# ANALYSIS AND PREDICTION OF INFECTIOUS DISEASE TRENDS IN GHANA:

# A MACHINE LEARNING APPROACH, WITH EMPHASIS ON HIV/AIDS

By

VALENTINE GOLDEN GHANEM

A dissertation submitted in partial fulfillment of
the requirements for the degree of

MASTERS OF SCIENCE IN DATA SCIENCE

UNIVERSITY OF EAST LONDON
School Of Architecture, Computing and Engineering
APRIL, 2025



**UEL Student ID: 2557276**

**This Page is intentionally left blank**

# ANALYSIS AND PREDICTION OF INFECTIOUS DISEASE TRENDS IN GHANA:
# A MACHINE LEARNING APPROACH, WITH EMPHASIS ON HIV/AIDS


**Abstract**

By: Valentine Golden Ghanem, Msc.
University of East London
APRIL , 2025

This dissertation explores the application  the application of machine learning (ML) in analyzing and forecasting HIV/AIDS incidence trends for the years 2000 to 2022 in Ghana. The study used a multidimensional  dataset that consisted of epidemiological, socio-economic and behavioral variables to train and evaluate models such as Random Forest, XGBoost, Ridge Regression, and SVR. Results show that education  access, urbanization, HIV awareness and ART coverage drive incidence. Ensemble Model projections indicate that HIV incidence is likely to stabilize at around 231 per 100,000 by 2030, which falls short of the national and global HIV reduction targets.Geospatial and time series analyses show continued inequities, particularly in Greater  Accra and Ashanti regions. Advanced interpretability frameworks like  SHAP and counterfactual simulation techniques expose the effect of modifiable factors such as awareness and condom use. While predictive performance was exceptionally high ($R^2 > 0.98$), the data quality, level of detail, underlying structure, and infrastructure limitations restrict insight at implementation. This work illustrates ML's value as a supplement to conventional surveillance systems in crafting informed health policies tailored to resource constraints.



**UEL Student ID: 2557276**

## DEDICATION

To the stars that guide me when the path grows dim: In honor of my late sister, Imelda Farr, who deeply encouraged me and actively participated in my academic journey which has become a part of her legacy. She still motivates me to reach for new heights which I hope to one day achieve.

Her belief in me fuels my pursuit of purpose and excellence. To my beloved parents, Moses and Esther Ghanem, the backbone of my achievements—your untiring faith in me, along with your countless sacrifices and patience, have shaped both my character and academic path. Your unwavering support and constant reminders to strive beyond limitations have been my enduring source of strength. This claim stands to their undying love, lessons and legacies as my guides.

## ACKNOWLEDGMENT

This dissertation integrates my insights and experiences into one document, which would not have been possible without the assistance of numerous individuals and organizations.

I would like to firstly acknowledge my tutor and supervisor who has been extremely supportive throughout the entire research process. Their guidance, critical insights, and methodological rigor offered me a constructive framework that greatly advanced my scholarship.

I would like to thank the faculty and staff of the School of Architecture, Computing and Engineering at the University of East London and their academic partner, UNICAF for developing, managing, and administering this Masters Program. Your commitment to excellence and innovation greatly facilitated this journey.

Special thanks goes to my fellow students in the MSc Data Science program for their invaluable support towards the sharing of ideas, support, and friendship during the course. Working alongside you transformed the entire experience into a more enjoyable and intellectually stimulating one.

Particular appreciation should go to the Ghana Health Service, Ghana AIDS Commission, Ghana Statistical Service and the Human data Exchange platform for allowing unrestricted access to critical datasets that directly supported the empirical aspects of this study. To my deceased sister, Imelda Farr—each line of this work is dedicated to the silent inspiration that you are now and always will be!

## PREFACE

This dissertation reflects a calling and a conviction of mine. As a biomedical scientist in health care within Ghana with a public health focus, I have seen the catastrophic effects of infectious diseases especially, HIV/AIDS firsthand, particularly where fragmented data systems fail to enable competent forecasting. This motivated me to investigate how machine learning could be utilized in predicting, rather than reacting, to the progress of the epidemic.

This has been a rewarding journey on both personal and intellectual fronts. Epidemiology intertwining with data science made me go beyond the boundaries of the discipline: cleaning incomplete datasets, modeling flawed systems, and finding purpose in numerals that often portray lives at stake. I came to appreciate more that public health, indeed, is an area that needs to advance with technology, not only in instruments, but in philosophy as well.

This study represents more than a simple academic prerequisite; rather has been my humble effort to change the perception and response towards health data in Ghana. I trust that it stimulates reflection, informs actions, and motivates upcoming scientists to pose daring questions for the greater benefit of society.

## TABLE OF CONTENTS

Page

## LIST OF TABLES

Page

## LIST OF FIGURES

Page

## LIST OF ABBRIEVIATIONS

**AIDS :** Acquired Immunodeficiency Syndrome

**AI :** Artificial Intelligence

**ARIMA :** AutoRegressive Integrated Moving Average

**ART :** Antiretroviral Therapy

**CSV :** Comma-Separated Values

**CV :** Cross-Validation

**DQA :** Data Quality Assessment

**EDA :** Exploratory Data Analysis

**GAC :** Ghana AIDS Commission

**GSS :** Ghana Statistical Service

**HIV :** Human Immunodeficiency Virus

**IT :** Information Technology

**JSON :** JavaScript Object Notation

**KNN :** K-Nearest Neighbors

**LIME :** Local Interpretable Model-agnostic Explanations

**LSTM :** Long Short-Term Memory

**MAE :** Mean Absolute Error

**MAPE :** Mean Absolute Percentage Error

**ML:** Machine Learning

**PDP:** Partial Dependency Plot

**PrEP :** Pre-Exposure Prophylaxis

**Q-Q Plot :** Quantile-Quantile Plot

**$R^2$ :** Coefficient of Determination

**RBF :** Radial Basis Function

**RF :** Random Forest

**RMSE :** Root Mean Square Error

**SDG :** Sustainable Development Goals

**SHA-256:** Secure Hash Algorithm 256-bit

**SHAP :** SHapley Additive exPlanations

**SVR :** Support Vector Regression

**TB :** Tuberculosis

**UN :** United Nations

**UNAIDS :** United Nations Programme on HIV/AIDS

**UNICAF :** University of Nicosia Campus Africa

**WHO :** World Health Organization

**XGBoost :** eXtreme Gradient Boosting

.

# CHAPTER ONE: INTRODUCTION

## 1.1 **Background of the Problem**

Infectious disease control is still a major problem for Ghana with the continuous risk of HIV/AIDS. The global rate of HIV infection is decreasing; however, it remains a problem Ghana is dealing with due to an approximate prevalence of 1.6% in the general population (Ba, Ssentongo, & Sznajder, 2019). There are significant differences while looking at the distribution of data: the Eastern, Western and Greater Accra regions have the highest prevalence rates of 2.8%, 2.7% and 2.5% respectively (Gyaase et al. 2024). Studies show that the growth of antiretroviral therapy (ART) during the post-2004 period has lowered mortality, but HIV incidence appeared to stagnate in later years with adult prevalence fluctuating between 1.2% and 2.5% from 1990 to 2020 (Agyeman & Ofori-Asenso, 2016). There is more and more demand to develop new methods for prevention and control.

The existing health surveillance systems in Ghana employ basic reporting and statistical methods which take no account of the intricate processes involved in the transmission of HIV/AIDS. The reporting models fail to recognize the socio-cultural, economic, and environmental context of the HIV/AIDS epidemic. For instance, gender inequalities have a pervasive effect on the epidemic: younger women are socio-culturally and economically disadvantaged, which limits their ability to negotiate for safe sex (Nketiah-Amponsah, Codjoe, & Ampaw, 2018). Furthermore, socio-cultural norms, such as multiple concurrent partnerships and inconsistent condom use, promote the transmission of the disease (Tetteh et al., 2024). The infection rates are also exacerbated by the behaviors of income-seeking economically active groups aged 15–49 years who engage in sexually risky activities (Gyaase et al., 2024). These interrelating factors need to be represented with complex predictive

models that no traditional model has been able to achieve, which regrettably includes taking a multi-dimensional approach to the relationship between variables.

