Department of Electrical and Computer Engineering
North South University

Directed Research

MACHINE LEARNING APPROACHES FOR IMMUNE SYSTEM ANALYSIS AND MODELING

MOHAMMOD ABDULLAH BIN HOSSAIN	**2022351042**
RISTIAN UDDIN	**2021188042**
FARZANA ERIN	**2021564042**
ZOBAER AHAMMOD ZAMIL	**2021796042**

Faculty Advisor:

Dr. Shahnewaz Siddique

Associate Professor

ECE Department

Spring, 2024

LETTER OF TRANSMITTAL

04th September, 2024
To
Dr. Rajesh Palit
Professor and Chairman
Department of Electrical and Computer Engineering

Through,
Dr. Shahnewaz Siddique
Assistant Professor
Department of Electrical and Computer Engineering
North South University Dhaka, Bangladesh

Subject: Submission of Directed Research Report on 'Machine Learning Approaches for Immune System Analysis and Modeling'.

Dear Sir,
Conveying the utmost complement to your position and honor that we would like to approach you with the following fact that we are very happy and grateful at this moment of submission of our Directed Research on 'Immune System Analysis and Modeling'. In partial fulfilment of the requirement for the degree program, Bachelor of Science in Electrical and Computer Engineering.

We have used all the data and information in this report that was found most relevant and correct as per best of our judgment. It will be worthwhile to mention this field work has immensely helped us to gather knowledge that might have been required long time for us to acquire. We sincerely hope this report will standard of our judgment. We will be always available for any inquiry or clarification regarding the report.

Sincerely,
Zobaer Ahammod Zamil (2021796042)
North South University

Farzana Erin (2021564042)
North South University

Ristian Uddin (2021188042)
North South University

Mohammod Abdullah Bin Hossain (2022351042)
North South University

APPROVAL

Mohammod Abdullah Bin Hossain (2022351042), Ristian Uddin (2021188042), Farzana Erin (2021564042) and Zobaer Ahammod Zamil (2021796042) from Electrical and Computer Engineering Department of North South University, have worked on the Directed Research Project titled "**MACHINE LEARNING APPROACHES FOR IMMUNE SYSTEM ANALYSIS AND MODELING**" under the supervision of Dr. Shahnewaz Siddique partial fulfillment of the requirement for the degree of Bachelors of Science in Engineering and has been accepted as satisfactory.

Supervisor's Signature

…….

Dr. Shahnewaz Siddique

Associate Professor

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

Chairman's Signature

…….

Dr. Rajesh Palit

Professor

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

DECLARATION

This is to declare that this project/directed research is our original work. No part of this work has been submitted elsewhere, partially or fully, for the award of any other degree or diploma. All project-related information will remain confidential and shall not be disclosed without the formal consent of the project supervisor. Relevant previous works presented in this report have been properly acknowledged and cited. The plagiarism policy, as stated by the supervisor, has been maintained.

Students' names & Signatures

1. Zobaer Ahammod Zamil

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

2. Mohammod Abdullah Bin Hossain

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

3. Ristian Uddin

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

4. Farzana Erin

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

ACKNOWLEDGEMENTS

The authors would like to express their heartfelt gratitude towards their project and research supervisor, Dr. Shahnewaz Siddique Associate Professor, Department of Electrical and Computer Engineering, North South University, Bangladesh, for his invaluable support, precise guidance and advice pertaining to the experiments, research and theoretical studies carried out during the course of the current project and also in the preparation of the current report.

Furthermore, the authors would like to thank the Department of Electrical and Computer Engineering, North South University, Bangladesh for facilitating the research. The authors would also like to thank their loved ones for their countless sacrifices and continual support.

ABSTRACT

MACHINE LEARNING APPROACHES FOR IMMUNE SYSTEM ANALYSIS AND MODELING

The rapidly changing healthcare landscape has enhanced the facilitation of HIV detection through artificial intelligence even more. This breakthrough is immense. Our results, based on sophisticated machine learning techniques, indicate the paramount importance of AI for enhancing diagnostic accuracy. CD4 and CD8 were a data set justifying the testing of validity in different realms using different algorithms: XGBoost, K-Nearest Neighbors, Decision Trees, and Random Forests, and on top of that, Logistic Regression. This was followed by advanced stages of pre-processing and the conceptualization of the algorithms. Tremendous gains in these matrices with proper hyper-parameter tuning under the purview of making each one sound and optimal have been realized. More specifically, the models developed consider Gradient Boosting and LightGBM models, with achievements where the testing accuracies are 86.68% and 86.21%, respectively, with the slightest difference between the training and cross-validation accuracies. Explainable AI interpretations by SHAP values tend to help in significant features of CD4 and CD8, thus beefing interpretability and reliability. Besides that, all other results obtained from the performance metrics have shown that the separation of classes of models has been done with significant values of AUC. Such results underscore the need for further fine-tuning and validation of this tool to ensure its robustness for practical purposes. We contribute to this growing research by using machine-learning models to identify HIV. Specifically, the reasons for touting the relevance of this paper were because we have done careful validation of model performance and the need for intelligent features.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 Background and Motivation

The CD4/CD8 ratio is increasingly recognized as a critical biomarker in the management and understanding of HIV/AIDS. This ratio, calculated by dividing the number of CD4 T-helper cells by the number of CD8 cytotoxic T-lymphocytes, offers significant insights into the immune system's health and its response to HIV infection. The advent of antiretroviral therapy (ART) has revolutionized HIV treatment, transforming it from a deadly disease to a manageable chronic condition. However, despite the effectiveness of ART in suppressing viral loads and increasing CD4 counts, the CD4/CD8 ratio has emerged as an important marker for assessing the broader immune dysfunction that persists in many individuals [1]. Traditionally, the management of HIV focused on monitoring CD4 counts and viral load as primary indicators of immune function and treatment success. CD4 cells, also known as T-helper cells, play a crucial role in orchestrating the immune response by activating other immune cells, including CD8 cells, macrophages, and B lymphocytes. CD8 cells, on the other hand, are responsible for identifying and destroying infected cells, particularly those harboring viruses like HIV. As HIV progresses, it selectively infects and destroys CD4 cells, leading to a decline in CD4 counts and an eventual inversion of the CD4/CD8 ratio. This inversion, where CD8 cells outnumber CD4 cells, is a hallmark of advanced HIV infection and reflects the underlying immune dysfunction [4][13][14].

With the introduction of modern ART, most patients experience a significant increase in CD4 counts, and viral loads often become undetectable. However, even when CD4 counts return to normal levels (generally above 500 cells/μL), the CD4/CD8 ratio does not always normalize. Many individuals who start treatment with a low CD4/CD8 ratio continue to exhibit an abnormal ratio for years, despite effective viral suppression. This persistent imbalance is of particular concern because a low CD4/CD8 ratio is associated with increased immune activation and a higher risk of non-AIDS-related complications, such as cardiovascular disease, liver disease, kidney disease, malignancies, and neurocognitive disorders [1][2][15].

The importance of the CD4/CD8 ratio extends beyond its role as a simple marker of immune suppression. It has been increasingly recognized as a reflection of immune dysregulation, a condition where the immune system is in a state of chronic activation despite effective control of the virus. This ongoing immune activation is thought to contribute to the increased morbidity and mortality observed in people living with HIV (PLWH), even when traditional markers like CD4 count and viral load suggest successful treatment [1][3]. Studies have shown that a low CD4/CD8 ratio is predictive of poorer clinical outcomes, including a higher risk of severe non-AIDS events and overall mortality. This makes the CD4/CD8 ratio a valuable tool for identifying individuals at higher risk of complications, even when other indicators appear normal [3][5][16].

In recent years, the CD4/CD8 ratio has gained attention not only as a marker of disease progression but also as a potential guide for personalized HIV care. Researchers have

highlighted its utility in predicting adverse outcomes, such as poor responses to vaccines and an increased likelihood of developing cancer. These findings suggest that the CD4/CD8 ratio could be used to tailor treatment strategies more effectively, ensuring that patients at higher risk of complications receive more intensive monitoring and care [2]. However, the role of the CD4/CD8 ratio in clinical practice is still evolving. While some clinical guidelines recommend monitoring the ratio, there is no consensus on the optimal cutoff points or the best way to integrate this marker into routine care. Moreover, the variability in how the ratio responds to ART underscores the need for further research to understand the underlying mechanisms and to develop strategies that can more effectively restore immune balance in PLWH [3][17].

Machine learning and other advanced analytical techniques offer promising avenues for addressing these challenges. By analyzing large datasets, these tools can help identify patterns and predictors that are not immediately apparent, potentially leading to more accurate risk stratification and more effective treatment strategies. As the field of HIV research continues to evolve, the CD4/CD8 ratio is likely to play an increasingly important role in guiding both clinical practice and research efforts, helping to improve outcomes for people living with HIV [3].

1.2 Purpose and Goal of the Project

The purpose of this project is to explore the significance of the CD4/CD8 ratio as a crucial biomarker in the management and understanding of HIV/AIDS. Given that traditional metrics such as CD4 counts and viral load do not fully capture the extent of immune dysfunction in individuals with HIV, this project aims to delve into the role of the CD4/CD8 ratio in reflecting persistent immune dysregulation despite effective antiretroviral therapy (ART). By analyzing how the CD4/CD8 ratio correlates with clinical outcomes and the risks of non-AIDS-related complications, the project seeks to underscore its potential as a key indicator for more comprehensive HIV care.

The goal of this project is to leverage advanced analytical techniques, such as machine learning, to enhance the understanding and utility of the CD4/CD8 ratio in personalized HIV treatment. By identifying patterns and predictors associated with an abnormal CD4/CD8 ratio, the project aims to contribute to more accurate risk stratification and better-informed treatment strategies for people living with HIV. Ultimately, the project aspires to improve patient outcomes by integrating the CD4/CD8 ratio into routine clinical practice, thereby addressing the broader immune dysfunction that persists in many individuals despite successful ART.

Chapter 2 Research Literature Review

2.1 Existing Research and Limitations

The paper highlights the particular importance of CD4 count as a marker of disease progression and immune competence in HIV infection. There have been some previous applications of machine learning for prediction, but these had not previously considered predicting CD4 count changes. The investigators therefore developed SVM models using baseline input variables (such as HIV genome sequences, viral load and time since the initial CD4 measure) to predict categories that describe a 6-month window of the change in CD4 count following that measurement. The authors developed models using three different SVM kernel functions (linear, quadratic and radial basis function) with different subsets of input variables. The model with the highest accuracy used sequence data, current viral load, and time since the initial measurement as inputs to predict one of three CD4 count change categories: 85%, increasing or stable; 84% decreasing but remaining above 350 cells/μL; and 78% declining below 350 cells/μL following the initial measure. This study suggests that there is potential to use machine learning techniques to predict CD4 count changes in response to antiretroviral therapy for clinical decision making/resource allocation purposes but acknowledges several limitations including sample size and notes that further work should be done to refine this modelling approach and validate it in other datasets. [6]

The research conducted by Tang et al. (2015) explores the clinical relevance of the CD4 lymphocyte ratio as a quantitative marker of immunologic health in HIV-1 infected individuals, particularly within an African cohort. The study establishes the stability of the CD4 ratio during the early years of infection and investigates its association with various factors including viral load, CD4+ T-cell count, HIV-1 subtype, and host genetics. The findings indicate that a CD4 ratio above 1.0 early in the infection correlates with delayed disease progression, irrespective of other factors. Moreover, the study identifies low viral load and specific HLA variants as predictors of a higher CD4 ratio, underscoring its potential as a robust marker for monitoring immunologic health in HIV-1 patients [7]

Jeffrey Daniel Jenks1, Martin Hoenigl discusses the limitations of CD4+ cell count and HIV RNA viral load in predicting non-AIDS related morbidity and mortality in people living with HIV (PLWH) on antiretroviral therapy (ART) with suppressed viral load. This research work has been concerned with the investigation of the CD4:CD8 lymphocyte ratio as an indicator of immune system health during HIV-1 infection. On that note, data was collected from 499 African seroconverters for analysis purpose. The investigation revealed that CD4:CD8 ratio did not show any marked change in the first three years of infection and could actually be used to predict the progression of diseases since it was found out that those who had values higher than 1.0 were more likely to progress slowly into severe immunodeficiency stages. When comparing CD4:CD8 rates with viral load or CD4 slope little was seen; meaning other words about disease condition's statuses are produced by it. The two main factors associated with high CD4:CD8 ratios over 1.0 include low-level viremia (viral load) and the present of HLA-A74:01 allele. Interestingly enough, however, other HLA polymorphisms associated

with viral loads and actual trajectories in CD4 did not seem to have any significant effects on the ratio between these two types of lymphocytes within an organism's body.The CD4:CD8 ratio represents a valid barometer for monitoring the state of our immune systems while infected with HIV-1 hence its value would be immensely high in medical care including making sure that patients who are often at risk receive antiviral drugs. In order to fully understand how the immune system regulates HIV-1 infection, the authors recommend further research into why certain HLA variants demonstrate preferable levels of the CD4:CD8 ratio. [8]

