Development of machine learning algorithms to predict viral load suppression among HIV patients in Conakry (Guinea)

Dègninou Yehadji (degninou.yehadji@fulbrightmail.org)
Médecins Sans Frontières

Geraldine Gray
Technological University Dublin

Carlos Arias Vicente
Médecins Sans Frontières

Petros Isaakidis
Médecins Sans Frontières, Southern Africa Medical Unit

Abdourahimi Diallo
Médecins Sans Frontières

Saa André Kamano
Médecins Sans Frontières

Thierno Saidou Diallo
Programme National de Lutte contre le VIH/Sida et les Hépatites (PNLSH)

Research Article

Keywords: HIV, antiretroviral therapy, viral load, machine learning, prediction, classification, algorithm, support vector machine, Random forest, Naive Bayes, Logistic regression

Posted Date: May 16th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2912310/v1

Abstract

Background

Viral load (VL) suppression represents a key to the end of the global HIV epidemic. It is critical for healthcare providers and people living with HIV (PLHIV) to be able to predict viral suppression. This study was conducted to explore the possibility of predicting viral suppression among HIV patients using machine learning (ML) algorithms.

Methods

Anonymized data were used from a cohort of HIV patients managed in eight health facilities in Conakry (Guinea). The data pre-processing steps included variable recoding, record removal, missing values imputation, grouping small categories, creating dummy variables and oversampling (only applied to the training set) of the smallest target class. Support vector machine (SVM), logistic regression (LR), naïve Bayes (NB), random forest (RF) and four stacked models where developed. The optimal parameters of the algorithms were determined with 03 cross-validation. The 30% of the sample was held as a test set to perform model evaluation. Techniques implemented to determine the most predictive variables were applied on LR, RF, and NB (with analysis of variance, ANOVA).

Results

LR was found to be the most optimal model to detect VL suppression and non-suppression. It produced an area under the curve (AUC) of 83%, with 74% and 78% sensitivity and specificity, respectively. In other words, it can correctly detect 74% of suppressed VL and correctly detect 78% of non-suppressed VL. With LR, Gender, Prior antiretroviral therapy (ART), Method into ART, Cotrimoxazole prophylactic therapy (CPT) at ART start, Second Line treatment, Last pre-ART CD4, Last ART CD4, Stage at ART start, Age, and Duration on ART were identified as the most predictive variables for VL suppression.

Conclusion

This study demonstrated the capability to predict VL suppression but has some limitations. The results are dependent on the quality of the data and are specific to the Guinea context and thus, there may be limitations with generalizability. Future studies may be conducting a similar study in a different context and develop the most optimal model into an application that can be tested in a clinical context.

Introduction

Human immunodeficiency virus (HIV) has become one of the global health and development challenges since its recognition and report of first cases in the 1980s, and its impacts include social, cultural, psychological, economic and political issues [1]. Despite the global mobilization to end the HIV epidemic, there are remaining challenges that limit the impact of the efforts. The United Nation's program on HIV/AIDS (UNAIDS) established the 90-90-90 strategy, aiming for 90% of people living with HIV (PLHIV) to

be aware of their status, 90% of those diagnosed to initiate antiretroviral therapy (ART), and 90% of those on ART to have viral loads (VL) suppressed below levels of detection, by 2020 [2]. These goals show that viral suppression represents a key to the end of the global HIV epidemic. The aim of clinical management of HIV is long-term viral suppression. Given the importance of viral suppression in HIV clinical management and epidemic control, it would be of a great utility to be able to predict it among PLHIV through the continuum of care: i) HIV diagnosis, ii) linkage to HIV medical care, iii) receipt of HIV medical care, iv) retention in medical care, and v) achievement and maintenance of viral suppression [3]. For healthcare providers as well as PLHIV, being able to predict viral suppression could help comply with treatment and possibly, adjust treatment to prevent virologic failure. That is where ML could contribute to a better monitoring of patients under ART.

Several studies have been conducted to identify factors of viral suppression among PLHIV. These studies identified factors such as gender, marital status, age, added body mass index, treatment regimen, clinical stage of the infection at the time of ART initiation, duration on ART, treatment adherence, active Tuberculosis, initial fasting glucose, alcoholism, smoking, facility type, baseline CD4 count, and recent CD4 count, availability of a daily caregiver, social isolation, high stigma, and belief that there is a cure for the acquired immunodeficiency syndrome (AIDS) [4–14]. These factors can be grouped into several categories such as social and demographic, behavioural, structural and clinical factors, and provide an orientation for the choice of variables to include in a predictive model for viral load suppression.