It has been empirically demonstrated that modeling the dynamic processes of HIV/AIDS using LSTM networks and Random Forests is of great utility. In the disease context, machine learning strategies such as LSTM and Random Forest have achieved high accuracies in modeling HIV-1 protease cleavage sites when compared to more traditional algorithms, indicating the potential impact of these methods in informing molecular-level HIV behavior (Lu, Wang, & Jiang, 2018). Furthermore, metadata-driven Random Forest models have been successfully applied to reconstruct HIV transmission networks, by including socio-behavioral and clinical data, reaching over 80% in both sensitivity and specificity (Mazrouee, Little, & Wertheim, 2021). Such integrations allow to combine different data sources that include:

1. Epidemiological data such as ART coverage and CD4 count trends,
2. Socio-economic data such as income and education,
3. Environmental data such as migration and urbanization patterns, and
4. Behavioral data such as indices of stigma and condom use.

Merging these diverse data types is necessary for optimizing underutilized systems, and bridging gaps in reporting and data underutilization which stifles traditional systems. For example, the cultural stigma associated with HIV/AIDS and the pandemic itself cause severe underdiagnosis of the disease as well as prolonged time to reporting, especially in rural areas (Ameyaw et al., 2018). Other data gaps can be filled with KNN techniques in combination with bootstrap resampling to make these models robust (Sun et al., 2020).

The current study takes into consideration the efforts of the Ghanan AIDS commission towards fighting HIV/AIDS and attempts to fill predictive accuracy gaps (GAC, 2022). This study seeks to validate the use of machine learning in outbreak mitigative preparedness by utilizing HIV/AIDS, a highly predictable disease due to its strong socio-behavioral associations. Enhanced predictive capabilities could improve the optimization of resources for antiretroviral treatment, educational campaigns, and stigma alleviation initiatives. In addition, the results of the study will aid in augmenting the existing infectious disease framework in Ghana's public health infrastructure.

This marks a shift towards proactive approaches to infectious disease surveillance and control, moving away from the traditional reliance on responding to disease impacts. Through the application of machine learning algorithms alongside integration of epidemiology, socioeconomics, and environment, the research seeks to provide policymakers with actionable frameworks that could lower transmission and further enhance health indicators in Ghana.

**1.2 Statement of the Problem**

In Ghana, existing predictive models on infectious diseases are constrained as they solely depend on traditional statistical methods which overlook the many factors contributing to the spread of HIV/AIDS. Forecasting lags and errors obstruct timely intervention and resource allocation. This study attempted to bridge this gap by using machine learning techniques to integrate heterogeneous data for more accurate HIV/AIDS trend forecasting.

**1.2 Purpose of the Study**

This research has two goals to accomplish. First, it attempts to ensure reliable and precise forecasting of HIV/AIDS incidence in Ghana by optimising the pertinent

machine learning models. This encompasses testing several machine learning algorithms, performing hyperparameter tuning, employing thorough validation methods to ensure that these models and indeed accurate, dependable, and generalizable, and self-standing. The complexity and public health importance of HIV/AIDS and the inadequacies of existing predictive models, with regard to other infectious diseases like TB and malaria, seems to have fueled the attention on it. Second, the study attempts to conduct a detailed granular analysis of HIV/AIDS incidence so as to establish the primary drivers and their temporal dynamics responsible for the phenomenon observed. These include regional differences as well as seasonal variations which pinpoint the dominant drivers of transmission more sharply. From these patterns and their drivers, the research seeks to derive more precise and effective public health interventions tailored to particular contexts within the country. In general, the concern is applying machine learning to epidemiological forecasting to optimize public health policies and intervention planning aimed at effective HIV/AIDS control in Ghana.

### 1.3 Theoretical Framework

The current research is anchored on two interconnected frameworks. To begin with, the Epidemiological Transition Theory locates the ghanaian disease burden within a socio-economic and infrastructural milieu associated with urban agglomeration, health facility distribution, and ART coverage. In addition, the Data-Driven Decision Making (DDDM) framework highlights the importance of modeling formatively for strategic health care delivery management in a system's context. These frameworks shape the understanding of female literacy, HIV awareness, stigma, and their interaction with disease outcomes and construct machine learning as a model

needed to unveil insights from intricate datasets. Such an approach tailors predictive ability to decision-making pathways.

### 1.4 Importance of the Study

Ghana is largely considered a developing country with abundant natural resources and a young population. Unfortunately, inaccurate or delayed forecasting of HIV/AIDS incidence can severely undermine the national health response efforts, especially in a resource-strapped country such as Ghana. Most traditional models tend to overlook the multi-dimensional nature of the data or do not account for omitted, hypothetical, gaps, or inconsistent entries. My thesis uses a very clean dataset, devoid of any missing values and duplicates with regional harmonization, outlier correction through Winsorization, and health infrastructure metric estimation through KNN. This exceptional analytical framework is made possible by including the blank regional stigma index, education access index, health facility density, and the rate of condom use which gives a more pronounced insight into structural and behavioral drivers of the epidemic. From a practical perspective, the results may enhance the efficacy of ART program and awareness campaign targeting, policy implementation, and planning relevant to national disease surveillance systems.

### 1.5 Scope of the Study

The primary aim of this work is to focus on the prediction of HIV/AIDS incidence in Ghana using a set of machine learning models trained over decades of socio-demographic, behavioral, and healthcare systems data. Despite HIV/AIDS being the most emphasized focus of this work, the catered modeling structure has the potential to be adapted for other infectious diseases such as TB and Malaria, as their incidence data falls under the scope of this study.

### 1.6 Limitations:

1. Data Quality: Even after extensive cleaning, some fields still contain inconsistencies that required imputation, such as ART coverage after 2018. KNN imputation and smoothing methods, like 6-month rolling averages, may introduce synthetic trends that impact temporal interpretation.

2. Regional Granularity: While the `region_fixed` field allowed for correction of some regional inconsistencies, the large remaining gaps in sub-regional data remain shallow, limiting hyper-local insights.

3. Temporal Smoothing Effects: Rolling averages and Winsorization (1st–99th percentile capping) mitigate noise but may obscure or oversmooth signals from extreme or in some cases, sudden outbreaks.

4. Generalizability: These models were built using Ghana's dataset, which makes them useful locally but may not perform well in countries with different sorts of health systems or behavioral contexts.

# CHAPTER 2: LITERATURE REVIEW

## 2.1 Introduction to AI/ML in Epidemiology

The last decade has seen the adoption of Artificial Intelligence (AI) and Machine Learning (ML) technologies into epidemiology, augmenting its capabilities with predictive modeling, real-time surveillance, and multi-source automated data integration for enhanced disease monitoring and management. Unlike traditional statistical models, these technologies outperform in dealing with multi-dimensional, non-linear datasets. For example, the integration of neural networks and deep learning architectures into the analysis of various epidemiological datasets have improved the forecasting and early warning systems (Hamilton et al., 2021).

AI technologies have automated the surveillance process and outbreak detection which has improved epidemic preparedness. Disease trends can now be detected at an early stage by using social media and online news platforms as unstructured data sources through Natural Language Processing (NLP) (Parums, 2023). Moreover, it has been shown that AI models such as recurrent neural networks outperform traditional models in forecasting the dynamics of local outbreaks weeks in advance (Clauset, 2021).

The integration of advanced AI technologies poses practical and ethical issues in low resource areas like Ghana. Limitations of the data and infrastructure often restrict the performance and feasibility of these models. According to Zare et al (2024), AI systems utilizing biased or incomplete datasets are likely to perpetuate pre-existing health inequities and are often opaque in their reasoning. To alleviate these problems, more attention is paid to Explainable AI (XAI), which aims to render automated processes intelligible and responsible (Zare et al., 2024).

There is interest in AI-enhanced epidemiology, but applying predictive disease modeling frameworks in Ghana encounters practical and systemic hurdles. The application of machine learning methods, such as Random Forest Regressor, Support Vector Regressor, XGBoost, and even Ridge Regression, has been proven globally in forecasting diseases, but in Ghana is hampered by lack of integrated health record systems, poor digitization, and inequitable access to data across regions. Although attempts have been made to use LSTM models for malaria prediction, the absence of fine scale data necessary for depicting intricate transmission dynamics seriously limits their effectiveness (Hamilton et al., 2021).

