Predicting Mother-To-Child HIV Transmission Among Mother-Baby Pairs in Kenya: A Focused Comparison of Random Forest and XGBoost Models

Odhiambo, Ashley Achieng

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Statistical Science of Strathmore University

Strathmore Institute of Mathematical Sciences
Strathmore University
Nairobi, Kenya

June, 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 dissertation contains no material previously published or written by another person except where due reference is made in the dissertation itself.

Name: **Odhiambo, Ashley Achieng**

Signature: .....

Date: April 1, 2025

Approval

The dissertation of Odhiambo, Ashley Achieng was reviewed and approved by the following:

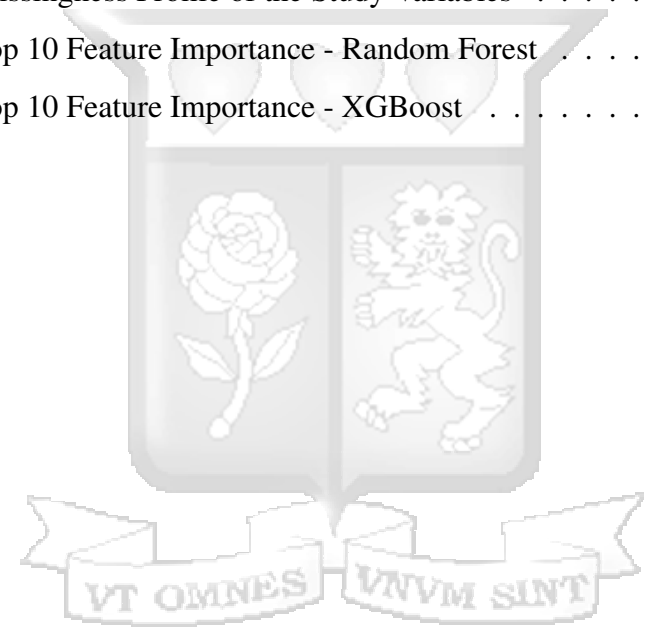
Prof. Bernard Omolo

Supervisor, Oguna-Omolo

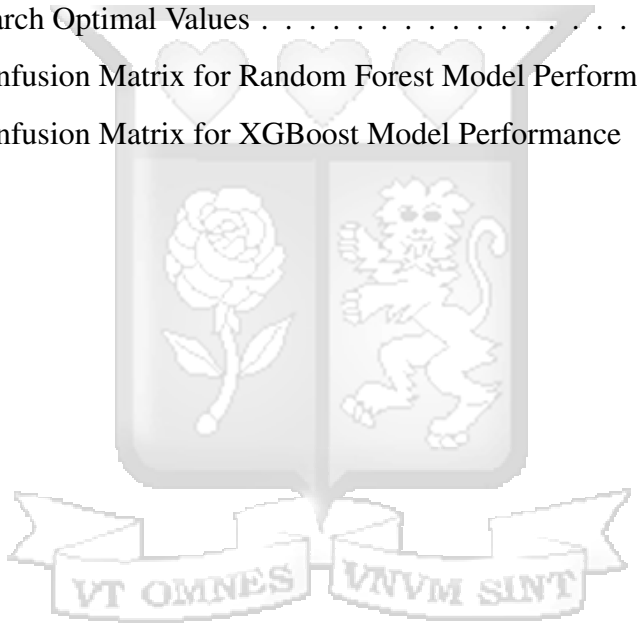
Institute of Mathematical Sciences, Strathmore University.

Table of contents

List of figures



List of tables



List of abbreviations

AHD	Advanced HIV Disease
ART	Antiretroviral Therapy
EDA	Exploratory Data Analysis
EMR	Electronic Medical Records
FN	False Negative
FP	False Positive
HIV	Human Immunodeficiency Virus
KNN	K-Nearest Neighbors
MAR	Missing Completely at Random
MDI	Mean Decrease Impurity
mHealth	Mobile Health
MICE	Multivariate Imputation by Chained Equations
MNAR	Missing Not at Random
MTCT	Mother-to-Child Transmission
NDW	National Data Warehouse
PMTCT	Prevention of Mother-to-Child Transmission
PWLHIV	Pregnant Women Living with HIV
SMOTE	Synthetic Minority Over-Sampling Technique
STI	Sexually Transmitted Infections
TN	True Negative
TP	True Positive
UNAIDS	Joint United Nations Programme on HIV/AIDS
XGBoost	eXtreme Gradient Boosting

Acknowledgement

I express my sincere appreciation to The Almighty for His guidance, strength, and grace throughout this journey.

My deepest gratitude goes to my supervisor Prof. Bernard Omolo for his invaluable support, technical contribution and mentorship while writing this dissertation.

I also appreciate Strathmore Institute of Mathematical Sciences (SIMS), my lecturers, and colleagues for their insights and encouragement.

Furthermore, I wish to thank my parents and sibling for their everyday support while working on this dissertation.

I sincerely appreciate the Palladium Group, KeHMIS Project for granting data access that made this research possible. Their support and commitment to knowledge sharing have been invaluable in shaping the findings of this study.

In addition, I acknowledge my partner Dr. Joseph Alulu and friends Benedette Otieno, Olivia Anyango, and Sally Woka for their value addition and continuous motivation to my work.

Lastly, I extend my gratitude to all who contributed in any way to the successful completion of this work.

Dedication

This dissertation is dedicated to my parents, Ms. Betty M'maitsi and Mr. Amos Kosanya, my brother Tyron Otieno, and my cousin Michael Kagollah who have been of great inspiration throughout my entire academic journey.

Chapter 1

Introduction

1.1 Background to the Study

HIV remains one of the most pressing global public health concerns, with the latest estimates indicating that approximately 39.9 million people are living with the virus worldwide (WHO, 2024). The epidemic's burden disproportionately affects specific populations, notably women and adolescent girls, who constitute a significant proportion of new infections. In 2023, sub-Saharan Africa (SSA) was particularly impacted, with 62% of all new HIV infections occurring among girls and women (Joint United Nations Programme, 2024). This underscores the complex social, economic, and healthcare disparities that continue to fuel the epidemic.

Among those most affected by HIV are pregnant women living with HIV (PWLHIV), who face unique health challenges (USAID, 2024) risking not only their own well-being but also that of their children. Each year, approximately 1.3 million girls and women living with HIV become pregnant, highlighting the urgent need for interventions that address the risks associated with pregnancy and HIV co-management (WHO, 2024). Mother-to-child transmission (MTCT) remains the primary route for HIV transmission to infants, accounting for most paediatric HIV cases worldwide (Gill et al., 2020) and can occur at various stages - during pregnancy, labour and delivery, or breastfeeding WHO (2024). Despite significant global efforts and strides made in preventing mother-to-child transmission (PMTCT) through antiretroviral therapy (ART), breastfeeding alternatives, and other interventions, cases of HIV transmission from mother to child continue to arise (White AB et al., 2024). According to UNAIDS, new HIV infections among children have significantly reduced over the years, yet an estimated 120,000 cases still occurred in 2023 (Joint United Nations Programme, 2024).

In Kenya, HIV places a significant burden on public health, with approximately 1.4 million people living with the virus, a substantial number of whom are women of reproductive age (Joint United Nations Programme, 2024). Kenya has made significant progress in reducing MTCT of HIV through the implementation of PMTCT programs. Among these initiatives is the Beyond Zero campaign, launched in 2014 by the Former First Lady of Kenya, as a national effort to enhance maternal and child health (UNAIDS, 2014). Nonetheless, challenges persist.

In recent years, researchers have focused on leveraging machine learning to enhance clinical decision-making in forecasting HIV transmission from different perspectives (Fieggen et al., 2022; Kagendi and Mwau, 2023; Li et al., 2022). Predictive models offer considerable potential in optimizing resource allocation, identifying care gaps, and informing targeted interventions (Maskew et al., 2022). For example, the study by Chaula and Justo (2022) highlighted the utility of the Random Forest algorithm in identifying MTCT risk by effectively capturing patterns in clinical and demographic data. Building on this, gradient boosting methods, such as eXtreme Gradient Boosting (XGBoost), may offer enhanced predictive power by iteratively refining predictions to capture more nuanced relationships within the data (Kharkar, 2023). Despite these advancements, the application of machine learning models in predicting MTCT risk among PWLHIV in Kenya remains underexplored.

This study aimed to explore the application of machine learning models in predicting MTCT risk, with a specific focus on identifying the most effective predictive model(s) to support Kenya's PMTCT initiatives. Building on the work of Chaula and Justo (2022), this research compares the best-performing model from their study, Random Forest, with eXtreme Gradient Boosting (XGBoost). By focusing on machine learning, the study sought to develop a more accurate and individualized approach to MTCT risk prediction in Kenya, ultimately aiming to reduce paediatric HIV incidence and strengthen the nation's HIV prevention strategies.

1.2 Problem Statement

MTCT of HIV remains a significant public health challenge in Kenya, despite ongoing efforts to prevent it (Tuthill et al., 2024). A major obstacle is the current reliance on broad epidemiological and demographic data to identify high-risk pregnancies, which does not fully account for the complex interactions between maternal health, clinical factors, and socioeconomic conditions that influence the transmission likelihood. While clinical guidelines provide a general framework for assessing MTCT risk (Ministry of Health, Kenya, 2012), they are not tailored to individual cases, limiting the precision of interventions. Accurately identifying women at high risk of transmitting HIV to their infants is crucial for improving PMTCT efforts as targeted interventions based on individualized risk predictions could significantly reduce MTCT rates.

Machine learning models, such as Random Forest and XGBoost, have shown promise in healthcare by analyzing complex datasets and identifying patterns that may not be immediately apparent through traditional methods. These models can capture non-linear relationships and adapt to changing data patterns, potentially offering more precise predictions of MTCT risk (Kharkar, 2023; Maskew et al., 2022). However, the use of machine learning in MTCT prediction in Kenya remains underexplored. In the study by Chaula and Justo (2022), Random Forest emerged as the best-performing model for predicting MTCT risk. This research aimed to build upon their work by exploring whether XGBoost, with its ability to iteratively refine predictions and capture more complex, non-linear relationships, could provide an even more accurate prediction of MTCT risk. By comparing these two models, this research endeavored to offer a more precise, individualized approach to MTCT risk prediction, ultimately improving the targeting of interventions and contributing to reducing paediatric HIV transmission in Kenya.

1.3 Research Objectives

1.3.1 Main Objective

The main objective of this study was to compare the predictive performance of the Random Forest and XGBoost models in predicting MTCT risk among PWLHIV in Kenya.

1.3.2 Specific Objectives

1. To optimize the hyperparameters of Random Forest and XGBoost models to improve the prediction accuracy of MTCT risk.
2. To evaluate and compare the predictive performance of Random Forest and XGBoost in assessing MTCT risk.
3. To identify key factors influencing MTCT risk by analyzing feature importance in the context of Random Forest and XGBoost models.

1.4 Justification of the Study

Existing methods for assessing MTCT risk in Kenya are predominantly based on broad demographic and epidemiological profiles, which fail to account for the individual-level factors that significantly influence transmission. This gap limits the precision of interventions, leaving many high-risk pregnancies unidentified and under served.

Emerging machine learning models have demonstrated the potential to enhance risk prediction by analyzing complex datasets and uncovering nuanced, nonlinear relationships. Unlike traditional methods, these models can adapt to diverse data patterns (Ali et al., 2024), enabling more accurate identification of women at high risk for MTCT.

