Predictive Modelling to Identify Patients at High Risk of Virological Failure in Kenya

Benedette Adhiambo Otieno

Submitted in total fulfilment of the requirements for the Master of Science in Statistical Science of Strathmore University

Institute of Mathematical Sciences
Strathmore University
Nairobi, Kenya

February 2025

Declaration

I declare that this work has not been previously submitted and approved for award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Student's name: Benedette Adhiambo Otieno

Signature: 

Date: February 15, 2025

Approval

The thesis of Benedette Otieno was reviewed and approved by the following:

Prof. Bernard Omolo
Supervisor, *Oguna-Omolo*
Institute of Mathematical Sciences, Strathmore University.
.

Dr. Godfrey Madigu
Dean,
Institute of Mathematical Sciences, Strathmore University.

Dr. Bernard Shibwabo
Director,
Office of Graduate Studies, Strathmore University.

Abstract

Virological failure (VF) remains a significant challenge in HIV treatment, necessitating the development of accurate risk prediction models. Traditional models have faced criticism for their limited ability to capture the complexity of patient data, often relying on a narrow set of variables. More robust approaches, such as machine learning and statistical modeling, have demonstrated improved predictive performance by integrating diverse clinical and demographic factors. Existing models have incorporated variables such as treatment adherence, baseline viral load, CD4 count, regimen type, and socio-economic determinants. In this study, a predictive model was developed to identify patients at high risk of VF in Kenya. Key variables were selected based on their clinical relevance, and the model was validated to ensure reliability. Performance was evaluated using metrics AUCPR(precision-recall) and AUCROC(receiver operating characteristic), providing insights into its effectiveness in identifying high-risk patients.

Three models—XGBoost Sparse, XGBoost Simple, and Random Forest (RF) Simple—were trained and evaluated using an imbalanced dataset, where the unsuppression rate was 9%. XGBoost Sparse outperformed other models, achieving the highest AUC-PR lift (4.93) and an AUC-ROC of 0.804, demonstrating its superior ability to detect virological failure. SHAP analysis revealed key predictors, including recent unsuppressed rate, previous viral load history, treatment failure, ART regimen optimization, and BMI. These findings emphasize the importance of targeted interventions, such as regimen optimization, adherence support, and socioeconomic assistance, to improve treatment outcomes. The study highlights the potential of ML-driven decision support tools in enhancing HIV care and reducing VF rates.

KEYWORDS: Predictive modelling, Virological failure, Electronic medical records, HIV treatment, Machine Learning, XGBoost, Random Forest, AUC-PR, AUC-ROC

Table of Contents

List of abbreviations

EMR: Electronic Medical Record NASCOP: National AIDS and STIs Control Programme XGBoost: eXtreme Gradient Boosting RF: Random Forest WHO: World Health Organization PLHIV: People living with HIV. MICE: Multiple imputations with chained equations AUCPR: Area Under the Precision Recall Curve VF: Virological Failure SHAP: Shapley Additive Explanations ART: antiretroviral therapy	PrEP: Pre-Exposure Prophylaxis IHME: Institute of Health Metrics and Evaluation MICE: Multiple imputations with chained equations PEPFAR: President's Emergency Plan for AIDS Relief AI: Artificial Intelligence KenPHIA: Kenya Population-based HIV Impact Assessments PEP: Post-Exposure Prophylaxis ROC: Receiver Operating Characteristics

List of Figures

List of Tables

Chapter 1

Introduction

1.1. Background of the study

HIV/AIDS remains a significant public health challenge globally, with Sub-Saharan Africa bearing the highest burden of the epidemic. According to the UNAIDS Data 2021 report, substantial progress has been made in reducing new infections and increasing access to antiretroviral therapy (ART); however, challenges persist, particularly in ensuring sustained viral suppression among people living with HIV (PLHIV) (UNAIDS, 2021). VF, defined as the inability to achieve or maintain viral suppression despite ART, remains a critical barrier to achieving the 95-95-95 targets, which aim for 95% of PLHIV to know their status, 95% of diagnosed individuals to be on ART, and 95% of those on ART to achieve viral suppression (UNAIDS, 2021).

Kenya, like many other countries in Sub-Saharan Africa, has made significant strides in HIV prevention and treatment. The UNAIDS Data 2021 report highlights that while ART coverage has expanded, gaps remain in adherence and retention, contributing to VF. Factors such as treatment interruptions, drug resistance, and socioeconomic barriers impede the long-term effectiveness of ART, increasing the risk of disease progression and transmission (UNAIDS, 2021). Structural challenges such as inadequate healthcare infrastructure, stigma, and limited access to viral load monitoring exacerbate these issues.(UNAIDS, 2021). The National AIDS and STI Control Programme (NASCOP, 2022) provides Kenya-specific guidelines for HIV programming, with a focus on key populations who are disproportionately affected by HIV. These guidelines emphasize a targeted, data-driven approach to HIV prevention, treatment, and care, including strategies to improve ART adherence and reduce VF among high-risk groups. NASCOP highlighted the importance of routine viral load monitoring, differentiated service delivery models, and community-based interventions to enhance treatment outcomes.(NASCOP, 2022).

The World Health Organization (WHO, 2020) recommends routine viral load monitoring as the gold standard for assessing treatment effectiveness and identifying patients at risk of VF. Unlike CD4 cell count, which was previously used to monitor immune function, viral load testing provides a direct measure of HIV replication, making it a more accurate indicator of treatment response. WHO emphasizes the need for regular viral load testing at six months after ART initiation and every 12 months thereafter to ensure timely identification of treatment failure and guide necessary interventions such as enhanced adherence counseling or regimen switching(World Health Organization, 2020). Supporting this, Shoko and Chikobvu (2019)

demonstrate the superiority of viral load over CD4 count in predicting mortality among HIV patients on ART. Their study highlights that patients with persistently high viral loads face a significantly higher risk of disease progression and death, even when CD4 counts remain relatively stable (Shoko & Chikobvu, 2019)

Regular virological monitoring is essential for maintaining viral suppression and ensuring long-term treatment success (Barnabas et al., 2017). Routine viral load testing, combined with timely clinical interventions, has significantly improved patient outcomes by enabling early detection of treatment failure and facilitating necessary adjustments to enhance adherence and optimize therapy (Barnabas et al., 2017). However, identifying individuals at risk of VF remains a challenge. Traditional statistical methods have been valuable in assessing treatment response at a population level, but they often lack the precision needed for individualized risk prediction, highlighting the need for more advanced predictive approaches (Bisaso et al., 2018).

Machine learning is a branch of artificial intelligence that uses algorithms to find patterns and forecast outcomes in massive datasets(Allwein & Schapire, 2000). By leveraging ML techniques, researchers can enhance decision-making processes by identifying patterns in large datasets that may not be easily detected using traditional statistical methods. Majumder et al. (n.d.) explored the application of machine learning in predicting HIV viral load hotspots in Kenya, demonstrating the potential of data-driven approaches in public health. Their study utilized real-world data to develop predictive models capable of identifying geographic areas with high viral load prevalence, enabling targeted interventions.(Majumder et al., n.d.). These models are revolutionizing healthcare by leveraging big data to enhance disease prediction and optimize treatment strategies.

Bisaso et al. (2018) explored the application of machine learning (ML) techniques in predicting HIV treatment outcomes, focusing on viral load suppression and CD4 status among PLWH. Seboka et al. (2023) utilized artificial intelligence (AI) models, including decision trees, support vector machines, and neural networks, to predict viral load and CD4 counts among patients on ART in Gedeo Zone public hospitals. Their findings indicated that ML models outperformed traditional statistical methods in accurately classifying patients' viral suppression status, with neural networks demonstrating the highest predictive accuracy. Furthermore, Bisaso et al. (2018) compared logistic regression-based ML models for predicting early virological suppression in ART-initiating patients, concluding that ML approaches significantly enhanced prediction accuracy compared to conventional clinical risk assessment methods. Both studies emphasize that ML-based predictive models can improve patient monitoring and treatment decisions, allowing for early identification of individuals at risk of treatment failure and enabling timely interventions to optimize ART outcomes.

Drain et al. (2019) emphasized the challenges of accessing viral load (VL) testing in resource-limited settings and its implications for HIV management. In many low-income regions, conventional laboratory-based VL testing is hindered by inadequate infrastructure, high costs, long turnaround times, and logistical barriers, leading to delays in clinical decision-making. Limited access to timely VL testing compromises the ability to monitor treatment effectiveness, detect virological failure early, and implement necessary interventions to prevent drug

resistance. Point-of-care VL testing is essential for bridging this gap, allowing real-time patient monitoring, enhancing retention in care, and strengthening a more effective and sustainable HIV response in these settings (Drain et al., 2019).

As a low- and middle-income country (LMIC), Kenya faces persistent challenges in HIV prevention, care, and treatment (NASCOP, 2022). To address these gaps, innovative, evidence-based approaches like hotspot mapping have been instrumental in identifying high-risk areas and optimizing resource allocation for targeted interventions. While traditional statistical methods have provided valuable insights at a population level, they often fall short in predicting individual risks(Majumder et al., n.d.). The integration of big data and machine learning offers a transformative opportunity to enhance HIV programs by enabling real-time decision-making.

This study seeks to contribute to the existing body of knowledge by applying a ML model for analyzing viral load trends among PLHIV. The proposed model will be applied to provide accurate and reliable viral load predictions using historical data.

1.2. **Problem Statement**

VF continues to pose a major challenge in managing patients on ART for HIV/AIDS. Predicting patients at risk of VF can significantly improve clinical decision-making, enabling timely interventions to prevent treatment failure and drug resistance. Although various ml-based predictive models have been developed, gaps remain in the existing literature that need to be addressed.

Comprehensive prediction models that consider several parameters impacting virological failure are lacking in existing research. Many studies ignore the holistic approach that considers clinical, immunological, virological, and behavioral aspects in favor of focusing on specific predictors like viral load, CD4 count, or adherence levels. For predictions to be accurate and trustworthy, a comprehensive predictive model that focuses on a wide range of factors and their interactions is necessary.

Secondly, the cross-sectional nature of many current studies makes it difficult for them to record temporal fluctuations and dynamic determinants. Longitudinal data and sophisticated machine learning methods, including ensemble models or deep learning, can improve prediction accuracy and offer a more thorough knowledge of the hazards associated with virological failure.

While significant research has been conducted on HIV treatment monitoring and virological failure prediction, limited studies have explored the application of several ML models to enhance predictive accuracy. This study aims to apply a hybrid ML framework to analyze and predict virological failure among HIV-positive individuals across diverse populations ensuring their reliability and usefulness in clinical practice.

1.3. Research Objectives

1.3.1. General Objective

The main objective of this study was to apply a ML method in identifying patients who are at high risk of VF in Kenya.

1.3.2. Specific Objectives

1. Apply multiple predictive models to classify patients into VF risk categories.
2. Optimize the hyperparameters of different models and compare the performance of each model using appropriate metrics to identify the best-performing model.
3. Assess the impact of different features or predictors on the predictive performance of the best model.