The ratio of CD4/CD8 cells is a key indicator of immunological health, particularly in people infected with HIV-1. In particular, while CD8 T cells had been recognized much later compared to their CD4 counterparts, this did not diminish their importance in combating intracellular infections and controlling chronic diseases. In HIV patients, the CD4/CD8 ratio serves as an indicator of the immune system status which indicates how good balance these two populations maintain. A falling rate often suggests disease progression while an increasing or stable one implies infection control measures are working. This is not only important for its role as a diagnostic measure but also for its therapeutic importance because it shows about immune reconstitution and effectiveness of antiretroviral therapies. More research on the CD4/CD8 ratios could potentially help improve management of HIV-1 infection through understanding more about the therapeutic potential of CD8 T cells. [9]

Over time, the research concerning CD8 T cells and their crucial role in clinical settings, particularly for HIV-1 diagnostic and monitoring of immunity, has extensively expanded. Earlier investigations utilized flow cytometry to establish standard laboratory control values for CD4 and CD8 T lymphocytes, thus facilitating accurate immune function assessments. A distinct decrease in CD4 T cells accompanied by an altered CD4/CD8 ratio is a decisive signal of ailment advancement and aids forecasting in persons infected by HIV-1. Studies have persistently confirmed that CD4 as well as CD8 levels remain stable throughout life while deviations from these norms point towards possible impairment within the immune system itself. These parameters have been monitored mainly through the application of flow cytometric techniques such as during diagnosing and managing HIV-1 disease which frequently shows lower counts of both CD4 cells and distorted ratios of CD4/CD8. Such information highlights significance of these two groups of lymphocytes not only at healthy state but also when facing various pathologies hence providing essential clinical practice/ research insights. [10]

The study titled "Machine Learning: Selected Variables Associated with CD4 T Cell Recovery" examines the factors that influence CD4 T cell recovery in patients undergoing antiretroviral therapy (ART). CD4 T cells are crucial in maintaining immune function, and their recovery is essential for improving the health outcomes of HIV patients. Despite the effectiveness of ART, not all patients experience optimal CD4 T cell recovery, which necessitates a deeper understanding of the variables that influence this process. The study employs a machine learning (ML) approach to identify key variables associated with CD4 T cell recovery. By leveraging ML algorithms, the researchers aim to uncover patterns and relationships within the data that traditional statistical methods might miss. The study's

methodology involves selecting a set of clinical, demographic, and genetic variables and applying various ML models to assess their influence on CD4 T cell recovery. The study identifies several key variables that significantly impact CD4 T cell recovery, including baseline CD4 count, age, duration of HIV infection, and adherence to ART. The ML models also highlight the importance of genetic factors, such as specific gene polymorphisms, in influencing immune recovery. One of the primary findings is that baseline CD4 count is a strong predictor of recovery, which aligns with existing literature. Additionally, the study suggests that older age and longer duration of HIV infection are associated with poorer recovery outcomes, indicating the importance of early intervention. The study appears to focus on cross-sectional data, which may not fully capture the dynamics of CD4 T cell recovery over time. [11]

The CD4/CD8 ratio is looked into in the paper titled “The Revival of an ‘Old’ Marker: CD4/CD8 Ratio” as it relates to its importance in HIV management. The authors discuss the effect of modern antiretroviral therapy on turning HIV into a chronic condition, leading to persistent inflammation and immune activation even when there is suppression of viral loads. This review highlights how, apart from predicting AIDS-related outcomes, the CD4/CD8 ratio can also be used as a biomarker for non-AIDS comorbidities, such as cardiovascular diseases or cancers. A discussion is provided about how the reverse of this ratio is associated with immune senescence and chronic inflammation thereby making it an important tool for following up HIV patients especially in evaluating their risk for severe non-AIDS events. They propose that this biomarker may serve a crucial function in directing more individualized patient care; they advocate for incorporating it into routine clinical practice to identify people who might need closer observation and possibly earlier intervention. [19]

The article dubbed "Consequences of CD4/CD8 Ratios in Human Immunodeficiency Virus Infections" researches the significant focus of the CDs’ existence as a parameter for effective HIV management. In particular, they assert that there is a drastic decrease in CD4 T-cells and an increase in CD8 T-cells resulting to an inverted CD4/CD8 ratio due to this infection. Immunological derangement, chronic inflammation as well as immune senescence are closely associated with this imbalance. This paper highlights how important forecasting AIDS-related and non-AIDS related comorbidities such as heart diseases, cancer, total deaths and many others, including those taking ART, largely depend on this ratio. It further states that if one has had a consistently low CD4/CD8 Ratio despite subjected to Antiretroviral therapy (ART) then he/she is likely to suffer more from multiple illness conditions leading up to earlier death hence calling for its routine use for monitoring disease course and response assessment. Also, it implies that the CD4/CD8 ratio may serve as an alternative marker of HIV reservoir where ART efficiency will be evaluated and future therapy aimed at lessening the volume of virus reservoirs could be examined on it. Generally, why-not practicing continuous monitoring of CD4/CD8 Ratio in HIV positive patients’ facilities so as to enhance management quality among these individuals? [20]

CD4/CD8 ratio predictive value and the outcome in ART-treated HIV patients were analyzed—a study based on earlier studies regarding immune markers and serious non-AIDS event risk. For sure, previous studies provided conflicting conclusions regarding such a

prediction ability of the CD4/CD8 ratio for various clinical endpoints, which could be most probably because of different methodologies, threshold levels possibly applied for the ratio, and endpoint measures. In this light, the research proffered is unique to consider the study of data in the CoRIS cohort through the use of solid statistical techniques so that confounding variables and selection bias are minimized ably, and the prognostic meaning of the CD4/ CD8 ratio, CD4 count, and CD8 count is established two years after the institution of ART. Results showed that a CD4/CD8 ratio below 0.3 and a CD8 count of 1500 cells/μL or more are strong predictors of the development of SNAE, including major cardiovascular events and non-AIDS-defining malignancies, in the following five years, independently of CD4 count. The study suggests that these markers, notably the CD4/CD8 ratio, could be very instrumental in point out patients at high risk and may, therefore, benefit from enhanced monitoring and any appropriate targeted interventions within a harmonized approach toward the assessment of these immune markers. [21]

The present study aims to describe the relationship of the CD4/CD8 ratio with risk for non-AIDS-defining severe events in HIV-infected patients under combination antiretroviral therapy and to determine whether the ratio predicts these events beyond that provided by CD4+ cell count. Earlier studies have identified the CD4/CD8 ratio as an immune activation marker, indicating morbidity and mortality with respect to aging and chronic HIV infection. A low CD4/CD8 ratio was a risk factor for non-AIDS events, including cardiovascular events, kidney disease, and bacterial infections, according to studies by Serrano-Villar et al. in 2014 and Mussini et al. in 2015. However, results are conflicting; some studies, such as Hatano et al. in 2013, found no significant association on adjustment for CD4+ counts. It contributes to this debate by showing that, while the CD4/CD8 ratio did not add predictive value in most cases for NADEs, a ratio below 0.5 was significantly associated with a higher risk of non-AIDS-defining cancers, highlighting the potential role for this marker in identifying those patients who could benefit from more intensive cancer prevention and screening strategies. In agreement with the Swiss HIV cohort and previous studies, low CD4/CD8 ratios were linked to an increased risk for cancer. Such evidence speaks to a particular sort of immune dysfunction that might make HIV-infected patients prone to malignancies in spite of effective ART. [22]

The research seeks to understand the effect of ferroptosis in individuals infected with human immunodeficiency virus on continuous combined antiretroviral therapy. Specifically, it measures the labile-bound iron pool in CD4+ and CD8+ T cells against some key markers of immune status. These results demonstrate the association of high levels of the labile-bound iron pool in CD8+ T cells with both T cell counts and a reduced CD4/CD8 ratio and hence suggest ongoing ferroptosis, even when viral replication is undetectable. This suggests that targeting ferroptosis with treatment could improve immune function in HIV-positive individuals on cART. The study also points out that further research aimed at firmly establishing a labile-bound iron pool as an indicator of prognosis and the damaging effects of cART use on iron balance is needed. [23]

Chapter 3 Methodology

3.1 System Design

This study aims to demystify the factors influencing the progression to AIDS by harnessing the combined power of machine learning and Explainable AI. By exploring a rich dataset that includes clinical, demographic, and treatment-related variables, particularly focusing on CD4 and CD8 cell counts and their ratios, our goal is to uncover critical patterns and predictors associated with the onset of AIDS. The following objectives are designed to address key research questions, providing a comprehensive approach to understanding and predicting AIDS progression.

1. **Comprehensive Identification of Predictive Factors for AIDS Progression:**
    - **Research Question:** *"What are the key factors that significantly influence the progression to AIDS, and how do these factors interact with each other?"*

    The first objective is to conduct an in-depth analysis of the various factors that contribute to the progression to AIDS. Through advanced feature selection techniques like Recursive Feature Elimination with Cross-Validation (RFECV), we aim to identify the most influential variables from our dataset, which includes 23 diverse factors such as age, treatment history, CD4 and CD8 cell counts, and more. This process will not only isolate the most critical predictors but will also provide insights into the complex interrelationships between these factors, offering a nuanced understanding of how they collectively impact AIDS progression.
2. **Development and Optimization of Predictive Models for AIDS Progression:**
    - **Research Question:** *"How effectively can machine learning algorithms predict the progression to AIDS, and what are the primary indicators contributing to the model's accuracy?"*

    The second objective focuses on creating and refining predictive models that can accurately forecast the likelihood of an individual progressing to AIDS. By leveraging an ensemble of machine learning algorithms, we aim to overcome challenges such as data imbalance and to improve key performance metrics like accuracy, precision, and recall. Special emphasis will be placed on understanding the role of CD4 and CD8 cell counts and their ratios in enhancing the predictive power of these models. This objective underscores our commitment to developing reliable, high-performing models that can be used in clinical settings to identify individuals at higher risk and facilitate early intervention.
3. **Enhancement of Model Transparency and Trust with Explainable AI:**
    - **Research Question:** *"How can Explainable AI techniques improve the transparency and trustworthiness of AI models used for predicting AIDS progression, and how does this transparency impact their acceptance in clinical practice?"*

    The third objective is to make the decision-making processes of our predictive models more transparent and understandable to healthcare professionals and stakeholders. By incorporating Explainable AI techniques, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations), we aim to clarify how the models arrive at their predictions, particularly in relation to key factors like CD4 and CD8 cell counts. This increased transparency is crucial for building trust in AI-driven tools, ensuring that they are not only accurate but also interpretable and actionable in real-world clinical environments. The ultimate goal is

to enhance the adoption and effective use of these models in healthcare, leading to better-informed decision-making and improved patient outcomes.

In pursuit of these objectives, our methodology integrates state-of-the-art machine learning techniques with a strong emphasis on clinical relevance and interpretability. This holistic approach is designed to advance the understanding of AIDS progression and to contribute to the development of practical tools that can assist in the timely identification and treatment of individuals at risk of developing AIDS.