Exploring their application in Kenya could provide effective predictive tools within PMTCT programs for identifying at-risk pregnancies. By addressing this critical gap, the study

will be supporting Kenya's efforts to achieve the UNAIDS 95-95-95 targets, Sustainable Development Goal 3's aim to end the AIDS epidemic, and the National AIDS Strategic Framework objectives (National Syndemic Diseases Control Council, 2021; UNAIDS, 2021; United Nations, 2023). The findings will contribute to reducing paediatric HIV incidence, improving maternal and child health, and advancing Kenya's commitment to a future free from new paediatric HIV infections.

1.5 Significance of the Study

This study has the potential to transform PMTCT strategies in Kenya by identifying more precise, individualized predictors of MTCT risk, thereby improving targeted intervention accuracy. Results from this study will contribute to the growing body of evidence on the role of predictive analytics in global health, aiding policymakers, healthcare practitioners, and researchers in enhancing paediatric HIV prevention efforts.

1.6 Expected Output

1. Optimized hyperparameters for both Random Forest and XGBoost models, leading to improved prediction accuracy in forecasting MTCT risk.
2. A comparative analysis of the predictive performance of Random Forest and XGBoost models, including performance metrics such as precision, recall, and F1 Score, to determine the most effective model for MTCT risk prediction.
3. A detailed list of key factors identified as most influential in predicting MTCT risk, derived from the feature importance analysis of the Random Forest and XGBoost models.

1.7 Dissemination and Utilization of Findings

The findings of this study will be presented as part of a master's dissertation at Strathmore University. Subsequently, the results may be disseminated through the university's online repository and library catalogue, academic blogs, conference presentations, and journal publications. The researcher hopes that the results will inform policy decision-making, support the development of targeted intervention programs, and further contribute to similar research in the use of machine learning models for predicting HIV transmission and related health outcomes.

Chapter 2

Literature Review

2.1 Overview of Mother-to-Child HIV Transmission

MTCT of HIV is the primary transmission route for paediatric HIV (Gill et al., 2020), with most paediatric cases occurring in sub-Saharan Africa, where HIV prevalence is highest (Joint United Nations Programme, 2024). There are three main periods during which HIV can be transmitted from mother to child (National Institute of Health, 2024):

1. During pregnancy (intrauterine transmission) – The virus can cross the placenta during pregnancy, though less common than transmission during labour and delivery. This can occur if the maternal viral load is high or if the mother has untreated HIV infection. The risk of transmission increases if she has a recent HIV infection or has advanced HIV disease (AHD).

2. During labour and delivery (intrapartum transmission) – Majority of MTCT cases occur during childbirth where the baby is exposed to the mother's blood and vaginal fluids, which may contain the virus. Factors such as prolonged labour, the use of invasive procedures, and maternal viral load at the time of delivery significantly influence the risk of transmission. The risk is higher if the mother is not on Antiretroviral Therapy (ART) at the time of delivery.

3. During breastfeeding (postpartum transmission) – HIV can also be transmitted to the infant through breastmilk, even if the mother is not symptomatic. The risk of transmission is highest when the mother has a high viral load, particularly during the early months of breastfeeding.

Without any intervention, an estimated 15–30% of PWLHIV are likely to transmit the virus to their babies during pregnancy and childbirth. Breastfeeding further raises the transmission probability by 10–15%, with the level of risk influenced by clinical factors as well as the pattern and length of breastfeeding (Joint United Nations Programme, 2024). Global efforts have reduced this risk to less than 5% when ART is used effectively (Mugwaneza et al., 2018), yet barriers remain particularly in low-resource settings, highlighting the need for effective PMTCT interventions.

2.2 The Role of ART in Preventing MTCT

ART remains a cornerstone of PMTCT, as it helps to maintain low viral loads in HIV-positive pregnant and breastfeeding women, significantly reducing the probability of transmission (Amin et al., 2021). It comprises a combination of drugs that target different stages of the HIV lifecycle and helps prevent disease progression (Pan American Health Organization, 2020). The "Option B+" approach, which involves "lifelong ART for all HIV-positive pregnant women", has been shown to be particularly effective. Research conducted in Malawi and Uganda demonstrated significant reductions in transmission rates following the implementation of Option B+ policies, marking a milestone in PMTCT programs (CDC, 2013; Koss et al., 2017).

However, studies indicate persistent challenges with ART adherence, impacting PMTCT outcomes. For example, a study in Rwanda showed that mothers who did not receive ART during pregnancy experienced higher MTCT rates (Mugwaneza et al., 2018). Similarly, South African research revealed that mothers with inconsistent ART adherence have higher rates of transmission, even within structured PMTCT programs (Mutabazi et al., 2020). These results highlight the importance of focused efforts to enhance ART adherence.

2.3 Socioeconomic and Cultural Barriers to Effective PMTCT

The effectiveness of PMTCT programs can be significantly hampered by socioeconomic and cultural barriers. These not only reduce the accessibility and quality of healthcare services for pregnant women but also perpetuate inequalities that heighten the risk of HIV transmission to their infants.

2.3.1 Access to Healthcare Services

According to Rural Health Information Hub, many resource-limited settings, particularly in rural areas, suffer from a lack of healthcare infrastructure, which limits pregnant women's ability to get tested for HIV and receive appropriate ART. Furthermore, Women in rural areas may be required to travel significant distances to reach healthcare facilities. A study in Nigeria highlights this issue, revealing inequalities in maternal healthcare services between rural and urban women (Okoli et al., 2020). These challenges can delay diagnosis, treatment initiation, and follow-up care, ultimately increasing the likelihood of vertical transmission.

2.3.2 Financial Barriers

Even when ART is provided for free by governments or through international programs, associated costs such as transportation, caregiving responsibilities, and missed working days can be prohibitive for many women. For those who are financially vulnerable, such costs can deter them from seeking care or continuing treatment U.S. CDC and Ministry of Health, Kenya (2013). Additionally, financial constraints also extend to the management of HIV beyond medication. Nutritional requirements for PWLHIV, the need for regular health check-ups, and postnatal care all incur costs that may be unsustainable for economically disadvantaged families. This financial vulnerability can discourage women from adhering to ART regimens, increasing the risk of MTCT and contributing to poorer health outcomes for both mother and child.

2.3.3 Lack of Social Support

Social support plays a critical role in the successful implementation of PMTCT programs. However, many women face significant gaps in support from their families or communities. For example, male partners are often uninvolved in antenatal care and PMTCT efforts (Ngangue et al., 2021). This lack of involvement can result in women shouldering the burden of managing HIV-related care alone, which can be emotionally overwhelming and logistically challenging. Without adequate support, women are less likely to disclose their HIV status or adhere to ART regimens. The absence of a supportive network can also deter women from attending regular clinic visits, leading to lapses in care that increase the risk of MTCT. Community-driven programs that foster awareness and encourage partner involvement in PMTCT are critical for overcoming these barriers.

2.3.4 Gender Inequality

Gender inequality remains a pervasive barrier to PMTCT, particularly in patriarchal societies where women have limited autonomy over healthcare decisions. In some cases, women need their partner's permission to access health services, which can delay or completely obstruct their ability to seek PMTCT interventions. Also, gendered power dynamics affect women's ability to negotiate safer sexual practices, such as condom use, or insist on testing and treatment for themselves and their partners (Moses et al., 2009). These dynamics contribute to the continued spread of HIV within households and communities, undermining PMTCT goals. Addressing gender inequality through education, community empowerment, and legal reforms is essential for improving maternal and child health outcomes.

2.3.5 Stigma and Discrimination

Many women fear being ostracized by their families or communities if their HIV status becomes known. This fear often discourages them from getting tested, enrolling in PMTCT programs, and adhering to ART regimens. Healthcare settings are not immune to stigma.

Reports of discriminatory treatment by healthcare providers can deter women from seeking care, in turn undermining their trust in the healthcare system (Logie et al., 2011). Reducing stigma through community education and enforcing anti-discrimination policies is critical to ensuring that women feel safe accessing PMTCT services.

2.3.6 Breastfeeding Practices

Breastfeeding is a culturally significant practice in many societies, but it poses a risk for HIV transmission from mother to child if not managed appropriately Joint United Nations Programme (2024). EBF for the first six months, combined with maternal ART, is recommended to reduce the risk of MTCT WHO (2010). However, cultural norms that favour mixed feeding can undermine these recommendations. Mixed feeding increases the likelihood of intestinal inflammation in infants, which provides a pathway for HIV transmission. Efforts to promote EBF in these settings have faced resistance, necessitating a culturally sensitive approach to PMTCT.

2.4 Clinical and Behavioural Predictors of MTCT

A range of clinical and behavioural factors impact MTCT rates, making personalized and evidence-based approaches to PMTCT essential. Understanding these predictors is critical for the overall success of PMTCT programs.

2.4.1 Maternal Viral Load

High maternal viral load levels are strongly correlated with increased risk MTCT of HIV. Achieving viral suppression through ART reduces this risk substantially. Evidence suggests that when maternal viral load is undetectable (<50 copies/mL), the likelihood of MTCT approaches zero. In the absence of ART, viral loads exceeding 10,000 copies/mL are associated with significantly higher transmission rates. Furthermore, studies indicate that

transient increases in viral load, or "blips," during pregnancy can also elevate MTCT risk, emphasizing the need for continuous viral suppression throughout the perinatal period (Amin et al., 2021; Myer et al., 2017).

2.4.2 Adherence to ART

Suboptimal adherence to ART undermines the efficacy of treatment, leading to detectable viral loads and an increased risk of MTCT. Studies show that adherence levels above 80-90% are necessary to maintain viral suppression (O'Halloran Leach et al., 2021), emphasizing the need for robust adherence support mechanisms during pregnancy and breastfeeding.

2.4.3 Mode of Delivery

The mode of delivery can influence the likelihood of MTCT, particularly in settings where maternal viral load is not adequately suppressed. Elective caesarean delivery has been shown to reduce intrapartum transmission of HIV compared to vaginal delivery in cases where viral suppression is not achieved. The protective effect of caesarean delivery is attributed to reduced exposure of the infant to maternal blood and genital secretions during labour. However, in women with suppressed viral loads, caesarean delivery does not offer additional benefits over vaginal delivery and may expose women to unnecessary surgical risks (Amin et al., 2021). Consequently, the decision on mode of delivery should be individualized, based on maternal viral load and overall health status.

2.4.4 Infant Prophylaxis

Antiretroviral regimens, such as daily nevirapine or zidovudine, administered to infants born to HIV-positive mothers significantly reduce, but not eliminate, the risk of acquiring the virus. Studies have demonstrated that extended infant prophylaxis during the breastfeeding period can reduce the risk of HIV transmission (Amin et al., 2021). WHO recommends that infant prophylaxis continues for at least 6–12 weeks postpartum, depending on breastfeeding

status and maternal viral load, to provide a protective buffer against transmission WHO (2010). Consistent adherence to these prophylactic regimens, coupled with early postnatal HIV testing, ensures optimal outcomes for HIV-exposed infants.

2.4.5 Maternal Health Status and Co-Infections

AHD is associated with increased transmission rates. Women with AHD often face additional challenges, such as malnutrition and opportunistic infections, which can compromise ART efficacy and adherence. Co-infections, particularly sexually transmitted infections (STIs), exacerbate the risk of MTCT. STIs like syphilis, herpes simplex virus, and chlamydia cause genital inflammation and increase maternal viral load in genital secretions. Additionally, conditions such as malaria and tuberculosis during pregnancy can impair maternal immunity and elevate the risk of HIV transmission to the child (Amin et al., 2021). Screening and treating co-infections during antenatal care are therefore critical for reducing MTCT risk.