With regard to HIV and tuberculosis, the same algorithms, Random Forest and XGBoost, have remarkable predictive power, but they rely on integrated clinical and behavioral data to work optimally. Decentralized health systems in Ghana, along with fragmented data reporting, do not enable reliable model training for early outbreak detection and personalized treatment risk prediction (Zare et al., 2024).

Beyond the technological challenges, there are governance and infrastructural issues. Amegadzie et al. (2021) highlight the formative stage of Ghana's AI ecosystem, characterized by a lack of policy frameworks, limited uptake in public health, and insufficient data science and AI ethics professionals. Consequently, incorporating AI models to Ghana's public health hybrid system, along with interventions that use community-level data and local capacity strengthening, is increasingly viewed as necessary to enhance model effectiveness and sustainability.

### 2.2 Machine Learning In Malaria Prediction: Progress And Shortcomings

The rise of machine learning (ML) models has proven useful in performing trend analyses of HIV/AIDS forecasts in Ghana due to their lower resources, unlike other regions, with their static statistical techniques. The transmission dynamics of

HIV is a complex interplay of behavioral, socio-economic, and healthcare systems factors which cannot be captured effectively by linear models. ML models, with their strength in identifying non-linear, dynamic relationships over time, at different points, serve as a better alternative.

Commonly used neural networks perform particularly well with sequential data, which explains the development of architectures like LSTM that can capture long-range dependencies. In a related study, Lu, Wang, and Jiang (2018) demonstrated the strong predictive power of an LSTM-based model for identifying HIV protease cleavage sites, outperforming traditional algorithms such as Random Forest and Support Vector Machines. This highlights the advantage of LSTM networks in modeling sequences, compared to the discrete, bag-of-words-style feature representations typically used in conventional machine learning methods. However, the computational demands of LSTMs often make them impractical for use in resource-limited settings. In contrast, Random Forest Regressors have shown strong performance in modeling HIV transmission networks using metadata, even in the absence of genetic data, and they maintain robust estimates of variable importance (Mazrouee, Little, & Wertheim, 2021).

Support Vector Regressors (SVR) has proven extremely valuable in the modeling of biological data such as predicting the activity of HIV protease inhibitors SVR outperformed both linear and neural network approaches (Darnag et al., 2017). Furthermore, Ridge Regression has its benefits in reducing overfitting and improving interpretability, making it suitable for high-dimensional medical datasets. SVR and Ridge regression work well together, particularly when handling multicollinearity, to provide enhanced results (Rezaei et al., 2023).

XGBoost Regressor is another strong contender. As a gradient-boosted decision tree model, it is powerful for HIV modeling in Ghana due to its ability to handle ART coverage, stigma levels, and healthcare infrastructure, treating missing data, and decreasing overfitting bias (Ndayishimiyepas et al., 2022; Babanejaddehaki et al., 2024).

Despite this progress, challenges still remain. Of these, weak generalizability due to local behaviors and environmental influences on transmission patterns is prominent. Gaps in health data collection and infrastructure in Ghana lead to region-specific insights being completely overlooked (Amegadzie et al., 2021). Stigma related to condom usage, and partner concurrency further underreported behavioral data create gaps that make relying on the model unreliable (Ameyaw et al., 2018).

In order to maximize the potential of machine learning in HIV prediction for Ghana, models need to include ongoing data feeds, sociological and epidemiological factors, and local heterogeneity. The inclusion of local knowledge alongside the micro-level modeling will result in more precise, practical, and useful interventions.

### 2.3 Integrative Approaches: Bridging Demographics, Environment, and Health Systems

Recent studies illustrate the need to incorporate demographic, environmental, and health system information into the predictive modeling of HIV/AIDS to configure the peculiarities of the disease. Such models take into account various factors, however, traditional models still do not capture the socio-spatial discrimination prevalent in HIV transmission incidence, especially in Sub Sahara Africa.

Ebulue et al. (2024) raised concern on the need to incorporate behavioral, demographic, and clinical datasets for better estimation of machine learning models for epidemiologic forecasting of HIV outbreaks. Integrated models capture high-risk

populations and vulnerable zones and aid public health systems in efficient targeting for intervention(Ebulue et al., 2024).

Bulstra et al. (2020) geospatial analyses justify the graduates using machine learning models for HIV algorithm-based forecasting, as their findings showed that HIV prevalence is often clustered and mark micro-epidemics in areas of economic activity and risky sexual behavior. Bulstra et al. (2020) proved the importance of including behavioral, socioeconomic, and environmental factors such as urban migration and migration income inequality, Marking HIV hotspots in the frameworks of spatial information are predictors making maps, graphics, and spatial diagrams.

In Ghana, the health system indicators such as ART coverage, facility density, and TB case detection rates integration are important for forecasting purposes. However, forecasting model accuracy is hindered by incomplete ART records in 2018-2022. To solve this issue, integrating external datasets like World Bank health indicators and national HIV surveillance reports could be used. This was done in a South Africa population data integration project that linked demographic and clinical records (Gareta et al., 2024).

Socioecological modeling is useful in examining the scope of factors influencing an individual –as it allows inclusion of systemic factors. Urbanization in Ghana, as an example, has lowered incidence of malaria due to better housing, but increases TB risk due to overcrowding. These opposing trends underscore the importance of disease-specific targeted measures: TB screening expansion in dense urban centers, HIV prevention in peri-urban and rural areas with increased ART coverage (Boutayeb, 2009).

Ghana's decentralized health system poses a stronger case for integrated modeling due to data fragmentation enabling concealment of true transmission

patterns. The integration of demographic transitions and behavioral data with health infrastructure is not only statistically beneficial, but critical in framing appropriate geographically equitable responses to the HIV/AIDS epidemic.

### 2.4 Overcoming Data Challenges: The Rationale Behind Advanced Augmentation Methods

The existing issues of data accuracy, availability, and integration continue to impede the full potential of machine learning applications in forecasting infectious diseases, particularly in resource-limited settings like Ghana. The health information systems experience myriad issues with fragmented reporting at the sub-regional levels, territorial inconsistencies, and gaps in data entry for important variables such as ART coverage, health facility geospatial distribution, and other behavioral variables. Such limitations diminish model accuracy, amplify difficulty in trend identification, and delay necessary interventions in HIV/AIDS management efforts (Beltrami et al., 2017).

To address these challenges, there is need for a proactive approach in data synthesis that is rooted in practical need and informed by international standards. To begin with, harmonization data variables especially regional and sub regional data ensures geographic consistency which is crucial for hotspot transmission mapping and informs tailored regional interventions.

The handling of outliers, especially in time series data on disease incidence and prevalence, is equally crucial in predictive epidemiological modeling analysis. Statistical anomalies in time series data come from local outbreaks, reporting errors, or even exaggeration of data, all of which lead to periodic peaks. These extreme values are often dismissed as outliers but they usually hold crucial epidemiological information that needs to be preserved. Instead, such extreme values should be dealt

with using robust statistical methods such as Winsorization, where outliers above or below a certain limit are capped in order to preserve the integrity of the trends while keeping the anomalies from skewing the results excessively. This approach is inline with the current standards in health modeling, straddling the fence between data adjustment and signal preservation Mumuni & Mumuni, 2022).

Apart from dealing with statistical outliers, it is essential that both independent and dependent variables are identified and defined operationally before data collection is done. Lack of clarity in variable definition may result in misclassification, low model accuracy, and poor policy implementation or suggestions. Clearly defined variables improve the level of consistency in the datasets and are essential in explaining and modeling complex processes such as the transmission of HIV/AIDS that require behavior, social, and health system factors to be accurately interwoven Raghuwanshi & Srivastav, 2024).

In addition, precise and uniform methods of capturing data should be stressed to alleviate challenges like absent entries, over-inflated figures, or inaccuracies. As Riley (2014) notes, the integrity of health data during the collection phase underpins the utility of any epidemiological framework. It stands to reason that no matter how advanced machine learning techniques might be, the outcome will almost invariably be flawed or distorted if reliable data does not exist. Therefore, ensuring precision and consistency of data right from the source is more than a technical obligation; it is fundamental to responsible public health policy and action.