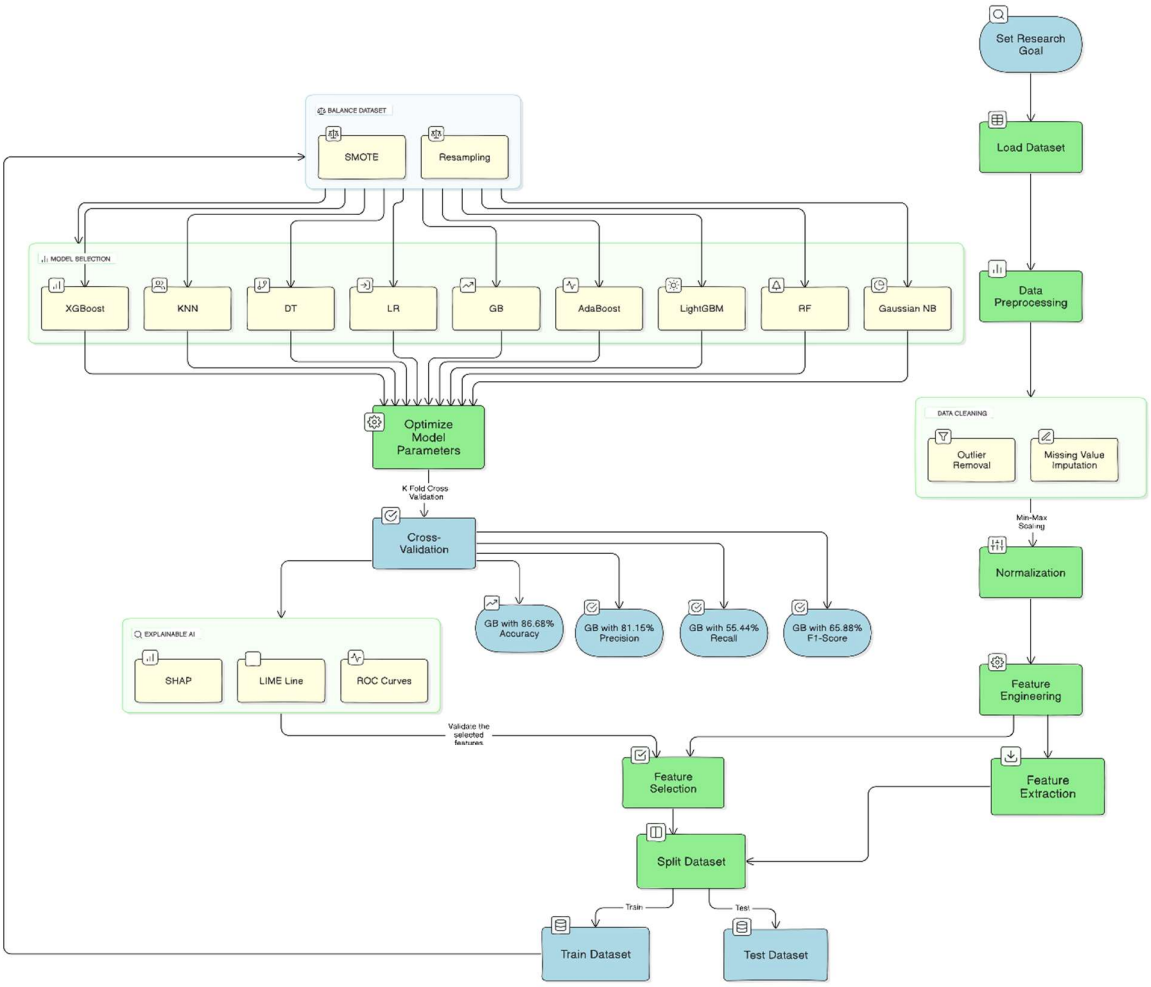

Figure 3.1.1: *Workflow Diagram*

3.2 Dataset Collection and Composition

For our research project titled "Demystifying AIDS Predictions: The Synergy of Explainable AI and Machine Learning," we utilized a robust dataset sourced from Kaggle [12]. This dataset plays a pivotal role in helping us analyze the influence of various clinical, demographic, and treatment-related factors—particularly CD4 and CD8 cell counts—on the progression to AIDS. Below, we detail the source, scope, and composition of the dataset used in this study.

3.2.1 Source and Accessibility

The dataset utilized in this study was sourced from Kaggle, a well-known platform that hosts a vast array of datasets contributed by both users and moderators. The dataset in question includes critical variables pertinent to understanding AIDS progression. This dataset contains healthcare statistics and categorical information about patients who have been diagnosed with AIDS. It was initially published in 1996. The ease of access to this dataset makes it a valuable resource for researchers and practitioners alike, facilitating reproducibility and further exploration in the field of AIDS research.

3.2.2 Data Scope and Period

The dataset encompasses a comprehensive range of data points related to the clinical and demographic profiles of individuals potentially progressing to AIDS. It includes 2,139 samples, each representing a unique individual, with 23 distinct features that cover various aspects such as treatment history, CD4 and CD8 cell counts, age, weight, and behavioral factors. Although the dataset does not specify the exact period over which the data was collected, it provides a broad snapshot of factors influencing AIDS progression. This extensive dataset offers a solid foundation for training machine learning models, ensuring that the models can capture a wide range of scenarios and individual profiles.

3.2.3 Dataset Composition

The dataset is composed of detailed records for each individual, including both clinical and demographic attributes. These features are integral to our analysis as they provide a multi-faceted view of each patient's health status and treatment history.

Features	Description	Type
Time	Time to failure or censoring, indicating the period until the progression to AIDS or until the last recorded data point.	Quantitative
Treatment Indicator (trt)	Categorizes the type of treatment received by the patient (0 = ZDV only; 1 = ZDV + ddI, 2 = ZDV + Zal, 3 = ddI only).	Categorical
Age	The age of the patient at baseline, providing a critical demographic context.	Quantitative
Weight (wtkg)	The patient's weight at baseline, which can be an indicator of overall health status.	Quantitative
Hemophilia (hemo)	Indicates whether the patient has hemophilia (0 = no, 1 = yes).	Categorical
Homosexual Activity (homo)	Captures the patient's history of homosexual activity (0 = no, 1 = yes).	Categorical
IV Drug Use (drugs)	Records whether the patient has a history of intravenous drug use (0 = no, 1 = yes).	Categorical
Karnofsky Score (karnof)	A score representing the patient's ability to carry out daily activities, ranging from 0 to 100.	Quantitative
Non-ZDV Antiretroviral Therapy (opri)	Indicates prior antiretroviral therapy before the start of the study (0 = no, 1 = yes).	Categorical
ZDV in the 30 Days Prior (z30)	Indicates whether the patient received ZDV treatment in the 30 days prior to the study (0	Categorical

	= no, 1 = yes).	
Pre-175 Anti-Retroviral Therapy (preanti)	Number of days on anti-retroviral therapy before the study baseline.	Quantitative
Race	The racial background of the patient (0 = White, 1 = Non-White).	Categorical
Gender	The gender of the patient (0 = Female, 1 = Male).	Categorical
Antiretroviral History (str2)	Indicates whether the patient was naive or experienced with antiretroviral treatments (0 = naive, 1 = experienced).	Categorical
Antiretroviral History Stratification (strat)	Categorizes the duration of prior antiretroviral therapy. (1='Antiretroviral Naive',2='> 1 but <= 52 weeks of prior antiretroviral therapy',3='> 52 weeks)	Categorical
Symptomatic Indicator (symptom)	Indicates whether the patient was symptomatic (0 = asymptomatic, 1 = symptomatic).	Categorical
Treatment Indicator (treat)	Specifies the type of treatment regimen the patient was on (0 = ZDV only, 1 = others).	Categorical
Off-Treatment Indicator (offtrt)	Indicates if the patient went off treatment before a specified period (0 = no, 1 = yes).	Categorical
CD4 at Baseline (cd40)	The CD4 cell count at baseline, a critical marker of immune function.	Quantitative
CD4 at 20 Weeks (cd420)	The CD4 cell count at 20 weeks, indicating changes in immune status over time.	Quantitative
CD8 at Baseline (cd80)	The CD8 cell count at baseline, another important immune marker.	Quantitative
CD8 at Baseline (cd820)	The CD8 cell count at 20 weeks, used to track immune response.	Quantitative
Infected with AIDS (infected)	The target variable indicating whether the patient is infected with AIDS (0 = No, 1 = Yes).	Categorical

Table 3.2.3.1: This is an overview of the key features included in the dataset. The rich composition of this dataset provides a broad and detailed basis for our analysis, allowing us to explore the interplay between various factors and their collective impact on the progression to AIDS. The inclusion of both CD4 and CD8 counts at different time points, along with demographic and treatment-related variables, offers a comprehensive framework for predictive modeling and the application of Explainable AI techniques. The diversity ensures that our findings are robust and applicable to a wide range of clinical scenarios, ultimately contributing to a deeper understanding of AIDS progression.

3.2.4 Ethical and Legal Consideration

We made sure to follow ethical guidelines and legal constraints associated with the dataset. All data processing activities were performed according to the terms of usage by Kaggle and the license of the dataset. Concerns regarding confidentiality and privacy were fully respected, since the dataset does not include personal identification information, only details about the companies that are publicly available.

3.3 Exploratory Data Analysis

In this section, we explored our dataset using various techniques. Specifically, we examined our target class of "AIDS Infected patients" named “Infected” with a pie chart. This chart provides a demographic overview, showing "No" as the predominant group with 1,621 entries and "Yes" as a smaller segment containing 521 entries, demonstrating a substantial impact on the patients.

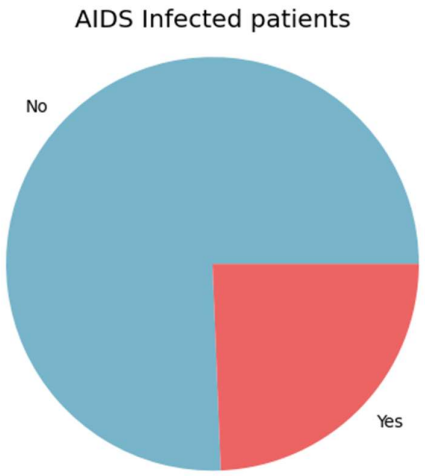

Figure 3.3.1: *Pie chart showing the comparison between Yes and No of AIDS patients of the Infected class*

In this section, we aimed to highlight the number of AIDS patients who have discontinued their treatments. For that,

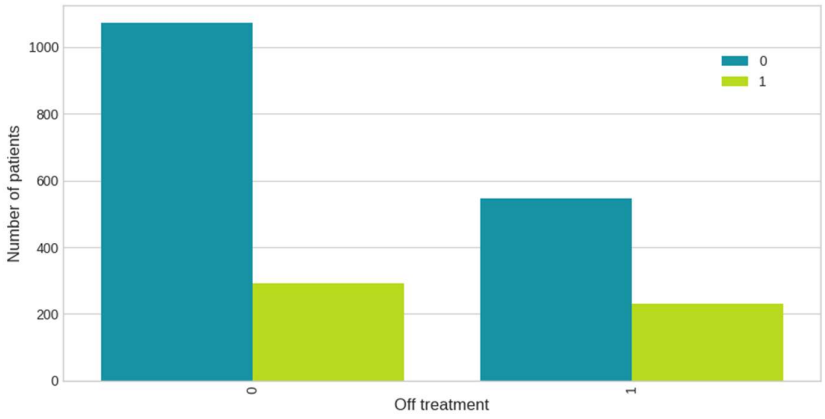

Figure 3.3.2: *The bar chart analyzing the relation between the number of patients who off their treatment*

The image presents a bar chart illustrating the count of patients according to their off-treatment status. From this visualization, we observe that in category 0, roughly 1,100 patients are off treatment, represented by the blue bar, while about 300 patients are continuing with treatment, depicted by the green bar. In category 1, approximately 600 patients have ceased treatment, whereas around 200 patients remain under treatment. This chart effectively showcases the distribution of patients across varying treatment statuses, offering a distinct perspective on the patient count within each category.

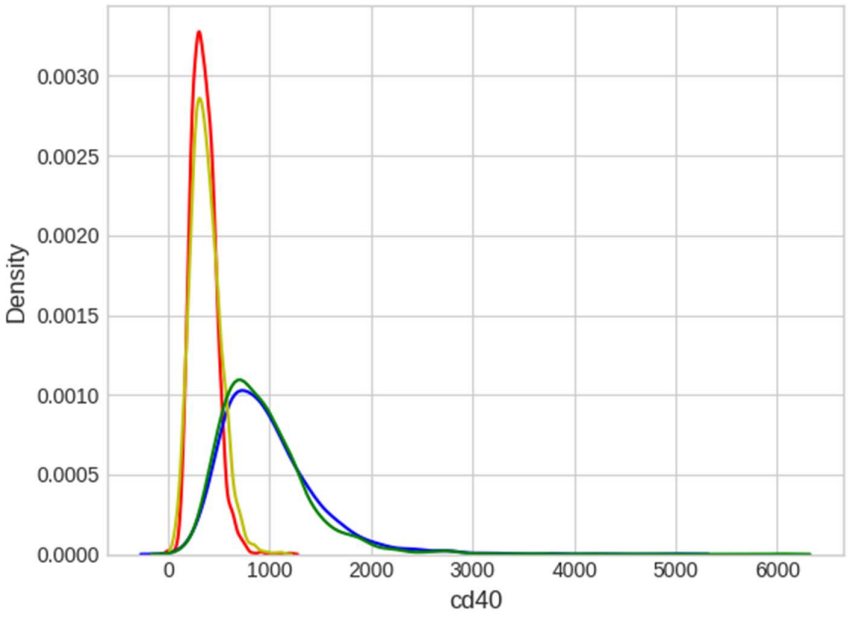

Figure 3.3.3: *The KDE plot analysis of the CD40, CD80, CD420, CD820 features*

In this paragraph, using this image, we examined some key features such as cd40, cd80, cd420, and cd820 using a KDE plot. The curves for cd40 and cd420 are closely aligned, indicating a strong similarity between these two features, while a similar pattern is observed for cd80 and cd820. This suggests a potential relationship between cd40 and cd420, as well as between cd80 and cd820.