2.5 PMTCT Programs in Kenya

Kenya has made remarkable progress in its PMTCT of HIV initiatives by adopting innovative and nationally scaled interventions. A flagship initiative is the Beyond Zero Campaign, which has effectively integrated PMTCT services within broader maternal and child healthcare frameworks. This initiative has been instrumental in improving access to ART for pregnant and breastfeeding women. By embedding PMTCT within comprehensive maternal health-care services, the campaign ensures HIV-positive mothers receive holistic care, including nutritional support, regular clinical monitoring, and adherence counselling for lifelong ART, in line with global standards (UNAIDS, 2014). These efforts ensure that mothers and their infants receive seamless and equitable care within a decentralized healthcare system.

According to NASCOP (2014), Kenya's PMTCT strategy is closely aligned with the World Health Organization's "Option B+" guidelines, which recommend lifelong ART for all HIV-

positive pregnant and breastfeeding women. This approach has been critical in reducing maternal viral loads and preventing vertical transmission.

Community-driven initiatives have also played a transformative role in supporting PMTCT outcomes. Community health workers and peer support groups have been pivotal in bridging service delivery gaps, especially in resource-limited settings. These programs provide education, psychosocial support, and ART adherence reinforcement to HIV-positive mothers. Evidence suggests that peer-driven models are particularly effective in reducing stigma and addressing socio-behavioural barriers that hinder the success of PMTCT interventions (Haldane et al., 2019).

Despite these achievements, significant challenges persist, particularly in rural and remote areas where healthcare infrastructure and resources remain inadequate. In Kenya, as of 2020, PMTCT coverage was at 94% nationally. However the national MTCT rate increased alarmingly between 2015 and 2020 from an 8.3% to 10.8% (Elizabeth Glaser Pediatric AIDS Foundation, 2021), far above UNICEF (2024) global elimination target of less than 5% for breastfeeding populations.

Such findings underscore the need for individualized targeting strategies that go beyond one-size-fits-all approaches. Tailored interventions could address specific hurdles faced by HIV-positive mothers in rural settings. For example, personalized adherence counselling, home-based treatment delivery, the use of digital health tools like SMS reminders, and even culturally tailored interventions that consider local beliefs and practices could be integrated into existing PMTCT programs. These strategies, combined with robust monitoring systems, can help point out mothers at higher risk of non-adherence and provide them with focused support.

By addressing unique socio-economic and geographic challenges through individualized targeting, Kenya can strengthen ART adherence and retention in care, ultimately reducing the risk of MTCT and advancing progress toward the elimination of vertical HIV transmission.

2.6 Predictive Modelling in PMTCT

The use of predictive modelling in public health has significantly advanced efforts to combat diseases such as HIV. In the context of PMTCT, predictive models have emerged as essential tools to identify at-risk populations. Both traditional statistical models and machine learning approaches offer unique advantages in this domain, yet the latters utility in PMTCT remains insufficiently explored.

Traditional statistical models have long been the cornerstone of risk prediction in healthcare, particularly valued for their interpretability and ability to provide clear, quantifiable relationships between predictors and outcomes. However, these models have limitations in capturing the non-linear, complex interactions within the data, prompting researchers to explore machine learning methods. Machine learning approaches excel in scenarios involving large, heterogeneous datasets and can effectively capture these complex patterns, making them ideal for forecasting such as predicting MTCT risk.

2.7 Machine Learning Approaches

2.7.1 Random Forest

Random Forest is an ensemble learning technique that aggregates several decision trees to enhance prediction accuracy and minimize overfitting. It works by randomly selecting subsets of the data and features to train individual decision trees, which are then aggregated to make the final prediction (Schott, 2019). The mathematical formulation of Random Forest is as follows:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(x) \tag{2.1}$$

where the final prediction is given by $\hat{y}$, T represents the total number of trees, and $f_t(x)$ is the prediction from tree t.

Research has consistently demonstrated the efficacy of Random Forest as a powerful predictive tool in forecasting MTCT of HIV. For instance, Chaula and Justo (2022) utilized Random Forest to identify high-risk MTCT pregnancies. Their findings revealed an impressive accuracy of 99%, rendering it the best-performing model among the five assessed in their study.

2.7.2 eXtreme Gradient Boosting

eXtreme Gradient Boosting (XGBoost) is a gradient boosting model that optimizes the performance of decision trees by focusing on misclassified instances in each iteration. It improves upon Random Forest by progressively modifying weights to reduce loss. The objective function to be optimized for the k^{th} boosted tree can be expressed as:

$$\hat{\text{Obj}}^{(k)} = \sum_{i=1}^{N} L(y_i, \hat{y}_i^{(k)}) + \sum_{k=1}^{K} \Omega(f_k) \tag{2.2}$$

where $L(y_i, \hat{y}_i^{(k)})$ is the loss function and $\Omega(f_k)$ represents the regularization term to prevent overfitting (Kharkar, 2023).

A study by Li et al. (2024) utilized XGBoost to forecast the length of stay and risk of long-term hospitalization among individuals living with HIV through a retrospective analysis. The model was effective in identifying significant predictors, with systemic multiple opportunistic infections being the most important variable. However, the K-Nearest Neighbor (KNN) outperformed it to achieve an R^2 value of 0.68 compared to a 0.42 by the XGBoost.

2.8 Research Gap

While both machine learning models have demonstrated considerable success in various predictive healthcare applications, direct comparisons between the two in the context of PMTCT are notably scarce. Existing literature highlights Random Forest as a leading model,

with studies such as Chaula and Justo (2022) showcasing its impressive predictive capabilities in identifying high-risk pregnancies for MTCT. However, despite the success of Random Forest, XGBoost, a more recent and sophisticated gradient boosting technique, remains underexplored in MTCT prediction. XGBoost has proven superior in many domains due to its ability to iteratively correct errors, thereby capturing complex relationships in data more effectively than Random Forest in some cases (Fatima et al., 2023). Given the rapidly evolving nature of HIV risk factors, it is crucial to investigate whether XGBoost could offer enhanced performance over Random Forest in predicting MTCT risk. This study aimed to fill this critical gap by directly comparing the two models in the context of MTCT prediction, providing valuable insights that could improve the precision of risk assessment and, consequently, the effectiveness of PMTCT interventions. By doing so, this study seeks to offer evidence-based recommendations that could transform how PMTCT strategies are designed and implemented, ultimately improving health outcomes for both mothers and children in high-risk settings like Kenya.

The Kenya National Data Warehouse (NDW) offers a unique opportunity to explore these methodologies. With its rich, longitudinal data on maternal and child health, the NDW enables a robust assessment of the comparative effectiveness of Random Forest and XGBoost in predicting MTCT of HIV. This comparative analysis has the potential to identify the most suitable methods for enhancing PMTCT strategies thereby generating actionable insights to optimize MTCT risk prediction and improve the effectiveness of PMTCT programs in Kenya.

Chapter 3

Methodology

3.1 Study Design and Population

This study used a retrospective cohort design to evaluate the predictive capabilities of the Random Forest and XGBoost models in forecasting mother-to-child HIV transmission among mother-baby pairs in Kenya. The population of interest comprised PWLHIV alongside their infants born within the study period, January 2019 to December 2022. Each mother was observed for 24 months postpartum, enabling the assessment of the impact of various clinical, demographic, and behavioural factors.

3.1.1 Inclusion Criteria

The cohort's inclusion criteria encompassed mothers who had given birth within the specified study period and whose infants had a recorded 24-month outcome. In this case, infant follow-up included a HIV test result at 24 months, confirming the absence or presence of HIV.

Maternal HIV diagnosis was not limited to the pregnancy period; mothers diagnosed prior to, during, and post-delivery were also included, ensuring a comprehensive analysis of the risk factors associated with MTCT of HIV.

3.2 Data Source

Data for this study was sourced from the NDW, which acts as a comprehensive repository of health data consolidated from health facilities utilizing Electronic Medical Records (EMRs) and implementing HIV programs countrywide. By leveraging this data source, the study aimed to capture a broad spectrum of mother-baby pairs and associated health outcomes within Kenya.

3.3 Sample Size Calculation

An initial sample size calculation was conducted based on a population proportion approach, targeting approximately 460 mother-baby pairs, accounting for a 10% estimated MTCT rate, a 95% confidence level, and a 3% margin of error. However, due to the nature of machine learning models, which benefit from larger datasets to improve prediction accuracy, the study utilized all available data from NDW, within the mentioned study period. This decision ensured that the machine learning models had access to as much information as possible, was essential for understanding the intricate connections between different factors affecting MTCT risk. This allowed the models to maximize their predictive power and generalizability, which is essential for producing accurate and individualized MTCT risk predictions.

3.4 Data Pre-processing

3.4.1 Multicollinearity Analysis

To assess and address multicollinearity within the study variables, Cramer's V was employed due to the categorical nature of the data, where values closer to 1 indicated a stronger association. Pairs of variables with a Cramer's V value above 0.80 were considered to have high multicollinearity and were flagged for removal. This step ensured that highly correlated

features did not negatively impact the stability of the models, preventing inflated variance in the coefficient estimates and improving model generalizability.

3.4.2 Data Splitting

Data splitting is a crucial step in model validation, where the dataset is divided into two distinct sets: training and testing. In this study, the dataset was split using a 70-30 ratio, with seventy percent of the data being assigned to the training set, while the remaining thirty percent was used for testing. While the process is random within predefined groups, it is important to note that this split was not entirely random due to the use of stratified partitioning. This ensured that the distribution of the outcome variable was preserved in both the training and testing sets - avoiding bias and ensuring that the model is equally sensitive to both outcomes. This method is particularly important in cases where the outcome variable is imbalanced, such as the MTCT rates, which tend to be skewed towards the negative class (HIV-negative infants) (Gen David L, 2023).

3.4.3 Addressing Class Imbalance

Given that this study aimed to predict MTCT, the class distribution in the dataset was highly imbalanced, with a greater number of negative cases compared to positive cases. To mitigate the effects of this imbalance, upsampling was performed on the training dataset. This technique involved randomly duplicating instances from the minority class (Positive) to balance the class distribution, ensuring that the model did not become biased toward the majority class.

The upsampling process was carried out only on the training dataset to prevent information leakage from the testing set. This balanced dataset was then used to train both the Random Forest and XGBoost models, allowing for a more reliable comparison of their performance on balanced data.

The training set was then used to develop and train the machine learning models. This set allowed the models to learn the patterns and relationships within the data that would support accurate predictions. The testing set was reserved for evaluating the performance of the trained models. This separation ensured that the models were assessed on unseen data, providing a more accurate measure of their ability to generalize.

3.4.4 Handling Missingness

Given the inherent challenges of missing data in health-related datasets, a comprehensive assessment of missingness was conducted to identify the underlying mechanisms and determine the extent of missingness in the data. Two distinct types of missingness were identified: Missing at Random (MAR) and Missing Not at Random (MNAR).

For the cases exhibiting MNAR, where missing values were found to occur in a manner that was related MTCT, a threshold of more than 50% missingness was set for exclusion, removing all variables exhibiting MNAR. This threshold was chosen to ensure that any records with substantial missing data would not disproportionately impact the integrity of the analysis, preventing bias and potential skewing in the results.