1.4. Justification of the study

Various factors contribute to VF, including poor adherence to ART, drug resistance, suboptimal ART regimens, co-infections and comorbidities, malnutrition, socioeconomic barriers to healthcare access, and treatment interruptions due to stockouts or migration. Frequent shortages of viral load test kits, particularly in resource-limited settings, further complicate timely diagnosis and intervention. These challenges highlight the need for effective strategies to identify patients at risk of virological failure early, enabling clinicians and policymakers to implement timely and targeted interventions to improve treatment outcomes.

Healthcare policymakers and clinicians require effective decision-support tools to optimize treatment regimens and allocate resources efficiently, ensuring that at-risk patients receive timely interventions. Furthermore, ML based predictive models can enhance HIV care by improving virological monitoring, reducing treatment failure rates, and mitigating the emergence of drug resistance. Given the growing interest in leveraging artificial intelligence in healthcare, developing and validating predictive models for VF is crucial in strengthening HIV management strategies(Rajula et al., 2020).

The results of this study are expected to provide valuable insights for healthcare providers, policymakers, and researchers by offering a robust framework for early risk detection and intervention in ART programs. These findings will also contribute to the ongoing efforts to enhance HIV treatment outcomes, optimize resource allocation, and improve the quality of care for PLHIV.

1.5. **Significance of the study**

This study is expected to guide healthcare providers and policymakers in developing data-driven approaches for identifying patients at risk of virological failure among PLHIV. By leveraging machine learning models, clinicians will gain a better understanding of patients likely to experience treatment failure, enabling them to implement timely interventions such as adherence counseling or treatment regimen adjustments to improve patient outcomes.

1.6. **Expected Output**

In this study, we expected to achieve the following results:

- A best performing ML model and apply it in predicting virological failure.
- Identify the best socio demographic and clinical characteristics contributing to VL from the best performing ML model

Chapter 2

Literature Review

2.1. **Introduction.**

This chapter provides a comprehensive review of the literature on VF prediction in PLHIV. It examines key measures of disease progression, including viral load and CD4 count, and their roles in monitoring treatment outcomes. It explores various ML models that have been developed for predicting virological failure, including tree boost models, deep learning models and traditional statistical models highlighting their strengths and limitations. Finally, it identifies gaps in existing studies and explains how the current research builds upon previous work to improve predictive accuracy and clinical decision-making

2.2. **Measures of HIV Disease Progression**

Viral load is a measure of the amount of HIV in a person's blood. It is measured as the number of copies of the virus per milliliter (copies/mL) of blood. Individuals with lower CD4 counts and higher viral loads are at greater risk of AIDS-related complications and mortality. (Emery et al., 2008).

Viral load testing is a critical tool in HIV management, as it helps determine how well ART is working. A low or undetectable viral load (<=200 copies/mL) indicates that treatment is effective in suppressing the virus, reducing the risk of disease progression and transmission. High viral(>200copies/mL) load suggests treatment failure, poor adherence, or drug resistance, requiring timely intervention to prevent virological failure. Regular viral load monitoring ensures that PLHIV receives optimal care and maintain long-term viral suppression.(NASCOP, 2022).

Sustained viral suppression through ART leads to improved life expectancy, bringing it closer to that of the general population. Lohse et al. (2007) highlighted the effectiveness of ART in improving survival outcomes for people living with HIV. The study underscored the role of early diagnosis and treatment initiation in enhancing long-term health outcomes(Lohse et al., 2007).

2.3. **Predictive Modelling for Virological Failure**

Osman and Yizengaw (2020) conducted a cross-sectional study at Jimma University Medical Center to assess virological failure among pediatric HIV/AIDS patients. They collected data from 262 children on ART, using structured questionnaires and patient records. VF was defined as a viral load ≥1,000 copies/mL after at least six months of ART, measured through plasma

viral load testing. Logistic regression analysis was used to identify risk factors, with odds ratios (ORs) and confidence intervals (CIs) reported. The study found an 11% VF rate, with significant risk factors including low weight at ART initiation (AOR = 4.17, 95% CI: 1.32–13.12) and advanced WHO clinical stage (AOR = 3.24, 95% CI: 1.16–9.08). The authors emphasized the need for early clinical monitoring and adherence support to reduce virological failure rates (Osman & Yizengaw, 2020). Ahoua et al. (2009) conducted a cross-sectional study in rural northwestern Uganda to identify risk factors associated with virological failure and subtherapeutic ARV drug concentrations among HIV-positive adults undergoing) ART. The study evaluated immunovirological, pharmacological, and adherence outcomes in patients who had been on fixed-dose ART combinations for 12 and 24 months. VF was defined as having an HIV RNA level exceeding 1,000 copies/ml. The findings revealed that 25% of patients at 12 months and 28% at 24 months experienced VF. Risk factors significantly associated with VF included poor adherence to ART, diagnosis of tuberculosis after ART initiation, subtherapeutic concentrations of non-nucleoside reverse transcriptase inhibitors (NNRTIs), presence of general clinical symptoms, and a lower weight compared to baseline (Ahoua et al., 2009). While Ahoua et al relied on cross-sectional analyses, our study incorporates a predictive modeling approach using historical features that can be integrated into clinical decision support systems to enhance early identification and intervention for patients at risk of VF. Furthermore, by utilizing a diverse set of features, including temporal and locational attributes, our study provides a more holistic view of risk factors, offering a novel contribution to the field of HIV treatment monitoring and optimization.

Andarge et al. (2022) conducted a retrospective cohort study to assess the incidence, survival time, and associated factors of VF among adult HIV/AIDS patients on first-line ART at St. Paul's Hospital Millennium Medical College. The study followed 477 patients over a period of five years and reported an overall virological failure incidence rate of 4.9 per 100 person-years. Key predictors of VF identified in the study included poor adherence to ART, low baseline CD4 count, tuberculosis co-infection, and prolonged duration on ART. The study utilized Kaplan-Meier survival curves and Cox proportional hazard models to estimate time-to-virological failure and assess the impact of different risk factors on treatment outcomes. The findings highlighted that patient with poor adherence had a significantly higher risk of VF, reinforcing the importance of adherence support interventions in HIV care. Our study differs from Andarge et al in methodology, scope, and analytical approach. While their study employed traditional statistical methods, including survival analysis, to estimate time to VF, our research leverages machine learning techniques to predict VF based on a combination of clinical, demographic, and adherence-related variables. By integrating machine learning models, our study aims to enhance predictive accuracy and offer real-time risk stratification for healthcare providers(Andarge et al., 2022).

While traditional statistical methods offer interpretability and a solid statistical foundation(Rajula et al., 2020), their effectiveness can be limited when dealing with complex, nonlinear relationships in large and high-dimensional datasets. Machine learning methods, such as random forests or support vector machines, are often considered when faced with such challenges(Kamal et al., 2021). Nevertheless, traditional statistical approaches remain valuable in many healthcare research contexts, including the prediction of VF in HIV

treatment(Meshesha et al., 2020). Our study differentiates itself by adopting a machine learning-based predictive modeling approach, leveraging algorithms such as XGBoost and random forests to enhance VF detection. Unlike traditional methods, which typically rely on predefined relationships between variables, our approach allows for the automatic identification of the most influential predictors without imposing strict parametric assumptions. Additionally, we integrate temporal and locational attributes, expanding the scope of risk factor analysis beyond conventional clinical and demographic predictors. Our study aims to improve predictive accuracy by using ML methods while maintaining a level of interpretability necessary for clinical decision support.

Kamal et al. (2021) conducted a study in Lausanne, Switzerland, to predict virologic outcomes among HIV-infected adults using a Random Forest machine learning algorithm. The study analyzed data from 187 HIV-positive individuals on combined antiretroviral therapy (cART) who were monitored electronically for adherence. VL was defined as a viral load >50 copies/mL. The model incorporated various predictors, including adherence patterns, demographic factors, and clinical variables. The evaluation metrics included sensitivity, specificity, and the AUC-ROC. The results showed that the Random Forest model achieved an AUC-ROC of 0.84, indicating good predictive performance. Key predictors of virologic failure included poor adherence, lower baseline CD4 count, and higher baseline viral load. Robbins et al. (2010) conducted a study in an HIV clinic to develop a predictive model for virologic failure among patients on ART. The study included 1,064 patients who initiated ART between 2000 and 2007, with virologic failure defined as a viral load >400 copies/mL after six months of therapy. The researchers used logistic regression as their modeling approach and evaluated performance using sensitivity, specificity, and the AUC-ROC. The results showed that 21% (224 patients) experienced virologic failure. Key predictors included higher baseline viral load (adjusted odds ratio [AOR] 2.4, 95% CI: 1.8–3.2), lower baseline CD4 count (AOR 1.5, 95% CI: 1.2–2.1), and suboptimal adherence (AOR 3.1, 95% CI: 2.3–4.5). The final model achieved an AUC-ROC of 0.74, indicating moderate predictive performance. The two studies showed that predictive models could help identify high-risk patients early, allowing for targeted interventions to improve treatment outcomes.
(Kamal et al., 2021; Robbins et al., 2010). However, our study differs by employing more advanced machine learning techniques, including XGBoost and Random Forest, which can capture complex, nonlinear relationships among predictor variables without requiring explicit assumptions about data distribution. Additionally, while robin et al primarily relied on traditional clinical and demographic predictors, our approach expands the feature set to include temporal and locational attributes, which may offer new insights into virological failure risks.

Yashik & Maurice (2012) explored the use of AI and expert systems to predict a single HIV drug resistance measure based on three international interpretation gold standards. Their study focused on improving the accuracy of drug resistance classification for individuals on highly active antiretroviral therapy (HAART)(Yashik & Maurice, 2012). The researchers found that the AI model demonstrated high concordance with the gold standards, improving classification accuracy in drug resistance prediction. The results indicated that AI could effectively automate and enhance the interpretation of HIV drug resistance, reducing inconsistencies observed in

manual assessments. However, specific numerical results, such as accuracy rates, sensitivity, and specificity, were not explicitly stated.

Voux and Maskew (2021) applied ML algorithms to predict retention and viral suppression in South African HIV treatment cohorts. Analyzing data from 445,636 patients for retention and 363,977 for VL suppression, they used logistic regression, random forests, and gradient boosting models. Their models achieved an AUC of 0.69 for predicting attendance at the next scheduled visit and 0.76 for predicting VL suppression. Key predictors for both outcomes included prior late visits, number of prior VL tests, time since the last visit, number of visits on the current regimen, age, and treatment duration. For retention, additional predictors were the number of visits at the current facility and details of the next appointment date, while for VL suppression, the range of previous VL values was also significant. (Voux & Maskew, n.d.). Our study builds on this approach by utilizing XGBoost and Random Forest models, which offer enhanced interpretability and efficiency in handling imbalanced datasets. While Voux & Maskew's primarily focused on demographic and clinical characteristics, our approach incorporates temporal and locational attributes, providing a more holistic view of virological failure risk. By expanding the range of predictive features, our study aims to refine risk stratification and optimize targeted interventions for patients at high risk of treatment failure.

