.7464	0.8510	0.8116	0.6418	0.7168	0.8255	0.7627
	Cosine KNN	0.7548	0.8365	0.8046	0.6731	0.7330	0.8197	0.7673
	Cubic KNN	0.7356	0.8173	0.7816	0.6538	0.7120	0.8021	0.7471
	Weighted KNN	0.7680	0.7981	0.7852	0.7380	0.7608	0.8214	0.7723
Ensemble	Boosted Trees	0.8786*	0.8702	0.8723	0.8870	0.8796	0.9390	0.8770
	Bagged Trees	0.8690*	0.8654	0.8663	0.8726	0.8695	0.9302	0.8683
	Subspace Discriminant	0.8329	0.8990	0.8837	0.7668	0.8211	0.8942	0.8456
	Subspace KNN	0.7897	0.8029	0.7975	0.7764	0.7868	0.8709	0.7916
	RUS Boosted Trees	0.8702*	0.8678	0.8684	0.8726	0.8705	0.9082	0.8697
Neural Network	Narrow Neural Network	0.8209	0.8125	0.8156	0.8293	0.8224	0.8477	0.8196
	Medium Neural Network	0.8065	0.8149	0.8117	0.7981	0.8048	0.8541	0.8078
	Wide Neural Network	0.8077	0.8149	0.8122	0.8005	0.8063	0.8725	0.8088
	Bilayer Neural Network	0.7993	0.8053	0.8029	0.7933	0.7981	0.8390	0.8002
	Tri-layered Neural Network	0.8005	0.7957	0.7976	0.8053	0.8014	0.8354	0.7998
Kernel	SVM Kernel	0.7813	0.7957	0.7896	0.7668	0.7780	0.8638	0.7833
	Logistic Regression Kernel	0.7849	0.7909	0.7883	0.7788	0.7836	0.8692	0.7857

* indicates trained model(s) with the top 3 performance.

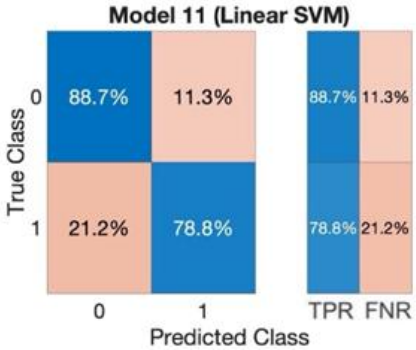


Figure 3 Confusion matrix for Linear SVM evaluated on the independent test dataset (1,307 patients) using all 23 predictors

4.2 Model Generalization and Selection for Feature Selection Analysis

All successfully trained models were evaluated on the independent testing dataset (1,307 patients) to assess generalization performance (Table 3). The testing phase revealed substantial performance degradation for several high-training-accuracy models, indicating overfitting. Most notably, the three top-performing models during training showed marked accuracy reductions when tested on unseen data: Boosted Trees decreased from 87.86% to 82.40% (Δ = −5.46%), RUS Boosted Trees from 87.02% to 80.15% (Δ = −6.87%), and Bagged Trees from 86.90% to 81.95% (Δ = –4.95%).

In contrast, several models demonstrated superior generalization stability. Coarse Tree achieved the highest testing accuracy at 84.05%, while Linear SVM obtained 82.25% accuracy with minimal degradation from its training performance (85.00% training vs. 82.25% testing; Δ = –2.75%). Table 3 presents complete performance metrics (accuracy, precision, recall, specificity, F1-score) for all models on the testing dataset.

To select the optimal model for subsequent feature selection analysis, we prioritized generalization stability over raw training accuracy, as models that generalize well are more likely to perform reliably in clinical settings. Linear SVM demonstrated the most consistent performance across training and testing phases, with performance degradation of less than 3% across all metrics. The model achieved 82.25% accuracy, 87.40% specificity, 85.95% precision, 77.10% recall, 81.29% F1-score, and 88.46% AUC on the testing dataset (Figure 3). The high specificity (87.40%) indicates a strong ability to correctly identify surviving patients, while moderate recall (77.10%) suggests some false negatives (deceased patients incorrectly predicted as survivors). Based on its balanced performance and excellent generalization, Linear SVM was selected for feature selection analysis.

Table 3 Testing performance of models evaluated on independent test dataset (n=1,307) using all 23 predictors

Model	Details	Accuracy	Specificity	Precision	Recall	F1 Score	AUC	Average of Performance Metrics
Tree	Fine Tree	0.7680	0.7930	0.7821	0.7430	0.7621	0.7769	0.7715
	Medium Tree	0.7985	0.8350	0.8220	0.7620	0.7909	0.8145	0.8044
	Coarse Tree	0.8405*	0.7860	0.8070	0.8950	0.8487	0.8407	0.8321
Discriminant	Linear Discriminant	-	-	-	-	-	-	-
	Quadratic Discriminant	-	-	-	-	-	-	-
Binary GLM Logistic Regression	Binary GLM Logistic Regression	0.8215	0.8530	0.8431	0.7900	0.8157	0.8766	0.8269
Efficient Logistic Regression	Efficient Logistic Regression	0.7975	0.8710	0.8488	0.7240	0.7814	0.8683	0.8103
Efficient Linear SVM	Efficient Linear SVM	0.7990	0.8650	0.8445	0.7330	0.7848	0.8792	0.8104
Naïve Bayes	Gaussian Naïve Bayes	-	-	-	-	-	-	-
	Kernel Naïve Bayes	0.7420	0.8360	0.7980	0.6480	0.7152	0.8186	0.7560
SVM	Linear SVM	0.8225*	0.8740	0.8595	0.7710	0.8129	0.8846	0.8318
	Quadratic SVM	0.8105	0.8400	0.8300	0.7810	0.8047	0.8723	0.8154
	Cubic SVM	0.7605	0.7880	0.7757	0.7330	0.7537	0.8196	0.7643
	Fine Gaussian SVM	0.6140	0.2760	0.5680	0.9520	0.7115	0.6781	0.6025
	Medium Gaussian SVM	0.8045	0.8280	0.8195	0.7810	0.7998	0.8630	0.8083
	Coarse Gaussian SVM	0.8115	0.8800	0.8610	0.7430	0.7976	0.8878	0.8239

Table 3 Cont.