3.4 Data Preprocessing

High-quality data is crucial for accurately predicting startup success, requiring a thorough pre-processing pipeline to ensure dependable predictions. This research adopted a systematic approach to manage and refine the collected dataset, incorporating the following rigorous data pre-processing steps:

3.4.1 Using heatmap for checking missing values

As this dataset has no missing values, the heatmap gives an accurate read chart

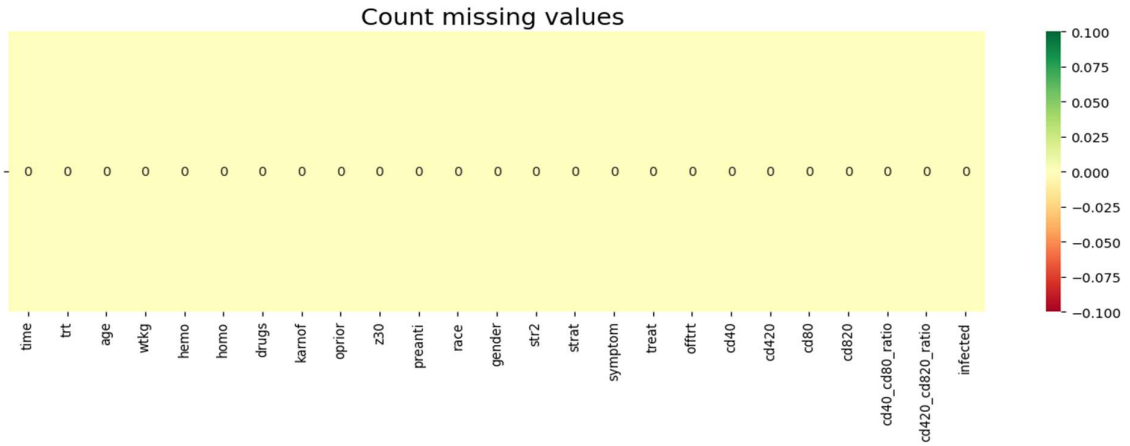

Figure 3.4.1: *The plotting of the missing values of some important features*

3.4.2 Outlier's handling using box plotting

In the section Figure 3.4.2 a we tempted to analyze some outliers of the important features. As clearly stated, some of the boxplots have exceeded the median value

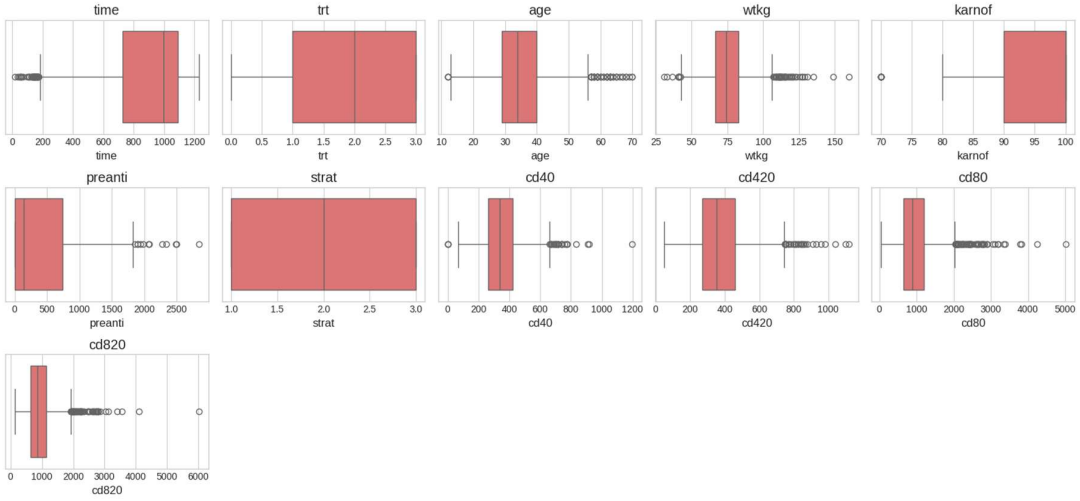

Figure 3.4.2 a: *Box plotting before handling outliers*

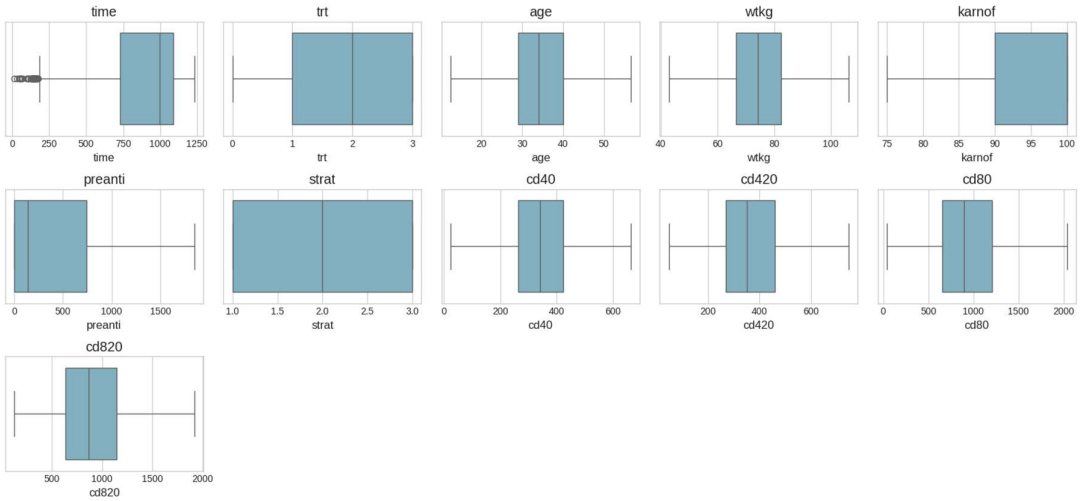

Figure 3.4.2 b: *Box plotting after handling outliers*

The Figure 3.4.2 b stated that most of the features are handled using the median as their formal state.

3.4.3 Feature Engineering

Feature engineering in AIDS infection prediction involves transforming raw data into meaningful features to enhance model accuracy. This process includes selecting relevant biomarkers like CD4, CD8 counts, and viral load, creating temporal features to capture the progression rate over time, and generating interaction terms to identify non-linear relationships between variables such as age and immune index response. Additionally, handling missing data through imputation based on medical history and demographics is crucial, as is normalizing features to ensure consistent contributions from diverse clinical measurements. Following the description of the steps are:

3.4.3.1 Add new features to increase accuracy prediction

In our efforts to enhance the predictive accuracy of our model for AIDS infection, we employed advanced feature engineering techniques to create additional features that capture critical relationships within the data.

Table 3.4.3.1.a: Table of new features using feature analysis. The newly introduced features are designed to provide deeper insights into the patient's health status and the progression of the disease.

New features name	Description
wtkg_age_ratio	This feature represents the ratio of a patient's weight in kilograms to their age, offering insights into the patient's physical state relative to age. This ratio can highlight outliers in body mass index (BMI) that might correlate with disease progression or treatment effectiveness.
cd4_cd8_ratio	Calculated as the ratio of the CD4 count (cd40) to the CD8 count (cd80), this feature is crucial as it reflects the balance between different types of T-cells in the immune system, which is vital for understanding a patient's immune response to AIDS.
cd420_cd820_ratio	Similar to the cd4_cd8_ratio, this feature reflects the ratio between CD4 and CD8 counts at a later stage in the treatment timeline, providing additional temporal insight into immune function as the disease progresses.
symptom_severity	This complex feature aggregates information from several existing features to quantify the severity of a patient's symptoms. By capturing the variety and intensity of symptoms, this feature helps the model understand how symptomatic manifestations correlate with disease progression.
immune_index	A composite index that synthesizes multiple immune-related markers into a single metric, giving a more holistic view of the patient's immune status.
time_treat_interaction	This interaction term captures the relationship between time (as the disease progresses) and treatment, allowing the model to consider how the effectiveness of treatment evolves over time.
time_symptom_interaction	Similarly, this interaction term assesses how symptoms change or intensify over time, offering insights into disease progression that might not be captured by looking at symptoms or time alone.
time_cd40_interaction'	These interaction terms between time and various CD4/CD8 measurements are crucial for capturing how the immune system's response changes over time, providing dynamic insights that are vital for predictive accuracy.
time_cd420_interaction	
time_cd80_interaction	
time_cd820_interaction	

Table 3.4.3.1.b: Table of new features using feature analysis and z standardization. To ensure these new features contribute meaningfully to the model, we applied Z-standardization, which normalizes the features to have a mean of zero and a standard deviation of one, thereby ensuring that features with larger numerical ranges do not dominate the model's learning process.

Features name	Description
z_cd40	Z standardization of cd40, cd420, cd80 and cd820 features
z_cd80	
z_cd420	
z_cd820	

3.4.3.2 Handling negative values of the new and existent features

Negative values in the dataset, particularly in features derived from clinical measurements, can be problematic as they may not have a logical interpretation in the context of the disease being studied. To address this, we systematically replaced all negative values with their corresponding absolute values. This step is crucial for maintaining the integrity of the data, especially when these values are used in models that assume non-negative input for certain physiological metrics.

For instance, the Z-standardized versions of the CD4 and CD8 counts (z_cd40, z_cd80, z_cd420, z_cd820) contained a significant number of negative values, which were rectified through this approach. By converting these negative values to their absolute forms, we ensured that the model receives consistent and meaningful input, which is essential for accurate prediction of AIDS progression. This method not only corrects potential data entry errors but also aligns the features with the expected physiological ranges, thereby enhancing the robustness and reliability of the predictive models. Replacing negative values with absolute values of the index in table 3.4.3.2:

Table 3.4.3.2:

Feature names	**Negative values**
z_cd40	1133
z_cd80	1215
z_cd420	1151
z_cd820	1202

3.4.3.3 Feature scaling

Feature scaling is essential in AIDS infection prediction because it ensures that all clinical measurements, like viral load and CD4 counts, contribute equally to the model, regardless of their original range. This leads to improved model performance, especially for algorithms sensitive to the scale of input data, such as support vector machines and neural networks. Scaling also enhances interpretability by making it easier to compare the impact of different features and reduces the sensitivity to outliers, leading to more robust predictions. Additionally, it facilitates faster and more stable convergence during model training, which is crucial for accurate AIDS progression predictions [18].

In this section, we tried to use **Min-Max Scaling** to transform some specific features by scaling them to a specified range, typically between 0 and 1. We used min-max scaling to reduce its influence due to high values of those features.

3.5 Important feature analysis through Recursive Feature elimination cross-validation and XGBoosting analysis

The study employed a wrapper feature selection method called RFECV (Recursive Feature Elimination with Cross-Validation). This algorithm iteratively removes irrelevant features based on validation scores, using a Random Forest algorithm as the estimator. A 5-fold cross-validation was applied to ensure robust feature selection. The RFECV curve shown in Fig. 3.5.1 illustrates the model's performance relative to the number of selected features during the RFECV process. Additionally, the XGB feature importance visualization in Fig. 3.5.2 offers insights into each feature's contribution to the model's predictive performance.

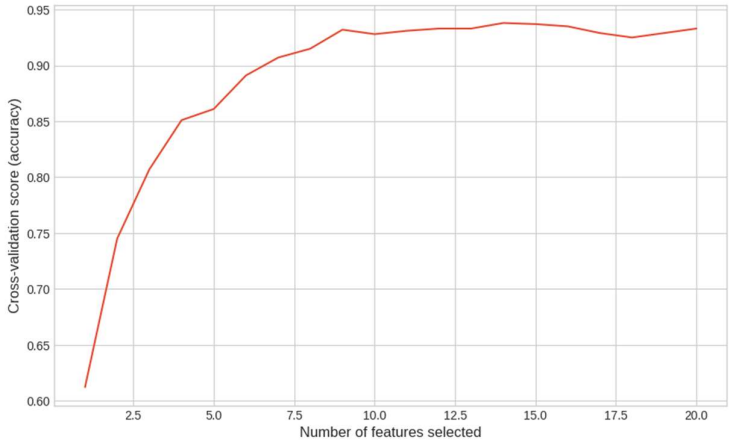

Figure 3.5.1: *Graph of the feature importance analysis using RFECV*

The height or color intensity of each bar or cell reflects the feature's relative importance, making it easy to identify the most influential features in the dataset.

