R.I.P HIV: Predicting antibody neutralization against HIV using Machine Learning

SUBMITTED BY

Lilian Nogueira

Intro

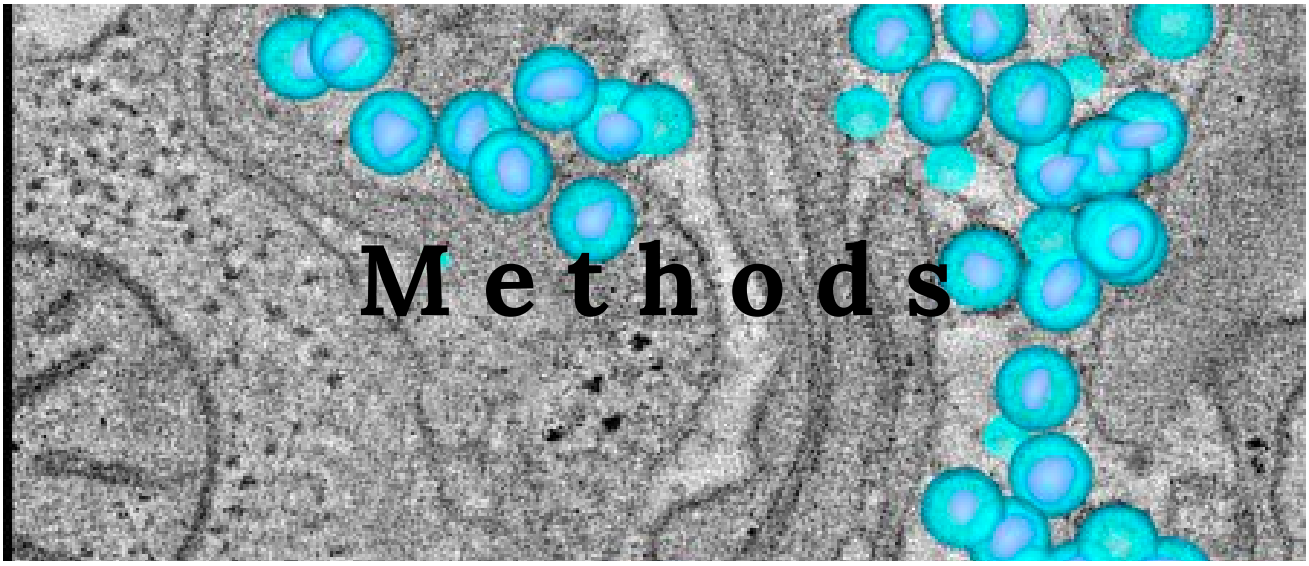
Broad neutralizing antibodies (bNAbs) are a type of antibodies that can recognize and neutralize a wide range of HIV strains by targeting conserved regions on the HIV envelope protein, offering potential for the development of effective therapies and vaccines against the virus.

By identifying and characterizing these potent antibodies, this work has the potential to advance HIV treatment and prevention strategies, improve public health outcomes, contribute to scientific knowledge, and create economic value through the development of novel therapies and potential commercial opportunities.

Applying data science techniques in the field of identifying and characterizing bNAbs is advantageous due to the complexity of HIV-related data, the need for big data analysis, the potential for machine learning and predictive modeling, and the ability to integrate diverse data sources.

Some machine learning models and techniques have been employed to address various aspects of the problem.

Gradient Machine Boost model was already used to integrate features such as sequence information, antibody properties, and viral characteristics to predict the neutralization potency of antibodies against specific HIV strains. The two models we are using in this project (Extreme Gradient Boost and Neural Networks) will be a new way to evaluate this problem.



Data Source:

HIV SEQUENCE DATABASE - CATNAP

https://www.hiv.lanl.gov/components/sequence/HIV/neutralization/index.html

Data Cleaning and EDA

The data downloaded from the CATNAP are three different tables, containing data from antibodies, viruses strains and assays done. The viral genome sequences are stored in a Fasta file (type of file are generated by a sequencer after sequence a genome).

After selecting the features to use in this project, the final data frame have the following columns: ANTIBODY, VIRUS, BINDING TYPE, SEQUENCE AND IC50.