Model	Details	Accuracy	Specificity	Precision	Recall	F1 Score	AUC	Average of Performance Metrics
KNN	Fine KNN	0.6435	0.6580	0.6478	0.6290	0.6383	0.6433	0.6446
	Medium KNN	0.6550	0.8150	0.7279	0.4950	0.5893	0.7595	0.6732
	Coarse KNN	0.7100	0.8390	0.7830	0.5810	0.6670	0.7632	0.7283
	Cosine KNN	0.6730	0.8030	0.7338	0.5430	0.6241	0.7620	0.6882
	Cubic KNN	0.6195	0.8010	0.6876	0.4380	0.5351	0.7196	0.6365
	Weighted KNN	0.6910	0.7340	0.7090	0.6480	0.6771	0.7613	0.6955
Ensemble	Boosted Trees	0.8240*	0.8190	0.8208	0.8290	0.8249	0.9070	0.8232
	Bagged Trees	0.8195	0.8490	0.8395	0.7900	0.8140	0.9031	0.8245
	Subspace Discriminant	0.7980	0.8720	0.8498	0.7240	0.7819	0.8701	0.8109
	Subspace KNN	0.7855	0.8150	0.8046	0.7620	0.7827	0.8637	0.7925
	RUS Boosted Trees	0.8015	0.8410	0.8274	0.7620	0.7933	0.8313	0.8080
Neural Network	Narrow Neural Network	0.7410	0.7390	0.7400	0.7430	0.7415	0.7711	0.7408
	Medium Neural Network	0.7730	0.7080	0.7416	0.8380	0.7869	0.8210	0.7651
	Wide Neural Network	0.7515	0.7700	0.7612	0.7330	0.7468	0.8167	0.7539
	Bilayer Neural Network	0.7260	0.8140	0.7743	0.6380	0.6996	0.7658	0.7381
	Tri-layered Neural Network	0.7635	0.7370	0.7502	0.7900	0.7696	0.8028	0.7602
Kernel	SVM Kernel	0.7180	0.7790	0.7483	0.6570	0.6997	0.8037	0.7256
	Logistic Regression Kernel	0.6760	0.7140	0.6905	0.6380	0.6332	0.7709	0.6796

* indicates trained model(s) with the top 3 performance.

Table 4 Feature importance rankings from four selection algorithms (MRMR, χ^2, ANOVA, Kruskal-Wallis) for all 23 predictors

No.	MRMR		χ^2		ANOVA		Kruskal Wallis	
1	time	0.3259	time	201.8875	time	225.3463	time	200.6717
2	race	0.1283	cd420	58.5601	cd420	69.744	cd420	73.0819
3	hemo	0.0187	karnof	20.4128	karnof	10.3323	cd40	21.3650
4	cd40	0.0186	cd40	18.9532	cd40	18.1431	karnof	20.9315
5	drugs	0.0185	z30	16.3283	z30	16.5460	z30	16.3105
6	strat	0.0105	str2	15.7498	str2	15.9491	str2	15.7326
7	treat	0.0104	preanti	14.0453	strat	13.4058	preanti	14.3845
8	symptom	0.0059	strat	13.8332	preanti	10.1935	strat	12.9096
9	cd420	0.0052	treat	8.8138	treat	8.8675	treat	8.8046
10	karnof	0.0050	trt	6.3411	drugs	6.3282	drugs	6.3066
11	age	0.0019	drugs	6.3130	offtrt	4.9914	offtrt	4.9808
12	offtrt	0.0012	offtrt	4.9857	cd80	4.6436	trt	4.5468
13	preanti	0.0008	symptom	4.3718	trt	4.4348	symptom	4.3676
14	oprior	0.0007	wtkg	2.6088	symptom	4.3744	cd80	3.6126
15	z30	0.0006	cd80	1.8004	age	2.1225	hemo	1.6455
16	str2	0.0004	hemo	1.6469	hemo	1.6446	age	1.3311
17	gender	0.0004	oprior	1.1714	oprior	1.1695	oprior	1.1705
18	homo	0	race	0.6639	cd820	0.8749	race	0.6634
19	cd80	0	gender	0.6298	race	0.6627	cd820	0.6364
20	wtkg	0	age	0.2666	gender	0.6287	gender	0.6293
21	cd820	0	cd820	0.1752	wtkg	0.1977	wtkg	0.1980
22	trt	0	homo	0.0696	homo	0.0695	homo	0.0696
23	zprior	0	zprior	0	zprior	0	zprior	0

4.3 Reducing the Number of Predictors Through Feature Selection Methods

To identify the minimum set of predictors required for accurate survival prediction, we applied four feature selection algorithms Minimum Redundancy MRMR, χ^2, ANOVA, and the Kruskal-Wallis test to rank all 23 predictors by their prognostic importance.

