Predicting AIDS Using Machine Learning in HIV Clinical Trials

Karen Lee
ICS 435 Project
Spring 2024

The dataset consists of healthcare statistics collected from patients diagnosed with acquired immunodeficiency syndrome (AIDS), initially published in 1996[1]. This dataset includes input features such as treatment indicators, baseline health metrics (e.g., weight, age), and risk factors (e.g., drug use, hemophilia). The dataset is composed of several categorical and continuous variables that provide a view of each patient's status and treatment history at the time of data collection.

Originally, the data was collected to compare the efficacy of nucleoside monotherapy versus combination therapy in managing HIV-1 infection in adults with CD4 cell counts between 200 to 500 per cubic millimeter. This randomized, double-blind study aimed to determine whether combination therapy could more effectively slow the progression of HIV disease, improve survival rates, and reduce the incidence of AIDS-defining illnesses compared to monotherapy alone. However, for the purposes of this project, the data is used to detect whether a patient was diagnosed with acquired immunodeficiency syndrome (AIDS) given the input features (see Table 1). The output is a binary classification indicating whether the patient is infected with AIDS.

ML can handle interactions between many variables and learn non-linear patterns from large datasets, which traditional statistical methods might struggle with. The function being approximated by the model is a mapping from the input features space to expressing the likelihood of AIDS infection.

The data that was published in the UC Irvine Machine Learning Repository[2] contains a total of **2139 subjects** with **23 features** including the target variable–whether they were diagnosed with AIDS–for this project. In total, to train the models, 22 features were used to predict whether a subject had AIDS. This published dataset is what was used for this project.

---

[1] 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.

[2] https://archive.ics.uci.edu/dataset/890/aids+clinical+trials+group+study+175 using the pip install ucimlrepo

Attribute	Description	Value / Scale
time	Time to failure or censoring	Continuous (time)
trt	Treatment indicator	0 = ZDV only 1 = ZDV + ddI 2 = ZDV + Zal 3 = ddI only
age	Age at baseline	Continuous (years)
wtkg	Weight at baseline	Continuous (kg)
hemo	Hemophilia status	0 = No 1 = Yes
homo	Homosexual activity	0 = No 1 = Yes
drugs	History of IV drug use	0 = No 1 = Yes
karnof	Karnofsky score	0-100
oprior	Non-ZDV antiretroviral therapy pre-175	0 = No 1 = Yes
z30	ZDV in the 30 days prior to 175	0 = No 1 = Yes
preanti	Days pre-175 anti-retroviral therapy	Continuous (days)
race	Race	0 = White 1 = Non-white
gender	Gender	0 = Female 1 = Male
str2	Antiretroviral history	0 = Naive 1 = Experienced
strat	Antiretroviral history stratification	1 = Antiretroviral Naive 2 = > 1 but <= 52 weeks of prior therapy 3 = > 52 weeks
symptom	Symptomatic indicator	0 = Asymptomatic 1 = Symptomatic
treat	Treatment indicator	0 = ZDV only 1 = Others
offtrt	Indicator of off-treatment before 96+/-5 weeks	0 = No 1 = Yes
cd40	CD4 at baseline	Continuous (count)
cd420	CD4 at 20+/-5 weeks	Continuous (count)
cd80	CD8 at baseline	Continuous (count)
cd820	CD8 at 20+/-5 weeks	Continuous (count)
infected	Is infected with AIDS	0 = No 1 = Yes

Table 1. List of features in the dataset

I chose to use three model types: Random Forest (RF); Support Vector Machine (SVM); and Neural Network (NN). These models are known to perform well when handling a combination of categorical and continuous data as well as incorporating non-linear relationships; there was the added benefit of not needing to scale any of the continuous variables to form approximately normal or uniform distributions. No transformations, padding, imputation, or augmentation was applied to the data. One-hot encoding of categorical variables was attempted, but this did not appear to enhance performance.