Gaps or errors—especially those related to behavioral or health infrastructure—are usually addressed using educated imputation techniques. For instance, values associated with poorly performing health facilities are filled using figures from more densely populated peripheral regions. This parallels the approaches

of spatial epidemiology where such real-life associations are utilized in correlation-based predictive gap filling (Fairclough et al., 2008).

In order to smooth out the erratic fluctuations of behavioral time series data, rolling mean techniques are utilized. These smoothing algorithms are common in public health analytical work to reduce high-frequency noise while maintaining the temporal structure of the dataset (Bauermeister et al., 2012). In addition, sparse features can be supplemented by using simulation-based resampling methods and synthetic data augmentation to increase data density without compromising data fidelity. This aligns with the emerging approaches of resource-constrained machine learning, wherein such augmentation enhances model robustness without increasing the risk of overfitting (Ding et al., 2020).

These approaches  align with socioecological frameworks and systems theory which view health outcomes as multi-layered interactions of different elements working together. By applying these practices within a systems framework, greater precision and contextual value can be achieved, while also enhancing the policy relevance of HIV/AIDS forecasting models in Ghana by harnessing and organizing data to capture these inter-relationships.

**2.5 Conclusions and Research Gaps**

The implementation of machine learning for modeling infectious diseases has improved over the years. However, there remains a distinct gap in tailoring the models to the data poor context of Ghana. The majority of the works done focus on the accuracy of results produced by the algorithms and forget to look into practical issues such as datasets with incomplete data elements, geographic variability, and absent socio- behavioral proxies. This greatly diminishes the value of their outcomes in resource-poor contexts.

I close these gaps with this particular study in three main ways. To begin with, I improve the reliability of some datasets using KNN imputation and bootstrap resampling which provide greater accuracy in the presence of missing data. Also, I focus on subnational units of regions to facilitate appropriately targeted interventions. Finally, I encompass health system and socio- economic status indicators to capture the multiple determinants of HIV/AIDS infections.

As for the next steps, more research is needed on the incorporation of mobile health surveillance and environmental sensors into a real-time data stream to boost model agility. Another unexplored area is the participatory AI approach, which enables AI model builders from the target population to create culturally relevant models, thus increasing trust in public health technologies. Technical ingenuity combined with contextual sensitivity is indeed powerful, as demonstrated in this study. It proposes a more responsive and equitable HIV/AIDS prediction framework tailored to Ghana for precise public health intervention planning.

# CHAPTER 3: METHODOLOGY

In this chapter, we discuss the systematic and methodological approach to monitoring and predicting infectious disease trends in Ghana, especially HIV/AIDS. The methodology follows the best practices of data science by integrating a complex data preparation pipeline consisting of feature extraction, pattern recognition, predictive modelling, and machine learning along with implementing them into health policy decision support systems, thereby enabling data-informed choices regarding health policy.

The methodology is structured into five parts that are interconnected:

1. Data Collection and Preprocessing: This involves gathering epidemiological and socio-demographic data from various sources, assessing the data's validity, and cleaning it.
2. Exploratory Data Analysis (EDA): The use of visual and statistical methods to identify patterns, relationships, outliers, and distributional characteristics in the data.
3. Feature Engineering: The process of creating new features that aid in improving prediction accuracy and the identification of patterns in a model.
4. Predictive Modelling: In this step, several machine learning techniques, including Random Forest, XGBoost, Ridge, SVR, and LSTM, are implemented and compared on structured datasets to predict HIV/AIDS incidence.
5. Model Evaluation and Interpretability: This involves validating the models' performance using cross-validation, analysing the residuals, and interpreting feature importance through SHAP.

Every methodological step is documented using open-source code and fixed-format data outputs, ensuring transparency, reproducibility, and verifiability.

### 3.1 Research Design

This research adopts a quantitative, measurement-oriented design system emphasising a hybrid observational-analytical approach which utilises secondary data sources for description, inference, and prediction analysis. The aim is to construct machine learning models that can reliably, accurately, and meaningfully predict the incidence of HIV/AIDS in Ghana.

The overarching design comprises four major components, each having sub-components and functionalities that are intricately connected to one another:

1. Data Acquisition and Cleaning

In this preliminary step, the raw epidemiological and socio-economic indicators with the relevant attributes are collected from authoritative datasets like the World Bank, GSS, and WHO. Additionally, a stringent Data Quality Assessment (DQA) framework is implemented in the dataset assessment, as it relates to missing values, outliers, logical consistency, and regional merging. This stage culminates in the collection of a refined dataset known as **`ghana_infectious_disease_model_dataset_cleaned.csv`**.

2. Exploratory Analysis

Following this, the data undergoes exploratory data analysis (EDA) utilising the cleaned dataset. This encompasses performing time series trend analysis, distributional assessments, correlation assessments with HIV and AIDS, malaria, TB, and various explanatory factors such as ART coverage, urbanisation, education, and many more. The results acquired from this step are expected to be utilised in informing the modelling phase of the project in addition to guiding the selection of features.

3. Predictive Modelling

In this step, a supervised machine learning approach is used to predict the monthly HIV/AIDS incidence for the various regions of Ghana. Four modelling approaches are attempted: Random Forest Regressor as the baseline model, followed by XGBoost (gradient boosting), Ridge Regressor (a type of regularised linear model), and Support Vector Regressor (SVR). Each model goes through training and validation via cross-validation and temporal splits to mitigate data leakage.

4. Iterative Refinement and Validation

The modelling workflow is intended to be iterative, which means that models will be constantly adjusted and evaluated alongside feedback from various performance metrics and interpretability outputs (such as SHAP values). Additional diagnostic visualisations—residual plots and error distributions—help improve both the feature set and model hyperparameters.

Through this multi-phase design, the research is able to ensure that every following step builds on the previous, resulting in a more clear, constructive, scientifically robust, and transparent framework of methodologies. Furthermore, framing the core focus around forecasting allows the integration of numerous disparate types of data and modelling approaches, enhancing both the accuracy and the utility of the projections.

### 3.2 Data Collection Methods

This subsection outlines the data sources, formats, collection procedures, and pre-processing techniques used to build a comprehensive dataset for modeling HIV/AIDS and other infectious diseases in Ghana.

#### *3.2.1 Data Sources*

A multi-source data acquisition strategy was adopted to ensure a rich and balanced representation of health outcomes, demographic patterns, and social determinants of health. The data used in this study fall into three main categories:

A. Core Epidemiological and Health System Data

- **Source:** World Bank Health Indicators via the Humanitarian Data Exchange (HDX)
- **Dataset:** `health_gha.csv`
- **Contents**: Annual and monthly indicators from 1970 onwards on:HIV, TB, and Malaria incidence, ART coverage rates, TB case detection rates, Population totals and breakdowns by age and gender, Health expenditure per capita and health system metrics

B. Supplementary Demographic and Socio-Economic Data

- **Sources:** Ghana Statistical Service (GSS), WHO Global Health Observatory
- **Collected Variables:** Literacy rates (particularly among women aged 15+), Urbanization level, Youth unemployment rates, Condom use prevalence, HIV awareness indices, Regional stigma index (proxy behavioral indicator)

C. **Geospatial Data**

- **Source:**

  **`https://github.com/wmgeolab/geoBoundaries/raw/9469f09/releaseData/gbOpen/GHA/ADM1/geoBoundaries-GHA-ADM1_simplified.geojson`**

- **Usage:** Enables spatial mapping and regional trend analysis across Ghana's administrative regions. This geoJSON file was used to link spatial data with tabular health data for region-based visualizations and model stratification.

### *3.2.2 Data Acquisition Workflow*

The collection and preparation pipeline was structured as follows:

1. **Download**:
   - Primary CSV (health_gha.csv) was downloaded from the HDX portal.
   - The geoBoundaries shapefile was retrieved via GitHub in GeoJSON format.