4.3.1 Consensus Findings Across Feature Selection Methods:

Table 4 reveals strong consensus across the four algorithms regarding the most important predictors. All four algorithms unanimously identified time (days to event or censoring) as the most important predictor, consistently ranking it first. This confirms time as the critical factor in survival prediction. Beyond this unanimous first-place ranking, the algorithms showed substantial convergence on the top-ranked predictors.

For the second most important predictor, three algorithms (χ^2, ANOVA, Kruskal-Wallis) ranked CD4 count at 96 ± 5 weeks (cd420) second, while MRMR placed it fourth, as detailed in Table 4. Similarly, Karnofsky performance score (karnof) appeared in the top four across all algorithms. Table 4 shows that χ^2 and ANOVA shared a common third feature, karnof, whereas MRMR's third feature was hemo (hemophilia), and Kruskal-Wallis identified cd40 (baseline CD4). The fourth most important feature for MRMR, χ^2, and ANOVA was cd40 (baseline CD4 count), while Kruskal-Wallis identified karnof.

According to Table 4, the six most consistently highly ranked predictors across methods were: (1) time, (2) cd420 (CD4 at 96 weeks), (3) karnof (Karnofsky score), (4) cd40 (baseline CD4), (5) z30 (ZDV use 30 days pre-enrollment), and (6) str2 (treatment stratification indicator). Notably, both χ^2 and ANOVA demonstrated the highest concordance, identifying identical top six predictors in the same order, as shown in Table 4. MRMR rankings diverged slightly by prioritizing race and hemophilia status higher than the other three methods. Kruskal-Wallis ranks closely aligned with χ^2 and ANOVA for the top predictors. As illustrated in Table 4, the value of the last predictor was zero across all algorithms, indicating negligible importance.

4.3.2 Sequential Feature Addition Analysis:

Table 5 presents four algorithms: MRMR, χ^2, ANOVA, and Kruskal-Wallis. This linear SVM model was trained by adding each parameter in turn, using predictors with 1 to 10 features. In each algorithm, the predictor order differed, as outlined in Table 4. The highest accuracy percentile was 85.0% in the fifth feature selection, which included χ^2, ANOVA, and Kruskal-Wallis, and in the sixth feature selection, which included ANOVA and Kruskal-Wallis. Compared to the other algorithms, MRMR demonstrated lower accuracy. ANOVA was determined to be the most reliable algorithm due to its highest accuracy. According to Table 4, the first six predictors in both χ^2 and ANOVA were identical.

4.4 Final Optimized Model Performance

The final Linear SVM model, trained using only the six identified predictors (time, cd420, karnof, cd40, z30, str2), achieved performance superior to the complete 23-predictor model. Training Performance (5-fold cross-validation): The 6-predictor model achieved 85.0% accuracy during cross-validation training (Figure 4), with specificity of 88.0%, precision of 87.2%, recall of 82.0%, F1-score of 84.5%, AUC of 86.9%, and overall average performance of 85.6%. This represents a 2.75% improvement over the 23-predictor Linear SVM model's training accuracy (82.25%).

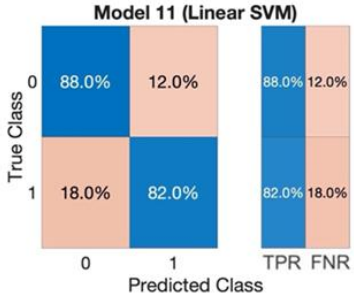


Figure 4 Training performance (5-fold cross-validation) of the optimized Linear SVM model using only 6 key predictors (time, cd420, karnof, cd40, z30, str2)

Table 5 presents the accuracy results of the Linear SVM model across each feature selection stage (1 to 10 predictors) for all four algorithms (MRMR, χ^2, ANOVA, and Kruskal-Wallis)

Feature Selection	MRMR	χ^2	ANOVA	Kruskal Wallis
1	74.6%	76.4%	68.9%	80.9%
2	81.7%	81.5%	81.5%	81.5%
3	81.7%	81.7%	81.7%	82.1%
4	81.6%	81.9%	81.9%	81.9%
5	81.6%	85.0%	85.0%*	85.0%
6	82.7%	84.9%	85.0%*	85.0%
7	82.9%	84.5%	84.0%	84.4%
8	82.8%	83.9%	83.9%	83.8%
9	83.7%	83.8%	83.8%	83.8%
10	83.5%	83.9%	83.3%	83.3%

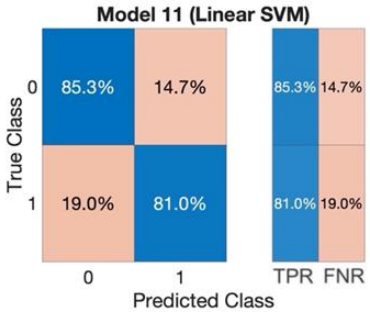


Figure 5 Testing performance of the optimized Linear SVM model using only 6 key predictors on the independent test dataset (1,307 patients)

After inputting the test dataset, we previously divided, the linear SVM model was trained with the six key predictors, as illustrated in Figure 5. The model achieved 83.15% accuracy (0.8315), with a specificity of 0.8530 and a precision of 0.8464. Its recall measure reached 0.8100, while the F1 score was 0.8278. The model maintained a substantial area under the curve (AUC) of 0.8692, resulting in an overall average performance rating of 0.8352.