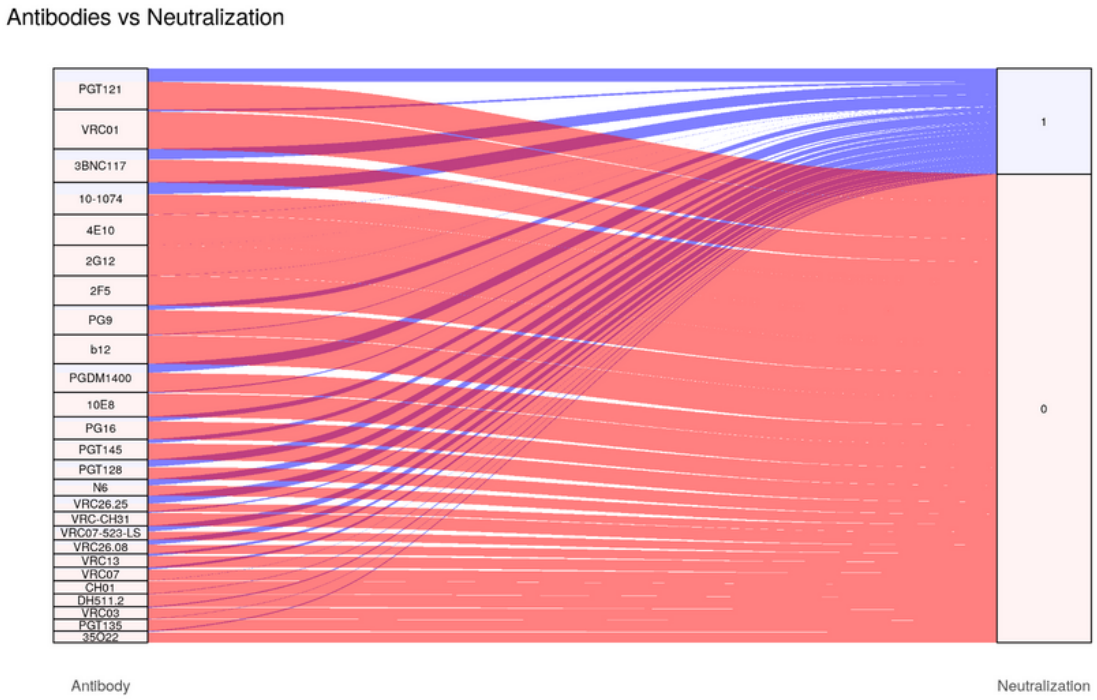
IC50 or half maximal inhibitory concentration (IC50) is a measure of the potency of a substance in inhibiting a given biological process or biological component by 50%. Lower values of IC50 means more potent inhibitory action. This is our target variable.

ANTIBODY and VIRUS columns are the IDs.

BINDING TYPE is the kind of binding the antibody forms with the virus. SEQUENCE contains the Envelope amino acid sequences. Envelope is a protein that forms the spike in the virus and it is where the antibody connects to the virus.

Our target variable is stored as a concentration and therefore, a float. Antibodies considered good virus neutralizers present IC50 values <= 0.5 microgram/ ml. In this case, I created a dummy version of the IC50 column called Neutralization, in the way that any IC50 values greater than 0.5 (non-neutralizer) receives 0 and IC50 values equal or lower than 0.5 receives 1.

Because we have many antibodies that were tested against only a few virus, I decided to keep only the antibodies that were tested against at least 300 virus, as we can check in the next plot.



As we can see below, the blue lines represents all the viruses neutralized (1) by a given antibody and the red ones represent not neutralized viruses (0). Since antibodies 4E10, 2G12 and 2F5 have close to zero number of viruses neutralized, I decided to remove them from the analysis as well to decrease the effect of a imbalanced dataset.

I also applied one-hot encoder for all my independent variables, generating a total of 20,823 features. The algorithm for one-hot encoder for amino acids was adapted from https://github.com/ronakvijay/Protein_Sequence_Classification.

Machine Learning Models

I first tried Extreme Gradient Boost (XGBoost). Gradient boosting is a technique that combines multiple weak learners sequentially, where each new model is trained to correct the errors made by the previous models.

XGBoost improves upon traditional gradient boosting algorithms by employing several optimizations and enhancements, resulting in faster training and better accuracy.

I also used Bayes Search for hyper parameters tuning. This approach uses stepwise Bayesian Optimization to explore the most promising hyper parameters in the problem-space. It gives the benefit that we can give a much larger range of possible values, since over time we automatically explore the most promising regions and discard the not so promising ones. Plain grid-search would need ages to stupidly explore all possible values.