Gallego et al. (2004) conducted a study to evaluate the correlation between rules-based interpretation and virtual phenotype interpretation of HIV-1 genotypes in predicting drug resistance. The study was conducted in Spain and analyzed HIV-1 genetic sequences from 100 HIV-infected individuals. The study period was not explicitly mentioned but aligns with the early 2000s when virtual phenotype testing was gaining traction. The researchers used two interpretation methods: a rules-based system (Stanford HIVdb) and a virtual phenotype system (Virco) to predict resistance to antiretroviral drugs. They assessed the agreement between these methods using correlation coefficients and percentage agreement rates. The results showed a strong correlation ($r = 0.85$) between the two interpretation methods across multiple drug classes, with 85% agreement for nucleoside reverse transcriptase inhibitors (NRTIs), 88% for non-nucleoside reverse transcriptase inhibitors (NNRTIs), and 79% for protease inhibitors (PIs). Despite high correlation levels, discrepancies occurred in specific cases where mutations led to differing resistance predictions(Gallego et al., 2004). While Gallego et al. (2004) examined genotype-based resistance prediction, our study leverages clinical, demographic, and adherence-related factors to predict virological failure. This distinction underscores the complementary nature of different predictive approaches in optimizing HIV treatment outcomes.

Fahey et al. (2022) conducted a study in Tanzania using machine learning to predict the risk of disengagement from HIV care based on routine electronic medical record (EMR) data. The study analyzed data from 2016 to 2021, including 73,115 individuals receiving HIV care. Several machine learning models were applied, including Random Forest, Gradient Boosting Machine (GBM), and Logistic Regression, with model performance evaluated using AUC-ROC and PR-AUC. The best-performing model, Random Forest, achieved an AUC-ROC of 0.78, demonstrating moderate predictive accuracy. Key predictors of disengagement included missed appointments, recent viral load results, CD4 count, and time since ART initiation(Fahey et al., 2022). While Fahey et al. (2022) focused on predicting disengagement from HIV care our study

specifically aims to predict virological failure among individuals already on ART. This distinction is crucial, as virological failure necessitates timely interventions to prevent disease progression and drug resistance. our study incorporates clinical, demographic, and adherence-related predictors, extending beyond behavioral and socioeconomic factors to enhance predictive accuracy.

In 2023, Mamo et al. applied machine learning techniques to predict VF among HIV-positive patients undergoing ART at the University of Gondar Comprehensive and Specialized Hospital in Ethiopia. The researchers utilized seven supervised classification algorithms, identifying the Random Forest classifier as the most effective, with an impressive AUC of 0.9989. Key predictors of VF included male gender, younger age, extended duration on ART, non-receipt of cotrimoxazole preventive therapy (CPT) and tuberculosis preventive therapy (TPT), secondary education level, specific ART regimens (TDF-3TC-EFV), and low CD4 counts, with CD4 count being the most significant factor(Mamo et al., 2023). Our study differs from Mamo et al.'s work in several aspects. While both studies employ machine learning to predict VF, we focus on a Kenyan cohort, thereby addressing potential geographical and population-specific variations in HIV treatment outcomes. Additionally, our research emphasizes the integration of temporal and locational attributes alongside clinical and demographic factors, aiming to capture a more comprehensive set of predictors. Methodologically, we compare the performance of XGBoost and Random Forest models, assessing their effectiveness in handling imbalanced datasets and their applicability in real-world clinical settings.

There is a growing body of literature on predictive modelling for virological failure in Kenya. Majumder et al. conducted a study in Kenya to develop a machine learning model aimed at predicting HIV VL hotspots, serving as an early warning system for health administrators. The study analyzed data from approximately 4 million VL tests across 4,265 health facilities. The study utilized demographic, clinical, and geographic data to train predictive models, including Random Forest, Gradient Boosting Machines (GBM), and Neural Networks. Model performance was evaluated using AUC-ROC and other relevant metrics. The results showed that the best-performing model achieved an AUC-ROC of approximately 0.85, indicating strong predictive capability. Key predictors identified included patient demographics, ART adherence levels, and regional healthcare access indicators. The study acknowledged certain limitations. Firstly, the reliance on routinely collected data may introduce inconsistencies or inaccuracies due to data entry errors or missing information. Secondly, the model's performance might be influenced by unmeasured confounding variables not captured in the available datasets. Lastly, while the model demonstrated good predictive accuracy, its generalizability to other settings or populations requires further validation. Majumder et al.'s study focused on predicting HIV viral load hotspots at a population level using machine learning, aiming to help health administrators identify high-risk geographic areas. In contrast, our study aims to predict individual patient risk of virological failure, allowing for targeted clinical interventions. Their study used aggregate real-world data from 4 million viral load tests across 4,265 health facilities, while ours leverages individual patient-level data, incorporating demographic, clinical, and treatment adherence factors. Additionally, while both studies apply machine learning, Majumder et al. employed spatial models for hotspot detection, whereas our study focuses on predictive modeling to assess patient-specific risks(Majumder et al., n.d.).

Chapter 3

Methods

3.1. Introduction

This chapter outlined the data set used and approach taken to apply and validate a predictive model aimed at identifying patients in Kenya who are at high risk of VF. The subsequent section explains the criteria for selecting key variables and the steps involved in building the XGBoost and Random Forest model. Validation techniques used to assess the model's reliability are then described, followed by an examination of the performance metrics applied in the evaluation process. Overall, this chapter provides a structured explanation of the methodology, ensuring clarity on how the model was designed, tested, and utilized to pinpoint high-risk patients.

3.2. Study Design and Population

The study design for this research was a retrospective cohort study. The study population consisted of HIV-infected patients who initiated ART at a health facility in Kenya. Data was sourced from EMRs, capturing patient socio-demographics, clinical characteristics, laboratory results, and medication history. Eligibility criteria included patients with at least one recorded viral load and exclude those who have never had their viral load taken and not documented. The study employed a multistage sampling method within administrative levels, which includes sampling from counties, facilities, and then individuals. This approach is suitable for our dataset, which covers 1.2 million PLHIV out of a national total of 1.3 million.

In the first stage, we randomly selected a sample of counties to ensure geographic diversity and representation of different administrative areas. In the second stage, facilities within these selected counties were sampled. This step ensured that a variety of healthcare settings are included, capturing the diversity of services and populations across regions. Finally, in the third stage, individuals were randomly sampled from within the selected facilities. The study included sample size of approximately 122,500 participants and had a margin of error of about 0.28%, ensuring high precision. The value p=0. was chosen because it represents the maximum variability in a population, leading to the most conservative (largest) sample size estimate.

To calculate the sample size, we denote:

- N as the total population size,
- n as the sample size,
- e as the margin of error.
- Z as the z-score

Given that N=1,200,000 and the desired margin of error is approximately 0.28%

$$n = \frac{Z^2 \times p(1-p)}{e^2} \tag{3.1}$$

$$n = \frac{(1.96)^2 \times 0.5(1-0.5)}{(0.0028)^2} \approx 122{,}500$$

3.3. **Data Sources**

3.3.1. **National Data Warehouse**

Data was obtained from the Kenya National Data Warehouse (NDW), an integrated longitudinal data repository for de-identified HIV data from over 2,000 facilities with Electronic Medical Records (EMR) in 45 of the 47 counties in Kenya. HIV delivery data is entered at EMR by healthcare workers during or after service provision. At facility level, the data is checked for duplicates and uploaded to the NDW monthly using a Data Warehouse Application Programming Interface. Variables of interest for this study were extracted from the NDW using Structured Query Language. Confidentiality was assured through fully anonymized patient identifiers by a hashing algorithm before extraction and investigators did not interact with the subjects.

3.3.2. **Geospatial Data**

The study utilized geographic information system coordinates for health facilities in Kenya to generate locational features using various publicly available geospatial data sets. Predictor variables from IHME, estimates from 2017, included estimates of HIV prevalence and prevalence of protective and risk factors such as condom use, sexual activity among youth, and multiple sexual partners. Predictor variables from WorldPop, estimates from 2020, included estimates of demographic and social factors including number of births, literacy rate, share of women that receive at least four ante-natal care visits before delivery, share of births with skilled birth attendance, and share of women who receive post-natal care within 48 hours of delivery. Data from Meta, estimates from 2020, included demographic density. We add facility-level features including ownership type (NGO, faith-based, public), service level (known as KEPH, or Kenya Essential Package for Health), and the number of patients actively on treatment

3.4. Variables

3.4.1. Outcome variables

The target variable in our study was virological failure, which we defined as a viral load measurement above the threshold level (200 copies/ml) on a clinical appointment.

3.4.2. Predictor variables

Table 1:Feature Generation to be included in the ML model

Feature Generation

VL History
Total number of viral loads taken
Share of viral loads that showed treatment failure
Share of viral loads that showed treatment failure in past three years
Most recent VL result
Share of consecutive unsuppressed VL
Share of >12months duration between two VLs

Visit History
History of Optimal/Suboptimal Adherence to ART and CTX
Stability Designation
Differentiated Care Status
Weight
Share DSD
BMI
Breastfeeding
Pregnancy
Share of Unscheduled Visits

Lateness History
Total touchpoints (clinical visits & pharmacy pickups)
Share of touchpoints for which patient was late by 1 day, 5 days, and 30 days
Share of five most recent touchpoints for which patient was late by 1 day, 5 days, and 30 days
Number of days between two most recent visits

Regimen History
Number of HIV regimens in past year
Status on Optimized HIV Regimen
Status on Other Regimens

Demographic Attributes
Gender
Age
Patient Source
Marital Status
Age at ART Start
Time on ART

Contextual Variables
Population Distribution
Prevalence of Social Factors such as Partner away, Multiple Sexual Partners, STI, among others also used in HTS and IIT application

3.5. Feature Engineering

XGBoost and Random forests models were trained using historical data from 2022 January through 2023 december to estimate the probability of VF at each patient visit, allowing these probabilities to adjust dynamically based on changes in patient conditions and variations in the timing of clinical encounters. By predicting VF risk for every appointment, the goal was to identify individuals at high risk of treatment failure at any stage of their care journey.

Data preprocessing involved identifying and handling missing values and inconsistencies, ensuring that only patients with complete outcome features were included. To enhance model performance, additional features were derived from existing data, incorporating demographic, clinical, and laboratory attributes. For instance, BMI was calculated from weight and height, while the patient's age at ART initiation was determined from their date of birth and treatment start date.

Since the XGBoost algorithm requires numerical input, categorical variables were transformed using one-hot encoding, converting qualitative data into binary dummy variables. This method ensures that each category is assigned a unique numerical representation, enabling the model to process and analyze the data effectively.

Following preprocessing and feature engineering, datasets were divided so that models learn patterns from one subset, the training set, and then generate predictions against a different subset, the test set, to generate an estimate of how accurately models will perform on new observations. We adopted a temporal split where we restricted observations in our training set between January 2022 and April 2023, and restricted observations in our test set to a later period between May through December 2023

3.6. **ML-Ready Dataset: Missing Data Imputation**

The level of completeness among generated features differed. Data quality plays a critical role of data in AI-driven prognostics and health management, missing data being as a major challenge that can compromise model accuracy and reliability. Cattaneo et al (2022) discussed various strategies for handling missing data, including imputation techniques such as mean, median, or mode substitution, which replace missing values with central tendency measures. To that end, we generated two versions of an ML-ready dataset. We assumed that data is missing at random and generate an imputed dataset using simple imputation as described in the table below.

Table 2: Imputation methods

Imputation Method	How it Works
Non imputed dataset (Sparse)	Training with missing values. XGBoost is designed to handle missing values gracefully. During training, it learns which branch (left or right) to assign to instances with missing values at each split. Other models are not robust enough.
Imputed dataset (Simple)	Mean imputation for categorical variables, while mode imputation for numeric variables.