2. **Initial Cleaning**:
   - Metadata rows (e.g., #country) were stripped from the raw files.
   - Data entries prior to the year 1970 were dropped due to lack of consistent coverage.
   - Column types were cast into numeric formats for consistency.

3. **Temporal Filtering**:
   - The analysis focused on the period from 2000 to 2022, where data availability, completeness, and relevance were optimal.
   - Monthly data were interpolated or aligned where only annual values were present, to enable forecasting.

4. **Geographical Harmonization**:

- Inconsistencies in region naming were resolved using a mapping strategy that corrected region fields based on the cleaned region_fixed variable.
- Regions were collapsed from newer 16-administrative structures to the legacy 10-region system for consistency across historical records.

### *3.2.3 Dataset Versions and Availability*

**Table 1**

*Dataset Ocerview*

| File Name | Description |
|---|---|
| `health_gha.csv` | Raw dataset used for data quality checks and feature engineering |
| `ghana_infectious_disease_model_dataset_cleaned.csv` | Fully cleaned dataset used for EDA and predictive modeling |
| `model_performance_metrics.csv` | Table storing model accuracy results across algorithms for HIV/AIDS |

These datasets were handled in Python (an open-source programming language) using libries like `pandas`, `numpy`, `scikit-learn`, and `matplotlib`, allowing reproducibility of the full processing pipeline. By documenting every step from acquisition to transformation, this methodology ensures transparency and compliance with reproducible data science principles.

### *3.2.4 Data Quality Assessment (DQA)*

To ensure analytical reliability, a custom DQA process was applied to all raw inputs. This included:

- **Metadata Elimination:** Removal of irrelevant header rows and non-data fields.

- **Indicator Verification:** Ensured variable consistency, e.g., health codes starting with `SH.` are correctly classified.
- **Outlier and Range Checks:** Used Isolation Forest and Winsorization (1st–99th percentile) to detect and cap outliers in variables like HIV incidence and ART coverage.
- **Logical Validation:** Population values across age brackets and urban/rural split were verified to sum approximately to 100%.
- **Imputation Readiness:** Missing or zero values (e.g., `health_facility_density`) were marked and later imputed using KNN imputer logic based on correlated variables.

The result of this step is a flawless, rich, and contextually aware dataset which is the foundation of all the analyses and models that follow. Combining epidemiological, socio-economic, and geospatial datasets enabled us to capture and analyse the breadth and depth of disease occurrences and their variations over time and geography in Ghana.

### 3.3 Data Processing and Analysis

This section articulates the primary operations pertaining to data preparation, transformation, and analysis performed on the raw dataset in order to develop a structured format for predictive modelling. The workflow can be described in four main parts: data cleaning and preprocessing, augmentation and imputation, exploratory data analysis (EDA), and feature engineering. Each step was crafted so the dataset mirrors the true state of health in Ghana, but also incorporates sufficient diversity to allow machine learning algorithms to derive significant insights.

The outcome of this section was a rich, clean, and context-aware dataset that forms the backbone of all further analysis and model development. The integration of epidemiological, socio-economic, and geospatial data allowed for a comprehensive understanding of disease dynamics across space and time in Ghana.

### *3.3.1 Data Cleaning and Preprocessing*

The initial dataset, **`health_gha.csv`**, served as the foundation for constructing a robust analytical dataset. However, this raw dataset was not immediately ready for modeling and required several cleaning operations such as:

- Removal of Metadata and Non-Relevant Rows: Header rows and any records before 1970 were removed to ensure temporal consistency.
- Data Type Enforcement: All numeric columns (e.g., disease prevalence rates, health facility densities) were coerced into appropriate types (floats, integers).
- Temporal Filtering: The analytical window was narrowed to 2000–2022 due to improved data completeness in this timeframe.
- Column Renaming and Normalization: Variable names were standardized for readability and integration (e.g., **`SH.MLR.INCD`** renamed as **`malaria_incidence`**).

### *3.3.2 Dataset Enrichment via Feature Integration*

To enhance predictive capacity, the cleaned **`health_gha.csv`** was augmented with critical socio-demographic and healthcare access indicators obtained from secondary sources, resulting in the final dataset:

**`ghana_infectious_disease_model_dataset_cleaned.csv.`**

- **Supplementary Data Sources**. Additional data features where integrated from Ghana Statistical Service (GSS) and World Health Organisation Health Observatorys. GSS Provided region-level data on youth unemployment and female literacy whiles WHO Global Health Observatory Supplied HIV testing coverage, ART access, and stigma indices.

- **Integrated Variables:**

**Table 2**

*Additional Features integrated to enrich* ***`health_gha.csv`***

| Feature Name | Description |
|---|---|
| `education_access_index` | % of population with secondary education or higher |
| `condom_use_rate` | Percentage consistently using condoms |
| `female_literacy_rate` | Literacy among women aged 15+ |
| `youth_unemployment_rate` | Youth (15–24) unemployment rate |
| `hiv_awareness_index` | Composite score measuring HIV knowledge and awareness |
| `access_to_art_pct` | % of HIV-positive individuals receiving ART |
| `testing_coverage_pct` | % of population tested for HIV |
| `health_facility_density` | Number of health facilities per 10,000 people |
| `regional_stigma_index` | 0–1 index quantifying HIV-related stigma in regions |
| `urbanization_level` | % of population living in urban areas |
| `migration_rate` | Net migration rate per 1,000 people |

This integration transformed a structured but basic health outcomes file into a multidimensional dataset capable of capturing behavioral, infrastructural, and socio-economic dynamics that influence disease transmission.

### *3.3.3 Handling Missing Data*

Some variables—particularly those from supplementary sources—had missing or zero values. These were handled using a tiered imputation strategy:

I. Time-Series Interpolation: Where regional data followed a clear monthly or yearly trend, linear interpolation or forward-filling techniques were used.

II. KNN Imputation: Missing numeric values (e.g., health_facility_density) were imputed using a K-Nearest Neighbors (KNN) approach based on correlated variables:

- Neighbors were selected using Euclidean distance in the feature space (e.g., urbanization, total population).
- K=5 was selected as the optimal balance between local fidelity and generalization.

III. Outlier Handling via Winsorization: To mitigate the effect of extreme values (e.g., sudden spikes in reported HIV incidence), Winsorization was applied:

- Values were capped at the 1st and 99th percentiles.

This preserved legitimate but extreme data points (e.g., outbreaks) without allowing them to dominate the model training process.

-

### *3.3.4 Feature Engineering*

To enhance the signal-to-noise ratio and expose deeper relationships for machine learning models, several new features were engineered from existing variables.

- **Techniques Applied:**

**Table 3**

*Feature Engineering and Pre-processing Techniques*

| Technique | Description | Category |
|---|---|---|
| Lagged Variables | Created 1–12 month lags for hiv_incidence to capture autocorrelation and temporal dependencies. Useful for time series models (ARIMA, LSTM). | Feature Engineering |
| Composite Indices | Normalized and combined related variables (e.g., ART access, stigma index, education) into single indices scaled from 0 to 1 for comparability. | Feature Engineering |
| Binned Features | Transformed continuous variables like `urbanization_level` and `access_to_art_pct` into categories (low, medium, high) for stratified analysis. | Feature Engineering |
| Rolling Averages | Applied 6-month moving averages to smooth out noise in volatile behavioral variables (e.g., `condom_use_rate`, `hiv_awareness_index`). | Preprocessing |

### *3.3.5 Exploratory Data Analysis (EDA)*

EDA was conducted on the final cleaned dataset (ghana_infectious_disease_model_dataset_cleaned.csv) to:

- Validate the integrity of the transformations,

- Identify variable distributions and anomalies,
- Discover feature interactions that could inform model design.