When comparing the outcomes of the optimized machine learning model in this study with a parallel prior survey conducted in Ethiopia which employed 10-fold cross-validation classification despite the difference in the size of the training dataset (our study used a much smaller dataset), it was notable that both models yielded similar accuracy percentages (84.20% in the Ethiopian study and 83.15% in our study), using an identical number of predictors (six) in the final model. Moreover, other key distinctions between the two studies were also essential to consider during the analysis. The attributes utilized in both studies differed significantly due to distinct healthcare circumstances and populations. The research carried out in Ethiopia estimated the risk of death within five years from the initiation of antiretroviral therapy for HIV patients (Endebu et al., 2025). In contrast, our study aimed to estimate the time until death for HIV patients; therefore, it reported different results.

Our linear SVM model achieved reasonably high accuracy using only approximately a quarter of the original attributes. Consequently, patients require fewer diagnostic tests and could avoid unnecessary medical assessments, reducing healthcare expenses and the time needed. It also minimized unnecessary hospital stays, treatments, and emergency visits. Therefore, the model supports end-of-life planning for patients and their families. Beyond patient care, the finalized machine learning model also facilitated decision-making and management processes in hospitals.

As shown in Table 6, the optimized 6-predictor model achieved comparable training performance (85.00%) and superior testing performance (83.15% vs 82.25%) compared to the full 23-predictor model, while requiring only 26% of the original clinical parameters. This 74% reduction in required predictors substantially decreases diagnostic burden, data collection costs, and model complexity without sacrificing and indeed slightly improving predictive accuracy (Δ improvement of +0.90%).

Table 6 Performance Comparison Summary

Model Configuration	Training Accuracy	Testing Accuracy	Δ (Generalization)	Number of Predictors
Full 23-predictor Linear SVM	85.00%	82.25%	−2.75%	23
Optimized 6-predictor Linear SVM	85.00%	83.15%	−1.85%	6
Improvement	0%	+0.90%	+0.90%	−74% reduction

5. Discussion

This study developed and validated machine learning models to predict survival time in AIDS patients, with a primary focus on identifying the minimum set of clinical predictors required to maintain high prognostic accuracy. Our findings demonstrate that accurate survival prediction can be achieved using only six clinical parameters representing a 74% reduction from the original 23 predictors while maintaining or even slightly improving predictive performance compared to the complete predictor set.

5.1 Summary and Interpretation of Main Findings

5.1.1 Superior Performance of Reduced Predictor Model

The most significant finding of this study is that the optimized 6-predictor Linear SVM model achieved superior testing performance (83.15% accuracy) compared to the 23-predictor model (82.25% accuracy) while requiring substantially fewer clinical inputs. This counterintuitive result that fewer predictors yield better performance can be attributed to three factors:

First, reduced overfitting: Models with fewer parameters have less capacity to memorize patterns in the training data, forcing them to learn more generalizable relationships. Our results clearly demonstrated this principle: ensemble models with high training accuracy (87−88%) exhibited substantial performance degradation on the test data (80−82%), whereas the streamlined Linear SVM with moderate training accuracy (85%) maintained consistent performance on unseen data (83.15%). The minimal performance gap ($\Delta = -1.85\%$) indicates excellent generalization.

Second, feature selection identified truly informative predictors: The convergence of four independent feature selection algorithms (MRMR, χ^2, ANOVA, Kruskal-Wallis) on the same six predictors provides strong evidence that these variables capture the most prognostically relevant information. The unanimous identification of "time" as the most important predictor, followed by consistent ranking of CD4 counts and Karnofsky score, suggests these parameters reflect fundamental biological and clinical aspects of disease progression.

Third, removal of noise and redundancy: The original 23-predictor set likely contained correlated variables and parameters with minimal independent prognostic value. By retaining only the six most informative predictors, we eliminated noise that could obscure meaningful patterns, leading to a more robust and interpretable model.

5.1.2 Clinical Significance of the Six Key Predictors

The six predictors identified time, CD4 count at 96 weeks (cd420), Karnofsky performance score (karnof), baseline CD4 count (cd40), ZDV use 30 days pre-enrollment (z30), and treatment stratification (str2) represent a clinically coherent and biologically plausible prognostic signature:

- Time to event or censoring emerged as the most crucial predictor across all algorithms, reflecting the fundamental relationship between observation duration and survival outcomes in time-to-event analysis. This finding aligns with standard survival analysis principles, where longer event-free time indicates better prognosis.
- CD4 counts (both baseline and at 96 weeks) represent immune function status and trajectory. CD4 count is the primary clinical marker of HIV disease progression, with values below 200 cells/mm³ defining AIDS diagnosis. The inclusion of both baseline and follow-up CD4 counts allows the model to assess not just initial immune status but also immunological response to treatment, capturing disease trajectory rather than static status.
- Karnofsky performance score measures functional capacity and overall health status. This patient-reported outcome reflects the real-world impact of disease on daily functioning and has been consistently associated with survival across many diseases, including AIDS. Its prominence in our model confirms that functional status provides prognostic information beyond laboratory parameters.

- ZDV (zidovudine) uses 30 days pre-enrollment represents treatment history and may serve as a proxy for disease duration, previous treatment exposure, and potentially treatment adherence. Patients with recent ZDV use may differ systematically from treatment-naïve patients in ways that affect prognosis.
- Treatment stratification indicator likely captures clinical categorization based on disease severity or risk factors at enrollment, incorporating the clinical judgment used to stratify patients in the original trial.