To prevent data leakage during imputation, the dataset was first split into training and test sets. Imputation statistics were then calculated only on the training set and applied to both training and test data, ensuring the model did not gain unintended future information from the test set.

3.7. **Model Training and Testing**

We employed multiple algorithms, including XGBoost and Random Forest regression, and trained various model variations for each. We use a modified version of 5-fold cross validation, called temporal cross validation. Temporal cross-validation respects the chronological order of observations. The training set consisted of past data, while the test sets contain future observations, preventing data leakage. To enhance robustness and generalizability, we experimented with different hyperparameter combinations—such as learning rate, number of trees, and maximum depth—alongside different imputation methods to identify the best-performing model.

3.7.1. **XGBoost**

XGBoost is a boosting algorithm that builds trees sequentially to correct the errors of previous trees. It is an extension of the gradient boosting algorithm that is optimized for speed and performance. The objective function in XGBoost consists of two parts; loss function which measures how well the model fits the training data and regularization term controls model complexity to prevent overfitting.

$$\mathcal{O}(t) = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i^{(t-1)}} + f_t(x_i)\right) + \Omega(f_t) \tag{3.2}$$

Where y_i is the actual target value $\widehat{y_i^{(t-1)}}$ is the prediction from the previous iteration $f_t(x_i)$is the new tree's prediction $\Omega(f_t)$ is the regularization term. To find the best weight for each leaf node, XGBoost minimizes the objective function:

$$w^* = -\frac{\sum_{i\in j} g_i}{\sum_{i\in j} h_i + \lambda} \tag{3.3}$$

Where j represents a leaf node $\sum \mathrm{g_i}$ and $\sum h_i$are sums of gradients and Hessians for all samples in that leaf. Each decision tree is a function of the input features and is trained to minimize the loss function for a given set of weights $w*$. The weights are updated at each iteration using gradient descent, which involves computing the gradient of the loss function with respect to the weights. XGBoost then selects the best split by maximizing the **gain** in loss reduction

$$Gain = \frac{1}{2}\left[\frac{(\sum_{i\in L} g_i)^2}{\sum_{i\in L} h_i + \lambda} + \frac{(\sum_{i\in R} g_i)^2}{\sum_{i\in R} h_i + \lambda} - \frac{(\sum_{i\in P} g_i)^2}{\sum_{i\in P} h_i + \lambda}\right] - v \tag{3.4}$$

Where L and R are the left and right child nodes, P is the parent, v is the pruning parameter.

3.7.2. **Random Forest**

The Random Forest algorithm is an ensemble learning method that constructs multiple decision trees to improve prediction accuracy and reduce overfitting. It operates through bootstrap aggregation (bagging) and random feature selection, enhancing its robustness for both classification and regression tasks. The prediction of a new data point is made by passing it through each tree in the forest and taking the majority vote of the predictions.

Given a dataset with observations and features, multiple bootstrap samples are created by randomly selecting samples with replacement. For each bootstrap sample, a decision tree is constructed as follows:

- At each node, a random subset of features is selected.
- The best feature is chosen based on impurity measure:

$$Gini = 1 - \sum_{i=1}^{c} p_i^2 \tag{3.5}$$

Where pi = Probability of an instance belonging to class i. c = Total number of classes, lower values indicate purer nodes. After training multiple trees, the final prediction is made by majority voting

$$\hat{y} = \arg\max_{c} \sum_{b=1}^{N} I(T_b(x) = c) \tag{3.6}$$

Where Tb(x) = Prediction of the b-th tree, N = Total number of trees and I(·) is indicator function (1 if true, 0 otherwise)
This rigorous approach allowed us to thoroughly explore the performance of different models and identify the optimal approach.

3.8. **Hyperparameter Tuning**

Grid-search was conducted to systematically explore different hyperparameter combinations and identify the optimal settings that maximized model performance. A predefined set of hyperparameter values for each model was selected -learning rate, maximum depth, number of rounds- and the model was trained and evaluated using each combination. Cross-validation was performed on the training set to assess performance, and the best-performing combination was

identified based on AUCPR. Finally, the optimal hyperparameters were used to train the final model on the full training data.

3.9. **Model Validation**

To evaluate model performance, we considered the Area Under the Precision Recall Curve, a metric that is particularly well-suited to instances of imbalanced data because it focuses on performance of the positive class. We also considered Area Under the Receiver Operating Characteristic Curve (AUC-ROC), the more traditional evaluation metric that reflects performance against both suppressed and non-suppressed cases.

Chapter 4

Results and Interpretation

4.1. **Introduction**

The performance of three machine learning models—XGB Sparse, XGB Simple, and RF Simple—was evaluated for their ability to detect virological unsuppression, which represented the minority class in this classification task. The AUC-PR was employed to specifically measure the models' effectiveness in identifying unsuppressed cases (true positives) while minimizing false positives. This metric is particularly useful in handling imbalanced datasets, as it prioritizes performance on the minority class rather than being skewed by the majority class. AUC-ROC was used to evaluate the overall discriminatory power of the models, assessing their ability to distinguish between suppressed and unsuppressed cases by analyzing the trade-off between sensitivity (true positive rate) and specificity (false positive rate) across all classification thresholds. The dataset exhibited an imbalanced distribution, with the minority class (unsuppressed cases) constituting only 9% of the total data. Women represented 67% of all observations and had a VL rate of 8.46%; compared to 33% for men and a VL rate of 9.2%. Patients over 15 represented 93.8% of observations with a VL rate of 7.63%, compared to 9.2% for patients under 15 and a VL rate of 9.09%..

4.2. **XGB Sparse**

The XGB Sparse model achieved a lift of 4.93 over a random classifier (0.444 / 0.09), meaning it is nearly five times more effective than random guessing at detecting unsuppressed cases. This high value reflects a well-balanced ability to maintain precision (fewer false positives) while achieving high recall (identifying most unsuppressed cases

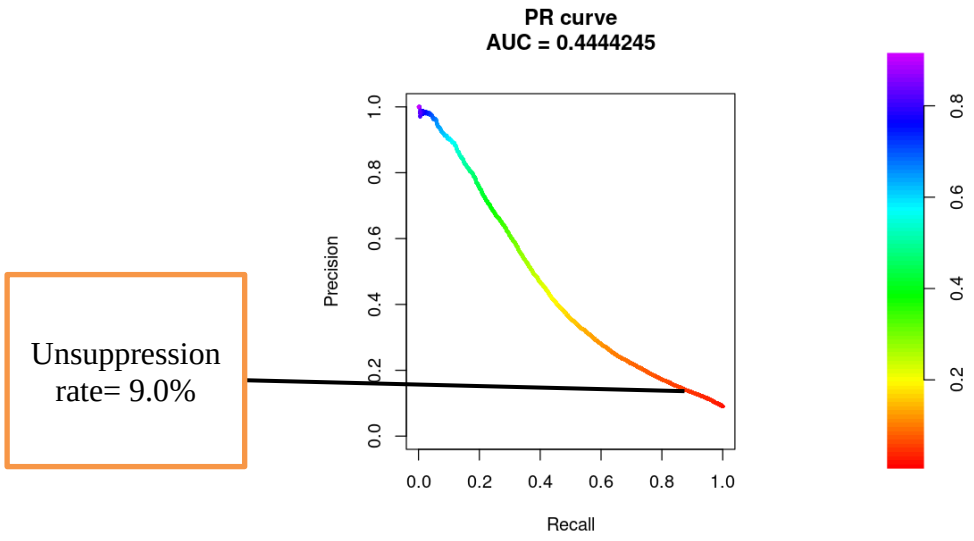


Figure 1: AUC PR for XGB Sparse

An AUC-ROC of 0.804 indicates strong overall discrimination between suppressed and unsuppressed cases.The high value suggests that the model is able to achieve high sensitivity and specificity across various thresholds, balancing false positive and false negative rates. This value demonstrates the model's robustness for general classification tasks and ensures reliability even when threshold tuning is required.

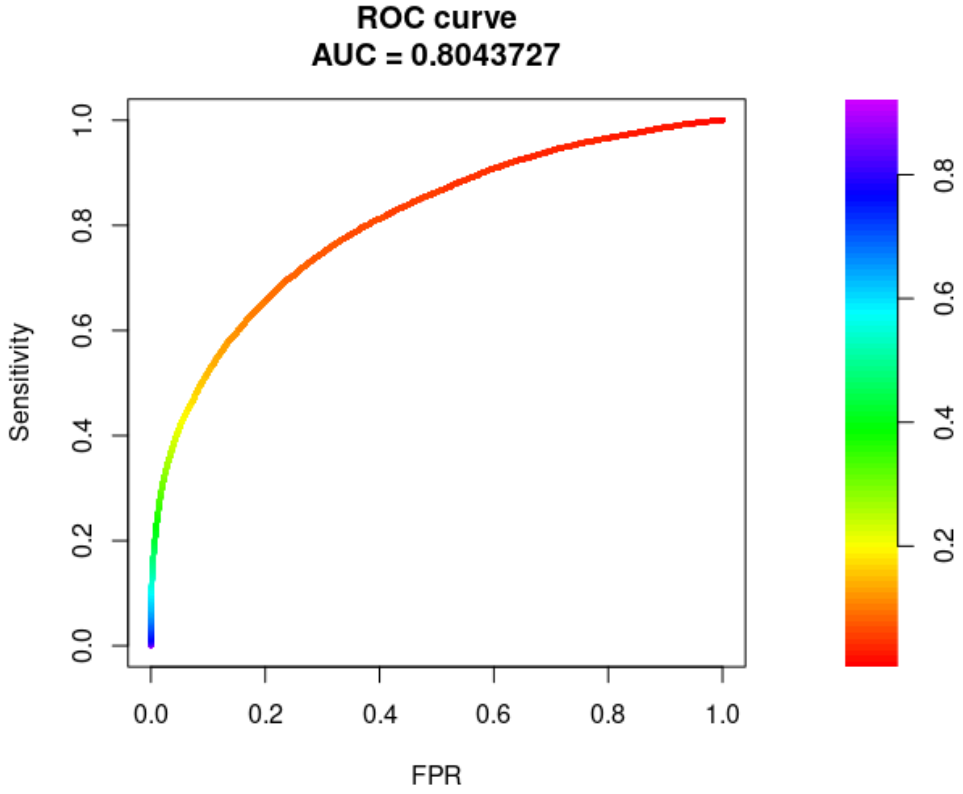


Figure 2: ROC curve for XGB Sparse

Using SHAP analysis, the table below shows 20 most important features to the XGBoost model from the original input of 64 predictor variables. The table the features ranked by their gain values, visualizes features that contribute the most to the predictive power of your XGBoost model. The twenty most important features were a variety of feature categories including unsuppression history, demographics (Age), clinical history (optimized regimen, weight ,time to next appointment, recent treatment failure and BMI), temporal attributes (timeon art), and locational attributes (hiv prevalence, poverty and circumcission).