With the growing availability of data in clinical setting, machine learning (ML) is being used for several purposes such as diagnosis, patient outcome prediction, personalized care, drug discovery, clinical trial, radiology and radiotherapy, smart electronic health records, and epidemic outbreak prediction. ML is categorized into supervised and unsupervised learning algorithms. Supervised learning algorithms are developed to predict or classify known outcomes with sets of predictors. When outcomes are unknown, unsupervised learning algorithms are used to partition samples into distinct groups where individuals of the same group have similar characteristics. Logistic regression (LR), decision trees (DT), boosted trees (BT), random forests (RF), naïve Bayes (NB), support vector machines (SVM), nearest neighbors (e.g. k-NN), and neural networks (NN) are some of the popular supervised learning algorithms [15]. Classification tasks are the cornerstone of ML applications in healthcare. They can be used to predict patient outcome, perform diagnosis or inform decisions about treatment. Given the variety of classification algorithms available, one of the challenges is to select the most suitable algorithms for healthcare datasets [16]. HIV clinical care and research are not outside the trend of ML applications in healthcare [17].

This study was conducted to explore the possibility of predicting viral suppression among HIV patients. Specifically, it aimed to determine the most predictive variables of VL suppression for PLVIH, the optimal parameters of the selected models, and determine whether VL suppression in a patient can be predicted with baseline and follow-up demographic and clinical data.

Methodology

Study dataset

The study was conducted on a cohort of HIV patients managed in 08 healthcare facilities supported by Médecins Sans Frontières (MSF), which is implementing a project aiming at the reduction of mortality and morbidity of PLHIV in Conakry (Guinea).

Table 1
List of variables extracted from the TIER.Net

Category	Variables
Demographic	
	Gender, Age At ART Start, Current Age
Clinical	
	Gender, Regiment schedule, Prior ART, Method into ART, Baseline CD4, Last Pre-ART CD4, Last ART CD4, Last Pre-ART Stage, Stage at ART Start, TB Rx Started, TPT Outcome, Regimen At Baseline, Last ART Prescription, Second Line Start Date, TB Status At Last Visit, CPT at ART Start, Duration on ART (months), Last ART VL Count
Structural	
	Facility

Disidentified patient data were extracted from the Three Interlinked Electronic Register (TIER.Net) used by the MSF's HIV/TB Project in Conakry, Guinea (Table 1). The TIER.Net is designed with modules to capture patient-level data on HIV counselling and testing (HCT), pre-ART and ART services [18]. The data set contains 20 variables with 30205 total records, including 20878 (69%) women, and 9327 (31%) men. The variables are mixed numerical and categorical.

Data cleaning and pre-processing

Some variables were recoded into new ones and some records were removed for not meeting criteria (Duration on ART less than 3 months, and missing VL). These actions were taken for data cleaning at this step before further data exploration and preparation (Table 2).

The above-described processes return a dataset of 21 variables with 13529 records. The majority of variables have full records, but some others, namely, Baseline CD4, Last Pre-ART CD4, Last ART CD4, Last Pre-ART Stage, Stage at ART Start, TB Rx Started, Age At ART Start, TB Status At Last Visit, CPT at ART Start, and Duration on ART have missing values ranging between 1% and 59%.

The cleaned dataset was split into training and test sets (test size = 30%), as recommended by Sidey-Gibbons & Sidey-Gibbons (2019) [19]. The following actions were taken during pre-processing: discretization of categorical variables, correction of class imbalance (89% in the positive class), and imputation of missing variables. Small categories were grouped into a single category for independent variables and categorical variables with non-numerical labels were discretized by converting into dummy

variables. For the target variable, class imbalance was corrected by performing a simple bootstrapping technique which consisted of oversampling the minority category ('VL Suppressed' = No; number of samples = 8392; random state = 5). Missing values imputation was performed with K-NN (K = 05) which was chosen for its capability to produce estimations close to the reality and preserve the associations in the dataset [20]. The training and the test sets were pre-processed separately to prevent information leakage from the training to the test set, and bootstrapping was performed on the training set only [21]. The data preparation processes returned a training set and a test set of 45 numerical variables with 16793 records in the training set and 4054 records in the test set.