For the remaining data, which exhibited MAR (where missingness was related to observed data but not dependent on unobserved data), Multiple Imputation by Chained Equations (MICE) was employed. MICE is a widely recognized and robust method for handling missing data, as it creates multiple plausible imputations based on observed patterns and relationships in the data (Gupta, 2023). By imputing missing values using MICE, the analysis is able to retain valuable information from all available observations while accounting for the uncertainty associated with missing values, ensuring the validity and reliability of subsequent analyses. This approach was chosen to strike a balance between minimizing data loss and mitigating bias, while maintaining the scientific rigor and generalizability of the findings. Importantly, MICE imputation was performed separately for training and testing datasets to prevent data leakage from the test set into the model development process.

3.4.5 Data Transformation

Data transformation plays a critical role in preparing raw data for analysis by converting them into a suitable format for analysis (Olamendy, 2024). In this study, one-hot encoding was applied to transform categorical variables into a numerical format that could be used effectively by machine learning algorithms, specifically XGBoost. XGBoost, being a gradient boosting algorithm, requires numeric data to build decision trees and cannot process categorical variables directly. Categorical variables, such as infant HIV status and treatment adherence, often represented distinct, unordered categories. If these variables were used in their raw form, XGBoost might have misinterpreted them by assuming an ordinal relationship where none existed. For example, it could have incorrectly treated "HIV positive" as numerically greater than "HIV negative", introducing bias into the model. By applying one-hot encoding, each category within a variable was transformed into its own binary feature, where the presence or absence of a category was indicated by a 1 or 0 respectively. This ensured that the model treated each category independently without imposing any artificial ordering, which was crucial for maintaining the integrity of the data and improving prediction accuracy.

3.5 Predictive Modelling Framework

In this study, two machine learning approaches were employed to predict MTCT: Random Forest and XGBoost. These algorithms were chosen for their proven effectiveness in handling complex datasets with numerous features and their ability to provide robust predictive models for classification tasks.

Random Forest is an ensemble learning technique that merges several decision trees, each trained on different random subsets of both data and features. This approach is especially useful for minimizing overfitting since it combines the predictions of multiple trees, each trained on a distinct segment of the data.

XGBoost is a gradient boosting model renowned for its scalability and effeciency, particularly when working with extensive datasets. Like Random Forest, XGBoost also builds

decision trees, but in a sequential manner where each tree attempts to correct the errors of its predecessors. This makes it a powerful tool for achieving high predictive accuracy.

3.6 Model Development

3.6.1 Hyperparameter Tuning and Cross Validation

Hyperparameter tuning is a critical step in the development of predictive models, as it directly influences their performance and generalizability. Unlike parameters learned from the data during model training, hyperparameters are set prior to training and dictate the model's structure and training behaviour. For this study, hyperparameter tuning was conducted using grid search and stratified 10-fold cross-validation for both Random Forest and XGBoost as shown in Figure 3.1. Grid search tested all predefined combinations of hyperparameter values, ensuring a comprehensive search for the optimal configuration.

For Random Forest, key hyperparameters such as the the number of features considered for splitting at each node (mtry), the minimum number of samples in a terminal node (nodesize), and the number of trees in the forest (ntrees) were tuned. These parameters were optimized to ensure that the model maintained a balance between underfitting and overfitting. Similarly, for XGBoost, the learning rate (eta), maximum depth of trees (max_depth), regularization parameter (gamma), fraction samples for the trees (colsample_bytree), and fraction training samples for each boosting round (subsample) were optimized using the same approach.

After the hyperparameter optimization process was completed for both models, the Random Forest and XGBoost models were compared based on their predictive performance in the context of predicting MTCT risk. The models' ability to distinguish between cases of HIV-positive and HIV-negative infants was crucial for the evaluation, as the study aimed to identify the most accurate and reliable model for forecasting the risk of MTCT.

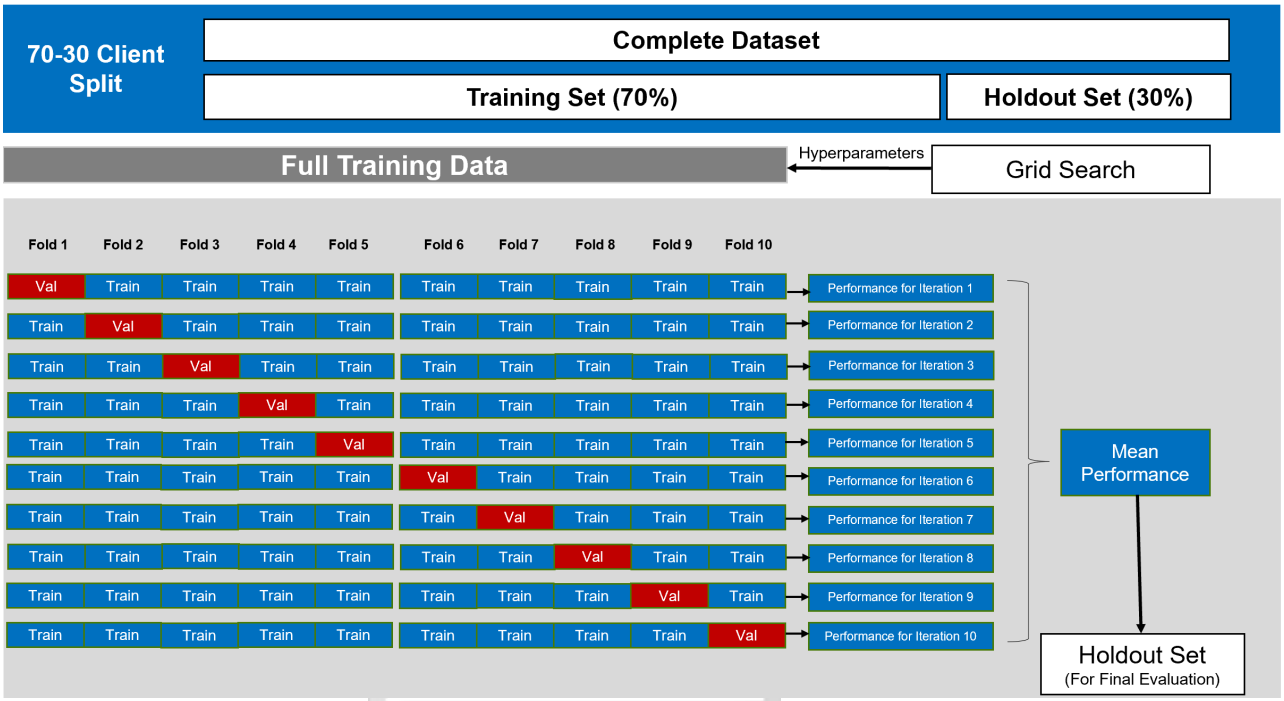


Figure 3.1: Model Training and Testing Using 10-Fold Cross Validation

3.6.2 Feature Selection

Feature selection was performed to identify the most influential predictors contributing to the model's predictive power, thereby improving both the model's performance and its interpretability. The goal of this step was to retain only the most relevant features, enhancing the model's ability to generalize to unseen data while ensuring that the analysis focused on actionable and meaningful variables.

To evaluate feature importance, Gini Importance was used. Gini Importance, also known as the Mean Decrease Impurity (MDI), measures each feature's contribution to reducing node impurity across all trees in the ensemble. This metric is particularly useful in decision tree-based models, as it quantifies the total decrease in node impurity (i.e., the Gini index) that results from splitting on a particular feature. The higher the Gini importance score of a feature, the more critical it was in improving the model's predictive accuracy. This method works by aggregating the reductions in Gini impurity for each feature across all decision

trees, providing an overall measure of feature importance (Terence Shin and Matthew Urwin, 2024).

The features with the highest Gini Importance scores were retained for further model optimization. These features were deemed the most influential in predicting the risk of MTCT. By reducing the dimensionality of the feature space, this step helped mitigate the risk of overfitting while maintaining the predictive integrity of the model. The final set of features selected were subsequently used in the final model training, ensuring that the model remained focused on the most significant predictors of MTCT risk.

3.7 Model Evaluation

Evaluating the performance of predictive models is essential to ensure they meet the desired accuracy and reliability for practical applications in PMTCT. Various metrics were employed to assess the performance of Random Forest and XGBoost, providing valuable insights into each model's strengths and areas for improvement (Brownlee, 2020).

For these classification tasks, metrics such as precision, recall, and F1-score were used to evaluate model performance. The AUC-ROC assessed the models' ability to discriminate between positive and negative classes, with a higher AUC indicating better performance in distinguishing high-risk from low-risk cases. However, this could be misleading when it comes to imbalanced datasets as in this case. As such, precision and recall come in to address this by evaluating the accuracy of positive predictions and the model's ability to capture all actual positive cases, respectively. Subsequently, the F1-score which combines precision and recall, will be useful when there is a trade-off between the two.

3.8 Statistical Analysis

Statistical analysis was conducted using R software, version 4.4.1, employing packages designed for machine learning. Exploratory data analysis precededed modeling to under-

stand data distributions, identify relationships between predictors and outcomes, and detect potential outliers. This step was helpful in informing the modeling process by ensuring that the data were suitable for analysis.

3.9 Ethical Consideration

Prior to commencing the research, ethical approval was obtained from the relevant institutional and national review boards to ensure compliance with all ethical standards. Data sourced from NDW was provided in a de-identified format to preserve individual privacy and confidentiality - adhering to legal and institutional regulations governing the handling of sensitive health information, particularly HIV-related data. Additionally, all findings herein are reported in aggregate form to prevent any risk of identifying individuals or specific health facilities.

Chapter 4

Results and Interpretation

4.1 Introduction

The purpose of this chapter is to present the results of the predictive performance of the Random Forest and XGBoost models in predicting MTCT. The models were evaluated based on their ability to classify HIV transmission status (Positive or Negative) effectively. In particular, effectiveness was assessed using precision, recall, and F1-score, which collectively provide a comprehensive understanding of their predictive capabilities.

The analysis is framed around the research questions established earlier, focusing on how each model performs in an imbalanced dataset context and how well they generalize to unseen data. To ensure optimal performance, both models underwent hyperparameter tuning aimed to identify the best-performing set of hyperparameters for each model to ensure that they were able to effectively provide reliable predictions. Additionally, the importance of individual features in predicting MTCT outcomes was explored, offering insights into which factors contributed most significantly to the model's decision-making process. This chapter thus aims to provide a clear and detailed assessment of both models, with implications for their potential use in real-world applications for HIV transmission risk prediction.

4.2 Data Pre-Processing Results

The assessment of multicollinearity using Cramer's V identified one pair of categorical variables with high correlation. Specifically, time of diagnosis and time of ART initiation, exhibited strong associations (Cramér's V = 0.89), as shown in 4.1, and were subsequently

removed to maintain model stability. This reduction in multicollinearity ensured that the predictive models would not suffer from inflated variance in coefficient estimates, thereby improving model reliability.