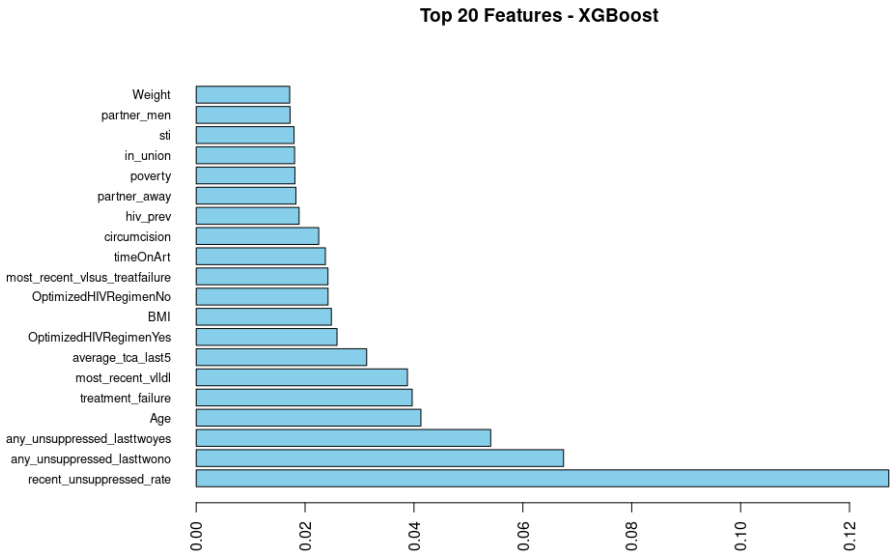


Figure 3: Feature Importance for best performing model (XGB sparse)

4.3. **XGB Simple**

The XGB Simple model achieves a **lift of 4.84** over a random classifier, slightly lower than the XGB Sparse model. While still highly effective, this suggests that imputing missing values may introduce minor inaccuracies compared to directly working with sparse data. The model maintains a strong balance between precision and recall, making it a viable alternative to XGB Sparse in situations where imputation simplifies preprocessing. The slight reduction in AUC-PR reflects the potential trade-offs of imputing missing values, particularly in imbalanced datasets.

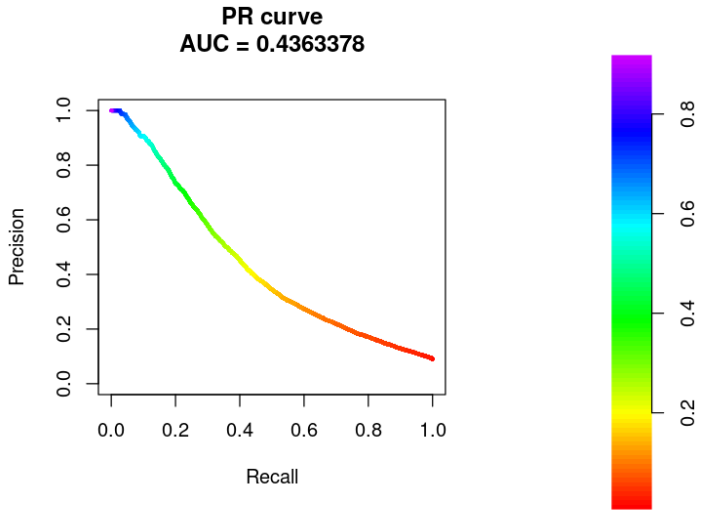


Figure 4: AUC PR for XGB Simple

The AUC-ROC of 0.799 is comparable to the XGB Sparse model, indicating that the XGB Simple model can still reliably distinguish between the two classes. The consistent AUC-ROC value demonstrates that imputing data does not significantly hinder the model's ability to handle class imbalance, though it slightly impacts precision and recall.

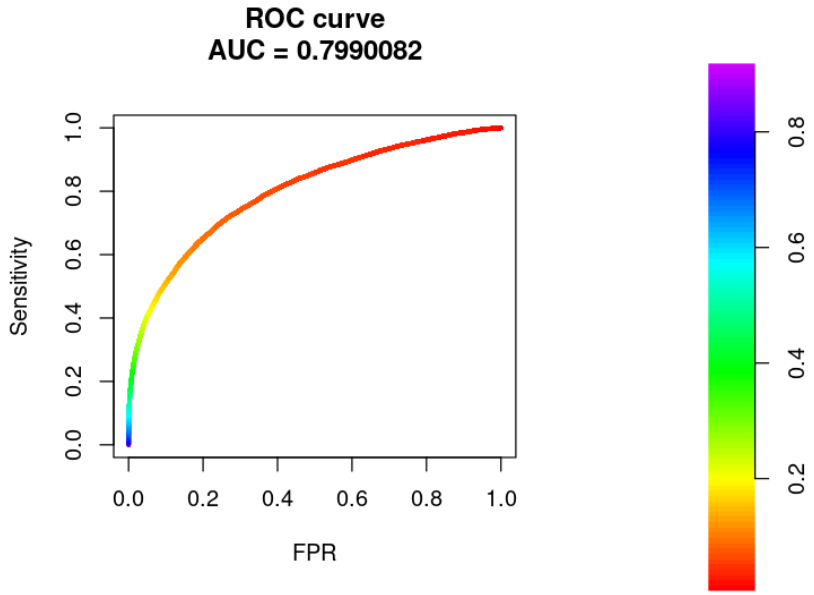


Figure 5: ROC Curve for XGB Simple

4.4. **RF Simple**

The RF Simple model achieves a lift of 3.22 over a random classifier (0.290 / 0.09). While still better than random guessing, its performance is significantly lower than both XGB models. The lower AUC-PR suggests that the model struggles to balance precision and recall effectively, leading to a higher rate of false positives or missed unsuppressed cases. The RF Simple model is less suited for detecting minority classes, particularly in scenarios where the minority class is critical to identify.

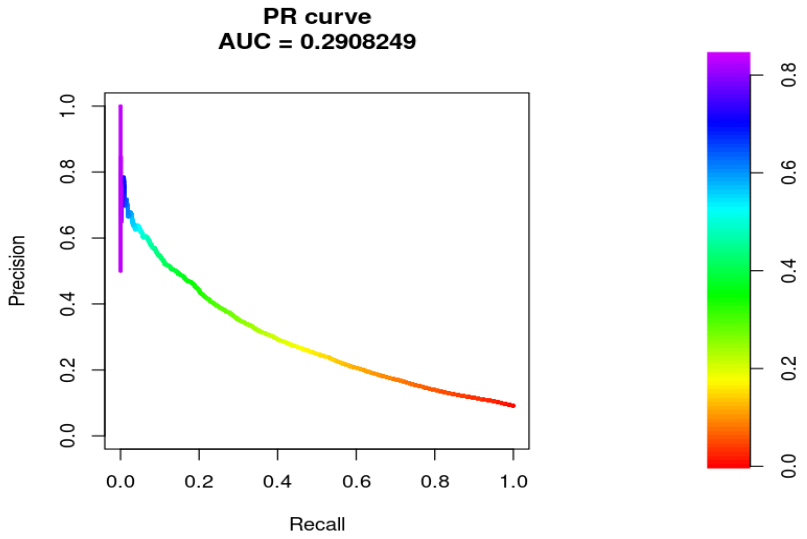


Figure 6: AUC PR for RF Simple

The AUC-ROC of 0.740 further highlights the model's limited ability to discriminate between suppressed and unsuppressed cases compared to XGBoost models. The lower value indicates that the RF model has less robust classification capabilities compared to XGBoost models.

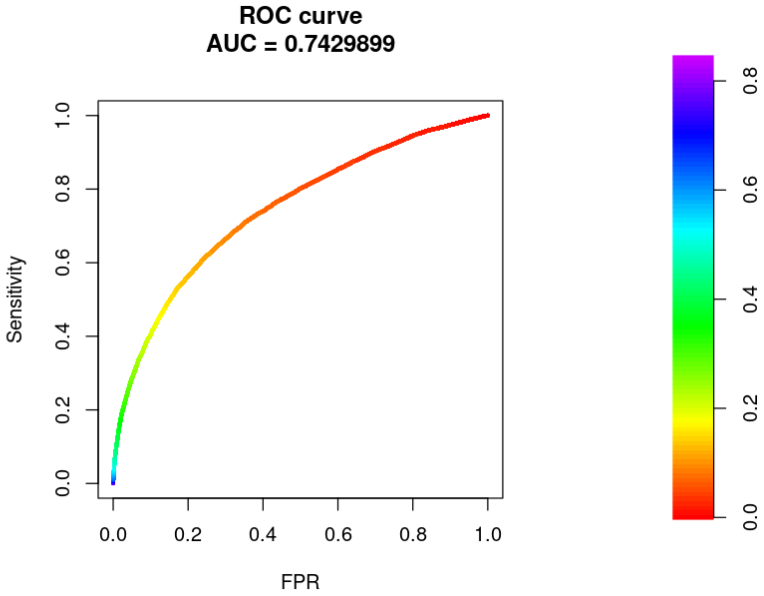


Figure 7: ROC Curve for RF Simple

4.5. Comparative Analysis

4.5.1. Performance by Metric

Among the three evaluated models, XGB Sparse demonstrated the strongest performance in detecting virological unsuppression. It achieved the highest AUC-PR with a lift of 4.93, indicating its superior ability to handle the minority class. In terms of overall discrimination, both XGB Sparse and XGB Simple performed comparably, with AUC-ROC values of 0.804 and 0.799, respectively. The RF Simple model, however, lagged with an AUC-ROC of 0.740, reflecting its limited capacity to distinguish between suppressed and unsuppressed cases.

4.5.2. Effect of Data Handling

The handling of missing data had a noticeable impact on model performance. The XGB Simple model, which used mean and mode imputation to handle missing values, showed a slight reduction in performance compared to the XGB Sparse model, which directly handled sparse data. Despite this, XGB Simple remained a strong alternative, especially for implementation in simpler machine learning pipelines. On the other hand, the RF Simple model exhibited considerable limitations, reinforcing its reduced effectiveness in scenarios where precise minority class detection is crucial.

4.5.3. Lift Analysis

The ability of the models to improve classification beyond random guessing was further assessed through lift analysis. The XGB Sparse and XGB Simple models both achieved a lift of approximately 4.9 times, underscoring their robustness in handling imbalanced datasets. In contrast, the RF Simple model achieved a lift of only 3.2 times, demonstrating its weaker ability to prioritize and correctly classify the minority class. This analysis further supports the superiority of the XGB models, particularly XGB Sparse, in identifying unsuppressed cases with high reliability.

Chapter 5

Discussion, Conclusion and Recommendations

5.1. **Discussion**

This study evaluated the performance of three machine learning models—XGB Sparse, XGB Simple, and RF Simple—in predicting virological failure among PLHIV in Kenya. The results demonstrated that both XGBoost models significantly outperformed the Random Forest model, with the XGB Sparse model achieving the highest predictive power. To enhance model prediction accuracy and generalizability, the selected models were used to develop and validate the best predictive model based on key predictors. The classifiers were trained using a stratified 5-fold cross-validation approach with default hyperparameter settings. Multiple experiments were conducted to determine the highest accuracy, comparing both imputed and non-imputed datasets. The results indicated that the imputed data yielded lower performance metrics. These findings align with previous studies that have applied machine learning to predict virological outcomes in HIV care.