Modelling

Four individual classifiers (SVM, RF, NB, and LR) and four stacked classifiers using combinations of the four classifiers were developed. Individual models' parameters tuning was performed with 03 cross-validation. After developing the SVM, RF, NB and LR algorithms, they were input into four other stacked classifiers, aiming at leveraging the performance of the individual classifiers. The output of three individual classifiers were stacked as inputs (classifiers) to the fourth one used as classifier to compute the final prediction (meta classifier): (inputs = [LR, NB, RF], meta classifier = SVM); (inputs = [LR, NB, SVM], meta classifier = RF); (inputs = [LR, SVM, RF], meta classifier = NB); (inputs = [NB, SVM, RF], meta classifier = LR).

LR, RF, and analysis of variance (ANOVA) for NB techniques were used to determine the top predictive variables of VL suppression. A LR summary was developed by creating and fitting a logit model. The LR summary presented the regression coefficients along with their p-values at 0.05 level of significance, the feature importance attribute was fitted on the RF classifier to determine the most predictive variables of VL suppression, and an ANOVA was run to assess differences in predictor's means by target class.

Evaluation

The performance of each of the 04 individual algorithms and the 04 stacked algorithms was measured on the test set using F-score, and area under the curve (AUC) as evaluation metrics. The F-score combines positive predictive value (precision) with sensitivity and is a relevant metric to assess the models' capability to predict the target positive class (VL Suppressed = 1). In clinical practice, predicting supressed viral load is equally important as predicting non-supressed viral load. Thus, in addition to F-score, AUC which also considers the negative class (VL Suppressed = 0), was considered in the models' evaluation. Determining the best performing model consisted in finding the best balance between F-score and AUC.

In summary, the machine learning pipeline developed for this study is presented in Fig. 1.

Results

The LR summary revealed that Gender_Male, Prior ART_Naive, Method into ART_New, CPT at ART Start_Yes, Second Line Rx_Yes, Last Pre-ART CD4, Last ART CD4, Stage at ART Start, Current Age,

Duration on ART (months) are the variables that are associated with VL Suppressed_Yes at $p < 0.05$ level of significance. With RF, Regiment schedule_6-Month, Duration on ART (months), Last ART CD4, Regiment schedule_Regular, Last Pre-ART CD4, Second Line Rx_Yes, Baseline CD4, Current Age, Age At ART Start, and Last ART Prescription_1T3E were found to be the 10 most predictive variables for viral load suppression. The ANOVA showed that the distribution of 32 variables by VL varied significantly at $p < 0.05$ level of significance (Table 3).

After parameters tuning, C = 40 (tested in the range [5, 60]) and gamma = auto (tested for scale and auto) was found to be the optimal parameters for SVM on the training set. The optimal maximum depth for RF equals 40 (tested in the range [5, 55]). LR parameters were kept by default, except maximum iteration, random state, and number of jobs which were purposefully set at 50, 5, and − 1, respectively.

The models' evaluation metrics showed that RF, ([LR, NB, RF], SVM), ([LR, SVM, RF], NB), and ([NB, SVM, RF], LR) are the top performing models based on F-score for the positive class (94%). SVM and ([LR, NB, SVM], RF) has similar results 93% F-scores. The NB and LR performance was lower with 89% and 84% F-scores respectively. Unsurprisingly given the level of class imbalance, AUC scores did not concur. Considering individual algorithms, SVM showed a poor performance with 57% AUC while RF, NB and LR performed higher with 82–83% AUC. The stacked algorithms performed worse with AUCs between 52% and 76%. Two of them – ([LR, NB, SVM], RF) and ([LR, SVM, RF], NB), did not performed better than random guesses, yielding 52% AUC (Table 4).

The purpose of this study was to find the best balance between the capability to predict the positive class (VL suppression, VL Suppressed_Yes = 1) as well as predicting the negative class (VL non suppression, VL Suppressed_Yes = 0). Consequently, AUC has been weighted in for its indication in discriminating between the target classes. Moreover, in a cohort where the VL Suppressed_Yes = 0 is the minority class, it is critical to select a model that can detect these non-supressed VL, meaning a model with a high specificity. Looking for the best balance between F-score and AUC, LR (F-score = 0.84; AUC = 0.83) is the algorithm that is optimal for predicting both classes.