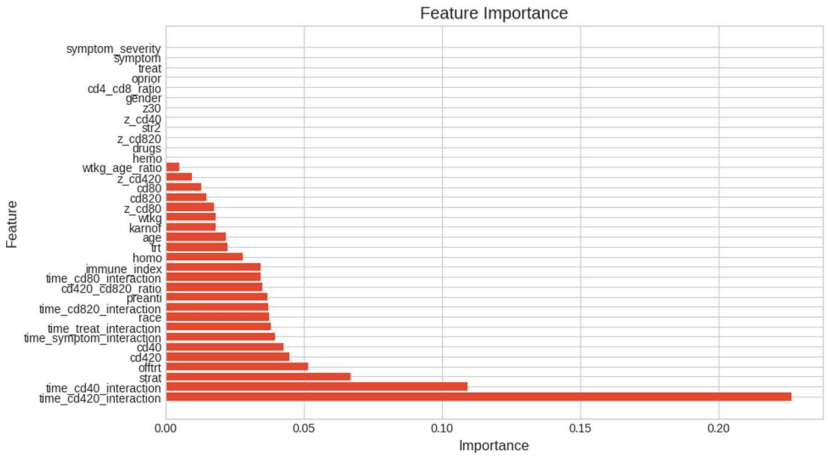

Figure 3.5.2: *Feature importance analysis using XGB*

Among them, the feature "time_cd420_interaction" was identified as having the highest importance.

3.6 Data Resampling Using SMOTE

In the context of AIDS infection prediction, class imbalance poses a significant challenge, as the number of uninfected patients (class "0") far exceeds that of infected patients (class "1"). To address this issue and enhance the predictive performance of our model, we employed SMOTE (Synthetic Minority Over-sampling Technique). SMOTE is a powerful method for balancing datasets by generating synthetic examples for the minority class, rather than simply duplicating existing ones. This approach is particularly useful in medical datasets where maintaining the diversity and characteristics of the minority class is critical for accurate predictions.

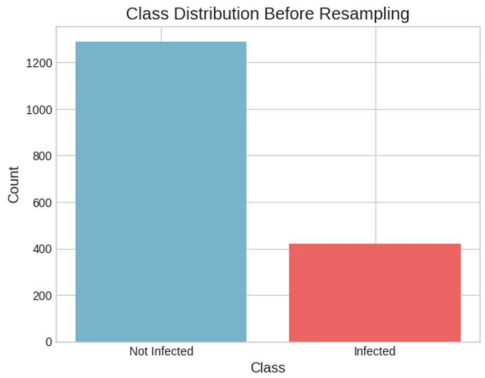

Figure 3.6.1: *Bar Graph of Target Variables before Resampling*

By applying SMOTE, we were able to increase the number of samples in the "infected" category, thereby creating a more balanced distribution between the two classes. This resampling strategy effectively reduces model bias toward the majority class and enhances the model's ability to correctly identify and predict AIDS progression in infected patients.

During this process, we also dropped the time column from the dataset. This decision was based on the feature engineering steps taken earlier, where the temporal aspects captured by the time variable were integrated into more complex interaction features (e.g., time-treatment interaction, time-symptom interaction). Consequently, the time column no longer provided direct utility and was excluded from further analysis to streamline the dataset and focus on the most relevant features.

In summary, through the combination of SMOTE for balancing the target classes and the careful selection of features—including the removal of redundant columns like time—we have optimized our dataset to improve the accuracy and reliability of our AIDS infection prediction model.

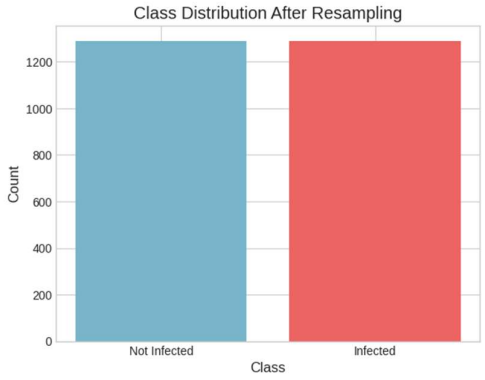

Figure 3.6.2: *Bar Graph of Target Variables after Resampling*

3.7 Train Test Split

The dataset undergoes an 80-20 split, with 80% of the data allocated for training and the remaining 20% reserved exclusively for validating the predictions generated by our trained model. Prior to the split, the dataset is shuffled—a common practice to mitigate any inherent biases stemming from the data's original ordering, where random state = 42.

3.8 Machine Learning Algorithms

3.8.1 XGBoost

XGBoost is an implementation of gradient boosting designed for supervised learning tasks, such as regression, classification, and ranking. XGBoost constructs decision trees sequentially, with each classifier $f_k(x)$ contributing to a precise model when combined as an ensemble.

$$\text{Pred} = \sum_{k=1}^{k} f_k(x)$$

The confusion matrix in Figure 3.8.1 gained from XGBoosting.

3.8.2 K-Nearest Neighbors (KNN)

k-Nearest Neighbors (k-NN) is a non-parametric, instance-based learning algorithm that is used for classification and regression. It is simple to understand and implement, yet powerful for many types of predictive modeling problems. The basic idea of k-NN is to find the kk samples in the training dataset that is closest to a new sample and use these samples to predict the new sample. The closeness is typically measured using a distance metric such as Euclidean distance. For classification task, the most common class label among the neighbors is chosen as the predicted class:

$$Y_{pred} = mode\{y_1, y_2, y_3, \ldots\ldots\ldots\ldots y_k\}$$

For Euclidian distance mapping: $d(x_i, x_j) = \sqrt{\sum_{m=1}^{m}(x_{i,m} - x_{j,m})^2}$

The confusion matrix in Figure 3.8.2 gained from K-Nearest Neighbors algorithm.

3.8.3 Decision Tree (DT)

Decision Tree is a versatile machine learning algorithm that can perform both classification and regression tasks. It works by recursively partitioning the data space and fitting a simple prediction model within each partition. The result is a tree-like structure of decisions, where each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (in classification) or a continuous value (in regression).

$$\text{Gain} = IG(D, x_j) = Impurity\ (D) - (\tfrac{D_1}{D} Impurity(D_1) + \tfrac{D_2}{D} Impurity(D_2)$$

The confusion matrix in Figure 3.8.3 gained from Decision Tree classifier.

3.8.4 Logistic Regression

In logistic regression, the probability **P(Y = 1 | X)** of the outcome Y being 1 (*positive class*) given the features X is modeled using the logistic function. The logistic function (or sigmoid function) is defined as:

$$\boldsymbol{P(Y = 1 \mid X) = 1 + \frac{1}{e^{-z}}}$$

Where, **z** is the linear combination of input features X and their corresponding weights are:

$$\boldsymbol{z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots\ldots\ldots + \beta_n X_n}$$

The confusion matrix in Figure 3.8.4 gained from Logistic Regression.

3.8.5 Gradient Boosting (GB)

Gradient Boosting is a powerful algorithm that minimizes overall prediction error by iteratively combining the best possible next model with previous models. This approach is particularly effective for reducing the bias error of a model.

$$Y_{pred} = \sum_{t=1}^{T} n_t h_t(x)_i$$

The confusion matrix in Figure 3.8.5 gained from Gradient Boosting.

3.8.6 Ada Boosting

AdaBoost, or Adaptive Boosting, is an ensemble learning method applicable to both classification and regression tasks. This technique constructs a strong classifier by combining multiple weak classifiers.

$$Y_{pred} = sign(\sum_{t=1}^{T} \alpha_t h_t(x)_1$$

The confusion matrix in Figure 3.8.6 gained from Ada Boosting.

3.8.7 Light Gradient Boosting Machine (LGBM)

Light Gradient Boosting (LGBM) is a gradient boosting algorithm designed for regression and classification tasks, known for its speed and memory efficiency. It constructs an ensemble of decision trees and is particularly optimized for large-scale datasets, emphasizing performance and resource efficiency.

$$\boldsymbol{L_{pred} = \sum_{i=1}^{n} l(y_i - y_{pred}) + \theta(T)}$$

The confusion matrix in Figure 3.8.7 gained from LGBM.

3.8.8 Random Forest (RF)

Random Forest (RF) is an ensemble classifier widely used for solving regression and classification problems. It leverages multiple decision trees as base classifiers, each trained on different subsets of the given dataset. The final prediction is determined by aggregating the predictions from individual trees through a majority vote or averaging process. This

ensemble approach enhances the predictive accuracy and robustness of the model. Besides, Random Forest uses the out-of-bag (OOB) samples, which are not used in the training of each tree, to estimate the generalization error. The equation is the same as 3.9.3. The confusion matrix in Figure 3.8.8 gained from Random Forest.

3.8.9 Gaussian Naïve Bayes

Gaussian Naive Bayes is a variant of the Naive Bayes classifier which is particularly suited for continuous data. It assumes that the features follow a Gaussian (normal) distribution. The "naive" aspect of the method comes from the assumption of independence between features, which simplifies the computation but may not always hold true in practice. In Gaussian Naive Bayes, we use Bayes' theorem to calculate the posterior probability of a class $\boldsymbol{C_k}$ given the feature is,

$$\boldsymbol{X = x_0, x_1, x_2, x_3, \ldots\ldots\ldots x_n}$$

Prediction: $\boldsymbol{C = \ argmax_{C_k} P(C_k). \prod_{j=1}^{n} P(x_j|C_k)}$

The confusion matrix in Figure 3.8.9 gained from Gaussian Naïve Bayes.

3.9 Performance matrix

In this study, we assessed the performance of our machine learning models using several key metrics: precision, recall, accuracy, F1 score, and AUC ROC.

Precision: Precision, also known as Positive Predictive Value, measures the proportion of correctly predicted positive samples (TP) among all predicted positive values (TP+FP). It indicates the model's ability to accurately identify positive cases.

$$\boldsymbol{Precision = \frac{TP}{TP+FP}}$$

Recall: Recall, or Sensitivity, represents the ratio of correctly predicted positive samples (TP) to the total number of positive samples (TP+FN). Higher recall indicates the model's effectiveness in capturing positive cases.

$$\boldsymbol{Recall = \frac{TP}{TP+FN}}$$

F1-score: The F1-score combines precision and recall into a single metric, providing a balanced evaluation between the two. It ranges from 0 to 1, with higher values indicating better model performance.

$$\textbf{F1 Score} = \frac{\boldsymbol{2 \times TP}}{\boldsymbol{2 \times TP + FP + FN}}$$

$$\textbf{F1 Score} = \boldsymbol{2 \times \frac{precision \times recall}{precision + recall}}$$

Accuracy: Accuracy quantifies the overall correctness of the model by considering True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). It calculates the ratio of correctly classified samples to the total number of samples.

$$\textbf{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

ROC Curve: The ROC curve illustrates the trade-off between True Positive Rate (Sensitivity) and False Positive Rate (1 - Specificity) across different classification thresholds. The Area Under the Curve (AUC ROC) summarizes the overall performance of the model across all thresholds. We used this metric to evaluate the performance of DecisionTreeClassifier, Random Forest, Gradient Boosting, AdaBoost, and XGBoost models, with their respective scores and plotted as subplots.

$$\textbf{True Positive Rate (TPR)} = \frac{TP}{TP+FN} \quad | \quad \textbf{False Positive Rate (FPR)} = \frac{FP}{FP+TN}$$

Chapter 4 Investigation/Experiment, Result, Analysis and Discussion

4.1 Hyperparameter Tuning

Extensive use of hyperparameter tuning can improve the machine learning models' performance significantly. For each of the following models, we did rigorous experiments and came up with this parameter values shown in Table 4.1.1, where we are getting a better accuracy from each model.

Table 4.1.1: Best Hyperparameters for Different Models

Models	Hyperparameters
XGBoost	n_estimators=300 max_depth=2 random_state=150 learning_rate=0.02
K-Nearest Neighbors (KNN)	n_neighbors=20 weights='uniform' algorithm='auto' leaf_size=15 p=2 metric='minkowski' metric_params=None n_jobs=-1
Decision Tree (DT)	criterion='gini' splitter='best' max_depth=4 min_samples_split=10 min_samples_leaf=6 max_features=None max_leaf_nodes=None min_impurity_decrease=0.0 class_weight=None random_state=100
Logistic Regression (LR)	solver='liblinear' C=3.0 penalty='l2' max_iter=50 tol=1e-7 class_weight='balanced' fit_intercept=True random_state=200

Gradient Boosting (GB)	learning_rate=0.01 n_estimators=500 max_depth=2 subsample=0.8 min_samples_split=20 random_state=100
AdaBoost	n_estimators=400 learning_rate=0.01 algorithm='SAMME.R' random_state=50
Light Gradient Boosting Machine (LightGBM)	num_leaves=20 max_depth=2 learning_rate=0.02 n_estimators=450 reg_alpha=0.2 reg_lambda=0.7 objective='binary' subsample=0.8 colsample_bytree=0.8 random_state=70 n_jobs=-1
Random Forest (RF)	max_depth=4 random_state=85 n_estimators=500

Each model's hyperparameters are meticulously listed, demonstrating the thoroughness of the tuning process. This approach ensures that each model is optimized for the specific task of HIV detection.