The second model I decided to employ in our analysis was Neural Networks. They consist of interconnected artificial neurons organized into layers. The layers include an input layer, hidden layers, and an output layer. Neural networks learn from data by adjusting the weights of connections between neurons to minimize a loss function

Results

XGBoost

The accuracy of this model was 84%.

When we have unbalanced datasets, like the one we are using, accuracy might yield misleading results, that is, when the numbers of observations in different classes vary greatly.

To solve that, we can use a confusion matrix to check the number of true positives and negatives and false positives and negatives.

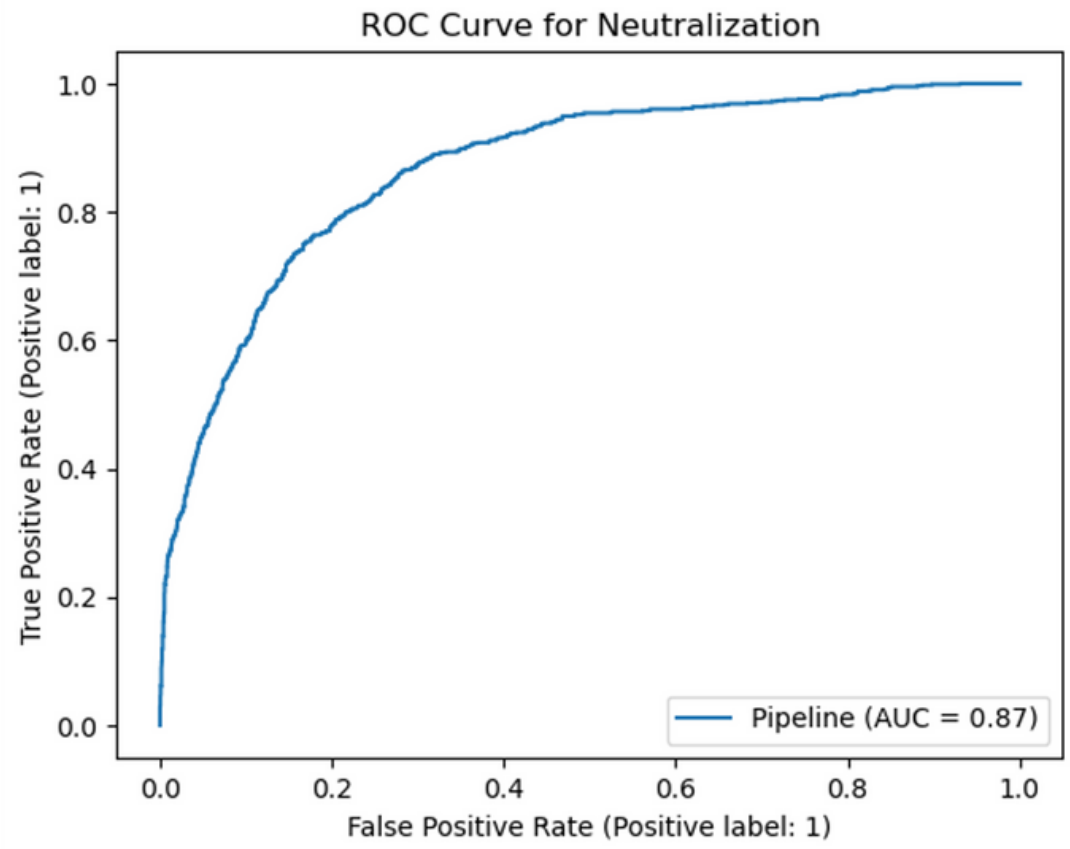
We can see that our model could predict 42% of the true neutralizers (352 out of 826). The precision for true predictions is 71% and recall is 57%.

Confusion Matrix	Predict Non-Neutralizer	Predict Neutralizer
Actual Non-Neutralizer	2379	189
Actual Neutralizer	352	474

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability.

This tells us how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. By analogy, the higher the AUC, the better the model is at distinguishing between neutralizer antibodies and non-neutralizer antibodies against HIV.

As we can see below, our model has AUC near to the 1 (0.87) which means it has a good measure of separability between neutralizers and non-neutralizers.