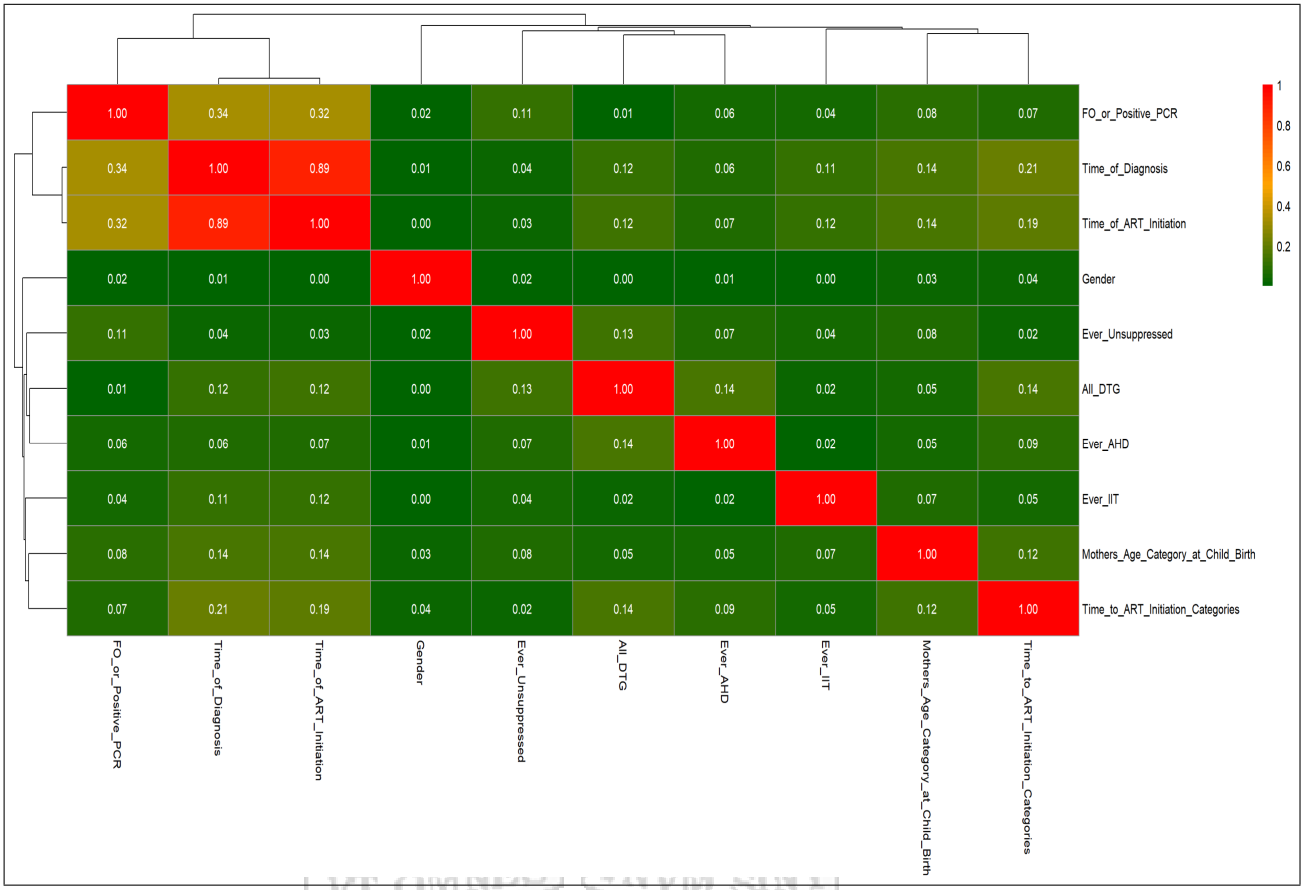


Figure 4.1: Cramer's V HeatMap

The dataset was then partitioned into training and testing sets using a 70:30 split, resulting in 8,923 instances in the training set and 3,823 instances in the testing set. Stratified partitioning was applied to maintain the distribution of the outcome variable, ensuring that the proportion of HIV-positive infants remained consistent across both sets at approximately 4%. This approach prevented bias and preserved the representativeness of the dataset.

A significant class imbalance was observed, with HIV-negative cases representing approximately 96% of the total cases, while HIV-positive cases accounted for only 4%. To address this, random upsampling of the minority class was performed in the training set, resulting in an equal representation of both classes. Post-upsampling, the training set contained 8,923 HIV-positive and 8,923 HIV-negative instances. Both models were trained on this balanced data, reducing the likelihood of bias toward the majority class.

A detailed missing data analysis, as seen on 4.2, revealed that 87% of variables exhibited missing values, with two primary mechanisms identified: MAR and MNAR. Variables exceeding 50% missingness and categorized as MNAR, such as exclusively breasfeeding (91%) , mode of delivery (79%), sub-optimal adherenace(61%), and lack of infant prophylaxis (56%), were removed to mitigate bias. The remaining missing values, primarily MAR, were imputed using MICE, which preserved the overall data structure. Post-imputation, distributions of key variables remained consistent with the original dataset, ensuring that no artificial patterns were introduced.

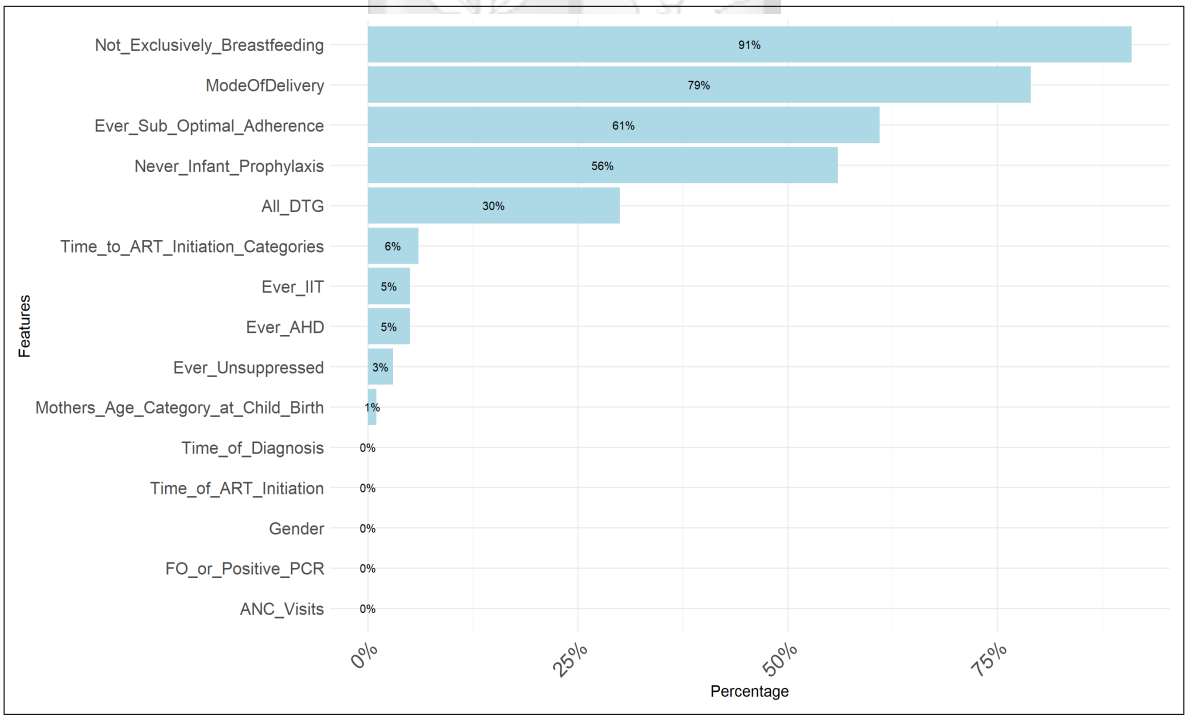


Figure 4.2: Missingness Profile of the Study Variables

Categorical variables were transformed using one-hot encoding to ensure compatibility with the XGBoost model, which requires numerical data for decision tree construction. This transformation expanded the dataset from 10 features to 31 features. Inasmuch as the encoding increased data dimensionality, it ensured that the data could be effectively used by XGBoost without introducing biases related to the interpretation of categorical variables.

Overall, these pre-processing steps optimized the dataset by mitigating multicollinearity, addressing class imbalance, handling missingness effectively, and transforming categorical variables into a suitable format. This ensured that the dataset was robust and ready for subsequent modeling and analysis.

4.3 Model Performance and Evaluation

4.3.1 Hyperparameter Tuning

Random Forest Hyperparameter Tuning

For the Random Forest model, several key hyperparameters were tuned to enhance its predictive accuracy as shown in Table 4.1 below:

Table 4.1: Hyperparameters for Random Forest Model: Description and Grid Search Optimal Values

Hyperparameter	Description	Grid Search Optimal Value
ntree	Number of trees to ensure stable predictions and reduce variance without excessive computational cost.	800
mtry	A value that provides a good balance between underfitting and overfitting, considering a moderate subset of features at each split.	12
nodesize	A reasonable threshold to ensure that nodes are not split too early, maintaining the model's generalizability while avoiding overfitting.	14
max_nodes)	The maximum number of terminal nodes allowed in the trees. Controlling this helps prevent overfitting by limiting the complexity of the trees.	34

XGBoost Hyperparameter Tuning

For the XGBoost model, a different set of hyperparameters were optimized, given its gradient boosting nature as shown in Table 4.2 below:

Table 4.2: Hyperparameters for XGBoost Forest Model: Description and Grid Search Optimal Values

Hyperparameter	Description	Grid Search Optimal Value
eta	This controls the step size at each iteration while moving toward a minimum. A lower learning rate improves the model's ability to converge but requires more trees.	0.1
max_depth	The maximum depth of the decision trees. Increasing this allows the model to capture more complex patterns but risks overfitting.	7
gamma	A regularization parameter that specifies the minimum loss reduction required to make a further partition on a leaf node.	0
colsample_bytree	Defines the fraction of features to sample for each tree.	0.8
subsample	This controls the fraction of training samples to use for each boosting round.	0.6

4.3.2 Model Performance Before and After Hyperparameter Tuning

The performance of both the Random Forest and XGBoost models improved significantly following hyperparameter tuning. Before optimization, the Random Forest model achieved an F1 score of 77%, while XGBoost performed slightly better with an initial F1 score of 90%. However, both models showed limitations in terms of recall and precision, particularly due to class imbalance in the dataset.

Hyperparameter tuning played a crucial role in refining the models' ability to balance precision and recall, a critical factor in MTCT prediction, where missing a positive case can have serious health consequences. By systematically adjusting the key parameters in each model, both models demonstrated significant improvements in classification performance.

After tuning, the Random Forest model's F1 score increased to 86%, with improvements in precision (from 84% to 98%) and recall (from 71% to 76%), indicating better identification of HIV-positive infants. The XGBoost model outperformed RF, achieving a final F1 score of 92%, with precision improving from 91% to 97% and recall increasing from 85% to 87%. These improvements ensured that the models could better capture complex patterns in the dataset, leading to more reliable predictions.

4.3.3 Classification Metrics

To evaluate the predictive performance of Random Forest and XGBoost in identifying MTCT risk, multiple classification metrics were analyzed. These included accuracy, precision, recall, and F1-score, each providing unique insights into the effectiveness of the models.

Overall Accuracy

Accuracy is the most commonly used metric in classification tasks, representing the proportion of correct predictions over the total number of cases. However, in imbalanced datasets, high accuracy does not necessarily equate to strong performance, as a model may achieve a high accuracy by simply predicting the majority class more frequently.

Random Forest and XGBoost both achieved an impressive accuracy of 99%, correctly classifying the vast majority of cases in the dataset. However, while it is a useful metric, it does not fully capture the models' effectiveness in detecting HIV-positive cases, particularly in an imbalanced dataset. To gain a more comprehensive understanding of their performance, precision, recall, and the F1-score were examined, providing deeper insight into each model's strengths and limitations in identifying true positive cases while minimizing false positives and false negatives.

Precision: Confidence in Positive Predictions

Precision refers to the proportion of correctly identified HIV-positive cases out of all the cases the model predicted as positive. A high precision value means that when the model predicts a positive case, it is likely to be correct, which is crucial in minimizing false positives — misclassifying HIV-negative infants as HIV-positive.