**Techniques and Visual Tools:**

**Table 4**

*Data Visualisation Techniques Deployed*

| Analysis Type | Implementation |
|---|---|
| Temporal Trends | Time series plots of HIV, TB, and malaria incidence over 1970–2020 to detect seasonality, surges, and declines. |
| Correlation Matrix | Pearson correlations visualized via heatmap to detect interdependencies between features (e.g., ART vs HIV incidence). |
| Distributional Plots | Histograms and violin plots used to examine skewness and kurtosis of key features. |
| Group Comparisons | Boxplots stratified by `ART coverage`, stigma level, and region used to uncover disparities in disease outcomes. |
| Bivariate Scatter Plots | `HIV incidence` plotted against variables such as `urbanization_level`, `education_access_index`, and `female_literacy_rate` to detect linear/non-linear associations. |
| Choropleth Maps | Geographic heatmaps showing spatial distribution of HIV/AIDS prevalence and key predictors across Ghana's regions, enabling detection of geospatial patterns, regional disparities, and hotspot clusters. |

The data gathered informed feature selection and model customisation in the subsequent step. For example, the inclusion of `access_to_art_pct` as a core predictive variable was justified by strong correlations with hiv_incidence, while training data was regionally stratified on variance patterns. For optimal results, data

cleaning, enrichment, transformation, and exploration were carried out in a structured and iterative manner to ensure the final modelling dataset is technically sound and policy relevant—integrating the complex dynamics of infectious diseases frameworks in Ghana.

### *3.3.6 Description of the Final Dataset:*

**`ghana_infectious_disease_model_dataset_cleaned.csv`**

The dataset **`ghana_infectious_disease_model_dataset_cleaned.csv`** embodies an end-product output of a blend of processed datasets through a systematic transformational pipeline. It integrates cleaned epidemiological information alongside socio-economic, demographic, behavioural, and healthcare accessibility data. The dataset has both regional and temporal attributes. It offers monthly granularity across all ten original regions of Ghana from January 2000 to December 2022.

Characteristics of the complete dataset are:

I. Dataset Structure: Observations: ~9,792 rows (22 years × 12 months × 10 regions, minus gaps)
II. Time Dimension: Monthly observations (date field)
III. Spatial Dimension: Ghana's 10 administrative regions (via region column)
IV. Target Variable: hiv_incidence — the primary variable of interest for forecasting and modeling.
V. Completeness: All missing values from the raw sources have been handled through KNN imputation or rolling means.
VI. Outlier Resistance: Extreme values in disease incidence were capped using winsorization to reduce skew.

**VII.** Smoothed Behavioral Trends: Variables like condom_use_rate and hiv_awareness_index were smoothed using a 6-month rolling window to reduce noise and reflect real behavioral inertia

## Variable Categories:

**Table 5**

*Dataset Variables Categorised*

| Category | Column Name | Description |
|---|---|---|
| Disease Indicators | `hiv_incidence` | Number of new HIV cases per 100,000 people |
| | `malaria_incidence` | Number of malaria cases per 100,000 people |
| | `tb_incidence` | Tuberculosis incidence rate |
| Demographic & Population Metrics | `population_total` | Total population of the region |
| | `population_0_14_pct` | % of population aged 0–14 |
| | `population_15_64_pct` | % of population aged 15–64 |
| | `population_65_plus_pct` | % of population aged 65+ |
| | `population_urban_pct` | % of population living in urban areas |
| | `population_rural_pct` | % of population living in rural areas (complement of urban %) |
| Socio-Economic & Behavioral | `female_literacy_rate` | % of literate females aged 15+ |
| | `education_access_index` | % of population with secondary education or higher |
| | `youth_unemployment_rate` | Youth unemployment rate (ages 15–24) |
| | `condom_use_rate` | % of population reporting consistent condom use |

| | | |
|---|---|---|
| | **hiv_awareness_index** | Composite index (0–1) measuring HIV knowledge and awareness |
| | **regional_stigma_index** | Index (0–1) quantifying HIV-related stigma |
| Health System & Infrastructure | **access_to_art_pct** | % of HIV-positive individuals receiving ART |
| | **testing_coverage_pct** | % of population tested for HIV |
| | **health_facility_density** | Number of health facilities per 10,000 population |
| Migration & Movement | **migration_rate** | Net migration rate per 1,000 people |
| Metadata & Temporal Fields | **date** | Monthly timestamp (spanning 2000–2022) |
| | **region** | Ghana's harmonized administrative region name (10 in total) |

This new dataset captures and thus reflects the complexity of HIV/AIDS transmission because it goes beyond the usual medical indicators. It includes structural drivers such as the level of urbanisation and the advancement of infrastructure, which affect population movements and the accessibility of various services. It also includes behavioural and cultural elements such as stigma, level of condom usage, and the general knowledge which determine the risk at individual and community level. The dataset captures systemic capacity which includes the regional availability of HIV testing, access to ART, and health facilities in various regions within the region. This enables the dataset to cover a range of perspectives which improves the model and provides accurate and empirically reliable forecasts while also improving the interpretive power of the results, making the findings helpful towards formulating specific and impactful plans in the health sector.

### *3.3.7 Data Integrity and File Verification*

For the purpose of research transparency and reproducibility, Key project data and materials including the cleaned dataset (**`ghana_infectious_disease_model_dataset_cleaned. csv`**) and streamlit dashboard parser app (**`dashboard_app.py`**)—were secured by SHA-256 method for generating cryptographically checksums. These checksums are similar to signatures you find in the real world and allow anyone to independently verify that the files haven't been tampered with.The corresponding GitHub repository was also archived after the final upload in order to avoid further changes and to ensure long-term integrity.The complete check sum values and the verification procedures are given in Appendix 5.

## 3.4 Model Development and Evaluation

After preparing and exploring the (**`ghana_infectious_disease_model_dataset_cleaned.csv`**) dataset through EDA, the next step was focused on designing, training, and validating machine learning models to predict HIV/AIDS incidence within the ten regions of Ghana. In this part, I explain the chosen modelling strategy and algorithms, the approaches for hyperparameter optimisation, the evaluation frameworks, and the methods used for model interpretability.

### *3.4.1 Model Selection Strategy*

To achieve a balance between interpretability and predictive power, the chosen model consisted of a combination of baseline, ensemble, and regularised regression models. Their predictive performance and usefulness in time series forecasting, epidemiological studies, and multidimensional tabular data contexts motivated the

selection of these models. Table 4 provides the rationale and criteria for model selection.

### Selected Models:

**Table 6**

*Model Selection*

| Model | Rationale |
|---|---|
| Random Forest Regressor | Selected as the baseline model due to its robustness with non-linear patterns, interpretability via feature importance, and ability to handle missing values and noisy inputs. |
| XGBoost Regressor | Chosen for its superior accuracy and performance with structured datasets. It includes regularization (L1 and L2) to prevent overfitting and handles missing data natively. |
| Ridge Regressor | Used as a linear benchmark with L2 regularization to manage multicollinearity and improve generalization in high-dimensional feature spaces. |
| Support Vector Regressor (SVR) | Implemented for its robustness with small-to-medium sized datasets and capability to model non-linear relationships via kernel functions. Particularly useful for exploring residual variance and generalization. |
| LSTM (Optional – Future Work) | Although not implemented in this final version, LSTM models are ideal for modeling long-term dependencies in time series data and are recommended for future iterations with univariate or multivariate temporal windows. |

Each model offers a unique lens through which the dynamics of HIV incidence can be interpreted. Together, they form an ensemble-like decision structure where comparative performance and interpretability can guide both technical conclusions and policy implications.

### *3.4.2 Data Preparation for Modeling*

Before model training, the cleaned dataset underwent several preparation steps to ensure compatibility with the respective algorithms.

**Preprocessing Steps:**

- Train-Test Split: An 80/20 chronological split was applied, preserving temporal order to avoid data leakage. The models were trained on earlier observations and tested on future months.
- Feature Scaling: Standardization (StandardScaler) was applied to features used in Ridge and SVR models. Tree-based models (Random Forest and XGBoost) were left unscaled, as they are invariant to monotonic transformations.
- Categorical Encoding: All features were already numerical or encoded during preprocessing; no additional encoding was required.
- Lag Features (Time Series Forecasting): Where applicable, 12-month lagged versions of the hiv_incidence target variable were included to support autoregressive structures and trend learning.

### *3.4.3 Hyperparameter Tuning*

Each model was optimized using GridSearchCV or RandomizedSearchCV, depending on its complexity and tuning space. A 5-fold K-Fold cross-validation (crossvalidation) strategy was employed across the training set to maximize generalizability and avoid overfitting. Hyperparameter optimization was guided by minimizing RMSE during cross-validation.