Discussion

The performance produced by the LR model developed in this project (AUC = 83%) is comparable to those obtained in other studies, where the AUCs varied between 63% and 83% [22–25]. For example, Revell et al. (2012) developed a RF model to predict VL reduction using data from North America, Western Europe and Australia. After excluding the genotype variable, implementing a model improvement strategy, and testing on data from Romania, the model produced an AUC equal to 83% [22]. Revell et al. (2013) also developed models that can predict VL suppression without a genotype and evaluated their applicability in resource-limited settings. The models were trained using data from well-resourced countries and evaluated data from well-resourced countries mixed with data from Southern Africa, India, and Romania. The models achieved an AUC of 76–77% with the test samples from well-resourced countries, 58%-65% with samples from Southern Africa, 63% with samples from India, and 70% with samples from Romania [23]. Petersen

et al. (2015) used data from US cohorts and applied a super learner algorithm for classifying virologic failure. The results showed that AUC was 78% and 79% for virologic for failures at > 1000 copies/ml or > 400 copies/ml thresholds respectively [24]. Kamal et al. (2021) developed a RF to predict viral rebound from medication adherence and clinical data in Switzerland, which produced an average AUC of 65% [25]. It can be observed that some of these models performed poorer (63%), while the top performing yielded exactly 83% AUC as in this project. Models tested by Revell et al. (2013) in different contexts produced lower performances as compared to testing with dataset from the setting where the training sets were collected [23].

The input variables used for these models are included this project with additional variables such as age, tuberculosis prevention and treatment, and health facility. The variables used in the study are also in line with those found in studies conducted by to identify factors of viral load suppression, to the exception of treatment adherence, marital status, initial fasting glucose, alcoholism, smoking, availability of a daily caregiver, belief that there is a cure for AIDS, social isolation, high stigma, and body mass index [4–14]. HIV genotype was not available in the dataset, but Revell et al. (2012) demonstrated that it is possible to develop models without it and obtain results that are as performing as those developed with it [22].

The results of this study have some limitations. The data used for model building were collected in the specific context of Conakry (Guinea). It has been demonstrated that performance may be reduced while testing in a different context, and consequently, results in this study may not be maintained if the models are evaluated in different settings [23]. Exploring the applicability of the results in different settings would be a relevant inquiry. Moreover, the results are dependent on the quality of the data used. The missing data imputation applied may have induced increased performance estimates of the models. TU Dublin ethical clearance was contingent on binning numeric attributes to ensure a greater level of anonymity, thus further reducing the quality of the data. The LR chosen in this study may also be affected by its specific limitations: it requires independence of observations, absence of multicollinearity among the independent variables, and linearity of independent variables and log odds, which cannot be assured with clinical data.

This study is limited to exploring ML algorithms with the most predictive variables, the optimal parameters, and the top performing model. The development phase of the CRISP-DM methodology was not addressed in this study, and thus the results are not readily usable in clinical practice.

Conclusion

SVM, RF, NB, LR and four stacked classifiers using combinations of the four individual ones were developed. Their evaluation showed that RF, ([LR, NB, RF], SVM), ([LR, SVM, RF], NB), and ([NB, SVM, RF], LR) yielded high performances based on F-score for the positive class (94%). Although the LR did not produced the highest F-score, it presented the optimal balance between F-score (84%) and AUC (83%). This means that the LR is useful in a way that it can be used to detect VL suppression and non-

suppression. With this model, the proportion of suppressed VL that can be detected is 74%; and the proportion of non-suppressed VL that can be detected is 78%.

The optimal parameters found were C = 40 and gamma = auto for SVM, Maximum depth = 40 for RF, and the default parameters were maintained for NB, and LR. The LR showed that Gender, ART status, CPT at ART start, Second line treatment, CD4 counts, Stage at ART start, Current age, and Duration on ART were the variables that are associated with VL suppression at p = 0.05 level of significance.

The possible future direction of this study is evaluating the models on data from different contexts to assess generalizability. Moreover, as LR is found to be the most relevant model, it can be developed into an application that can be tested in a clinical context.

Table 2
Data cleaning tasks performed on the original dataset

Procedure	Variable	Task description	Rational
Variable recoding			
	VL Suppressed (Target variable)	Use Last ART VL Count to create a binary variable indicating VL suppression (Yes / No)	A threshold of < 1000 RNA copies/ml is used to define suppressed viral load [26]
	Second Line Treatment	Use Second Line Start Date to create a binary variable indicating if patient is on second line treatment (Yes / No)	Reported date is indicative that patient is on second line treatment
	Baseline CD4, Last Pre-ART CD4 Count, and Last ART CD4 Count	Bin CD4 values into 100-unite ranges	CD4 < 200 cells/µL is the threshold of immunologic failure
	Age At ART Start, Current Age	Bin ages into 5-year age groups	5 years is a common interval for age groups creation
	Duration on ART (months)	Bin Duration on ART (months) into 6-month categories	Decision is taken to categorise at each semester
Record removal			
	VL Suppressed	Remove records with missing values	VL Suppressed is the target variable. Thus, only non-missing records will be kept in the final dataset
	Duration on ART (months)	Remove records with values < 3	The minimum timeline to expect viral suppression after ART initiation is 06 months [27]. Decision was made to consider 3 months after ART initiation
	Current Age	Remove records with values < 18	Because of ethical considerations, under-18 patients were not included
Missing values imputation			
	TPT Outcome	Fill missing values as No treatment	Missing TPT Outcome is indicative that patient is not under TPT treatment