Random Forest exhibited a precision of 98%, meaning that when it predicted an infant as HIV-positive, it was correct 98% of the time. XGBoost had a slightly lower precision of 97%, still demonstrating strong reliability in positive case predictions.

These values indicate that both models are highly precise and effective in ensuring that healthy infants are not unnecessarily subjected to additional medical interventions due to false-positive classifications.

Recall: Sensitivity in Identifying HIV-Positive Cases

Recall, also known as sensitivity, measures the model's ability to identify all actual HIV-positive cases. A higher recall means the model captures more true positive cases but may come at the expense of a higher false-positive rate. In the context of MTCT prediction, high recall is essential because missing an HIV-positive case (false negative) could have severe consequences, leading to an untreated infant who remains at risk of disease progression.

Random Forest demonstrated a recall of 76%, meaning that it correctly identified 76% of all actual HIV-positive infants but missed 24% of them. XGBoost on the other hand achieved a recall of 87%, significantly outperforming Random Forest in identifying a greater proportion of true positives.

This result highlights a key advantage of XGBoost: it is more effective at recognizing HIV-positive cases, ensuring that fewer at-risk infants go undetected. The 11% difference in recall between XGBoost and Random Forest indicates that XGBoost is a superior choice in scenarios where missing a positive case is highly consequential.

F1-Score: Balancing Recall and Precision

The F1-score is the harmonic average of precision and recall, offering a unified measure that balances both factors. Precision aims to minimize false positives, while recall prioritizes identifying all true positives. In medical contexts such as MTCT prediction, maintaining a balance between these two is essential.

Random Forest achieved an F1-score of 86%, indicating a good balance between recall and precision. However, XGBoost outperformed with an F1-score of 92%, demonstrating superior overall effectiveness in both minimizing false negatives and maintaining precision.

This higher F1-score suggests that XGBoost provides a better trade-off between recall and precision, making it the more reliable model for identifying HIV-positive infants while keeping false positives low.

4.3.4 Confusion Matrix Analysis

The confusion matrices for both models were analyzed to provide a more granular understanding of how each model classified the data. The confusion matrix details the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for each model, which are critical for calculating metrics like sensitivity and specificity.

As shown in Table 4.3, the Random Forest model correctly identified 110 true positive cases but misclassified 35 HIV-positive infants as negative (false negatives). It also demonstrated a high ability to correctly classify negative cases, with 3,676 true negatives and only 2 false positives. This resulted in a high precision score (98%) but a relatively lower recall (76%), indicating that while the model was highly selective in predicting positive cases, it failed to identify a significant number of actual HIV-positive infants.

Table 4.3: Confusion Matrix for Random Forest Model Performance

Actual\Predicted	**Positive**	**Negative**
Positive	110	35
Negative	2	3,676

Conversely, as detailed in Table 4.4, the XGBoost model outperformed Random Forest in recall, correctly identifying 126 true positive cases while misclassifying only 19 as negative. While XGBoost exhibited a slightly higher false positive count (3), it demonstrated superior recall (87%) and a higher F1-score (92%), making it more effective at identifying at-risk infants.

Table 4.4: Confusion Matrix for XGBoost Model Performance

Actual\Predicted	**Positive**	**Negative**
Positive	126	19
Negative	3	3,675

Overall, these results emphasize the trade-offs between precision and recall. While Random Forest minimizes false positives, it does so at the cost of a higher false negative rate, potentially missing crucial cases. XGBoost, on the other hand, prioritizes recall, making it a more suitable choice in clinical applications where identifying every possible HIV-positive case is crucial for timely intervention and treatment.

4.3.5 Feature Importance Analysis

Feature importance analysis was performed to determine which factors had the most significant impact on predicting MTCT in both models. For Random Forest, feature importance was assessed using the Gini index, which measures the impurity of each feature's splits. The five main drivers identified by Random Forest were timing of maternal ART initiation, maternal viral load, number of ANC visits attended, any interruptions in treatment within 24 months, and AHD. These features were crucial for the model's decision-making process, with mother's time of ART initiation standing out as the most influential (Figure 4.3).

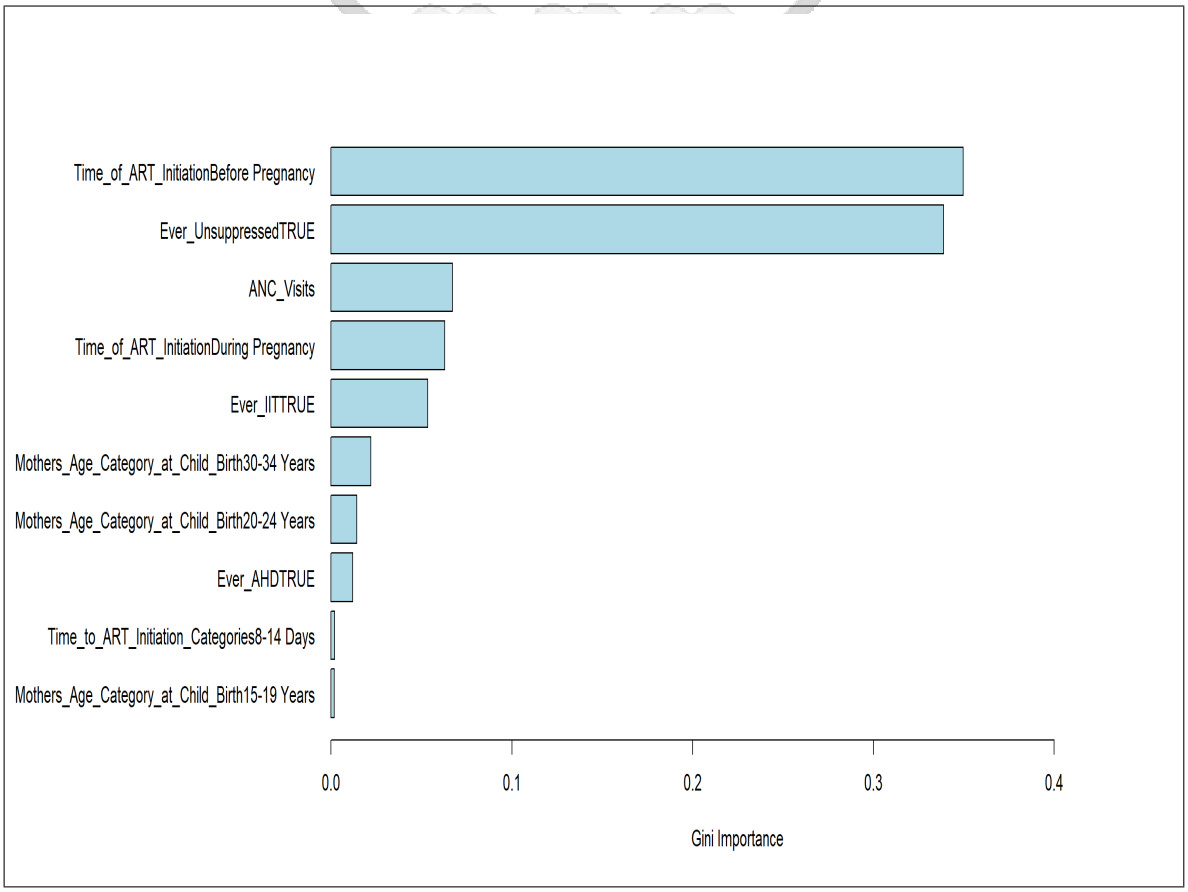


Figure 4.3: Top 10 Feature Importance - Random Forest

XGBoost's feature importance, derived from the model's boosting process, showed a similar trend, though there were slight differences in the ranking of some features. The five most influential features identified by XGBoost were the timing of maternal ART initiation, maternal viral load, the number of ANC visits attended, any interruptions in treatment within 24 months, and the mother's age at childbirth. Notably, XGBoost assigned greater importance to the number of ANC visits attended while placing slightly less emphasis on maternal viral load compared to Random Forest. This suggests that XGBoost may consider regular ANC attendance as a stronger predictor of MTCT risk, potentially reflecting the role of comprehensive prenatal care in preventing transmission (Figure 4.4).

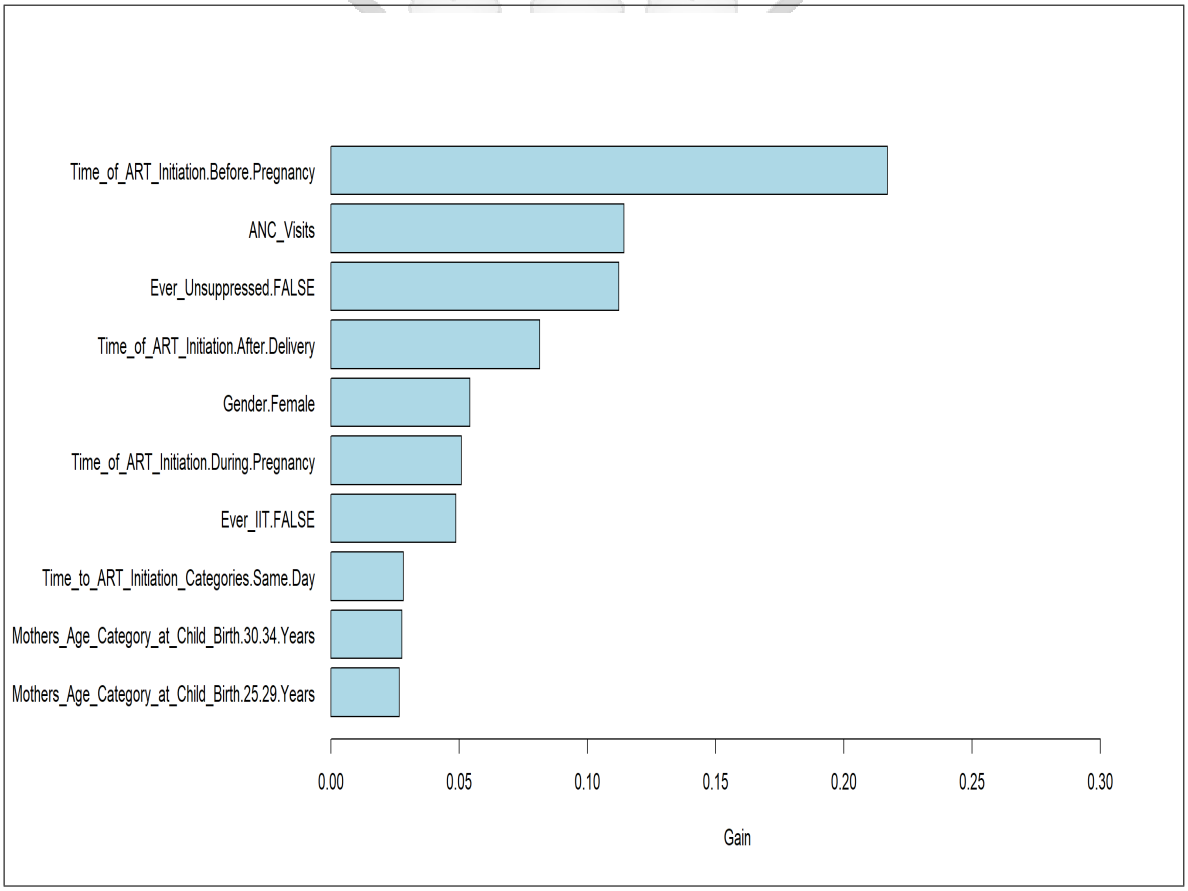


Figure 4.4: Top 10 Feature Importance - XGBoost

Despite these differences, both models largely identified similar key features, underscoring the importance of the timing of maternal ART initiation maternal viral load, and number of ANC visits attended in preventing MTCT. The consistency in the most important features across both models further strengthens the validity of these predictors.