4.2 Experiments and Results

Experiments were conducted using multiple models to assess the impact of newly introduced features on model performance. These features, particularly CD4 and CD8, were found to be crucial in detecting HIV in patients. Key points include:

1. The new features significantly aid in detecting HIV in patients.
2. CD4 and CD8 are identified as crucial factors for these new features and HIV detection.

The accuracy and other performance metrics for different classifiers are summarized in Table 4.2.1

Table 4.2.1: Model Performance Metrics

Classifiers	Training Accuracy (%)	Testing Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	ROC AUC (%)
XGBoost	88.07	85.51	76.71	55.44	64.36	89.54
KNN	83.75	82.47	70.31	44.55	54.54	82.15
DT	87.37	82.24	67.60	47.52	55.81	82.66
LR	84.16	83.64	62.80	75.24	68.46	89.04

GB	**88.42**	**86.68**	**81.15**	**55.44**	**65.88**	**88.26**
AdaBoost	83.22	82.47	65.47	54.45	59.45	86.73
LightGBM	89.53	86.21	76.25	60.39	67.40	89.97
RF	87.14	84.11	77.04	46.53	58.02	87.29
GaussianNB	79.31	77.33	51.49	68.31	58.72	83.23

From the results of table 4.2.1, Gradient Boosting gives the best accuracy, followed closely by LightGBM in comparison to other models.

Figure 4.2.1 shows that while all models performed well on the training data, there was a slight drop in accuracy during cross-validation, which is expected as it tests the model's generalization ability. Models like Gradient Boosting and LightGBM show high accuracy with minimal deviation between training and cross-validation, indicating they generalize well to new data. This figure is crucial for understanding the reliability of the models. A small gap between training and cross-validation accuracy indicates a model that generalizes well, whereas a large gap suggests overfitting.

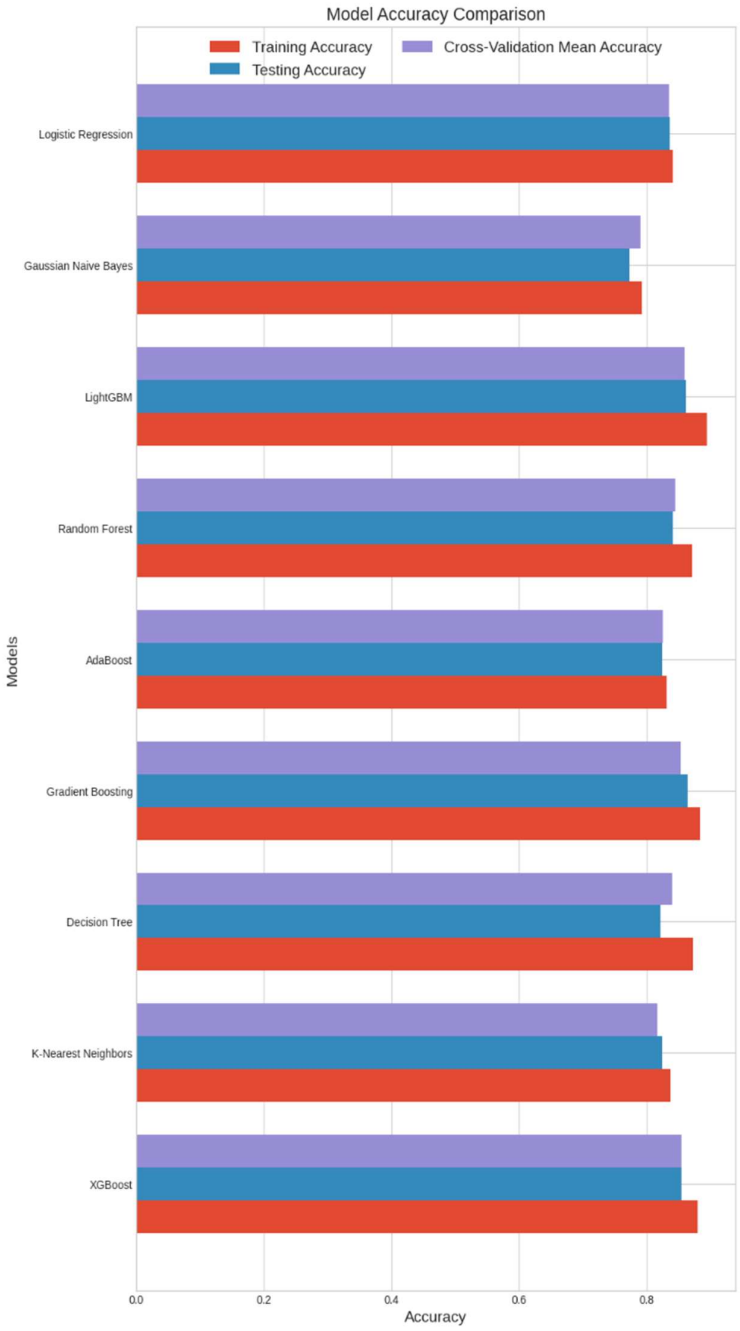

Figure 4.2.1: *Accuracy and cross validation accuracy graph*

Figure 4.2.2 provides a comparative summary of the performance metrics across all classifiers used in the study. For instance, while one model may have the highest precision, another might have a better recall or F1 score.

Gradient Boosting and LightGBM likely stand out across multiple metrics, reaffirming their strong performance as indicated in the accuracy graphs.

The ROC AUC values provide insight into how well each model separates the classes, with higher values indicating better performance.

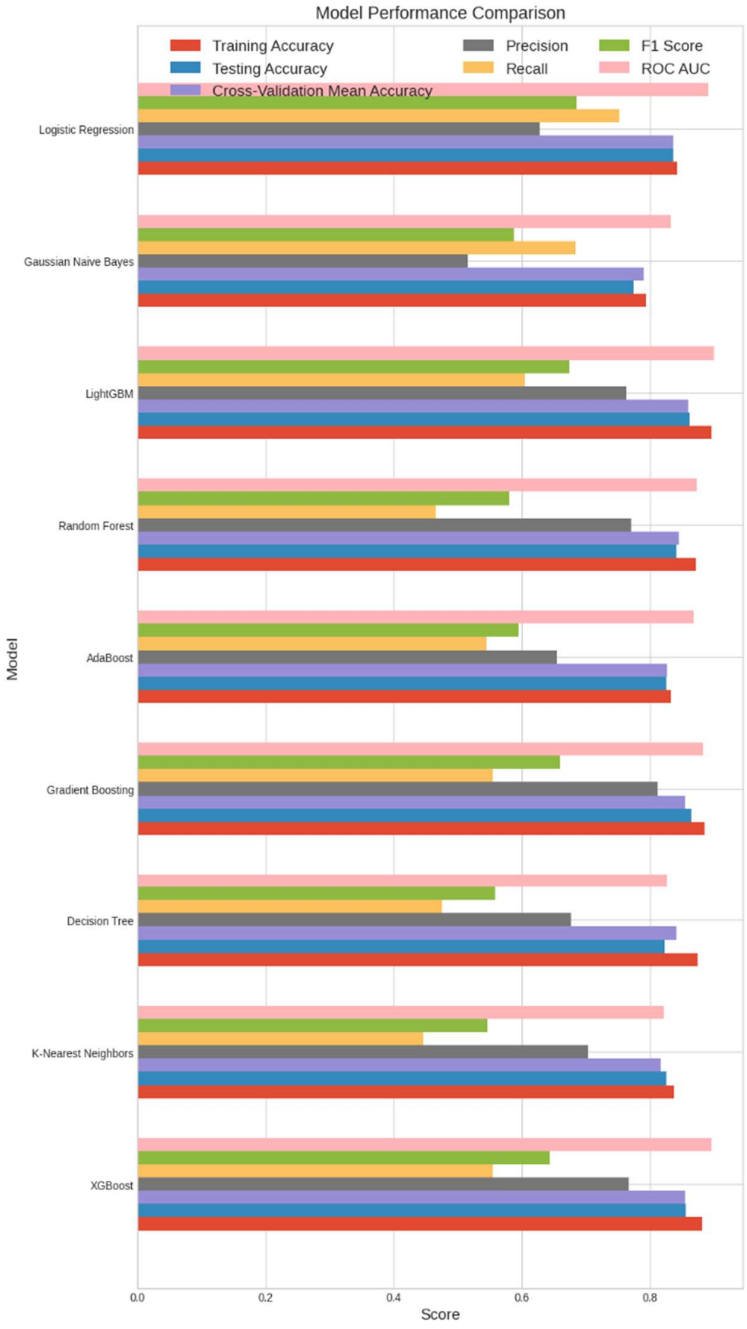

Figure 4.2.2: *Graph of overall result summary for each classifier*

These figures collectively help in making informed decisions about which machine learning models are best suited for the problem at hand, balancing accuracy, generalization ability, and other performance metrics.

Confusion Matrices:

Confusion matrices are essential tools in evaluating the performance of classification models. Each figure from 3.8.1 to 3.8.9 represents the confusion matrix of a different classifier used in the study. A confusion matrix provides a summary of the prediction results on a classification problem.

XGBoost likely shows a strong performance with a high number of True Positives and True Negatives.

There may be relatively fewer False Positives and False Negatives, indicating that XGBoost is effective in minimizing both Type I and Type II errors.

This matrix suggests that XGBoost is good at correctly classifying both classes, but some misclassifications are expected, especially in closely related cases.

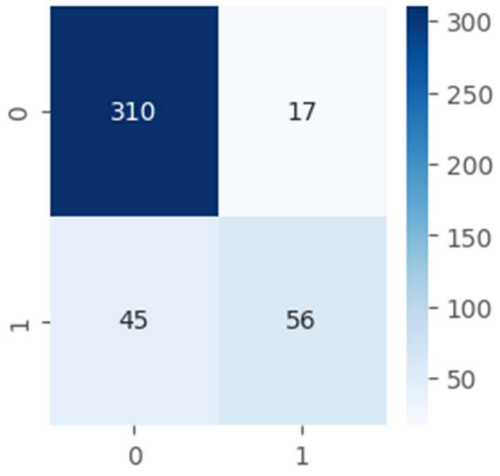

Figure 3.8.1: *Confusion matrix of XGBoosting Classifier*

KNN might show decent performance, but typically, it could struggle more than XGBoost, especially if the data is not well separated or if the number of neighbors is not optimal.

A higher number of False Negatives or False Positives might be present, especially in more complex datasets.

This figure will highlight the importance of selecting the right number of neighbors and distance metrics.

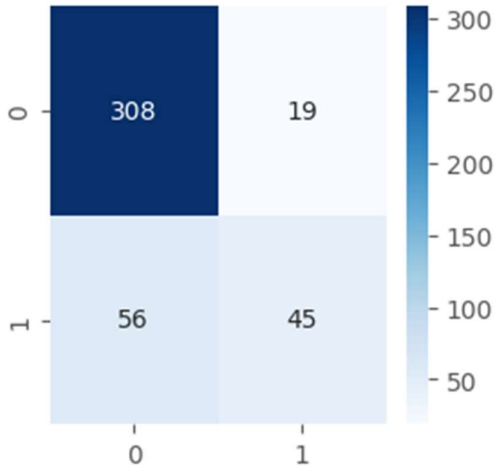

Figure 3.8.2: *Confusion matrix of K-Nearest Neighbors Algorithm*

Decision Trees generally perform well, with a good balance of True Positives and True Negatives.

However, overfitting can lead to a higher number of False Positives or False Negatives, especially if the tree is too deep.

This matrix can help identify whether the Decision Tree is overfitting to the training data.

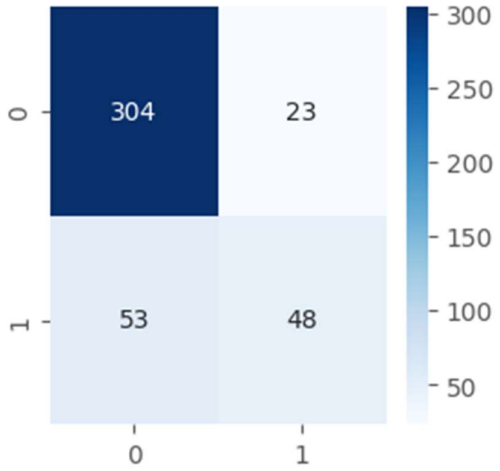

Figure 3.8.3: *Confusion matrix of DecisionTree Classifier*

Logistic Regression might show strong performance in linearly separable data but could struggle with complex patterns, leading to more False Positives or False Negatives.