Notably absent from the optimal predictor set are demographic variables (age, gender, race), behavioral factors (homosexual activity, IV drug use), and specific laboratory values (CD8 counts, weight), suggesting these contribute minimal independent prognostic information beyond the six selected predictors. This finding has practical implications: clinicians could make informed prognostic assessments that focus on immune function (CD4), functional status (Karnofsky), treatment history, and observation time, without requiring comprehensive demographic and behavioral histories.

5.1.3 Comparison of Model Approaches and Generalization Patterns

Our systematic evaluation of 34 models revealed important patterns regarding model complexity and generalization. Ensemble methods (Boosted Trees, Bagged Trees, RUSBoosted Trees) achieved the highest training accuracy (87−88%) but showed substantial overfitting, with testing accuracy dropping 5−7 percentage points. This pattern suggests that ensemble approaches, despite their theoretical advantages in reducing variance and bias, may be prone to overfitting in datasets of this size (832 training cases).

In contrast, simpler models, such as Linear SVM, demonstrated more consistent performance across training and testing phases. Linear SVM's strong generalization likely stems from its geometric approach to classification, which seeks an optimal separating hyperplane rather than memorizing complex decision boundaries. This finding suggests that for medical prediction tasks with moderate sample sizes, simpler models with strong regularization may be preferable to complex ensemble methods, prioritizing generalization over training accuracy.

The identification of Linear SVM as the optimal model also has practical advantages: SVMs are computationally efficient, have well-understood theoretical properties, and produce deterministic results (unlike some ensemble methods with stochastic components). These characteristics support reproducibility and clinical implementation.

5.2 Comparison to Previous AIDS Survival Prediction Studies

Our findings align with and extend previous research on machine learning for AIDS patient outcomes. Endebu et al. (2025) recently developed a machine learning model to predict loss to follow-up in HIV care in Ethiopia, achieving 84.20% accuracy with six predictors using 10-fold cross-validation. The convergence on six optimal predictors across two independent studies in different populations (U.S. clinical trial vs. Ethiopian routine care) and with different outcomes (survival time vs. loss to follow-up) provides strong external validation for the principle that HIV/AIDS outcomes can be accurately predicted using a small set of key clinical parameters.

However, significant differences exist between studies. Endebu et al. (2025) focused on predicting loss to follow-up within 5 years of ART initiation in a resource-limited setting, emphasizing healthcare engagement rather than mortality. Their predictor set likely included different variables relevant to treatment adherence and healthcare access. Our study addresses survival time prediction using data from a controlled clinical trial with standardized follow-up and comprehensive data collection. Despite these differences, both studies demonstrate that feature selection can identify minimal predictor sets without sacrificing accuracy, supporting the generalizability of this methodological approach across HIV/AIDS prediction tasks.

Shi et al. (2022) developed machine learning models to predict mortality in HIV patients with Talaromycosis, a specific opportunistic infection, using detailed clinical data. While their focus on disease-specific mortality in HIV patients relates to our work, their models required extensive clinical parameters specific to fungal infection diagnosis and treatment. Our broader focus on all-cause mortality in AIDS patients and successful reduction to six general clinical predictors represents a more generalizable approach applicable across diverse clinical settings, including resource-constrained environments where comprehensive diagnostic testing may be unavailable.

Previous studies applying machine learning to HIV have primarily focused on infection risk prediction (May et al., 2024; Volk et al., 2024) or diagnosis support rather than prognostic modeling for patients already diagnosed with AIDS. Our study fills a significant gap by demonstrating that survival prediction in AIDS patients can be achieved with high accuracy using a minimal, clinically obtainable predictor set.

5.3 Clinical Implications and Practical Applications

5.3.1 Resource Optimization in Clinical Settings

The 74% reduction in required predictors from 23 to 6 has substantial practical implications for clinical implementation, particularly in resource-constrained settings where comprehensive diagnostic testing may be limited by cost, availability, or infrastructure. Our findings suggest that clinicians could make informed prognostic assessments using only:

- Time since diagnosis or treatment initiation (readily available)
- CD4 counts at baseline and follow-up (standard HIV monitoring)
- Karnofsky performance score (simple clinical assessment)
- Basic treatment history (available from medical records)
- Treatment stratification category (clinical judgment)

This streamlined approach could enable prognostic assessment in settings lacking access to comprehensive laboratory panels, specialized testing, or detailed patient histories. The reduction in required data also decreases data collection burden, potentially improving data completeness and quality while reducing costs.

5.3.2 Treatment Planning and Resource Allocation

Accurate survival prediction using minimal clinical data could inform several clinical decisions:

Treatment intensification: Patients predicted to have poor survival outcomes might benefit from more aggressive treatment regimens, closer monitoring, or earlier referral to specialized care.

Clinical trial enrollment: Prognostic models could help identify appropriate candidates for clinical trials testing new interventions, ensuring trials enroll patients most likely to benefit or those at the highest risk.

Palliative care planning: For patients with poor predicted outcomes, early integration of palliative care services could improve quality of life and align care with patient preferences.

Healthcare resource allocation: In resource-limited settings, prognostic tools could help prioritize intensive interventions for patients most likely to benefit, optimizing population-level health outcomes.

However, we emphasize that our model provides population-level probability estimates and should inform not replace individualized clinical judgment. Prognostic predictions must be interpreted in the context of each patient's unique circumstances, preferences, and values.

