
Multi-Drug Resistance Prediction in HIV/AIDS Using Hybrid Machine Learning Ensemble

Titash Roy¹, Bijoyini Bagchi²

^{1,2}Department of Computer Science & Engineering, Institute of Engineering &
Management

metitash06@gmail.com , bijoyini.bagchi@iem.edu.in

Abstract.

HIV multi-drug resistance (MDR) leads to decreased efficacy of antiretroviral treatments. In this study, an ensemble-based soft voting approach involving two machine learning algorithms, XGBoost and LightGBM, is suggested for predicting HIV drug resistance. With the aid of mutations collected from the Stanford HIV Drug Resistance Database, the suggested model attained 95% accuracy, which is higher than the individual classifiers' performance.

Keywords. HIV, drug resistance, machine learning, XGBoost, LightGBM, ensemble learning.

1. INTRODUCTION

The Human Immunodeficiency Virus (HIV) continues to be a formidable public health problem worldwide, even with improvements in Antiretroviral Therapy (ART). While ART is quite effective in controlling the virus, HIV mutates at an alarming rate to cause Multi-Drug Resistance (MDR), thus posing greater chances of virological failure. It is, therefore, necessary to detect the presence of resistant HIV strain. [1]

The non-nucleoside reverse transcriptase inhibitor Efavirenz (EFV) is especially prone to mutations in the reverse transcriptase region of the HIV genome. The traditional method for genotypic resistance detection is time-consuming and expensive, yet accurate. Machine Learning offers an effective solution to predicting drug resistance using the mutation profiles. [2]

This research suggests a hybrid soft voting ensemble that utilizes both the XGBoost and LightGBM classifiers to predict resistance to EFV based on mutation information available in the Stanford HIV Drug Resistance Database. The proposed model performed with 95% accuracy, surpassing individual classifier performance.

This model incorporates both XGBoost and LightGBM through soft voting to increase accuracy and reached an accuracy of 95% in the HIVDB dataset.

2. LITERATURE REVIEW

Machine learning algorithms like random forest, deep learning, and ensemble learning have proven effective in predicting HIV drug resistance. However, many methods use only a single classifier, which has prompted the development of this hybrid approach.

Table I: Literature Review

Reference	Method	Key Finding
Shen et al.[3]	Random Forest, KNN	Achieved high prediction accuracy using genotype data
Cai et al.[4]	Weighted Machine Learning	Demonstrated effective mutation-based resistance prediction
Singh[5]	Machine Learning +ANRS	Improved prediction over rule-based methods.
Riveros Maidana et al.[7]	ML Benchmarking	Highlighted the importance of preprocessing and validation
Xing et al.[8]	Stacking Ensemble	Predicted resistance to multiple HIV drugs

3. METHODOLOGY

3.1. Dataset and Preprocessing

The data set used in this experiment was derived from the Stanford HIV Drug Resistance Database (HIVDB). There were two separate data sets which contain mutation profile information about NRTI and NNRTI drugs. These two data sets were merged with a common Sequence Identifier (SeqID) to have a combined data set used in predicting resistance. Positions P1 to P318 were encoded as binary features depending on whether there is mutation present or not, indicated by 1 and 0 respectively. EFV (Efavirenz) resistance was chosen as the target and transformed to a binary classification problem. Duplicated SeqIDs were dropped to avoid data leakage. The data set was split into training and test sets in a 80%-20% ratio through stratified sampling.

3.2. Proposed Hybrid Soft Voting Ensemble

This proposed model utilizes the combination of XGBoost and LightGBM through soft voting ensemble with equal weights. In contrast to individual boosting models, this model utilizes the combination of the strong points of the XGBoost and LightGBM, where the XGBoost is the use of regularized gradient boosting, while the LightGBM is the efficient leaf-wise learning technique. The probabilities from both classifiers are taken into account.:

$$P_{(ensemble)} = (P_{(XGBoost)} + P_{(LightGBM)}) / 2$$

A classification threshold of 0.39 was chosen after empirical evaluation to optimize the performance of the classifier in terms of predicting the outcome.