Seboka et al. (2023) used machine learning models, including artificial neural networks and decision tree-based classifiers, to predict viral load suppression status. Their study highlighted the superiority of tree-based models in handling imbalanced data, which is consistent with our findings that XGBoost models—particularly the sparsity-aware XGB Sparse—were the most effective at identifying unsuppressed cases(Seboka et al., 2023). While Seboka et al did not report lift scores, their models achieved high precision-recall performance, reinforcing the utility of boosting algorithms in HIV predictive modeling.

Ahoua et al. (2009) identified key risk factors for virological failure, including subtherapeutic antiretroviral concentrations and non-adherence. While their study relied on traditional statistical methods rather than machine learning, their findings on demographic and clinical predictors of virological failure align with the feature importance analysis in our study(Ahoua et al., 2009). Our SHAP analysis identified demographic variables such as age, as well as clinical factors including treatment regimen, weight, and scheduled clinic appointments (TCA), as key contributors to model predictions. This consistency suggests that machine learning models can effectively capture the relationships established by conventional epidemiological methods while enhancing predictive accuracy.

Bisaso et al. (2018) compared logistic regression-based machine learning models for predicting early virological suppression and found that tree-based ensemble models provided superior

discrimination. Our findings corroborate this, as XGBoost models consistently outperformed RF Simple in terms of both AUC-PR and AUC-ROC(Bisaso et al., 2018). However, Bisaso et al. (2018) reported that missing data handling played a crucial role in model performance. This is reflected in our results, where XGB Simple, which imputed missing values, had slightly lower performance than XGB Sparse, which directly leveraged sparsity.

Kamal et al. (2021) employed Random Forest models to predict virological outcomes using electronically monitored adherence data in a Swiss cohort. Their study achieved strong predictive performance but found that Random Forest models were sensitive to feature selection and data imbalances(Kamal et al., 2021). Our results indicate similar limitations, as the RF Simple model exhibited weaker discrimination (AUC-ROC = 0.740) and reduced minority class detection (lift = 3.22). This suggests that while Random Forest models are useful in certain contexts, boosting methods like XGBoost may be more robust for handling imbalanced datasets in virological failure prediction.

Robbins et al. (2010) developed a predictive model for virological failure in an HIV clinic setting using clinical and demographic factors. Their study demonstrated that combining longitudinal clinical data with statistical modeling improved prediction accuracy. Our findings similarly show that temporal attributes, such as time on ART, were among the most important predictors in our XGBoost models(Robbins et al., 2010). The strong performance of XGB Sparse in our study further supports the notion that advanced machine learning techniques, particularly those leveraging sparsity and feature importance ranking, can enhance virological failure prediction beyond traditional regression-based approaches.

Mamo et al. (2023) reported that the Random Forest classifier achieved an AUC of 0.9989, indicating exceptional predictive performance. In contrast, our study found that the XGB Sparse model attained an AUC-ROC of 0.804, while the RF Simple model reached 0.740. The higher AUC in Mamo et al.'s study suggests a superior model performance; however, this discrepancy may stem from differences in dataset characteristics, feature selection, or model tuning. However, both studies identified critical predictors of virological failure. Mamo et al. highlighted factors such as male gender, younger age, longer duration on ART, non-use of CPT and TPT, secondary education level, specific ART regimens (TDF-3TC-EFV), and low CD4 counts, with CD4 count being the most significant(Mamo et al., 2023). Similarly, our study recognized features including unsuppression history, age, optimized regimen, weight, time on ART, and demographic densities as influential. The overlap in predictors, particularly regarding age, duration on ART, and CD4 count, underscores their importance in virological failure risk assessment. While both studies employed machine learning techniques, Mamo et al. utilized seven supervised classification algorithms, with Random Forest emerging as the top performer. Our study, however, concentrated on three models: XGB Sparse, XGB Simple, and RF Simple, finding XGB Sparse to be the most effective. This variation in methodological approaches highlights the potential benefits of exploring multiple algorithms to identify the most suitable model for specific datasets.

Overall, our study contributed to the growing body of evidence supporting the application of machine learning in HIV care. The superior performance of XGBoost models highlights the

importance of selecting appropriate algorithms for imbalanced classification problems which mimic the real-world scenarios of outcome distribution.

5.2. **Limitations of the Analysis**

Despite the promising results of the predictive models in identifying patients at high risk of virological failure, several limitations must be acknowledged. The model's predictive power was heavily influenced by the quality and completeness of the input features(Cattaneo et al., 2022). Missing data was handled through imputation in the XGB Simple and RF Simple models, which slightly reduced performance compared to the XGB Sparse model. While XGBoost outperformed Random Forest in this study, the choice of algorithms was limited to tree-based models. Other machine learning approaches, such as deep learning or Bayesian networks, were not explored and may yield different insights. Furthermore, while SHAP analysis provided interpretability for XGBoost models, black-box nature of machine learning models remains a limitation for clinical implementation. Clinicians may require more interpretable models before integrating AI-driven decision support tools into routine HIV care

5.3. **Conclusion**

This study evaluated the performance of three machine learning models—XGB Sparse, XGB Simple, and RF Simple—in predicting virological failure among people living with HIV (PLHIV) in Kenya. The results demonstrate that XGBoost models, particularly the XGB Sparse model, are highly effective in identifying patients at high risk of virological failure, achieving an AUC-PR lift of 4.93 and an AUC-ROC of 0.804. Male, non-optimized regimen, younger age, shorter time on ART time on ART and recent unsuppressed viral load were significant features for predicting virological failure. These findings highlight the potential of machine learning in improving early detection and intervention for virological failure in Kenya, ultimately enhancing HIV treatment outcomes.

5.4. **Recommendations**

The SHAP analysis identified key predictors of virological suppression, including recent unsuppressed rate, any unsuppressed last two viral load tests, treatment failure, age, BMI, and ART regimen optimization. To improve treatment outcomes, targeted interventions should be implemented based on these findings. Standardized data collection protocols should be implemented to minimize missing values and inconsistencies. Explainable AI techniques should be further developed to enhance model interpretability for clinical decision-making. Predictive models should be integrated into electronic medical records (EMRs) to provide real-time decision support for clinicians.

References

Ahoua, L., Guenther, G., Pinoges, L., Anguzu, P., Chaix, M. L., Le Tiec, C., Balkan, S., Olson, D., Olaro, C., & Pujades-Rodríguez, M. (2009). Risk factors for virological failure and subtherapeutic antiretroviral drug concentrations in HIV-positive adults treated in rural northwestern Uganda. *BMC Infectious Diseases*, *9*. https://doi.org/10.1186/1471-2334-9-81

Allwein, E. L., & Schapire, R. E. (2000). Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers Yoram Singer. In *Journal of Machine Learning Research* (Vol. 1).

Andarge, D. E., Hailu, H. E., & Menna, T. (2022). Incidence, survival time and associated factors of virological failure among adult HIV/ AIDS patients on first line antiretroviral therapy in St. Paul's Hospital Millennium Medical College—A retrospective cohort study. *PLoS ONE*, *17*(10 October). https://doi.org/10.1371/journal.pone.0275204

Barnabas, G., Sibhatu, M. K., & Berhane, Y. (2017). Editorial: Antiretroviral therapy program in Ethiopia benefits from virology treatment monitoring. *Ethiopian Journal of Health Sciences*, *27*(1), 1. https://doi.org/10.4314/ejhs.v27i1.1S

Bisaso, K. R., Karungi, S. A., Kiragga, A., Mukonzo, J. K., & Castelnuovo, B. (2018). A comparative study of logistic regression based machine learning techniques for prediction of early virological suppression in antiretroviral initiating HIV patients. *BMC Medical Informatics and Decision Making*, *18*(1). https://doi.org/10.1186/s12911-018-0659-x

Cattaneo, L., Polenghi, A., Macchi, M., & Pesenti, V. (2022). On the role of Data Quality in AI-based Prognostics and Health Management. *IFAC-PapersOnLine*, *55*(19), 61–66. https://doi.org/10.1016/j.ifacol.2022.09.184

Drain, P. K., Dorward, J., Bender, A., Lillis, L., Marinucci, F., Sacks, J., Bershteyn, A., Boyle, D. S., Posner, J. D., & Garrett, N. (2019). *Point-of-Care HIV Viral Load Testing: an Essential Tool for a Sustainable Global HIV/AIDS Response*. http://cmr.asm.org/

Emery, S., Neuhaus, J. A., Phillips, A. N., Babiker, A., Cohen, C. J., Gatell, J. M., Girard, P. M., Grund, B., Law, M., Losso, M. H., Palfreeman, A., Wood, R., Gordin, F., Finley, E., Dietz, D., Chesson, C., Vjecha, M., Standridge, B., Schmetter, B., … Vacarezza, M. (2008). Major clinical outcomes in antiretroviral therapy (ART)-naive participants and in those not receiving ART at baseline in the SMART Study. *Journal of Infectious Diseases*, *197*(8), 1133–1144. https://doi.org/10.1086/586713

Fahey, C. A., Wei, L., Njau, P. F., Shabani, S., Kwilasa, S., Maokola, W., Packel, L., Zheng, Z., Wang, J., & McCoy, S. I. (2022). Machine learning with routine electronic medical record data to identify people at high risk of disengagement from HIV care in Tanzania. *PLOS Global Public Health*, *2*(9), e0000720. https://doi.org/10.1371/journal.pgph.0000720

Gallego, O., Martin-Carbonero, L., Aguero, J., Mendoza, C. de, Corral, A., & Soriano, V. (2004). Correlation between rules-based interpretation and virtual phenotype interpretation of HIV-1 genotypes for predicting drug resistance in HIV-infected individuals. *Journal of Virological Methods*, *121*(1), 115–118. https://doi.org/10.1016/j.jviromet.2004.06.003

Kamal, S., Urata, J., Cavassini, M., Liu, H., Kouyos, R., Bugnon, O., Wang, W., & 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 - Psychological and Socio-Medical Aspects of AIDS/HIV*, *33*(4), 530–536. https://doi.org/10.1080/09540121.2020.1751045

Lin, Y.-H., Peng, B., Kim, H., Chai, C., Wu, C., Copyright, fpubh, He, J., Li, J., Jiang, S., Cheng, W., Jiang, J., Xu, Y., Yang, J., & Zhou, X. (n.d.). *Application of machine learning algorithms in predicting HIV infection among men who have sex with men: Model development and validation*.

Lohse, N., Ann-Brit, ;, Hansen, E., Pedersen, G., Kronborg, G., Gerstoft, J., Sørensen, H. T., Vaeth, M., & Obel, N. (2007). *Survival of Persons with and without HIV Infection in Denmark, 1995-2005*. www.annals.org

Majumder, A., Yan, P., Chung, J., Kagendi, N., Masyn, S., & Mwau, M. (n.d.). *A Novel Machine Learning Approach to Predict HIV Viral Load Hotspots in Kenya Using Real-World Data*. https://ssrn.com/abstract=4252673

Mamo, D. N., Yilma, T. M., Fekadie, M., Sebastian, Y., Bizuayehu, T., Melaku, M. S., & Walle, A. D. (2023). Machine learning to predict virological failure among HIV patients on antiretroviral therapy in the University of Gondar Comprehensive and Specialized Hospital, in Amhara Region, Ethiopia, 2022. *BMC Medical Informatics and Decision Making*, *23*(1). https://doi.org/10.1186/s12911-023-02167-7

Meshesha, H. M., Nigussie, Z. M., Asrat, A., & Mulatu, K. (2020). Determinants of virological failure among adults on first-line highly active antiretroviral therapy at public health facilities in Kombolcha town, Northeast, Ethiopia: A case-control study. *BMJ Open*, *10*(7). https://doi.org/10.1136/bmjopen-2019-036223

NASCOP. (2022). *Ministry of Health NATIONAL GUIDELINES FOR HIV/STI PROGRAMMING WITH KEY POPULATIONS*.