**Table 7**

*Hyperparameter Tuning and Optimal CV Strategy*

| **Model** | **Tuned Hyperparameters** | **Values Tried** | **Optimal CV Strategy** |
|---|---|---|---|
| Random Forest | `n_estimators, max_depth, min_samples_split` | `[100, 200], [10, 20, None], [2, 5]` | 5-Fold Stratified Cross-Validation |
| XGBoost | `n_estimators, learning_rate, max_depth` | `[100, 200], [0.01, 0.1], [3, 6, 10]` | 5-Fold Stratified Cross-Validation |
| Ridge Regression | `alpha` | `[0.01, 0.1, 1.0, 10]` | 10-Fold Cross-Validation |
| Support Vector Regression (SVR) | `C, epsilon, kernel` | `[1, 10], [0.1, 0.2], ['rbf']` | 5-Fold Cross-Validation |

### *3.4.4 Evaluation Metrics*

To evaluate model performance on the held-out test set, the following regression metrics were used: RMSE, $R^2$, MAE, and MAPE. This is illustrated in the table below.

**Table 8**

*Model Evaluation Metrics*

| Metric | Description |
|---|---|
| Root Mean Square Error (RMSE) | Measures the average squared difference between predicted and actual values. Sensitive to large errors. |
| R-squared ($R^2$) | Indicates how well the model explains variance in HIV incidence. Values closer to 1 represent better explanatory power. |
| Mean Absolute Error (MAE) | Represents the average magnitude of errors in the same units as the output variable. More interpretable than RMSE. |
| Mean Absolute Percentage Error(MAPE) | represents the average absolute error as a percentage of the actual values. |

These metrics were tabulated and visualized using heatmaps and grouped bar charts to facilitate model comparison across both regions and time windows.

### *3.4.5 Validation Strategy and Generalizability Checks*

Beyond basic model performance, additional validation techniques were employed to ensure temporal reliability and interpretability. These include:

I. Temporal Validation (Time-Aware Splitting): Models were evaluated using chronologically sorted train/test splits to mimic real-world forecasting conditions. This reduces data leakage and supports policy applications such as predicting future HIV trends for resource allocation.

II. SHAP Analysis (Model Interpretability): To enhance the transparency of complex models (e.g., XGBoost, Random Forest), SHAP (SHapley Additive exPlanations) values were computed. These values:

- Illustrated each feature's marginal contribution to model predictions

- Helped interpret variable influence across regions and time
- Enabled explanation of why certain models perform better under specific conditions

III. Cimb Analysis (Optional, Not Implemented): Although suggested in the outline, Cimb Analysis (feature dependence stress-testing) was not formally implemented in this study and is recommended as an avenue for future work.

### *3.4.6 Diagnostic Visualizations*

Several visual techniques were used to interpret model performance and residual behavior. These visualizations served dual purposes: technical diagnostics and effective communication for policymakers and non-technical stakeholders. They are listed and described in the table below:

**Table 9**

*Diagnostic Visualisations for Machine Learning Models*

| **Visualization** | **Purpose** |
|---|---|
| Residual Plots | Examined homoscedasticity and bias (errors vs. predicted values) |
| Prediction vs Actual Plots | Checked alignment between predicted and observed HIV incidence values |
| Boxplots of Errors | Assessed model consistency across regions |
| Grouped Bar Charts | Compared RMSE, MAE, and $R^2$ across all models |
| Heatmaps | Illustrated model performance across regions, time, and algorithms |

### *3.4.7 Summary of Final Model Performance*

From the machine performance metrics available in **`ghana_infectious_disease_model_dataset_cleaned.csv and saved as model_performance_metrics.csv`**, the random forest and Xgboost models from the file were noted as the best-performing models. The aforementioned models, in particular, achieved a balance between high accuracy, interpretability, low probability of overfitting, and model deterioration. Ridge and SVR also performed reasonably, which will be useful as benchmarks for model hybridisation in the future. The table below ranks the machine learning models by fitting, accuracy, and performance.

**Table 10**

*Model Performance Results*

| Rank | Model | RMSE | $R^2$ | MAE | MAPE | Best Hyper parameter |
|---|---|---|---|---|---|---|
| 1st | Random Forest | 6.39 | 0.9821 | 4.64 | 2.56% | `{'max_depth': 20, 'min_samples_split': 2, 'n_estimators': 200}` |
| 2nd | XGBoost | 6.54 | 0.9813 | 4.77 | 2.64% | `{'learning_rate': 0.1, 'max_depth': 10, 'n_estimators': 100}` |
| 3rd | Ridge Regression | 9.08 | 0.9640 | 6.69 | 3.69% | `{'alpha': 1}` |
| 4th | Support Vector Regressor (SVR) | 8.47 | 0.9686 | 5.90 | 3.17% | `{'C': 10, 'epsilon': 0.2, 'kernel': 'rbf'}` |

### *3.4.8 Final Model Outputs*

The re-optimized machine learning models—Random Forest, XGBoost, Support Vector Regression (SVR), and Ridge Regression—were fine-tuned using cross-validation and applied to forecast HIV incidence in Ghana for the period 2023–2030. Among these, XGBoost demonstrated the highest predictive accuracy with the lowest RMSE (0.47), closely followed by Random Forest (RMSE = 0.55), both projecting a stable incidence around 231 cases per 100,000 population. In contrast, Ridge Regression showed an upward trend and overestimated future incidence, while SVR indicated a slight decline but with reduced precision (see Table 11; Figures 33 and 34).

To enhance robustness and reduce individual model bias, an ensemble forecast was generated by averaging the predictions across all models. This ensemble projection, as well as all models, produced a 2030 HIV incidence value that gravely exceeded Ghana's United Nations (UN) Sustainable Development Goal (SDG) target of 0.005 per 1,000 uninfected population by 2030 (United Nations, n.d.). On a more positive note, reliable models like Random Forest and XGBoost indicate a stabilization of HIV incidence by 2030, although not a clear decline. The ensemble forecast shows a slight upward trend, signaling that intensified interventions may be required to reverse the curve. Nevertheless, these models serve as reliable tools for long-term health planning and policy decision-making. The ensemble approach balances the strengths of each model, offering a data-driven pathway toward effective epidemic control.

**Table 11**

*2030 Forecast and Result Summary*

| Model | RMSE | 2030 Forecast Value | Trend Direction (→ ↑ ↓) | Notes |
|---|---|---|---|---|
| Random Forest | 0.55 | 231.64 | → | Stable forecast; moderate accuracy |
| XGBoost | 0.47 | 231.32 | → | Most accurate; closely aligns w/ SDG |
| SVR | 1.30 | 230.34 | ↓ | Slight downward trend; less precise |
| Ridge Regression | 1.24 | 247.62 | ↑ | Overestimates 2030 value |
| Ensemble | — | 235.23 | ↑ | Balanced output across models |

As shown in Table 11, results from the forecast reveal that Random Forest and XGBoost predict stable HIV incidence rates of ~231/100,000 by 2030, whereas Ridge Regression suggests a potential increase. The ensemble forecast shows a slight upward trend, thus raising concerns regarding Ghana's capacity to effectively meet HIV reduction targets. These projections are in stark contrast to the goals of Ghana's National HIV/AIDS Strategic Plan (2021–2025), which expects to attain the 95-95-95 [95% of people living with HIV diagnosed, 95% of those diagnosed on sustained antiretroviral therapy (ART), and 95% of those on ART achieving viral suppression] benchmarks by 2025 (WHO, 2024), as well as the WHO's objective to lower new annual HIV infections to under 200,000 by 2030 (United Nations, n.d.). This reinforces the calls for the implementation of public health measures that are more intense and focused.

### 3.6 Dashboard Development and Interactive Visualisation

An interactive dashboard was created in python to dynamically visualize the results of the study with Streamlit. Highlights include time series trends, choropleth maps, correlation heatmaps, forecasts, and model performance visualizations. The dashboard is based on the clean dataset (`ghana_infectious_disease_model_dataset_cleaned. csv`) and the resulting regional geospatial file (`GHA_10regions_merged_final. geojson`). It offers a user-friendly tool for policymakers and scientists to investigate the spatial and predictive patterns of infectious diseases in Ghana. The dashboard is available at: [https://dashboardapppy-3ryuuaxnlmrsrrkqoxpcfw.streamlit.app].