Data processing, model training, and model evaluation were conducted within a Google Colab. The data was first randomly split 90 to 10, the 10% (n = 214) being the test data. To evaluate the ML models, 5-fold cross validation was used on the remaining 90% (n = 1925)

of the data to ensure generalizability of the model and to prevent overfitting. RF and SVM models were trained using scikit-learn version 1.2.2; the NN, on Keras and TensorFlow version 2.15.0. The loss functions for RF and the NN were set to gini coefficient and binary cross entropy. For RF and SVM models, the hyperparameters were optimized by grid search and the NN, by hand tuning.

For the RF classifier, number of trees (n_estimators) {100, 150, 200, 250, 300}, maximum tree depth (max_depth) {6, 8, 10}, minimum samples of leaf nodes for an internal split (min_samples_leaf) {2, 4, 6}, and minimum samples required to be a leaf (min_samples_leaf) {1, 2} were explored. A total of 90 RF classifiers models were trained with 450 cross-validation fits. The model that yielded the highest validation AUROC was selected to be evaluated on the set aside test data.

For the SVM classifier the parameters that were searched were the type of kernel {'rbf', 'poly'} and C {0.1, 1, 10}. Preliminary exploration of SVM parameters revealed linear kernels taking too long to train thus were ejected during the grid search. A total of 6 SVMs were trained with 30 cross-validation fits. The model that yielded the highest validation AUROC was selected to be evaluated on the set aside test data.

In contrast to the other two classifiers, the NN was hand tuned to train an uncounted number of models (over 20) until resulting in the desired learning curve (Figure 1) by varying: the number of hidden layers; layer dimensions; l2 regularization and dropout layers, if overfitting; Leaky ReLUs to improve gradients if models failed to train; activation functions; learning rates; and batch sizes.

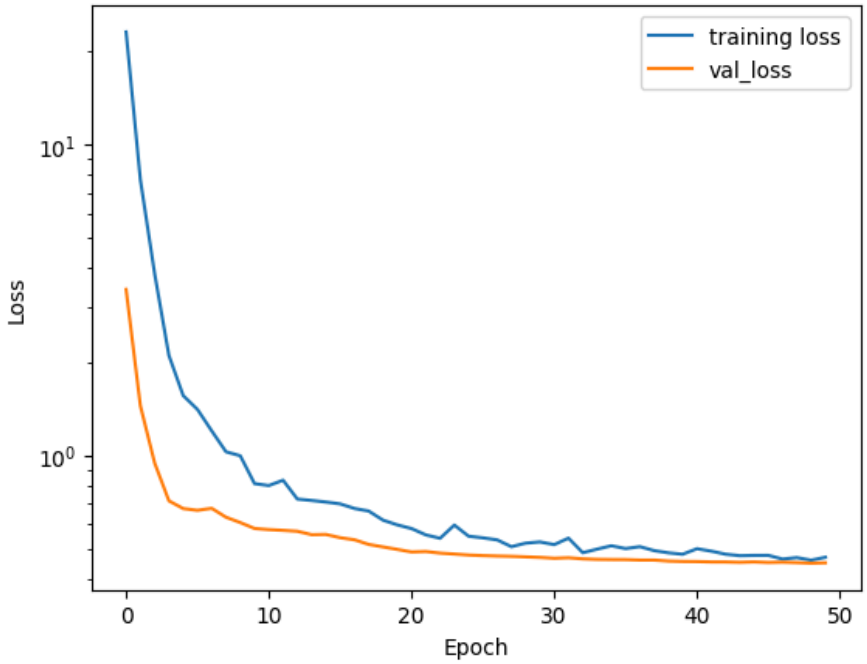

Figure 1. Learning Curve for the last iteration of the NN. "Loss" is binary cross entropy loss.