Osman, F. T., & Yizengaw, M. A. (2020). Virological Failure and Associated Risk Factors among HIV/AIDS Pediatric Patients at the ART Clinic of Jimma university Medical Center, Southwest Ethiopia. *The Open AIDS Journal*, *14*(1), 61–67. https://doi.org/10.2174/1874613602014010061

Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N., & Fanos, V. (2020). Comparison of conventional statistical methods with machine learning in medicine: Diagnosis, drug development, and treatment. *Medicina (Lithuania)*, *56*(9), 1–10. https://doi.org/10.3390/medicina56090455

Robbins, G. K., Johnson, K. L., Chang, Y., Jackson, K. E., Sax, P. E., Meigs, J. B., & Freedberg, K. A. (2010). Predicting virologic failure in an HIV clinic. *Clinical Infectious Diseases*, *50*(5), 779–786. https://doi.org/10.1086/650537

Seboka, B. T., Yehualashet, D. E., & Tesfa, G. A. (2023). Artificial Intelligence and Machine Learning Based Prediction of Viral Load and CD4 Status of People Living with HIV (PLWH) on Anti-Retroviral Treatment in Gedeo Zone Public Hospitals. *International Journal of General Medicine*, *16*, 435–451. https://doi.org/10.2147/IJGM.S397031

Shoko, C., & Chikobvu, D. (2019). A superiority of viral load over CD4 cell count when predicting mortality in HIV patients on therapy. *BMC Infectious Diseases*, *19*(1). https://doi.org/10.1186/s12879-019-3781-1

UNAIDS. (2021). *UNAIDS data 2021*.

Voux, D. L., & Maskew, M. (n.d.). *Machine learning to predict retention and viral suppression in South African HIV treatment cohorts*. https://doi.org/10.1101/2021.02.03.21251100

World Health Organisation. (2020).

Yashik, S., & Maurice, M. (2012). Predicting a single HIV drug resistance measure from three international interpretation gold standards Asian Pacific Journal of Tropical Medicine Drug resistance Antiretroviral therapy Highly active HIV Artificial intelligence Expert systems. In *Asian Pacific Journal of Tropical Medicine*. http://sierra2.

Appendices

Appendix A

R Code

```
library(dplyr)
library(caret)
library(xgboost)
library(PRROC)

# First, let's read in and combine input tables ----------------
vl <-
readRDS("~/Dropbox/VLPREP_CV/VL_samp_250k_v2.rds")
# # Add in GIS features ------------------------
library(readxl)
gis <- read_excel("Dropbox/July data/gis_hts_matrix.xlsx")
vl$SiteCode <- as.character(vl$SiteCode)
vl <- merge(vl, gis,by.x = "SiteCode", by.y = "FacilityCode",
all.x = TRUE) %>%
  dplyr::select(-SiteCode)
# 423,311 obs. 91 variables
set.seed(2231)
vl <- vl %>%
  group_by(key) %>%
  mutate(patient_group = sample(1:2, 1, prob = c(0.8,
0.2)))%>%
  ungroup

nrow(filter(vl, patient_group==1 & PredictionDate<"2023-01-
01"))
nrow(filter(vl, patient_group==2 & PredictionDate>="2023-
01-01"))

class<-vl%>%
  filter(patient_group==1 & PredictionDate>="2023-01-01")

group1<-filter(vl, patient_group==1 & PredictionDate<"2023-
01-01")
group2<-filter(vl, patient_group==2 &
PredictionDate>="2023-01-01")



# helper functions ----------------------------
encodeXGBoost <- function(dataset){
  # Need to one-hot encode all the factor variables
  ohe_features <- names(dataset)[ sapply(dataset, is.factor) |
sapply(dataset, is.character) ]

  dmy <- dummyVars("~ Gender + PatientSource +
MaritalStatus + PopulationType +
                   OptimizedHIVRegimen + Other_Regimen +
Pregnant + DifferentiatedCare +
                   StabilityAssessment +
most_recent_art_adherence + most_recent_ctx_adherence
+
                   most_recent_vl + Breastfeeding",
select(-patient_group) %>%
  arrange(PredictionDate) %>%
  filter(PredictionDate > "2022-05-31")

# Create folds
nfolds <- 10
cuts <- seq(min(train_data$PredictionDate),
max(train_data$PredictionDate), length.out = nfolds
+ 2)


grid_simple <- expand.grid(model = "xgboost",
                           sparsity = "simple",
                           eta = c(0.01, 0.1),
                           max_depth = c(6, 8, 10, 12),
                           cs = c(.3, .5, .7))

start <- Sys.time()

for(i in 1:nrow(grid_simple)){

  aucpr <- c()

  for(j in 1:nfolds){

    # Set train and validation
    tmp <-  train_data %>%
    group_by(key) %>%
    mutate(patient_group = sample(1:2, 1, prob =
c(0.7, 0.3)))

   train_tmp <- tmp %>%
     ungroup() %>%
     filter(patient_group == 1) %>%
     filter(between(PredictionDate, as.Date(cuts[1]),
as.Date(cuts[j+1]))) %>%
     select(-PredictionDate, - key)

   val_tmp <- tmp %>%
     ungroup() %>%
     filter(patient_group == 2) %>%
     filter(between(PredictionDate, cuts[j+1],
cuts[j+2])) %>%
     select(-PredictionDate, - key)

   train_tmp <- train_tmp %>%
select(intersect(names(train_tmp),
names(val_tmp)))
   val_tmp <- val_tmp %>%
select(intersect(names(train_tmp),
names(val_tmp)))
```

```
                data = dataset)
  ohe <- data.frame(predict(dmy, newdata = dataset))
  dataset <- cbind(dataset, ohe)

  dataset[, !(names(dataset) %in% ohe_features)]

}

Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
replaceWithMode <- function(dataset_calc, dataset_impute,
position){
  dataset_impute[, position] <- ifelse(is.na(dataset_impute[,
position]),
                                Mode(dataset_calc[,
position][!is.na(dataset_calc[,position])]),
                                dataset_impute[, position])
}
replaceWithMean <- function(dataset_calc, dataset_impute,
position){
  dataset_impute[ position] <- ifelse(is.na(dataset_impute[,
position]),
                                mean(dataset_calc[, position],
na.rm = TRUE),
                                dataset_impute[, position])
}

columns_to_exclude <- c("key", "PredictionDate",
"patient_group")

excluded_columns <- vl[, columns_to_exclude]

vl <- encodeXGBoost(vl[, !(names(vl) %in%
columns_to_exclude)])

# Bind the excluded columns back to vl
vl <- cbind(vl, excluded_columns)


# Let's do temporal cross validation first for sparse ----------
-------

train_data <- vl %>%
  ungroup() %>%
  filter(patient_group == 1) %>%
  select(-patient_group) %>%
  arrange(PredictionDate) %>%
  filter(PredictionDate >= "2021-01-01") %>%
  filter(PredictionDate <= "2022-08-30") #%>%
  # select(-PredictionDate,- key) %>%
  # encodeXGBoost()

test_data <- vl %>%
  ungroup() %>%
  filter(patient_group == 2) %>%
  select(-patient_group) %>%
  arrange(PredictionDate) %>%
  filter(PredictionDate >= "2022-09-01") %>%
  select(-PredictionDate, - key) #%>%
  #encodeXGBoost()
```

```
    # Impute
    train_tmp <- data.frame(train_tmp)
    val_tmp <- data.frame(val_tmp)

    for(k in which(sapply(train_tmp, class) %in%
c('character', 'factor'))){
      train_tmp[, k] <- replaceWithMode(train_tmp,
train_tmp, k)
      val_tmp[, k] <- replaceWithMode(train_tmp,
val_tmp, k)
    }

    for(k in which(sapply(train_tmp, class) %in%
c('numeric', 'integer'))){
      train_tmp[, k] <- replaceWithMean(train_tmp,
train_tmp, k)
      val_tmp[, k] <- replaceWithMean(train_tmp,
val_tmp, k)
    }

    # Encode
    train_tmp <- encodeXGBoost(train_tmp)
    val_tmp <- encodeXGBoost(val_tmp)

    train_tmp <- train_tmp %>%
select(intersect(names(train_tmp),
names(val_tmp)))
    val_tmp <- val_tmp %>%
select(intersect(names(train_tmp),
names(val_tmp)))

    dtrain <- xgb.DMatrix(data =
data.matrix(train_tmp[,which(names(train_tmp) !=
"Target")]),
                          label = train_tmp$Target)
    dval <- xgb.DMatrix(data =
data.matrix(val_tmp[,which(names(val_tmp) !=
"Target")]),
                        label = val_tmp$Target)
    watchlist <- list(train=dtrain, test=dval)

    set.seed(2231)
    xgb <- xgboost::xgb.train(data = dtrain,
                              eta = grid_simple[i, 3],
                              max_depth = grid_simple[i, 4],
                              colsample_bytree =
grid_simple[i, 5],
                              nround = 1000,
                              early_stopping_rounds = 50,
                              objective = "binary:logistic",
                              metric = 'aucpr',
                              eval.metric = "aucpr",
                              watchlist = watchlist,
                              verbose = 1
    )

    aucpr <- c(aucpr, xgb$best_score)
  }

  grid_simple$val_pr_auc[i] <- mean(aucpr)

}
```

```
# Let's do a single cross validation
# Create folds
nfolds <- 5
# cuts <- round(seq(1, nrow(train_data), length.out =
nfolds+2))
cuts <- seq(min(train_data$PredictionDate),
max(train_data$PredictionDate), length.out = nfolds + 2)
# cutgap <- cuts[2]/2

grid_sparse <- expand.grid(model = "xgboost",
                           sparsity = "sparse",
                           eta = c(0.01, 0.1),
                           max_depth = c(6, 8, 10, 12),
                           cs = c(.3, .5, .7))

start <- Sys.time()


for(i in 1:nrow(grid_sparse)){
  set.seed(2231)
  print(i)
  aucpr <- c()
  num_iter <- c()

  for(j in 1:nfolds){

    # Set train and validation
    tmp <-  train_data %>%
      group_by(key) %>%
      mutate(patient_group = sample(1:2, 1, prob = c(0.7,
0.3)))

    train_tmp <- tmp %>%
      ungroup() %>%
      filter(patient_group == 1) %>%
      filter(between(PredictionDate, cuts[1], cuts[j+1])) %>%
      select(-PredictionDate, - key, -patient_group) #%>%
      #encodeXGBoost()

    val_tmp <- tmp %>%
      ungroup() %>%
      filter(patient_group == 2) %>%
      filter(between(PredictionDate, cuts[j+1], cuts[j+2])) %>%
      select(-PredictionDate, - key,-patient_group) #%>%
      #encodeXGBoost()

    train_tmp <- train_tmp %>%
select(intersect(names(train_tmp), names(val_tmp)))
    val_tmp <- val_tmp %>%
select(intersect(names(train_tmp), names(val_tmp)))


    # train_tmp <- train_data[cuts[1]:cuts[j+1], ]
    # val_tmp <- train_data[(cuts[j+1]+1):(cuts[j+1]+cutgap), ]
    dtrain <- xgb.DMatrix(data =
data.matrix(train_tmp[,which(names(train_tmp) !=
"Target")]),
                          label = train_tmp$Target)
    dval <- xgb.DMatrix(data =
data.matrix(val_tmp[,which(names(val_tmp) != "Target")]),
                        label = val_tmp$Target)
    watchlist <- list(train=dtrain, test=dval)


end <- Sys.time()
print(end - start)

saveRDS(grid_simple, "xgb_vl_simple.rds")

# Let's try this for random forest now ----------

train_data <- vl %>%
  ungroup() %>%
  filter(patient_group == 1) %>%
  select(-patient_group) %>%
  arrange(PredictionDate) %>%
  select(-PredictionDate, - key)

test_data <- vl %>%
  ungroup() %>%
  filter(patient_group == 2) %>%
  select(-patient_group) %>%
  arrange(PredictionDate) %>%
  select(-PredictionDate, - key)

# Create folds
nfolds <- 10
cuts <- seq(min(train_data$PredictionDate),
max(train_data$PredictionDate), length.out = nfolds
+ 2)

grid_rf_simple <- expand.grid(model = "rf",
                              sparsity = "simple",
                              mtry = c(6,8, 10),
                              nodesize = c(20, 10, 5))

start <- Sys.time()

for(i in 1:nrow(grid_rf_simple)){
  print(i)
  aucpr <- c()

  for(j in 1:nfolds){
    print(j)
    # Set train and validation
    tmp <-  train_data %>%
      group_by(key) %>%
      mutate(patient_group = sample(1:2, 1, prob =
c(0.7, 0.3)))

    train_tmp <- tmp %>%
      ungroup() %>%
      filter(patient_group == 1) %>%
      filter(between(PredictionDate, cuts[1],
cuts[j+1])) %>%
      select(-PredictionDate, - key)

    val_tmp <- tmp %>%
      ungroup() %>%
      filter(patient_group == 2) %>%
      filter(between(PredictionDate, cuts[j+1],
cuts[j+2])) %>%
      select(-PredictionDate, - key)
```

```
   set.seed(2231)
   xgb <- xgboost::xgb.train(data = dtrain,
                        eta = grid_sparse[i, 3],
                        max_depth = grid_sparse[i, 4],
                        colsample_bytree = grid_sparse[i, 5],
                        nround = 1000,
                        early_stopping_rounds = 25,
                        objective = "binary:logistic",
                        #metric = 'aucpr',
                        eval.metric = "aucpr",
                        watchlist = watchlist,
                        verbose = 0
   )

   aucpr <- c(aucpr, xgb$best_score)
   num_iter <- c(num_iter, xgb$best_iteration)

 }

 grid_sparse$val_pr_auc[i] <- mean(aucpr)
 grid_sparse$num_iter[i] <- mean(num_iter)

}

end <- Sys.time()
print(end - start)# 19min

saveRDS(grid_sparse, "xgb_vl_sparse.rds")

train_data_final <- train_data %>%
 select(-PredictionDate) %>%
 select(- key) #%>%
 #encodeXGBoost()

dtrain <- xgb.DMatrix(data =
data.matrix(train_data_final[,which(names(train_data_final)
!= "Target")]),
                 label = train_data_final$Target)
set.seed(2231)
xgb <- xgb.train(data = dtrain,
            eta = 0.01,
            max_depth = 6,
            colsample_bytree = .3,
            nround = 165,
            objective = "binary:logistic",
            verbose = 1
)

val_predict <- predict(xgb,newdata = data.matrix(test_data[,
-which(names(test_data) == "Target")]))
fg <- val_predict[test_data$Target == 1]
bg <- val_predict[test_data$Target == 0]
prc <- pr.curve(scores.class0 = fg, scores.class1 = bg, curve
= T)
plot(prc)
roc <- roc.curve(scores.class0 = fg, scores.class1 = bg,
curve = T)
plot(roc)

# Now, let's do it with simple imputation ----------

train_data <- vl %>%
 ungroup() %>%
   train_tmp <- train_tmp %>%
select(intersect(names(train_tmp),
names(val_tmp)))
   val_tmp <- val_tmp %>%
select(intersect(names(train_tmp),
names(val_tmp)))

   # # Set train and validation
   # train_tmp <- train_data[cuts[1]:cuts[j+1], ]
   # val_tmp <-
train_data[(cuts[j+1]+1):(cuts[j+1]+cutgap), ]

   # Impute
   train_tmp <- data.frame(train_tmp)
   val_tmp <- data.frame(val_tmp)

   for(k in which(sapply(train_tmp, class) %in%
c('character', 'factor'))){
     train_tmp[, k] <- replaceWithMode(train_tmp,
train_tmp, k)
     val_tmp[, k] <- replaceWithMode(train_tmp,
val_tmp, k)
   }

   for(k in which(sapply(train_tmp, class) %in%
c('numeric', 'integer'))){
     train_tmp[, k] <- replaceWithMean(train_tmp,
train_tmp, k)
     val_tmp[, k] <- replaceWithMean(train_tmp,
val_tmp, k)
   }

   rf <- randomForest(
    as.factor(Target) ~ .,
    data = train_tmp,
    mtry = grid_rf_simple[i, 3],
    nodesize = grid_rf_simple[i, 4]
   )

   pred_val = predict(rf, newdata=val_tmp, type =
"prob")
   fg <- pred_val[val_tmp$Target == 1, 2]
   bg <- pred_val[val_tmp$Target == 0, 2]
   prc <- pr.curve(scores.class0 = fg, scores.class1
= bg, curve = T)

   aucpr <- c(aucpr, prc$auc.integral)
 }

 grid_rf_simple$val_pr_auc[i] <- mean(aucpr)

}

saveRDS(grid_rf_simple, "rf_vl_simple.rds")
```

```
  filter(patient_group == 1) %>%
  select(-patient_group) %>%
  filter(PredictionDate >= "2021-01-01") %>%
  filter(PredictionDate <= "2022-05-31")

test_data <- vl %>%
  ungroup() %>%
  filter(patient_group == 2) %>%
```