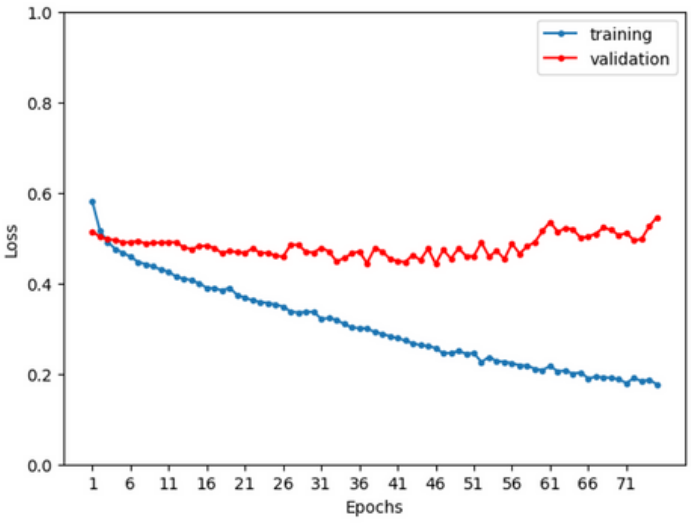
Neural Networks

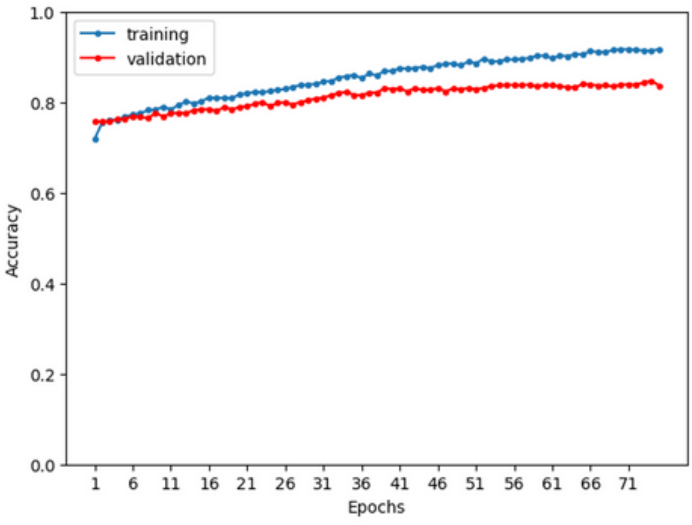
The test accuracy was 0.82. But as we can see in the confusion matrix below, the model is did not predict any true neutralizer.

This is happening because the default threshold is 0.99, very conservative.

Also we can see that our model is overfitting, according to the validation curve plotted bellow.

	Predict Non-Neutralizer	Predict Neutralizer
Actual Non-Neutralizer	2568	0
Actual Neutralizer	826	0