The final iteration of the NN (Figure 2) includes three dense layers with 16, 8, and 16 neurons respectively, using the GELU activation function, interspersed with dropout layers set at 10% and 30% to reduce overfitting. The model uses a sigmoid activation function in the output layer to output probabilities, is compiled with the Adam optimizer and binary cross entropy loss. This model was used to evaluate the set aside test data.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 16)                368

 dropout (Dropout)           (None, 16)                0

 dense_1 (Dense)             (None, 8)                 136

 dropout_1 (Dropout)         (None, 8)                 0

 dense_2 (Dense)             (None, 16)                144

 dropout_2 (Dropout)         (None, 16)                0

 dense_3 (Dense)             (None, 1)                 17

=================================================================
Total params: 665 (2.60 KB)
Trainable params: 665 (2.60 KB)
Non-trainable params: 0 (0.00 Byte)
```

Figure 2. Neural Network Model Summary

Results:

The models were evaluated on the held-out test dataset (n = 214). The performance of the best RF classifier was 0.96 AUROC (Figure 3).

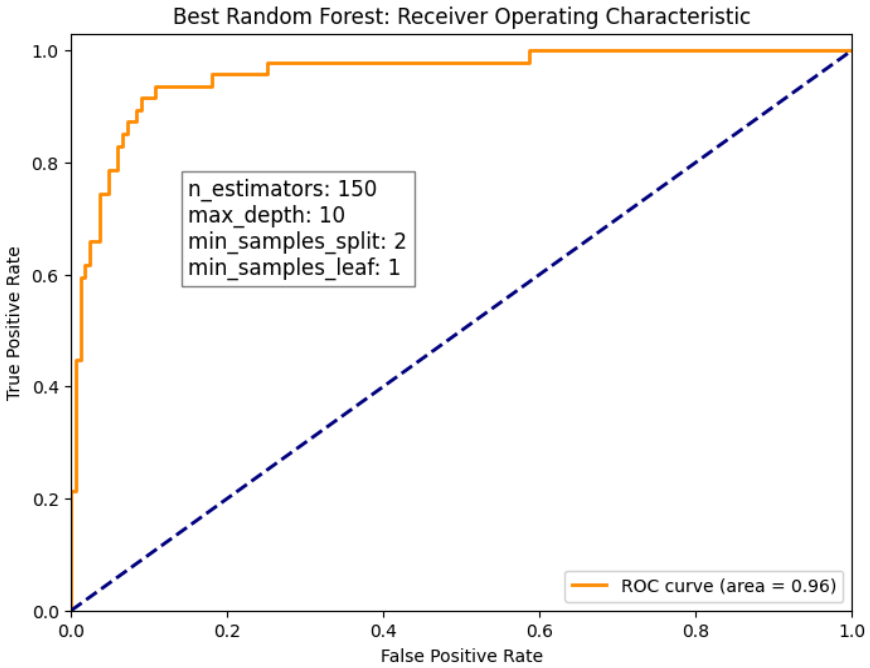

Figure 3. ROC of the best RF classifier

For the best SVM classifier, their performance on the test dataset was 0.94 AUROC (Figure 4).