5.3.3 Model Interpretability and Clinical Acceptance

An essential advantage of our 6-predictor Linear SVM model is its interpretability relative to complex ensemble methods or deep learning approaches. With only six inputs and a linear decision boundary, clinicians can understand which factors drive predictions and how changes in clinical parameters might affect prognosis. This transparency is crucial for clinical acceptance and trust, particularly in high-stakes medical decision-making.

The identified predictors also align with clinical understanding of AIDS progression: immune function (CD4 counts) and functional status (Karnofsky score) are already central to clinical assessment. The model thus formalizes and quantifies relationships that clinicians recognize intuitively, potentially increasing confidence in its predictions.

5.4 Study Limitations and Considerations

Several limitations warrant consideration when interpreting our findings:

5.4.1 Data Source and Generalizability

First, our study utilized data from a single clinical trial (ACTG 175) conducted in the 1990s. While this dataset provided high-quality, standardized data with complete follow-up, several factors may limit generalizability:

Temporal context: The data were collected during an earlier era of antiretroviral therapy, before the advent of highly active antiretroviral therapy (HAART) and modern treatment regimen. Treatment options, prognosis, and survival patterns have changed substantially since the 1990s. Our model's performance with contemporary patients receiving current-generation antiretrovirals remains unknown.

Population characteristics: Clinical trial participants may differ systematically from general AIDS patient populations. Trial enrollment criteria

(CD4 counts 200−500 cells/mm³) excluded patients with very advanced or early-stage disease, potentially limiting applicability to the full spectrum of AIDS patients.

Geographic and demographic scope: The trial was conducted in the United States with specific demographic characteristics. Model performance in other geographic regions, healthcare systems, or populations with different demographic compositions, comorbidity patterns, or healthcare access remains uncertain.

5.4.2 Methodological Considerations

Class imbalance handling: We employed random undersampling to address class imbalance, discarding 1,202 survivor cases to balance the training set. While this approach prevented majority class bias and improved computational efficiency, it sacrificed potentially informative data. Alternative methods, such as SMOTE or class weighting, might yield different results, though we chose undersampling to avoid introducing synthetic data artifacts.

Feature selection method dependence: While four feature selection algorithms converged on similar predictors, different selection methods or criteria might identify alternative predictor sets. Our choice of ANOVA as the primary algorithm was based on concordance with other methods and performance, but this represents one of several defensible choices.

Single train-test split: We employed a single 80:20 train-test split rather than multiple splits or nested cross-validation. While our testing set was substantial (1,307 patients) and independent, performance estimates might vary with different random splits. However, the consistent performance across 5-fold cross-validation training suggests results are reasonably stable.

Hyperparameter optimization: We relied on MATLAB's automated hyperparameter optimization with default settings. More extensive hyperparameter search or different optimization strategies might yield improved performance, though likely with diminishing returns given already-high accuracy.

5.4.3 Clinical Validation

Critically, our model has not been validated in prospective clinical use. Retrospective accuracy does not guarantee real-world utility or clinical impact. Prospective validation in contemporary AIDS patient cohorts, ideally across multiple sites and populations, is essential before clinical implementation. Such validation should assess not only predictive accuracy but also clinical utility, implementation feasibility, and impact on patient outcomes and healthcare delivery.

5.4.4 Missing Predictors and Data

Our analysis was constrained to the 23 predictors available in the ACTG 175 dataset. Contemporary clinical practice incorporates additional parameters that may improve prediction: viral load (not widely measured in the 1990s), comorbidity data, specific history of opportunistic infections, adherence measures, and social determinants of health. The absence of these variables represents both a limitation (potentially missing significant predictors) and a strength (demonstrating that useful predictions can be made with basic clinical data).

5.4.5 Model Simplicity Trade-offs

While our 6-predictor Linear SVM model offers advantages in simplicity and interpretability, more complex models (particularly deep learning approaches) might capture nonlinear interactions and complex patterns we did not assess. However, such complexity would sacrifice interpretability and likely require much larger training datasets than available in this study. Our choice prioritizes generalizability and clinical interpretability over potentially marginal gains in accuracy.

5.5 Future Research Directions

Several avenues for future research could address current limitations and extend this work:

5.5.1 Prospective Validation and Implementation Studies

The most critical next step is prospective validation in contemporary AIDS patient cohorts receiving modern antiretroviral therapy. Such validation should:

- Assess model performance with current-era treatment regimens
- Evaluate performance across diverse geographic, demographic, and healthcare settings
- Test clinical utility and impact on patient outcomes
- Identify any need for model recalibration or updating

Implementation research could assess the feasibility, acceptability, and impact of integrating this prognostic tool into clinical workflows, electronic health records, or clinical decision support systems.

5.5.2 Model Enhancement and Extension

Future work could extend our approach by:

- Incorporating additional predictors available in modern clinical practice (viral load, resistance testing, comorbidities, social determinants)
- Developing time-varying models that update predictions based on changing clinical status
- Creating risk stratification categories (low/ moderate/high risk) rather than continuous probability estimates
- Exploring nonlinear models or deep learning approaches while maintaining interpretability
- Developing ensemble approaches that combine multiple simple models

5.5.3 Mechanism Investigation

Our finding that 6 predictors outperform 23 suggests the presence of redundant or noisy variables. Future research could be conducted:

- Correlation structures among predictors
- Identification of which predictors provide redundant information
- Understanding why certain intuitively important variables (age, gender) provide minimal independent prognostic value
- Biological or clinical mechanisms underlying the prognostic importance of the six selected predictors

5.5.4 Comparative Effectiveness Research

Studies comparing ML-based prognostic predictions to existing clinical assessment methods (e.g., physician judgment, simple scoring systems) would establish the added value of our approach and identify optimal strategies for combining ML predictions with clinical expertise.