3.3. Experimental Setup

The model was tested using an 80:20 split of the data into training and testing sets and accuracy, precision, recall, and F1-score were used to evaluate its predictive ability.

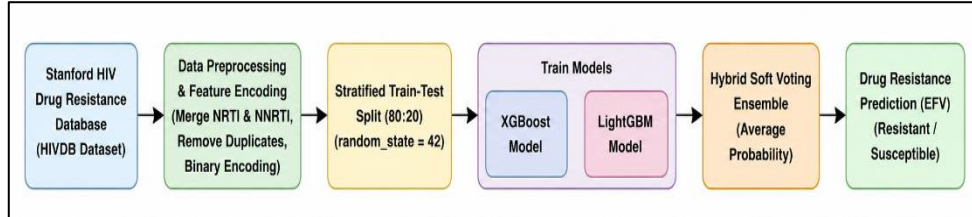


Fig.1. Process flow of the proposed Hybrid Soft Voting Ensemble for predicting HIV drugs resistance

4. RESULTS AND DISCUSSION

The Hybrid Soft Voting Ensemble (HSVE) approach has been analyzed by using Accuracy, Precision, Recall, and F1-Score. As evident from Table II, the proposed hybrid soft voting ensemble technique has performed very well and provided maximum accuracy (95.00%).

Table II: Performance comparison of the proposed model

Model	Accuracy (%)	Precision	Recall	F1 -Score
XGBoost	94.68	95.72	95.21	95.46
LightGBM	94.06	96.00	94.00	95.00
HSVE Model	95.00	95.74	95.74	95.74

The suggested classifier of the proposed HSVE exhibited the highest level of accuracy of 95.00% and had consistent improvements in the other measures namely the precision, recall, and f-measure when compared with the individual classifiers.

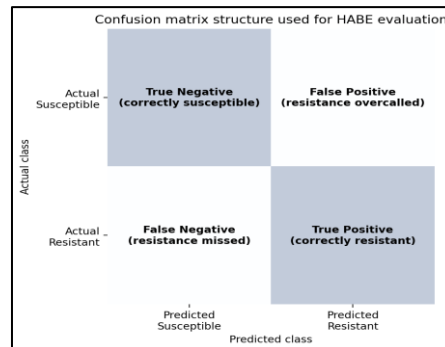


Fig.2. Hybrid soft voting ensemble confusion matrix

5. CONCLUSION

The hybrid soft voting ensemble is an efficient method for predicting the HIV drug resistance through the combination of XGBoost and LightGBM techniques. The model obtained 95.00% of accuracy, which was better than the performance of each individual classifier.

6. REFERENCE

- [1] H. M. Swinkels, A. D. Nguyen, T. Samandari, and P. G. Gulick, "HIV drug resistance and antiretroviral therapy," National Library of Medicine.
- [2] C. L. Boyce et al., "Genotypic resistance testing and HIV treatment."
- [3] C. H. Shen et al., "Automated prediction of HIV drug resistance from genotype data," BMC Bioinformatics, 2016.
- [4] Y. Cai, X. Yu, H. Sun, et al., "Predicting HIV drug resistance using a weighted machine learning method," 2021.
- [5] A. Singh, "Machine learning approaches to improve ANRS HIV drug resistance prediction system," Journal of Translational Medicine, vol. 15, p. 197, 2017.
- [6] D. F. Steiner et al., "Deep learning for HIV-1 drug resistance prediction," Bioinformatics, vol. 36, no. 7, pp. 2011–2017, 2020.
- [7] D. Riveros Maidana et al., "Benchmarking machine learning models for HIV protease inhibitor resistance prediction," Bioinformatics, 2025.
- [8] L. Xing et al., "Simultaneous prediction of multiple HIV drug resistances using stacking algorithms," arXiv:1901.10153, 2019.
- [9] S. Y. Rhee et al., "Stanford HIV Drug Resistance Database," Nucleic Acids Research, 2006.
- [10] Stanford University, "HIV Drug Resistance Database." Available: <https://hivdb.stanford.edu/>