VL = viral load; ART = antiretroviral therapy; TPT = Tuberculosis preventive treatment

Table 3
Summary of the most predictive variables by model

LR (p-value < 0.05)	RF (Top 10 variables)
Gender_Male	Regiment schedule_6-Month
Prior ART_Naive	Duration on ART
Method into ART_New	Last ART CD4
CPT at ART start_Yes	Last pre-ART CD4
Second line treatment_Yes	Second line treatment_Yes
Last pre-ART CD4	Regiment schedule_Regular
Last ART CD4	Baseline CD4
Stage at ART start	Current age
Current age	Age at ART start
Duration on ART	Last ART prescription_1T3E

Table 4
Summary confusion matrixes and evaluation metrics of the individual and stacked models developed

Model	Predicted negative	Predicted positive	F-score (positive class)	AUC
SVM				
Actual negative	23	434	-	-
Actual positive	75	3522	0.93	0.57
RF[αβ]				
Actual negative	138	319	-	-
Actual positive	140	3457	0.94	0.82
NB				
Actual negative	257	200	-	-
Actual positive	533	3064	0.89	0.82
LR				
Actual negative	358	99	-	-
Actual positive	924	2673	0.84	0.83
[LR, NB, RF], SVM[α]				
Actual negative	139	318	-	-
Actual positive	135	3462	0.94	0.76
[LR, NB, SVM], RF				
Actual negative	23	434	-	-
Actual positive	75	3522	0.93	0.52
[LR, SVM, RF], NB[α]				
Actual negative	16	441	-	-
Actual positive	8	3589	0.94	0.52
[NB, SVM, RF], LR[α]				
Actual negative	136	321	-	-
Actual positive	109	3488	0.94	0.72

(α) Top performing algorithms based on F-score. (β) Algorithm with the best balance between F-score and AUC.

Abbreviations

AIDS: acquired immunodeficiency syndrome

ANOVA: analysis of variance

AUC: area under the curve

CPT: cotrimoxazole prophylactic therapy

HIV: human immunodeficiency virus

LR: logistic regression

ML: machine learning

NB: naïve Bayes

PLHIV: people living with HIV

RF: random forest

SVM: support vector machine

VL: viral load

Declarations

Consent to participate

The data was routinely collected for an HIV cohort management by Médecins Sans Frontières in Conakry, which uses them for clinical and program monitoring purposes. Procedures were set, allowing patients to give approval for their data to be used in research. They were publicly and verbally informed that their data may be used for research, that they are free to opt out at any time, and their decision does not affect their access to care. At the time of data acquisition for this study, the program did not record any request to opt out.

Ethical statement

The study was conducted in in accordance with relevant guidelines and regulations on research involving human data, especially the Declaration of Helsinki. It fulfilled the exemption criteria set by the Médecins Sans Frontières Ethics Review Board for a posteriori analysis of routinely collected clinical data. Ethical

approvals were received from the Technological University Dublin and the Guinea National Ethics Committee for Health Research (128/CNERS/21).

Availability of data and materials

The data used in this study will not be published. An investigator interested in these data can submit a request to: Médecins Sans Frontières Belgium, Rue de l'Arbre Bénit 46, 1050 Bruxelles, Belgium. E-mail: dpo@brussels.msf.org

The codes developed for data cleaning, model development and evaluation, are available in the Zenodo repository, https://doi.org/10.5281/zenodo.7793245 [28].

Competing Interest

The authors declare that they have no competing interests.

Funding source

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

Dègninou Yehadji: Conceptualization, Methodology, Data Curation, Software, Formal analysis, Visualization; Writing - original draft; Writing - review & editing. **Geraldine Gray:** Methodology, Supervision, Writing - review & editing. **Carlos Arias Vicente:** Supervision, Validation, Writing - review & editing. **Petros Isaakidis:** Supervision, Validation, Writing - review & editing. **Abdourahime Diallo:** Data Curation, Writing - review & editing. **André Kamano Saa:** Data Curation, Writing - review & editing. **Thierno Saidou Diallo:** Writing - review & editing.