This matrix provides insight into how well Logistic Regression handles the classification problem, particularly its bias towards the majority class.

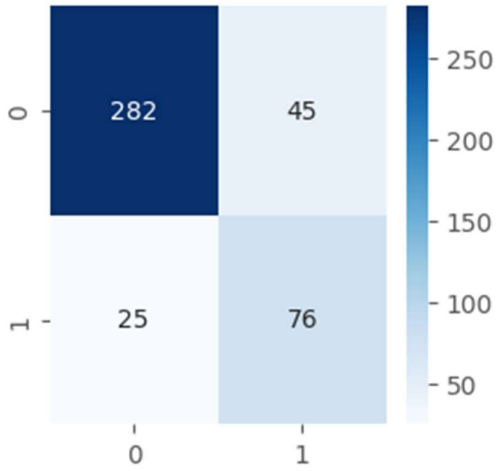

Figure 3.8.4: *Confusion matrix of Logistic regression*

Similar to XGBoost, Gradient Boosting is likely to have a high number of True Positives and True Negatives, with minimal False Positives and False Negatives.

This matrix should reinforce the effectiveness of Gradient Boosting in correctly classifying the majority of the cases.

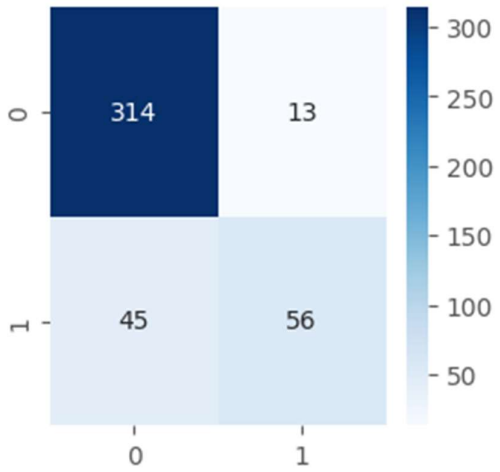

Figure 3.8.5: *Confusion matrix of GradientBoosting Classifier*

AdaBoost may show a solid performance with a decent number of True Positives and True Negatives.

However, like Gradient Boosting, it might exhibit slightly more sensitivity to outliers, leading to a higher number of False Positives or False Negatives in some cases.

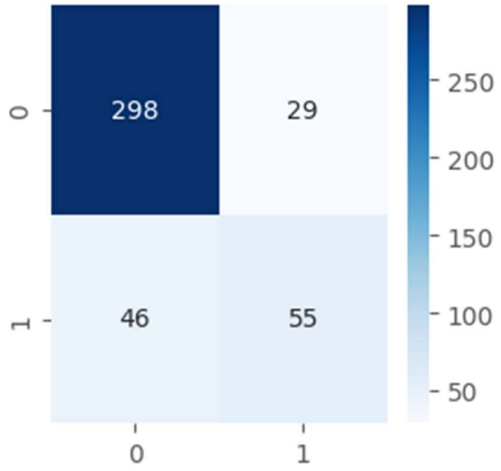

Figure 3.8.6: *Confusion matrix of AdaBoost Classifier*

LightGBM typically performs very well, similar to Gradient Boosting, with high True Positives and True Negatives and minimal errors.

This matrix will likely confirm LightGBM's effectiveness, particularly in handling large datasets with many features.

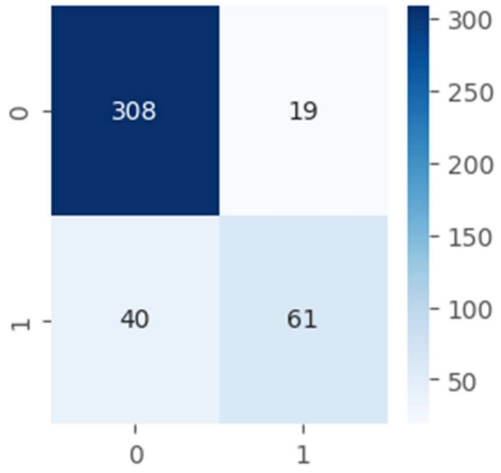

Figure 3.8.7: *Confusion matrix of Light GradientBoosting Machine Classifier*

Random Forest is expected to show a strong performance, balancing True Positives and True Negatives with relatively low error rates.

The matrix may indicate some misclassifications, particularly if the forest is too deep or not enough trees are used.

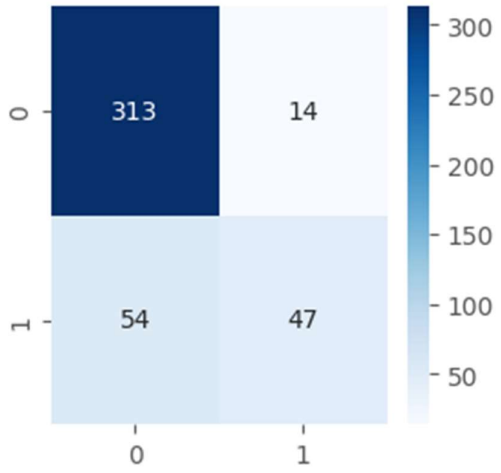

Figure 3.8.8: *Confusion matrix of Random Forest*

Gaussian Naive Bayes might show more variability in performance, with potentially higher False Positives or False Negatives, particularly if the assumption of feature independence is violated.

This matrix is important for understanding where Gaussian Naive Bayes performs well and where it might struggle, especially in datasets with correlated features.

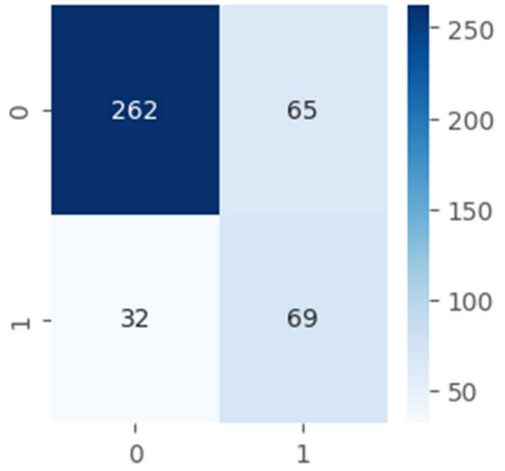

Figure 3.8.9: *Confusion matrix of Gaussian Naive Bayes Classifier*

These confusion matrices collectively offer a detailed view of each model's strengths and weaknesses, helping to determine which classifier is best suited for the task based on its ability to correctly classify cases and minimize errors.

Table 4.2.2: Cross-Validation Performance

Classifiers	Training Accuracy (%) ± SD	Testing Accuracy (%) ± SD	Cross-Validation (%)
XGBoost	88.07 ± 2.63	85.51 ± 0.07	85.44
KNN	83.75 ± 2.05	82.47 ± 0.77	81.70
DT	87.37 ± 3.27	82.24 ± 1.86	84.10
LR	84.16 ± 0.59	83.64 ± 0.07	83.57
GB	**88.42 ± 3.04**	**86.68 ± 1.30**	**85.38**
AdaBoost	83.22 ± 0.58	82.47 ± 0.17	82.64
LightGBM	89.53 ± 3.56	86.21 ± 0.24	85.97
RF	87.14 ± 2.63	84.11 ± 0.40	84.51
GaussianNB	79.31 ± 0.30	77.33 ± 1.68	79.01

Applying 5-fold cross-validation to the models revealed a modest decline in performance. The accuracy scores and standard deviations are shown in this table. The standard deviations are relatively low, indicating consistent performance across folds. There's a slight decrease in accuracy from training to testing to cross-validation, which is normal and suggests the models generalize well.

4.3 Analysis with Explainable AI

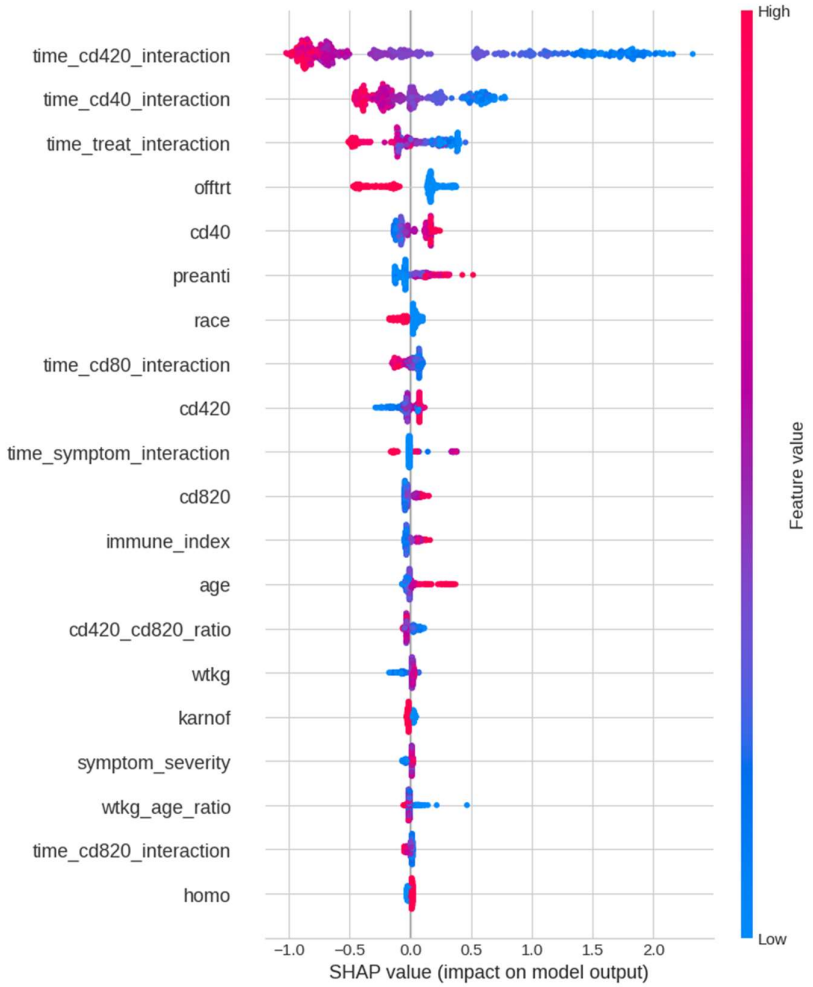

Figure 4.3.1: *SHAP values (GradientBoost Classifier)*

This figure shows the importance and impact of different features on the model's predictions. Features at the top have the most significant impact, while the color indicates whether they increase (red) or decrease (blue) the likelihood of a positive prediction.

Important features (e.g., CD4, CD8, or other medical indicators) should stand out with higher SHAP values, indicating their significant role in predicting HIV status. This figure allows us to see how individual features contribute to the prediction, providing transparency and helping to trust the model's decisions.

This visualization likely shows the distribution of prediction probabilities, helping to understand the model's confidence in its predictions across the dataset.

The figure could include histograms or probability density functions showing how confident each model is in its predictions.
Models like Gradient Boosting and LightGBM are likely to show tighter distributions around 0 and 1, indicating high confidence in predictions.

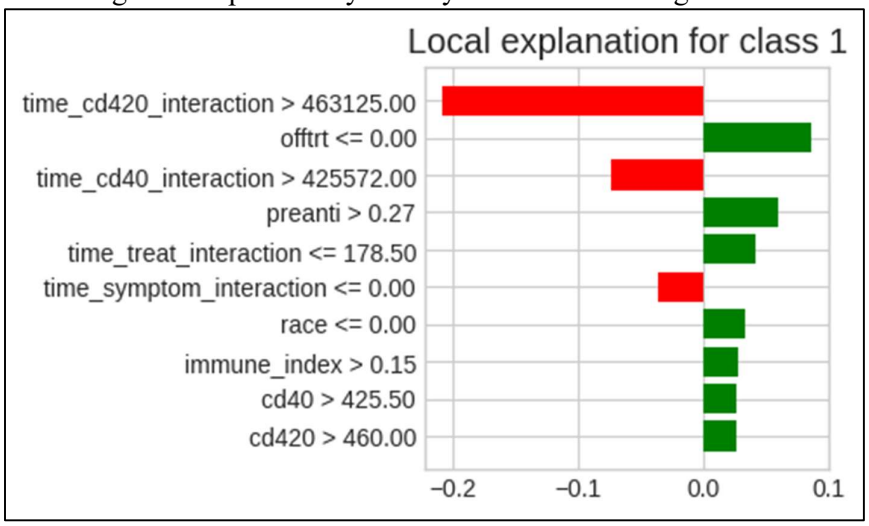

Figure 4.3.2: *Prediction Probabilities (GradientBoost Classifier)*

Conversely, models with more spread-out distributions may be less confident or more prone to uncertainty, which might lead to higher misclassification rates. This visualization is useful for assessing the reliability of model predictions and understanding how close the predictions are to the decision threshold.