Chapter 5

Discussion

5.1 Introduction

This section discusses the implications of these findings, particularly in relation to the importance of recall in clinical decision-making, the role of key predictive features such as maternal viral load, ART adherence, and ANC attendance, and the broader impact of machine learning in public health interventions. Additionally, the strengths and limitations of the study are considered, along with recommendations for integrating predictive modeling into existing PMTCT strategies.

5.2 Evalauating Model Performance

5.2.1 The Superiority of XGBoost in MTCT Prediction

The results of this study demonstrated that XGBoost consistently outperformed Random Forest across multiple key performance metrics, particularly in recall (87%) and F1-score (92%). These findings align with previous research indicating that gradient boosting methods, such as XGBoost, frequently surpass traditional ensemble-based bagging models in classification tasks involving complex, high-dimensional datasets (Chen and Guestrin, 2016). Unlike Random Forest, which builds trees independently and aggregates results, XGBoost optimizes its learning process through sequential boosting, where each new tree corrects errors made by previous iterations. This iterative approach likely contributed to its superior performance in predicting MTCT of HIV.

A key advantage of XGBoost was its higher recall rate (87%), which directly reflects its ability to identify true positive cases more effectively than Random Forest (76%). In the context of MTCT, recall is a crucial metric because failing to detect HIV-positive infants (false negatives) could result in missed opportunities for early medical intervention. Given that early administration of ART significantly reduces disease progression and improves survival rates among HIV-positive infants (Violari et al., 2008), a model with higher recall ensures that fewer HIV-positive cases go undiagnosed, thus improving clinical outcomes.

The findings of this study suggest that incorporating XGBoost into MTCT risk prediction frameworks could enhance the accuracy of early warning systems, thereby supporting data-driven decision-making in PMTCT programs.

5.2.2 The Role of Precision in Clinical Decision-Making

While XGBoost demonstrated superior recall, it is important to recognize the role of precision in clinical decision-making. Random Forest exhibited a slightly higher precision (98%) compared to XGBoost (97%), indicating that it was more conservative in classifying cases as HIV-positive. This conservatism is beneficial in minimizing false positives, where HIV-negative infants are incorrectly classified as HIV-positive. Reducing false positives is critical because it prevents unnecessary psychological distress for caregivers and avoids unwarranted medical interventions, such as additional ART initiation or prolonged clinical monitoring.

However, in the specific context of MTCT prevention, recall is typically prioritized over precision. This is because the consequences of missing an HIV-positive case (false negative) are significantly more severe than misclassifying a negative case as positive. Missing a HIV-positive infant delays life-saving ART initiation, increasing the risk of rapid disease progression, immune deterioration, and mortality. In contrast, while false positives may result in temporary emotional or medical burdens, they can be resolved through confirmatory testing and follow-up diagnostics.

This prioritization of recall over precision aligns with medical screening principles observed in other high-risk disease domains. For instance, cancer screening models often emphasize

recall because failing to detect a malignant tumor poses a greater health risk than incorrectly flagging a benign case(Rohit Kundu, 2022). Similarly, in infectious disease surveillance, a higher recall ensures that no potential cases are overlooked, thereby allowing for early containment and treatment interventions.

Given these considerations, XGBoost remains the more suitable model for MTCT prediction, as its superior recall minimizes the risk of undiagnosed HIV-positive infants, ensuring timely medical intervention. However, precision must still be accounted for, and incorporating additional post-prediction validation steps—such as confirmatory viral load testing or clinical follow-up protocols — could help mitigate concerns associated with false positives. By leveraging XGBoost's strengths in recall while addressing precision concerns through clinical verification, healthcare systems can enhance early HIV detection, optimize PMTCT strategies, and ultimately contribute to the global goal of eliminating pediatric HIV.

5.2.3 The Balanced Performance of F1-Score

The F1-score, which balances precision and recall, further supports the robustness of XGBoost in this predictive task. A higher F1-score suggests that XGBoost was better at managing the trade-offs between false positives and false negatives, ensuring a more reliable classification of cases. In medical predictive modeling, a high F1-score is desirable when dealing with imbalanced datasets, as it prevents model bias toward the majority class (HIV-negative cases in this case). The findings confirm previous research that supports the application of F1-score as a strong indicator of a model's real-world utility (Saito and Rehmsmeier, 2015).

5.3 Feature Importance and Clinical Implications

5.3.1 The Predictive Role of Maternal ART Initiation Timing

The timing of maternal ART initiation emerged as one of the most influential predictors of MTCT in both the Random Forest and XGBoost models. This aligns with established

research demonstrating that early ART initiation — preferably before conception or during the first trimester — dramatically reduces the risk of vertical transmission (Clinical Info, 2024b). Women who initiate ART late in pregnancy or during labor have significantly higher transmission risks due to insufficient viral suppression (Chagomerana et al., 2018).

This finding reinforces the need for universal early HIV testing and immediate ART initiation for all HIV-positive pregnant women, as recommended by the World Health Organization (WHO) Option B+ strategy (Darby et al., 2021). In resource-limited settings such as Kenya, where late presentation to ANC is common, machine learning models can play a transformative role in identifying high-risk pregnancies based on ART initiation timelines. Healthcare providers can then prioritize late presenters for more aggressive intervention strategies, such as enhanced adherence counseling and expedited viral load monitoring.

Machine learning algorithms could be integrated into EMRs to flag patients who initiate ART late, prompting immediate clinical follow-up. Such data-driven interventions could significantly reduce the burden of delayed treatment initiation, ensuring that more pregnant women achieve optimal viral suppression before delivery.

5.3.2 Maternal Viral Load as a Primary Risk Factor for MTCT

Both models confirmed that maternal viral load remains the most dominant predictor of MTCT, reinforcing decades of research linking high maternal viremia to increased perinatal HIV transmission (Thea et al., 1997). When a pregnant woman's viral load exceeds 1,000 copies/mL, the risk of transmission rises substantially, particularly during vaginal delivery and breastfeeding . In contrast, maintaining an undetectable viral load (<50 copies/mL) through consistent ART use reduces transmission risk (Chagomerana et al., 2018; Clinical Info, 2024a).

The strong predictive power of maternal viral load in both models underscores the critical need for routine viral load monitoring throughout pregnancy. Frequent testing ensures that viral rebound is detected early, allowing clinicians to modify treatment regimens accordingly.

Machine learning-powered risk stratification can enhance this process by flagging viral load fluctuations that indicate potential adherence issues or drug resistance.

5.3.3 The Impact of Antenatal Care (ANC) Attendance on MTCT Prevention

The number of ANC visits attended was identified as a significant predictor of MTCT risk, with XGBoost placing greater emphasis on this factor than Random Forest. Frequent ANC attendance is a well-established proxy for improved maternal health outcomes (Ndege et al., 2016), as it provides healthcare providers with opportunities to monitor HIV disease progression, optimize ART regimens, and reinforce adherence counseling.

The predictive strength of ANC attendance in the models suggests that public health efforts should prioritize increasing ANC uptake among HIV-positive mothers. Despite national guidelines recommending at least four ANC visits during pregnancy, many women in Kenya do not meet this threshold, often due to socioeconomic barriers, stigma, and lack of health facility access (Logie et al., 2011; Okoli et al., 2020). AI-based risk prediction models could identify pregnant women with low ANC attendance early on, prompting community health workers to conduct targeted outreach and home-based ANC services.

Additionally, digital health interventions, such as mobile phone reminders and telemedicine consultations, could play a pivotal role in increasing ANC attendance. Countries like Kenya, where mobile health (mHealth) initiatives have successfully improved ART adherence, could expand these efforts to enhance ANC engagement for HIV-positive expectant mothers.

5.3.4 The Consequences of Treatment Interruptions on MTCT Risk

One of the most concerning findings from the feature importance analysis was the strong predictive role of interruptions in ART treatment within 24 months before childbirth. Temporary discontinuation of ART — even for a short period — can result in rapid viral rebound, sig-

nificantly increasing the risk of intrauterine, intrapartum, and postpartum HIV transmission (Clinical Info, 2024a).

Research has shown that intermittent ART adherence can lead to drug resistance, reducing the effectiveness of first-line regimens and making viral suppression more difficult to achieve (SeyedAlinaghi et al., 2023). This highlights the urgent need to identify women experiencing ART interruptions and provide timely interventions to prevent virologic failure.

The incorporation of machine learning algorithms into ART adherence monitoring systems could allow for early identification of high-risk patients based on prescription refill patterns, viral load trends, and self-reported adherence data. Healthcare providers could then offer real-time support, such as:

1. Home-based ART delivery for mothers facing mobility challenges.
2. Psychosocial support programs to address stigma-related ART discontinuation.
3. Targeted adherence counseling and peer-led interventions to reinforce long-term commitment to ART.
4. Governments and public health institutions should also explore financial incentives and social support programs for HIV-positive mothers, as economic hardship remains a major driver of ART discontinuation in low-resource settings (Bassett et al., 2015).

5.4 Limitations and Future Directions

5.4.1 Data Imbalance and Potential Biases

A limitation of this study is the imbalance in the dataset, where there is a greater proportion of HIV-negative cases than HIV-positive ones. Although random upsampling was employed to address this imbalance, it is worth noting that upsampling may not always fully mitigate bias in predictions. Despite the upsampling, other techniques, such as synthetic minority over-sampling technique (SMOTE), could potentially enhance model generalizability further

(Chawla et al., 2002). Future research may explore additional methods to improve model robustness and reduce any residual bias stemming from class imbalance.

Additionally, the models were trained on Kenyan data, which may limit their applicability to other geographic regions with different socio-economic and healthcare infrastructures. External validation using datasets from diverse populations may benefit model robustness assessment.

5.4.2 Clinical Integration

The incorporation of machine learning models into clinical workflows presents several practical challenges, particularly in resource-limited settings. Many healthcare facilities in Kenya and similar regions lack the digital infrastructure needed to support real-time predictive analytics. Future efforts should prioritize the development of lightweight, mobile-compatible models that are adaptable to these environments, ensuring that they can be deployed effectively in low-resource healthcare settings.

References

Ali, A., Jayaraman, R., Azar, E., and Maalouf, M. (2024). A comparative analysis of machine learning and statistical methods for evaluating building performance: A systematic review and future benchmarking framework. *Building and Environment*, page 111268.

Amin, O., Powers, J., Bricker, K. M., and Chahroudi, A. (2021). Understanding viral and immune interplay during vertical transmission of hiv: implications for cure. *Frontiers in Immunology*, 12:757400.

Bassett, I. V., Wilson, D., Taaffe, J., and Freedberg, K. A. (2015). Financial incentives to improve progression through the hiv treatment cascade. *Current Opinion in HIV and AIDS*, 10(6):451–463.

Brownlee, J. (2020). How to calculate precision, recall, and f-measure for imbalanced classification-machine learning mastery, january 3, 2020.[-ganda//ukrainian scientific journal of information security, 2020, vol. 26, issue 3, pp. 139-144.]. *: https://machinelearningmastery. com/precision-recall-and-f-measure-forimbalanced-classification*.

CDC (2013). Impact of an innovative approach to prevent mother-to-child transmission of HIV–malawi, july 2011-september 2012. 62(8):148–151.