References

1. Mann, J. (1987). AIDS: the global challenge. *Development Communication Report*, (57), 7-8.
2. UNAIDS. (2014). *90-90-90: An ambitious treatment target to help end the AIDS epidemic*. Geneva, Switzerland: Joint United Nations Programme on HIV/AIDS (UNAID.
3. Gardner, E. M., McLees, M. P., Steiner, J. F., del Rio, C., & Burman, W. J. (2011). The Spectrum of Engagement in HIV Care and its Relevance to Test-and-Treat Strategies for Prevention of HIV Infection. *Clinical Infectious Diseases*, *52*(6), 793-800. https://doi.org/10.1093/cid/ciq243
4. Sinai, I., Bowsky, S., Cantelmo, C., Mbuya-Brown, R., Panjshiri, Y., & Balampama, M. (2019). *Adolescent HIV in Tanzania: Factors Affecting Viral Load Suppression and the Transition to Adult Care*. Washington, DC: Palladium, Health Policy Plus.

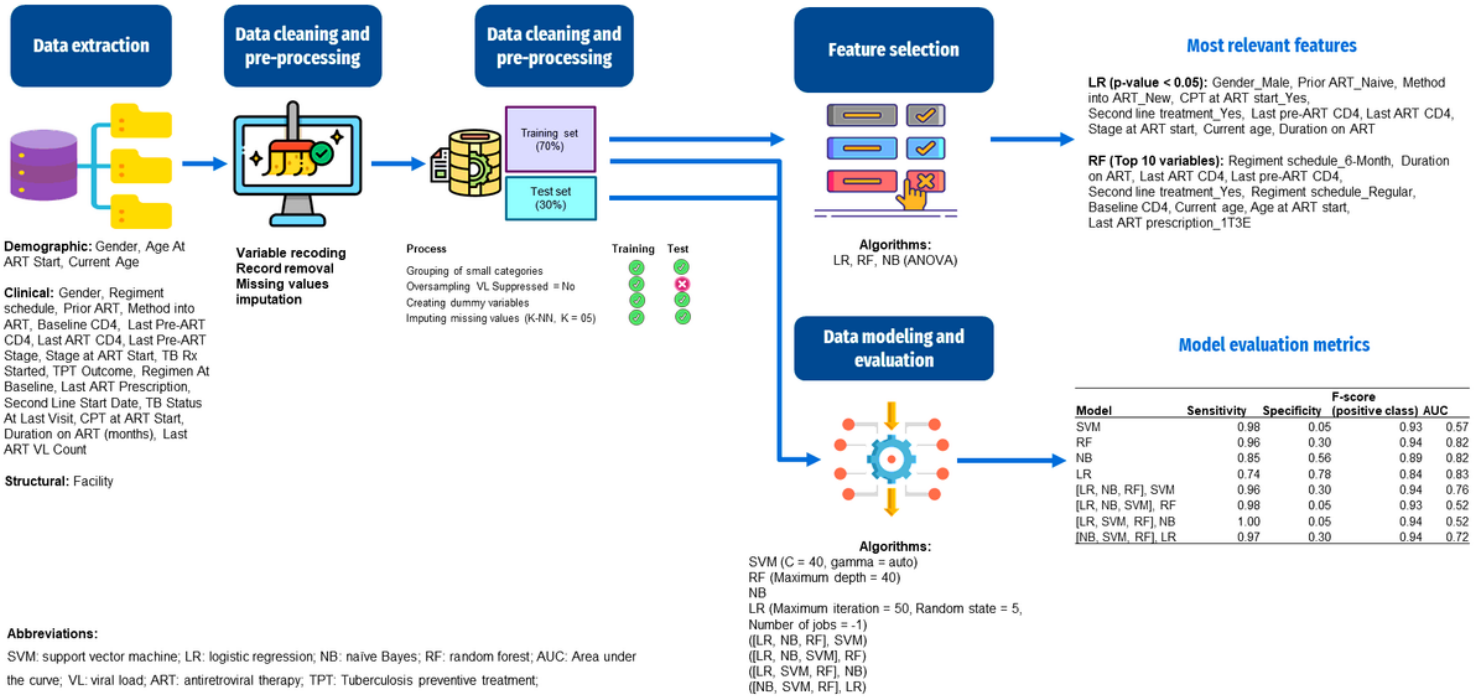
5. Ssemwanga, D., Asio, J., Watera, C., Nannyonjo, M., Nassolo, F., Lunkuse, S., … The Uganda HIV Drug Resistance Technical Working Group. (2020). Prevalence of viral load suppression, predictors of virological failure and patterns of HIV drug resistance after 12 and 48 months on first-line antiretroviral therapy: a national cross-sectional survey in Uganda. *Journal of Antimicrobial Chemotherapy*, *75*(5), 1280-1289. https://doi.org/10.1093/jac/dkz561
6. Njuguna, I., Neary, J., Mburu, C., Black, D., Beima-Sofie, K., Wagner, A. D., … John-Stewart, G. (2020). Clinic-level and individual-level factors that influence HIV viral suppression in adolescents and young adults: a national survey in Kenya. *AIDS*, *34*(7), 1065-1074. https://doi.org/10.1097/QAD.0000000000002538
7. Maina, E. K., Mureithi, H., Adan, A. A., Muriuki, J., Lwembe, R. M., & Bukusi, E. A. (2020). Incidences and factors associated with viral suppression or rebound among HIV patients on combination antiretroviral therapy from three counties in Kenya. *International Journal of Infectious Diseases*, *97*, 151-158. https://doi.org/10.1016/j.ijid.2020.05.097