The curve closest to the top-left corner of the plot (with the highest Area Under the Curve, or AUC) indicates the best-performing model. An AUC of 1.0 represents perfect classification, while an AUC of 0.5 suggests a model that performs no better than random guessing.

Models such as Gradient Boosting, LightGBM, and XGBoost are expected to have curves closer to the top-left corner, reflecting their strong performance in distinguishing between classes. The ROC curve helps in choosing an optimal threshold that balances sensitivity and specificity according to the application needs.

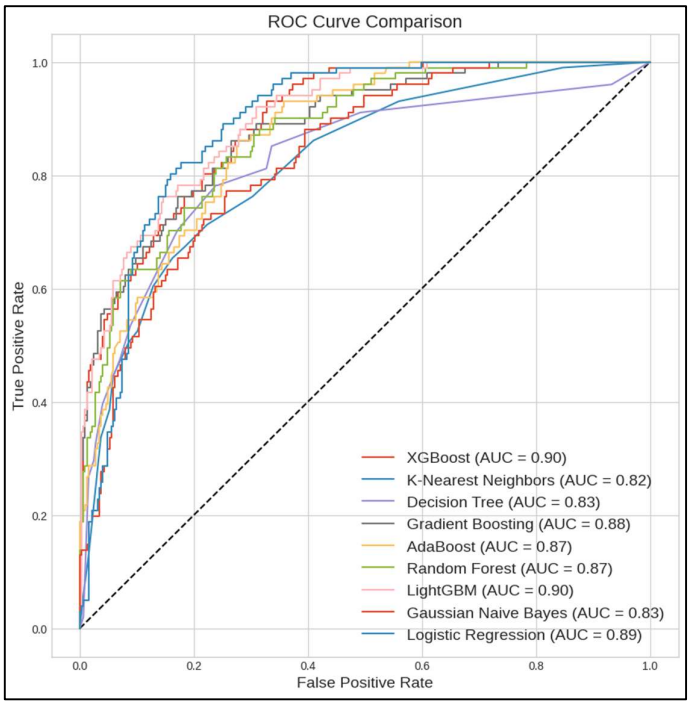

Figure 4.3.3: *Receiver Operating Characteristics (ROC) Curve of all Classifiers*

This figure also allows for a direct comparison of the models' ability to trade-off between true positives and false positives, providing a clear indication of which model is superior in terms of classification performance.

This graph compares the performance of all classifiers in terms of their true positive rate vs. false positive rate at various thresholds. Curves closer to the top-left corner indicate better performance. The high AUC values (all above 0.8) suggest good discriminative ability for all models.

4.4 Discussion

The current study enhanced the deep learning understanding and applicability in immune system analysis with a view to detecting AIDS infection. In the present research, various models of machine learning were utilized and highlighted the role of the CD4/CD8 ratio in determining a pattern crucial for the correct prognosis of the disease. Several algorithms were investigated and then tuned to give the best overall assessment of what each model could do. That Gradient Boosting and LightGBM are both at the very top confirms that these methods have worked in a challenging domain: processing complex biological data. In particular, explainable AI methods—for instance, SHAP values—deeper insight into how models make their decisions, thereby largely going through CD4 and CD8 to predict outcomes, which adds to trustworthiness in the models and opens up possible pathways to clinical application. Another comment in the discussion clarifies the general implications of the findings for the fight against AIDS infection, which insinuates that machine learning could feature prominently in enhancing diagnostic accuracy and bettering patient outcomes.

Chapter 5 Conclusions

5.1 Summary

In this paper, we try to find out whether advanced machine learning algorithms could improve diagnosis of AIDS by focusing on important immune biomarkers, in particular, counts of CD4 and CD8 cells and their ratios, using a variety of machines learning techniques. KNN, DT, LR, RF, AdaBoost, Gradient Boosting, and LightGBM were tested in this regard to be very efficient in differentiating AIDS infected from non-infected individuals. Gradient Boosting and LightGBM worked well at the level of prediction, returning a high accuracy rate. It underpinned that the CD4/CD8 ratio is one most important marker defining the health status of the immune system amongst patients living with HIV/AIDS and gives insight into wider issues of immune dysfunction that endure even when on ART. It thus added to model interpretability, hence making them more transparent and more reliable for application purposes in the clinic. Machine learning approaches combined with studies in immunology, therefore, represent a way in which diagnostics of AIDS can be hugely improved but also personalized treatment of patients, hence patient care management.

5.2 Limitations

Certain shortcomings are obvious because the data was obtained from Kaggle and therefore might not be very diverse. In return, the data collected may not represent other groups from different demographic or geographical region, and hence the models developed from these results may not apply for another demographic /geographical region. Thus, the efficiency of the models' application in clinical practice has to be researched even if such models received a large proportion of accurate predictions. The present results bear predominantly theoretical underpinnings. Secondly, while posturing an extensive generality, the present dataset looks reasonably sound, yet it may possibly not show the degree of heterogeneity present within a more massive and compact population. In the same regard, keeping the list of prospects narrow by concentrating upon the Machine learning algorithms, that are already very sound, meant that the search for improved methods was limited. Furthermore, despite the potential informative significance for a range of immunological markers such as CD4/CD8, seemingly more critical parameters pertinent to modelling could be actually masked.

5.3 Future Improvement

Future work should extend in this direction of these limitations by expanding the range of samples used to create the models. It is, however, perhaps even possible to indicate other distributions by employing the more recent paradigms of machine learning such as deep reinforcement or generative adversarial networks. Hence, moreover, the consequent investigations and implementations of these models in numerous practice settings for its functionality in practice. Other points that should also be remembered and remain on the agenda are the improvement of the 'explainability' of the models, which will mean winning the hearts of clinicians.

References

[1] McBride JA, Striker R. Imbalance in the game of T cells: What can the CD4/CD8 T-cell ratio tell us about HIV and health? PLoS Pathog. 2017 Nov 2;13(11):e1006624. doi: 10.1371/journal.ppat.1006624. PMID: 29095912; PMCID: PMC5667733.

[2] Keith Alcorn, Is the CD4/CD8 ratio a useful test for people with HIV?, 23 February 2024, Is the CD4/CD8 ratio a useful test for people with HIV? | aidsmap

[3] Ron R, Moreno E, Martínez-Sanz J, Brañas F, Sainz T, Moreno S, Serrano-Villar S. CD4/CD8 Ratio During Human Immunodeficiency Virus Treatment: Time for Routine Monitoring? Clin Infect Dis. 2023 May 3;76(9):1688-1696. doi: 10.1093/cid/ciad136. PMID: 36883584.

[4] Mauro Garcia, CD4/CD8 ratio, July 2024, CD4/CD8 ratio | aidsmap

[5] Parang Mehta. What to Know About the CD4:CD8 Ratio Test, reviewed by Dany P. Baby, MD, May 04, 2022, https://www.webmd.com/hiv-aids/what-to-know-about-cd4-cd8-ratio-test

[6] Singh, Yashik & Mars, Maurice. (2010). Support vector machines to forecast changes in CD 4 count of HIV-1 positive patients. Scientific Research and Essays. 5. 2384-2390.

[7] Tang J, Li X, Price MA, Sanders EJ, Anzala O, Karita E, Kamali A, Lakhi S, Allen S, Hunter E, Kaslow RA and Gilmour J (2015) CD4:CD8 lymphocyte ratio as a quantitative measure of immunologic health in HIV-1 infection: findings from an African cohort with prospective data. Front. Microbiol. 6:670. doi: 10.3389/fmicb.2015.00670

[8] Jenks JD, Hoenigl M. CD4:CD8 ratio and CD8+ cell count for prognosticating mortality in HIV-infected patients on antiretroviral therapy. J Lab Precis Med 2017;3:8.

[9] Masopust D, Vezys V, Wherry EJ, Ahmed R. A brief history of CD8 T cells. Eur J Immunol. 2007 Nov;37 Suppl 1:S103-10. doi: 10.1002/eji.200737584. PMID: 17972353.

[10] Bofill M, Janossy G, Lee CA, MacDonald-Burns D, Phillips AN, Sabin C, Timms A, Johnson MA, Kernoff PB. Laboratory control values for CD4 and CD8 T lymphocytes. Implications for HIV-1 diagnosis. Clin Exp Immunol. 1992 May;88(2):243-52. doi: 10.1111/j.1365-2249.1992.tb03068.x. PMID: 1349272; PMCID: PMC1554313.

[11] Romero-Rodríguez, D.P., Ramírez, C., Imaz-Rosshandler, I. *et al.* Machine learning-selected variables associated with CD4 T cell recovery under antiretroviral therapy in very advanced HIV infection. *transl med commun* **5**, 6 (2020). https://doi.org/10.1186/s41231-020-00058-x

[12] https://www.kaggle.com/datasets/aadarshvelu/aids-virus-infection-prediction/data

[13] Oeckinghaus, A. (2007). Regulation of Malt1 and Bcl10 function by the ubiquitin and SUMO systems: Implications for NF-κB signaling in lymphocytes.

[14] MicroBiology MCQs with Answer and Explnation | Chapter 39 | Lab Tests Guide. https://www.labtestsguide.com/microbiology-mcqs-with-answer-and-explnation-chapter-39

[15] Hema, M., Ferry, T., Dupon, M., Cuzin, L., Verdon, R., Thiébaut, R., Protopopescu, C., Leport, C., & Raffi, F. (2016). Low CD4/CD8 Ratio Is Associated with Non-AIDS-Defining Cancers in Patients on Antiretroviral Therapy: ANRS CO8 (Aproco/Copilote) Prospective Cohort Study. PLoS One, 11(8), e0161594.

[16] Bohórquez, J. A., Wang, M., Pérez-Simó, M., Vidal, E., Rosell, R., & Ganges, L. (2018). Low CD4/CD8 ratio in classical swine fever postnatalpersistent infection generated at 3 weeks after birth. Transboundary and Emerging Diseases. https://doi.org/10.1111/tbed.13080.

[17] Diagnostic Value of CD4/CD8 in Scrub Typhus in: The American Journal of Tropical Medicine and Hygiene Volume 106 Issue 3 (2022). https://www.ajtmh.org/abstract/journals/tpmd/106/3/article-p792.xml?rskey=fuiCJB&result=1

[18] R. I. Jennrich, "Scaling of Variables and Interpretation of Multiple Regression," The American Statistician, vol. 17, no. 2, pp. 42-44, Apr. 1963

[19] A. S. L. M. a. G. A. Giuseppe Bruno, "The Revival of an “Old” Marker: CD4/CD8 Ratio," in *PubMed*, 2017;19.

[20] G. Y. S. F. A. G. U. O. Emmanuel Ifeanyi Obeagu, "Implications of CD4/CD8 ratios in Human Immunodeficiency Virus infections," INTERNATIONAL JOURNAL OF CURRENT RESEARCH IN MEDICAL SCIENCES, vol. Volume 9, no. http://dx.doi.org/10.22192/ijcrms.2023.09.02.002, pp. 6-13, 2023.

[21] J. D.-Á. M. R. C.-S. R. R. J. A. I. E. B. Javier Martínez-Sanz, "Expanding HIV clinical monitoring: the role of CD4, CD8, and CD4/CD8 ratio in predicting non-AIDS events," eBioMedicine, vol. 95, no. https://doi.org/10.1016/j.ebiom.2023.104773, 2023.

[22] T. F. M. D. L. C. R. V. R. T. C. P. C. L. F. R. V. L. M. Mariam Noelie Hema, "Low CD4/CD8 Ratio Is Associated with Non AIDS-Defining Cancers in Patients on Antiretroviral Therapy: ANRS CO8 (Aproco/Copilote) Prospective Cohort Study," PLOS ONE, no. https://doi.org/10.1371/journal.pone.0161594, 2016.

[23] Y. T. M. A. R. D. L. G. D. V. I. A. N. Y. M. N. R. Emilova, "FERROPTOSIS IN CD4+ AND CD8+ T-CELLS IN THE SETTINGS OF HIV INFECTION," Problems of Infectious and Parasitic Diseases, vol. 51, no. http://dx.doi.org/10.58395/pmgvqy76, pp. 5-10, 2024.