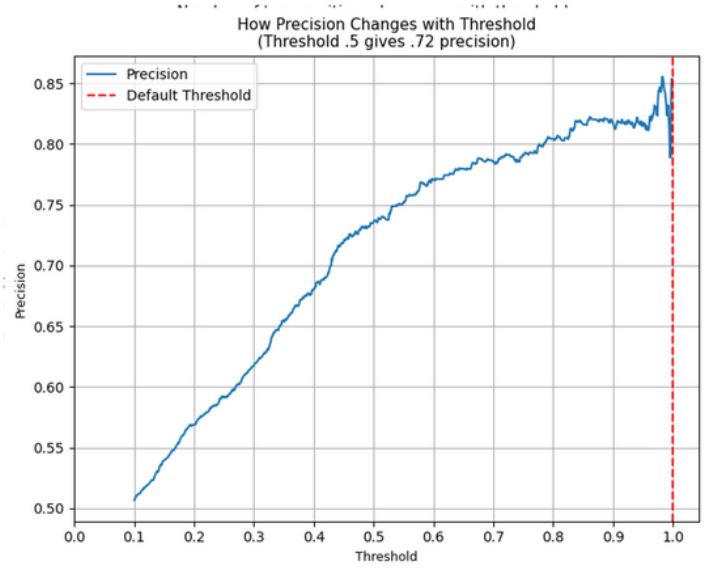
Neural Networks

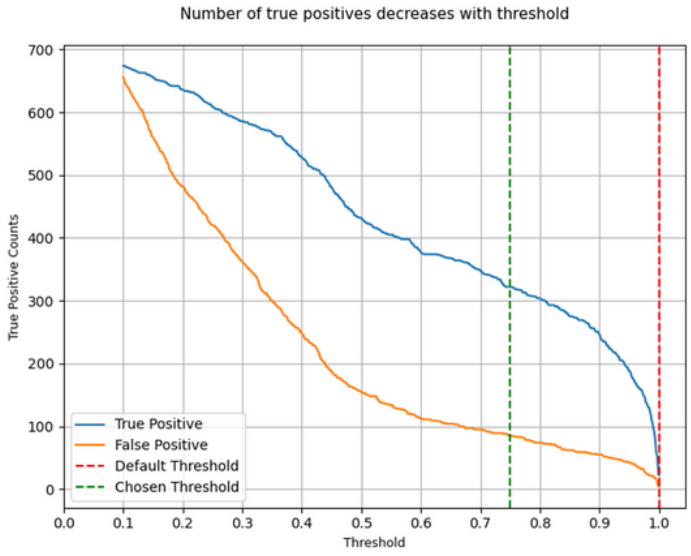
To see if we can improve our prediction, a function was created in order to investigate the precision and the threshold of the model.

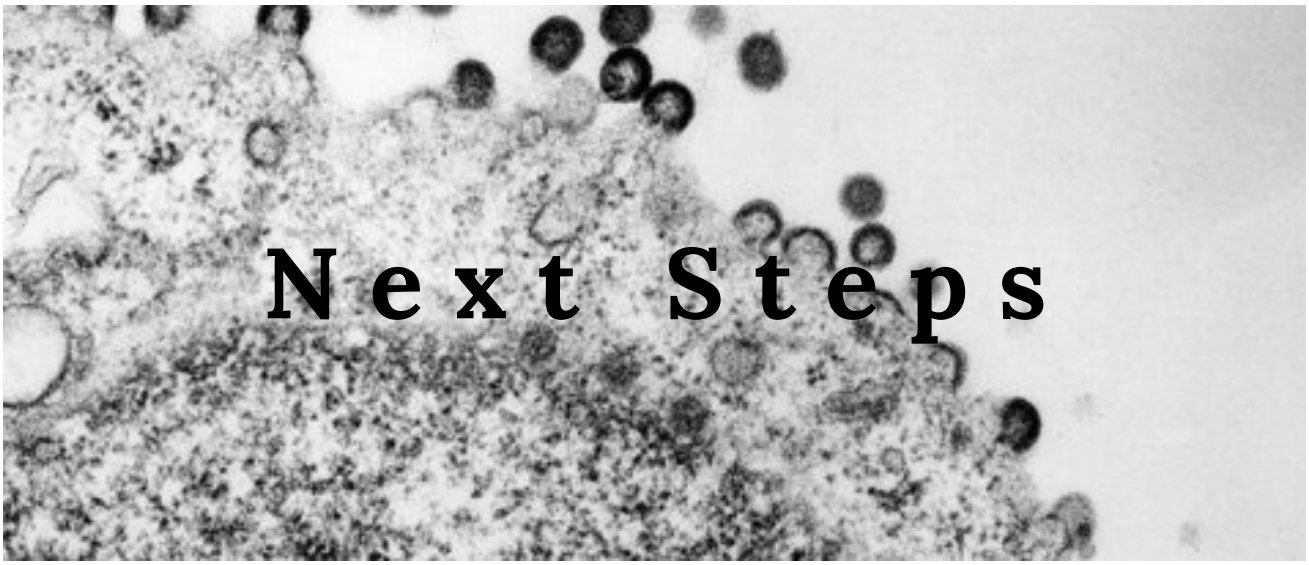
With the increase of the threshold, we can see that the precision increases. But if it is close to 1, also is not going to identify neutralizers as non-neutralizers. Between 0.7 and 0.8 is when our precision is high and able to separate classes.

Because our model now is less conservative to make predictions, we can separate better our classes. With a threshold of 0.75, we can get at least 369 correct predictions for true neutralizers out of 826 true neutralizers.

	Predict Non-Neutralizer	Predict Neutralizer
Actual Non-Neutralizer	2425	143
Actual Neutralizer	457	369







Next Steps

XG Boost is much simpler and easier to handle and to interpret than Neural Networks and we got similar results with both methods.

Next steps will be testing the model for single antibodies to check if we can make the predictions more precise.