8. Hicham, T., Ilyas, E., Tarik, H., Noureddine, B., Omar, B., Rachid, F., … Mohammed, B. (2019). Risk factors associated with unsuppressed viral load in HIV-1 infected patients at the first antiretroviral therapy in Morocco. *International Journal of Mycobacteriology*, *8*(2), 113-117. https://doi.org/10.4103/ijmy.ijmy_41_19
9. Desta, A. A., Woldearegay, T. W., Futwi, N., Gebrehiwot, G. T., Gebru, G. G., Berhe, A. A., & Godefay, H. (2020). HIV virological non-suppression and factors associated with non-suppression among adolescents and adults on antiretroviral therapy in northern Ethiopia: a retrospective study. *BMC Infectious Diseases*, *20*(1), 4. https://doi.org/10.1186/s12879-019-4732-6
10. Chhim, K., Mburu, G., Tuot, S., Sopha, R., Khol, V., Chhoun, P., & Yi, S. (2018). Factors associated with viral non-suppression among adolescents living with HIV in Cambodia: a cross-sectional study. *AIDS Research and Therapy*, *15*(1), 20. https://doi.org/10.1186/s12981-018-0205-z
11. Rangarajan, S., Colby, D. J., Giang, L. T., Bui, D. D., Hung Nguyen, H., Tou, P. B., … West, G. (2016). Factors associated with HIV viral load suppression on antiretroviral therapy in Vietnam. *Journal of Virus Eradication*, *2*(2), 94-101.
12. Lokpo, S. Y., Ofori-Attah, P. J., Ameke, L. S., Obirikorang, C., Orish, V. N., Kpene, G. E., … Tetteh Quarshie, S. (2020). Viral Suppression and Its Associated Factors in HIV Patients on Highly Active Antiretroviral Therapy (HAART): A Retrospective Study in the Ho Municipality, Ghana. *AIDS Research and Treatment*, *2020*, e9247451. https://doi.org/10.1155/2020/9247451
13. Sunkanmi, F., Paul, Y., Peter, D., Nsikan, A., Joseph, J., Opada, E., … James, N. (2020). Factors Influencing Viral Load Non-suppression among People Living with HIV (PLHIV) in Borno State, Nigeria: A Case of Umaru Shehu Ultra-Modern Hospital. *Journal of Advances in Medicine and Medical Research*, 98-105. https://doi.org/10.9734/jammr/2020/v32i330388
14. Bulage, L., Ssewanyana, I., Nankabirwa, V., Nsubuga, F., Kihembo, C., Pande, G., … Kiyaga, C. (2017). Factors Associated with Virological Non-suppression among HIV-Positive Patients on Antiretroviral

Therapy in Uganda, August 2014–July 2015. *BMC Infectious Diseases*, *17*(1), 326. https://doi.org/10.1186/s12879-017-2428-3