Health Economics Analysis

A cost-effectiveness analysis could quantify the economic value of reduced diagnostic requirements by comparing healthcare costs and outcomes between comprehensive assessment and our streamlined 6-predictor approach. Such analysis would inform policy decisions about resource allocation in different healthcare contexts.

Extension to Related Conditions

The methodological approach developed here systematic feature selection to identify minimal predictor sets could be applied to prognostic prediction in other chronic diseases, opportunistic infections in HIV, or other AIDS-related outcomes (treatment failure, loss to follow-up, quality of life). Comparative studies across conditions could identify common principles for effective medical prediction modeling.

6. Conclusion

This study demonstrates that accurate survival prediction in AIDS patients can be achieved using only six clinical predictors (time, CD4 counts at baseline and 96 weeks, Karnofsky performance score, ZDV use history, and treatment stratification), representing a 74% reduction from the original 23-predictor set. Notably, this streamlined 6-predictor Linear SVM model achieved superior testing performance (83.15% accuracy) compared to the complete 23-predictor model (82.25% accuracy), challenging the assumption that more data necessarily improves prediction. These findings have significant practical implications for resource-constrained clinical settings where comprehensive diagnostic testing may be unavailable or cost-prohibitive. By demonstrating that survival prediction requires only basic immune function markers (CD4 counts), functional status assessment (Karnofsky score), and treatment history all readily obtainable in routine clinical practice this work provides a foundation for developing practical prognostic tools that balance accuracy with feasibility. The methodological contribution extends beyond AIDS research: systematic feature selection successfully identified a minimal predictor set that outperformed the whole feature space, illustrating the value of parsimony in medical prediction modeling. This approach of prioritizing generalization over training accuracy through model simplification may inform predictive modeling across diverse medical applications.

However, important limitations, particularly the use of historical trial data from the 1990s and the absence of prospective validation in contemporary patient populations, necessitate cautious interpretation. Future research validating this approach with modern antiretroviral regimens and diverse patient populations will be essential to translate these findings into clinical practice and realize the potential of streamlined, accurate AIDS survival prediction for improving patient care and resource allocation.

7. Abbreviations

Abbreviation	Full Form
ACTG	AIDS Clinical Trials Group
AI	Artificial Intelligence
AIDS	Acquired Immunodeficiency Syndrome
ANOVA	Analysis of Variance
ART	Antiretroviral Therapy
AUC	Area Under the Curve
CD4	Cluster of Differentiation 4 (T-lymphocyte cells)
CD8	Cluster of Differentiation 8 (T-lymphocyte cells)
χ^2	Chi-Square
ECG	Electrocardiography
FN	False Negative
FP	False Positive
HAART	Highly Active Antiretroviral Therapy
HIV	Human Immunodeficiency Virus
HIV-1	Human Immunodeficiency Virus Type 1
HIV-2	Human Immunodeficiency Virus Type 2
HMIS	Health Management Information System
IV	Intravenous
KNN	K-Nearest Neighbors
MATLAB	Matrix Laboratory (software)
ML	Machine Learning
MRI	Magnetic Resonance Imaging
MRMR	Minimum Redundancy Maximum Relevance
ODbL	Open Database License
PCR	Polymerase Chain Reaction
PCP	Pneumocystis jirovecii Pneumonia
PrEP	Pre-Exposure Prophylaxis
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
RUS	Random Under-Sampling
SIV	Simian Immunodeficiency Virus
SMOTE	Synthetic Minority Over-sampling Technique
STI	Sexually Transmitted Infection
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
ZDV	Zidovudine (also known as AZT)

8. CRediT Statement

Suejit Pechprasarn: Conceptualization, Methodology, Investigation, Writing – Review & Editing, Project Administration, Supervision.

Punn Santichanyaphon: Methodology, Investigation, Formal Analysis, Writing – Original Draft, Writing – Review & Editing

Rawisara Triyaprasertporn: Methodology, Formal Analysis, Writing – Original Draft, Writing – Review & Editing

Achiraya Asawarachan: Methodology, Formal Analysis, Writing – Original Draft, Writing – Review & Editing, Visualization

9. Acknowledgements

We acknowledge and are grateful to the National Institute of Allergy and Infectious Diseases (NIAID), Katzenstein D, Hammer S, and the AIDS Clinical Trials Group Study 175 Study Team for providing us with the dataset obtained through the open-source Kaggle platform. We also sincerely thank Satriwithaya School and the Institute of Research, Rangsit University, for the research and publication opportunity.

We acknowledge that this research was primarily designed, written, and conducted by the authors, with AI tools used to assist in refining language and presentation.