Appendix B

Ethics

Strathmore UNIVERSITY

28th August 2024

Ms Otieno Benedette,
benedette.adhiambo@strathmore.edu

Dear Ms Otieno,

RE: Predictive Modelling to Identify Patients at High Risk of Virological Failure in Kenya

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2310/24.** The approval period is from **28th August 2024 to 27th August 2025.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

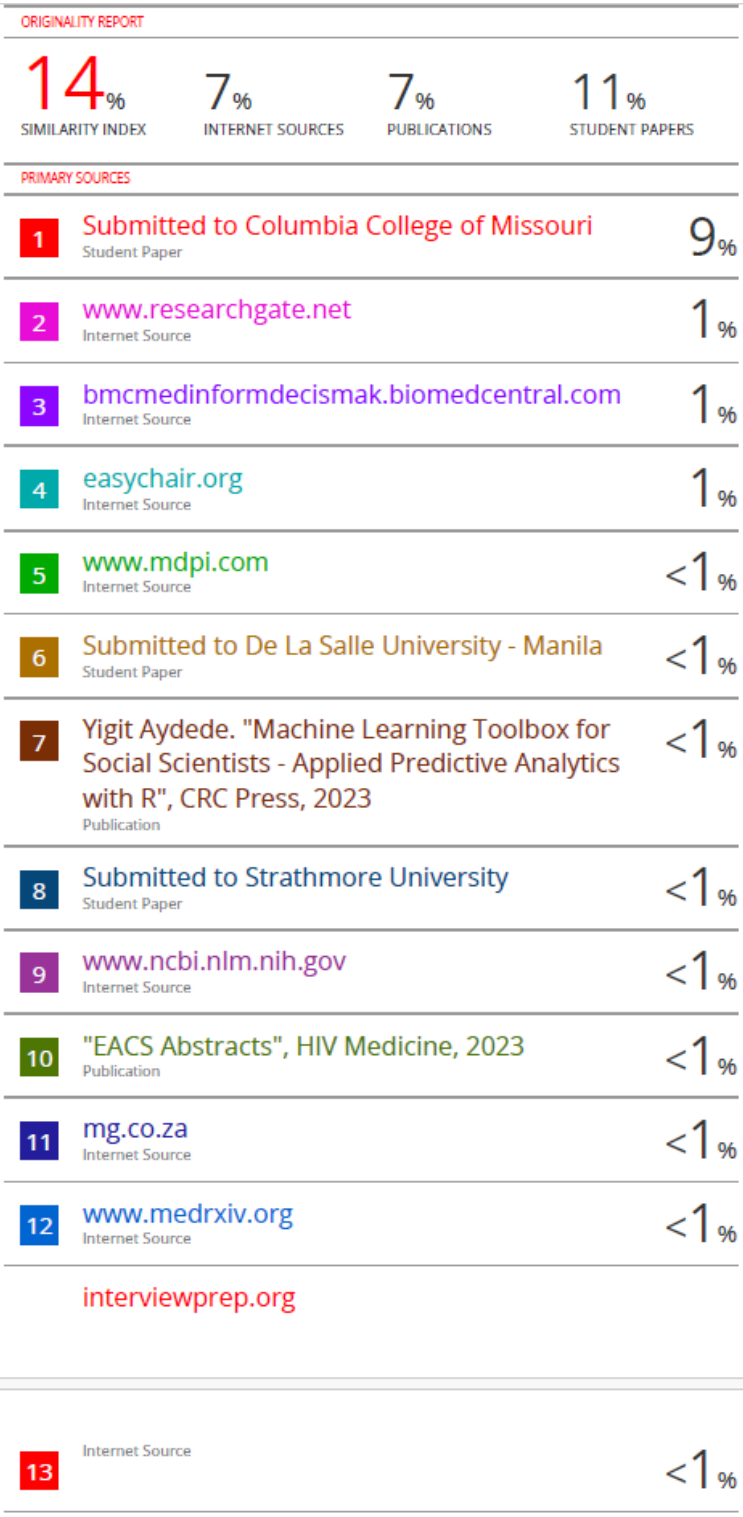
Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

Appendix C

Similarity Index



ORIGINALITY REPORT

14%	7%	7%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Columbia College of Missouri Student Paper	9%
2	www.researchgate.net Internet Source	1%
3	bmcmedinformdecismak.biomedcentral.com Internet Source	1%
4	easychair.org Internet Source	1%
5	www.mdpi.com Internet Source	<1%
6	Submitted to De La Salle University - Manila Student Paper	<1%
7	Yigit Aydede. "Machine Learning Toolbox for Social Scientists - Applied Predictive Analytics with R", CRC Press, 2023 Publication	<1%
8	Submitted to Strathmore University Student Paper	<1%
9	www.ncbi.nlm.nih.gov Internet Source	<1%
10	"EACS Abstracts", HIV Medicine, 2023 Publication	<1%
11	mg.co.za Internet Source	<1%
12	www.medrxiv.org Internet Source	<1%
	interviewprep.org	
13	Internet Source	<1%