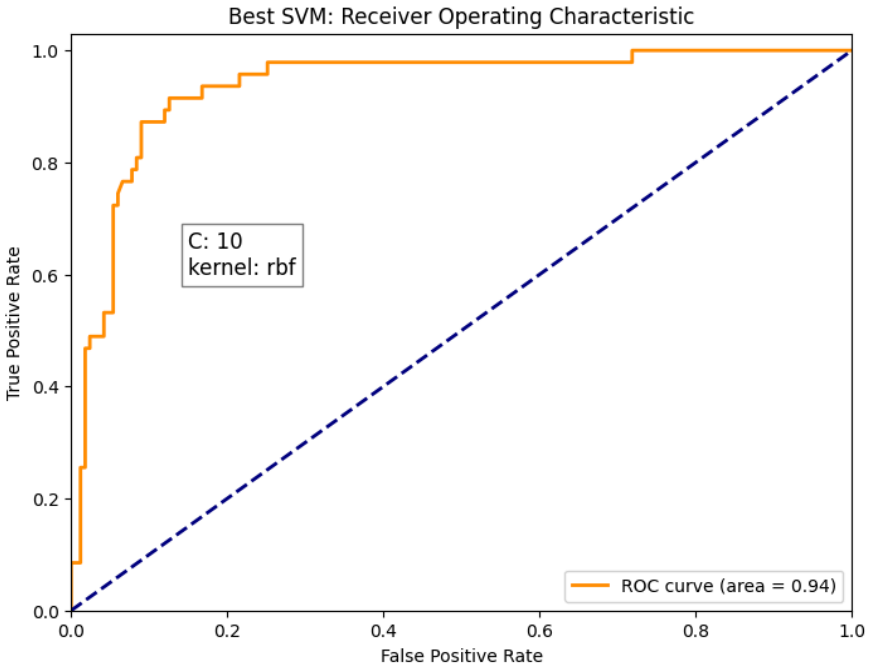

Figure 4. ROC of the best SVM classifier

For the last iteration of the NN, the performance on the test dataset was 0.89 AUROC (Figure 5).

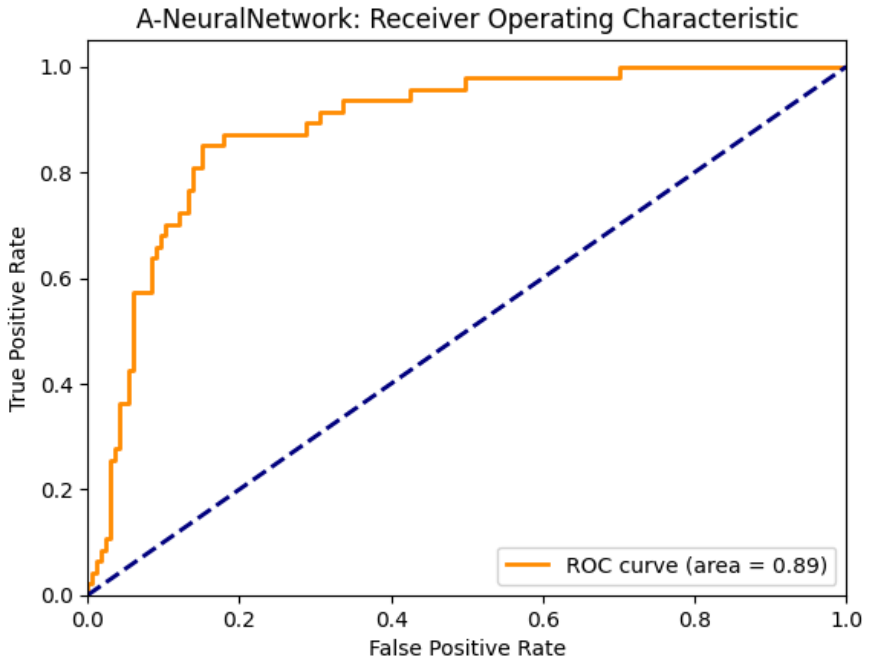

Figure 5. ROC of the final NN

The summary of model performance on the test dataset (Table 2) reveals that the best-performing RF classifier achieved an AUROC of 0.96. The next highest scores were recorded by the best SVM classifier at 0.94 AUROC, followed by the NN, which was an AUROC of 0.89.

Model	AUROC
Random Forest	0.96
Support Vector Machine	0.94
Neural Network	0.89

Table 2. Summary of performance for each ML model

I believe that these models would perform comparably on a dataset of HIV+ individuals with AIDS from the mid-1990s. However, their generalizability to current HIV+ populations are likely to be very limited. Modern advances in treatment, unavailable during the original study, have altered the age distribution of individuals with AIDS by delaying its onset due to improved treatment options. HIV now impacts a wider demographic: the virus has reached regions of the developing world where the treatments from this study might not be accessible. I would expect that these models would underperform on today's populations if they were applied using the same features on which they were originally trained.