10. References

Battistini Garcia, S. A., Zubair, M., & Guzman, N. (2023). *CD4 Cell Count and HIV*. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK513289/

Bukachi, S. A., Onono, J., Onyango-Ouma, W., Onyango, T., Jeptoo, M., Yussuf, B., ... & Richards, S. (2024). Opportunities, gaps, and challenges in the implementation of the One Health approach in Kenya. *One Health Cases*, 2024, Article ohcs20240019. https://doi.org/10.1079/onehealthcases.2024.0019

Chanthara, C., Khattiya, J., Roytrakul, S., et al. (2025). Plasma proteomic signatures in HIV-infected individuals post–SARS-CoV-2 infection. *BMC Infectious Diseases*. https://doi.org/10.1186/s12879-025-12307-1

Endebu, T., Taye, G., & Deressa, W. (2025). Development of a machine learning prediction model for loss to follow-up in HIV care using routine electronic medical records in a low-resource setting. *BMC Medical Informatics and Decision Making*, *25*(1), Article 192. https://doi.org/10.1186/s12911-025-03030-7

Gallo, R. C., & Montagnier, L. (1987). The chronology of AIDS research. *Nature*, 326, 435–436. https://doi.org/10.1038/326435a0

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., ... & Merigan, T. C. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, *335*(15), 1081-1090. https://doi.org/10.1056/NEJM199610103351501

Kumah, E., Boakye, D. S., Boateng, R., & Agyei, E. (2023). Advancing the global fight against HIV/Aids: Strategies, barriers, and the road to eradication. *Annals of Global Health*, *89*(1), Article 83. https://doi.org/10.5334/aogh.4277

Latt, P. M., Soe, N. N., King, A. J., Lee, D., Phillips, T. R., Xu, X., ... & Ong, J. J. (2024). Preferences for attributes of an artificial intelligence-based risk assessment tool for HIV and sexually transmitted infections: a discrete choice experiment. *BMC Public Health*, *24*(1), Article 3236. https://doi.org/10.1186/s12889-024-20688-2

Marcus, J. L., Sewell, W. C., Balzer, L. B., & Krakower, D. S. (2020). Artificial intelligence and machine learning for HIV prevention: Emerging approaches to ending the epidemic. *Current HIV/AIDS Reports*, *17*(3), 171-179. https://doi.org/10.1007/s11904-020-00490-6

Matharaarachchi, S., Domaratzki, M., & Muthukumarana, S. (2024). Enhancing SMOTE for imbalanced data with abnormal minority instances. *Machine Learning with Applications*, *18*, Article 100597. https://doi.org/10.1016/j.mlwa.2024.100597

May, S. B., Giordano, T. P., & Gottlieb, A. (2024). Generalizable pipeline for constructing HIV risk prediction models across electronic health record systems. *Journal of the American Medical Informatics Association*, *31*(3), 666-673. https://doi.org/10.1093/jamia/ocad217

Motomura, K., Chen, J., & Hu, W. S. (2008). Genetic recombination between human immunodeficiency virus type 1 (HIV-1) and HIV-2, two distinct human lentiviruses. *Journal of Virology*, *82*(4), 1923-1933. https://doi.org/10.1128/jvi.01937-07

Payagala, S., & Pozniak, A. (2024). The global burden of HIV. *Clinics in Dermatology*, 42(2), 119-127. https://doi.org/10.1016/j.clindermatol.2024.02.001

Pechprasarn, S., Srisaranon, N., & Yimluean, P. (2025). Optimizing diabetes prediction: an evaluation of machine learning models through strategic feature selection. *Journal of Current Science and Technology*, *15*(1), Article 75. https://doi.org/10.59796/jcst.V15N1.2025.75

Sahoo, C. K., Sahoo, N. K., Rao, S. R. M., & Sudhakar, M. (2017). A review on prevention and treatment of AIDS. *Pharmacy & Pharmacology International Journal*, *5*(1), Article 00108. https://doi.org/10.15406/ppij.2017.05.00108

Shi, M., Lin, J., Wei, W., Qin, Y., Meng, S., Chen, X., ... & Jiang, J. (2022). Machine learning-based in-hospital mortality prediction of HIV/AIDS patients with Talaromyces marneffei infection in Guangxi, China. *PLoS Neglected Tropical Diseases*, *16*(5), Article e0010388. https://doi.org/10.1371/journal.pntd.0010388

Tiribelli, S., Pansoni, S., Frontoni, E., & Giovanola, B. (2024). Ethics of artificial intelligence for cultural heritage: Opportunities and challenges. *IEEE Transactions on Technology and Society*, *5*(3), 293 – 305. https://doi.org/10.1109/TTS.2024.3432407

UNAIDS. (2024). *Global HIV & AIDS statistics-Fact sheet*. In *The urgency of now: 2024 global AIDS update*. Retrieved from https://www.unaids.org/en/resources/fact-sheet

Van Heuvel, Y., Schatz, S., Rosengarten, J. F., & Stitz, J. (2022). Infectious RNA: Human immunodeficiency virus (HIV) biology, therapeutic intervention, and the quest for a vaccine. *Toxins*, *14*(2), Article 138. https://doi.org/10.3390/toxins14020138

Volk, J. E., Leyden, W. A., Lea, A. N., Lee, C., Donnelly, M. C., Krakower, D. S., ... & Silverberg, M. J. (2024). Using electronic health records to improve HIV preexposure prophylaxis care: A randomized trial. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, *95*(4), 362-369. https://doi.org/10.1097/QAI.0000000000003376

Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data

mining. *Information*, *14*(1), Article 54. https://doi.org/10.3390/info14010054

Wu, X., Zhou, X., Chen, Y., Lin, Y. F., Li, Y., Fu, L., ... & Zou, H. (2024). Global, regional, and national burdens of HIV/AIDS acquired through sexual transmission 1990–2019: An observational study. *Sexual Health*, *21*(5), Article SH24056. https://doi.org/10.1071/SH24056