Chagomerana, M. B., Miller, W. C., Tang, J. H., Hoffman, I. F., Mthiko, B. C., Phulusa, J., John, M., Jumbe, A., and Hosseinipour, M. C. (2018). Optimizing prevention of hiv mother to child transmission: Duration of antiretroviral therapy and viral suppression at delivery among pregnant malawian women. *PloS one*, 13(4):e0195033.

Chaula, R. B. and Justo, G. N. (2022). A robust random forest prediction model for mother-to-child hiv transmission based on individual medical history. *Tanzania Journal of Engineering and Technology*, 41(3).

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Clinical Info (2024a). Pregnant people who have not achieved viral suppression on antiretroviral therapy | NIH.

Clinical Info (2024b). Use of antiretroviral drugs to prevent perinatal HIV transmission and improve health for pregnant people | NIH.

Darby, A., Jones, S. H., Hope, S., and Hiv, K. (2021). World health organization guidelines (option a, b, and b+) for antiretroviral drugs to treat pregnant women and prevent hiv infection in infants. *The Embryo Project Encyclopedia. The Embryo Project at Arizona State University, Tempe*.

Elizabeth Glaser Pediatric AIDS Foundation (2021). Innovations and impact toward the elimination of mother-to-child transmission in kenya - EGPAF.

Fatima, S., Hussain, A., Amir, S. B., Ahmed, S. H., and Aslam, S. M. H. (2023). Xgboost and random forest algorithms: an in depth analysis. *Pakistan Journal of Scientific Research*, 3(1):26–31.

Fieggen, J., Smith, E., Arora, L., and Segal, B. (2022). The role of machine learning in hiv risk prediction. *Frontiers in Reproductive Health*, 4:1062387.

Gen David L (2023). Five methods for data splitting in machine learning.

Gill, M. M., Natumanya, E. K., Hoffman, H. J., Okomo, G., Taasi, G., Guay, L., and Masaba, R. (2020). Active pediatric hiv case finding in kenya and uganda: A look at missed opportunities along the prevention of mother-to-child transmission of hiv (pmtct) cascade. *PloS one*, 15(6):e0233590.

Gupta, V. (2023). Unveiling the magic of MICE: (multiple imputation with chained equations).

Haldane, V., Chuah, F. L., Srivastava, A., Singh, S. R., Koh, G. C., Seng, C. K., and Legido-Quigley, H. (2019). Community participation in health services development, implementation, and evaluation: A systematic review of empowerment, health, community, and process outcomes. *PloS one*, 14(5):e0216112.

Joint United Nations Programme (2024). Global hiv aids statistics — fact sheet.

Kagendi, N. and Mwau, M. (2023). A machine learning approach to predict hiv viral load hotspots in kenya using real-world data. *Health Data Science*, 3:0019.

Kharkar, D. (2023). Unravelling the power of xgboost: Boosting performance with extreme gradient boosting. *Retrieved May*, 6:2024.

Koss, C. A., Natureeba, P., Kwarisiima, D., Ogena, M., Clark, T. D., Olwoch, P., Cohan, D., Okiring, J., Charlebois, E. D., Kamya, M. R., et al. (2017). Viral suppression and retention in care up to 5 years after initiation of lifelong art during pregnancy (option b+) in rural uganda. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 74(3):279–284.

Li, B., Li, M., Song, Y., Lu, X., Liu, D., He, C., Zhang, R., Wan, X., Zhang, R., Sun, M., et al. (2022). Construction of machine learning models to predict changes in immune function using clinical monitoring indices in hiv/aids patients after 9.9-years of antiretroviral therapy in yunnan, china. *Frontiers in Cellular and Infection Microbiology*, 12:867737.

Li, J., Hao, Y., Liu, Y., Wu, L., Liang, H., Ni, L., Wang, F., Wang, S., Duan, Y., Xu, Q., et al. (2024). Supervised machine learning algorithms to predict the duration and risk of long-term hospitalization in hiv-infected individuals: a retrospective study. *Frontiers in Public Health*, 11:1282324.

Logie, C. H., James, L., Tharao, W., and Loutfy, M. R. (2011). Hiv, gender, race, sexual orientation, and sex work: a qualitative study of intersectional stigma experienced by hiv-positive women in ontario, canada. *PLoS medicine*, 8(11):e1001124.

Maskew, M., Sharpey-Schafer, K., De Voux, L., Crompton, T., Bor, J., Rennick, M., Chirowodza, A., Miot, J., Molefi, S., Onaga, C., et al. (2022). Applying machine learning and predictive modeling to retention and viral suppression in south african hiv treatment cohorts. *Scientific reports*, 12(1):12715.

Ministry of Health, Kenya (2012). Guidelines, standards & policies portal.

Moses, A., Chama, C., Udo, S., and Omotora, B. (2009). Knowledge, attitude and practice of ante-natal attendees toward prevention of mother to child transmission (pmtct) of hiv infection in a tertiary health facility, northeast-nigeria. *East African journal of public health*, 6(2):128–135.

Mugwaneza, P., Lyambabaje, A., Umubyeyi, A., Humuza, J., Tsague, L., Mwanyumba, F., Mutabazi, V., Nsanzimana, S., Ribakare, M., Irakoze, A., et al. (2018). Impact of maternal art on mother-to-child transmission (mtct) of hiv at six weeks postpartum in rwanda. *BMC Public Health*, 18:1–11.

Mutabazi, J. C., Gray, C., Muhwava, L., Trottier, H., Ware, L. J., Norris, S., Murphy, K., Levitt, N., and Zarowsky, C. (2020). Integrating the prevention of mother-to-child transmission of hiv into primary healthcare services after aids denialism in south africa: perspectives of experts and health care workers-a qualitative study. *BMC health services research*, 20:1–18.

Myer, L., Phillips, T., McIntyre, J., Hsiao, N.-Y., Petro, G., Zerbe, A., Ramjith, J., Bekker, L.-G., and Abrams, E. (2017). Hiv viraemia and mother-to-child transmission risk after antiretroviral therapy initiation in pregnancy in cape town, south africa. *HIV medicine*, 18(2):80–88.

NASCOP (2014). National guidelines for hiv/sti programming with key populations.

National Institute of Health (2024). Preventing perinatal transmission of HIV.

National Syndemic Diseases Control Council (2021). Kenya AIDS strategic framework (KASF) II 2020/21-2024/25.

Ndege, S., Washington, S., Kaaria, A., Prudhomme-O'Meara, W., Were, E., Nyambura, M., Keter, A. K., Wachira, J., and Braitstein, P. (2016). Hiv prevalence and antenatal care attendance among pregnant women in a large home-based hiv counseling and testing program in western kenya. *PloS one*, 11(1):e0144618.

Ngangue, P., Fleurantin, M., Adekpedjou, R., Philibert, L., and Gagnon, M.-P. (2021). Involvement of male partners of pregnant women in the prevention of mother-to-child transmission (pmtct) of hiv in haiti: a mixed-methods study. *American journal of men's health*, 15(2):15579883211006003.

Okoli, C., Hajizadeh, M., Rahman, M. M., and Khanam, R. (2020). Geographical and socioeconomic inequalities in the utilization of maternal healthcare services in nigeria: 2003–2017. *BMC Health Services Research*, 20:1–14.

Olamendy, J. C. (2024). Data transformations: Encoding data for machine learning.

O'Halloran Leach, E., Lu, H., Caballero, J., Thomas, J. E., Spencer, E. C., and Cook, R. L. (2021). Defining the optimal cut-point of self-reported art adherence to achieve viral suppression in the era of contemporary hiv therapy: a cross-sectional study. *AIDS Research and Therapy*, 18(1):36.

Pan American Health Organization (2020). Antiretroviral therapy - PAHO/WHO | pan american health organization.

Rohit Kundu (2022). Precision vs. recall: Differences, use cases & evaluation.

Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432.

Schott, M. (2019). Random forest algorithm for machine learning. *Capital One Tech*.

SeyedAlinaghi, S., Afsahi, A. M., Moradi, A., Parmoon, Z., Habibi, P., Mirzapour, P., Dashti, M., Ghasemzadeh, A., Karimi, E., Sanaati, F., et al. (2023). Current art, determinants for virologic failure and implications for hiv drug resistance: an umbrella review. *AIDS research and therapy*, 20(1):74.

Terence Shin and Matthew Urwin (2024). Understanding feature importance in machine learning.

Thea, D. M., Steketee, R. W., Pliner, V., Bornschlegel, K., Brown, T., Orloff, S., Matheson, P. B., Abrams, E. J., Bamji, M., Lambert, G., et al. (1997). The effect of maternal viral load on the risk of perinatal transmission of hiv-1. *Aids*, 11(4):437–444.

Tuthill, E. L., Odhiambo, B. C., and Maltby, A. E. (2024). Understanding mother-to-child transmission of hiv among mothers engaged in hiv care in kenya: a case report. *International Breastfeeding Journal*, 19(1):14.

UNAIDS (2014). New "beyond zero campaign" to improve maternal and child health outcomes in kenya.

UNAIDS (2021). Understanding measures of progress towards the 95–95–95 HIV testing, treatment and viral suppression targets.

UNICEF (2024). Elimination of mother-to-child transmission - UNICEF DATA.

United Nations (2023). — SDG indicators.

U.S. CDC and Ministry of Health, Kenya (2013). The cost of comprehensive HIV treatment in kenya : report of a cost study of HIV treatment programs in kenya.

USAID (2024). Preventing vertical transmission (PVT) | global health.

Violari, A., Cotton, M. F., Gibb, D. M., Babiker, A. G., Steyn, J., Madhi, S. A., Jean-Philippe, P., and McIntyre, J. A. (2008). Early antiretroviral therapy and mortality among hiv-infected infants. *New England Journal of Medicine*, 359(21):2233–2244.

White AB, Mirjahangir JF, Horvath H, Anglemyer A, and Read JS (2024). Using antiretroviral medication to prevent transmission of HIV from mother-to-child during breastfeeding.

WHO (2010). Guidelines on HIV and infant feeding 2010: principles and recommendations for infant feeding in the context of HIV and a summary of evidence. In *Guidelines on HIV and infant feeding 2010: principles and recommendations for infant feeding in the context of HIV and a summary of evidence*.

WHO (2024). World health organization (WHO).

Appendix A

Similarity Report

Ashley Achieng

Thesis_Ashley_Achieng_01042025_V3.pdf

Strathmore University (Main Account)

Document Details

Submission ID
trn:oid:::2945:275684761

Submission Date
Apr 1, 2025, 12:20 PM GMT+3

Download Date
Apr 1, 2025, 12:23 PM GMT+3

File Name
Thesis_Ashley_Achieng_01042025_V3.pdf

File Size
943.2 KB

63 Pages

13,410 Words

79,179 Characters

turnitin Page 1 of 81 - Cover Page Submission ID trn:oid:::2945:275684761

25% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

256 Not Cited or Quoted 22%
Matches with neither in-text citation nor quotation marks

49 Missing Quotations 4%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

13% Internet sources
14% Publications
22% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Appendix B

Ethical Clearance Confirmation

Strathmore UNIVERSITY

17th February 2025

Ms Odhiambo Ashley,
ashley.achieng@strathmore.edu

Dear Ms Odhiambo,

RE: Predicting Mother-To-Child HIV Transmission among Mother-Baby Pairs in Kenya: A Focused Comparison of Random Forest and XGBoost Models

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2658/25.** The approval period is from **17th February 2025 to 16th February 2026.**

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