15. Mastoli, M. M. (2019). Machine Learning Classification Algorithms for Predictive Analysis in Healthcare, *06*(12), 5.
16. Weng, W.-H. (2020). Machine Learning for Clinical Predictive Analytics. In L. A. Celi, M. S. Majumder, P. Ordóñez, J. S. Osorio, K. E. Paik, & M. Somai (Éd.), *Leveraging Data Science for Global Health* (p. 199-217). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-47994-7_12
17. Bisaso, K. R., Anguzu, G. T., Karungi, S. A., Kiragga, A., & Castelnuovo, B. (2017). A survey of machine learning applications in HIV clinical research and care. *Computers in biology and medicine*, *91*, 366-371. https://doi.org/10.1016/j.compbiomed.2017.11.001
18. Osler, M., Hilderbrand, K., Hennessey, C., Arendse, J., Goemaere, E., Ford, N., & Boulle, A. (2014). A three-tier framework for monitoring antiretroviral therapy in high HIV burden settings. *Journal of the International AIDS Society*, *17*(1). https://doi.org/10.7448/IAS.17.1.18908
19. Sidey-Gibbons, J. A. M., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, *19*(1), 64. https://doi.org/10.1186/s12874-019-0681-4
20. Sania, A., Pini, N., Nelson, M. E., Myers, M. M., Shuffrey, L. C., Lucchini, M., … Fifer, W. P. (2020). *The K nearest neighbor algorithm for imputation of missing longitudinal prenatal alcohol data* (preprint). In Review. https://doi.org/10.21203/rs.3.rs-32456/v1
21. Mierswa, I. (2017, janvier 24). Avoiding Accidental Contamination of Data [3 Examples]. *RapidMiner*. Consulté 20 avril 2021, à l'adresse https://rapidminer.com/blog/learn-right-way-validate-models-part-4-accidental-contamination/
22. Revell, A. D., Ene, L., Duiculescu, D., Wang, D., Youle, M., Pozniak, A., … Larder, B. A. (2012). The use of computational models to predict response to HIV therapy for clinical cases in Romania. *Germs*, *2*(1), 6-11. https://doi.org/10.11599/germs.2012.1007
23. Revell, A. D., Wang, D., Wood, R., Morrow, C., Tempelman, H., Hamers, R. L., … Larder, B. A. (2013). Computational models can predict response to HIV therapy without a genotype and may reduce treatment failure in different resource-limited settings. *The Journal of antimicrobial chemotherapy*, *68*(6), 1406-1414. https://doi.org/10.1093/jac/dkt041
24. Petersen, M. L., LeDell, E., Schwab, J., Sarovar, V., Gross, R., Reynolds, N., … Bangsberg, D. R. (2015). Super Learner Analysis of Electronic Adherence Data Improves Viral Prediction and May Provide Strategies for Selective HIV RNA Monitoring. *Journal of acquired immune deficiency syndromes (1999)*, *69*(1), 109-118. https://doi.org/10.1097/QAI.0000000000000548
25. Kamal, S., Urata, J., Cavassini, M., Liu, H., Kouyos, R., Bugnon, O., … Schneider, M.-P. (2021). Random forest machine learning algorithm predicts virologic outcomes among HIV infected adults in Lausanne, Switzerland using electronically monitored combined antiretroviral treatment adherence. *AIDS care*, *33*(4), 530-536. https://doi.org/10.1080/09540121.2020.1751045

26. WHO. (2016). *Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach*. Geneva, Switzerland: World Health Organization. Consulté à l'adresse https://apps.who.int/iris/handle/10665/208825
27. Ali, J. H., & Yirtaw, T. G. (2019). Time to viral load suppression and its associated factors in cohort of patients taking antiretroviral treatment in East Shewa zone, Oromiya, Ethiopia, 2018. *BMC Infectious Diseases*, *19*(1), 1084. https://doi.org/10.1186/s12879-019-4702-z
28. Yehadji, D. (2023). Codes for development of algorithms for prediction of viral load suppression in an HIV patients cohort in Conakry, Republic of Guinea. https://doi.org/10.5281/zenodo.7793245

Figures



Model	Sensitivity	Specificity	F-score (positive class)	AUC
SVM	0.98	0.05	0.93	0.57
RF	0.96	0.30	0.94	0.82
NB	0.85	0.56	0.89	0.82
LR	0.74	0.78	0.84	0.83
[LR, NB, RF], SVM	0.96	0.30	0.94	0.76
[LR, NB, SVM], RF	0.98	0.05	0.93	0.52
[LR, SVM, RF], NB	1.00	0.05	0.94	0.52
[NB, SVM, RF], LR	0.97	0.30	0.94	0.72

Figure 1

Workflow of the machine learning for prediction of viral load suppression among HIV patients in Conakry

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalmaterialHIVVLSupressionCNKGN.docx