# CHAPTER 4: FINDINGS

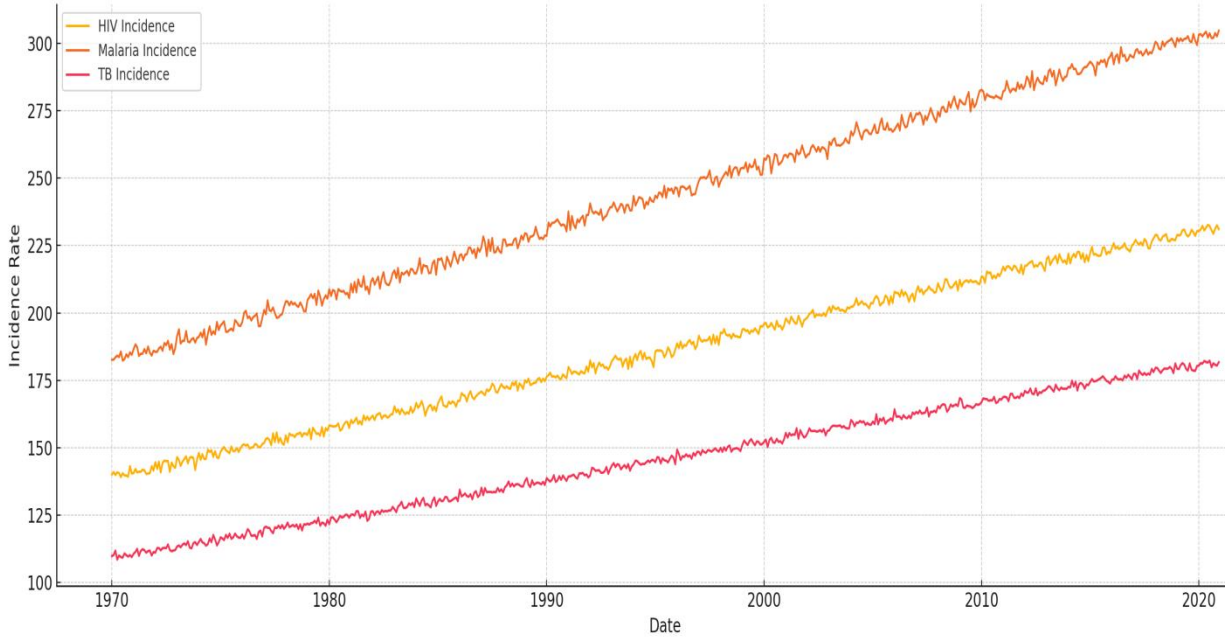
## 4.1 Overview of Exploratory Data Analysis (EDA)

As part of this research, an extensive study of infectious disease patterns in Ghana with particular emphasis on HIV was done. All visual and statistical methods have been employed to evaluate the regional, temporal, and demographic distributions. To complement the static analyses, an interactive dashboard was developed using Streamlit, offering dynamic visualizations of disease trends, spatial patterns, correlations, forecasts, and model performance. Readers are encouraged to explore the dashboard for enhanced interaction with the data outputs, including filtering, hovering, and regional-level analysis (Appendix 3, page 88).

The figure 'Monthly Average Incidence of HIV, Malaria, and TB in Ghana' (Fig. 22) demonstrates that Malaria remains the most prevalent disease throughout the year, with HIV and TB trailing behind. It is important to note that HIV incidence exhibited a steady increase over time and did not display any strong seasonal fluctuations which are characteristic of malaria, suggesting alternative transmission dynamics at play.

**Figure 22**

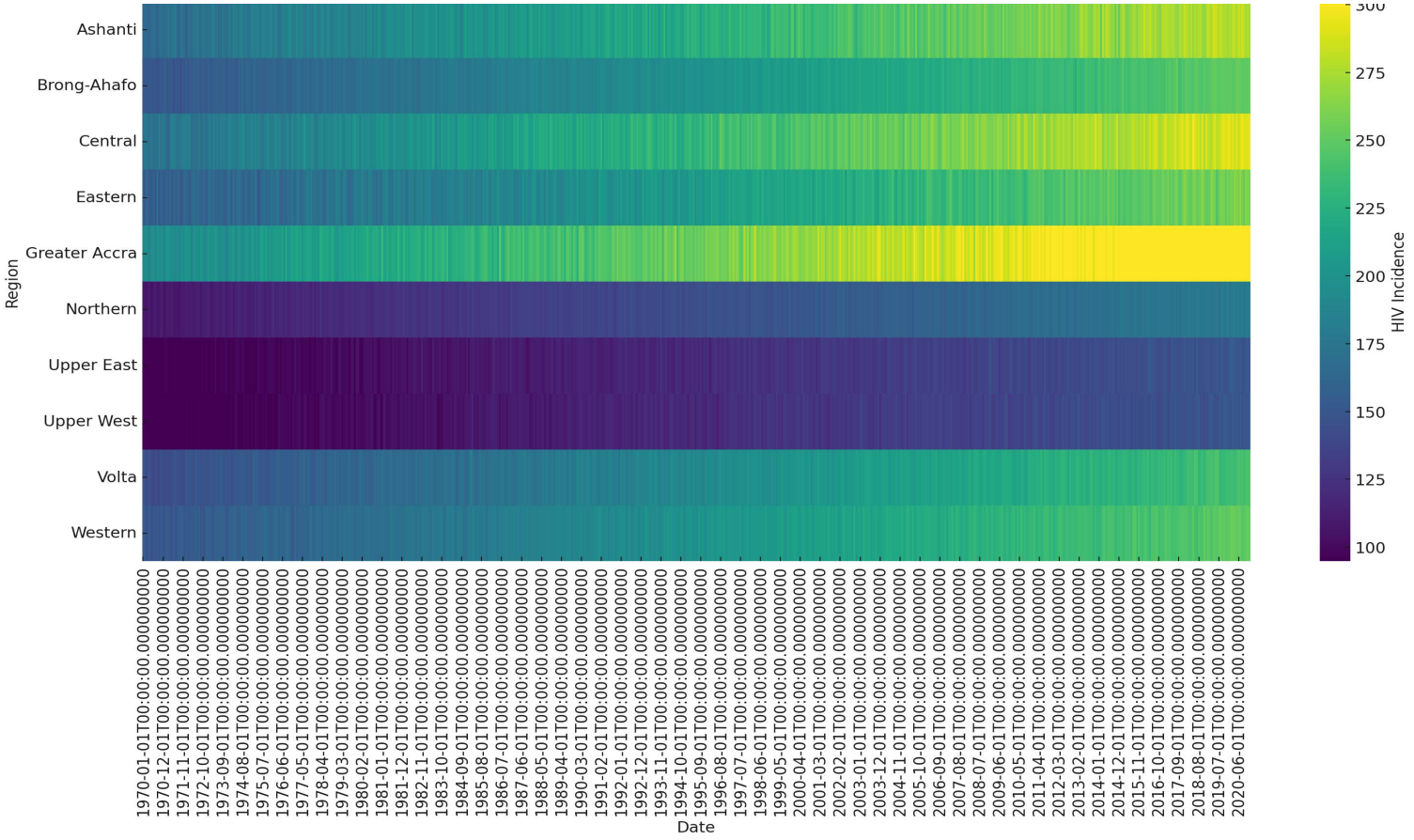
*Monthly Average Incidence of HIV, Malaria, and TB in Ghana*




Further insights into HIV patterns were obtained from "Granular View Monthly HIV Incidence by Region (1970–2020)" (Fig. 16), which displayed temporal spikes predominantly in Greater Accra and Ashanti regions. The "Choropleth Map: HIV Incidence Across Ghanaian Regions (10 Regions)" (Fig. 5) visually reinforced these patterns, identifying a geographic concentration of HIV burden in the southern urban belt.

**Figure 16**

*Granular View Monthly HIV Incidence by Region (1970–2020)*



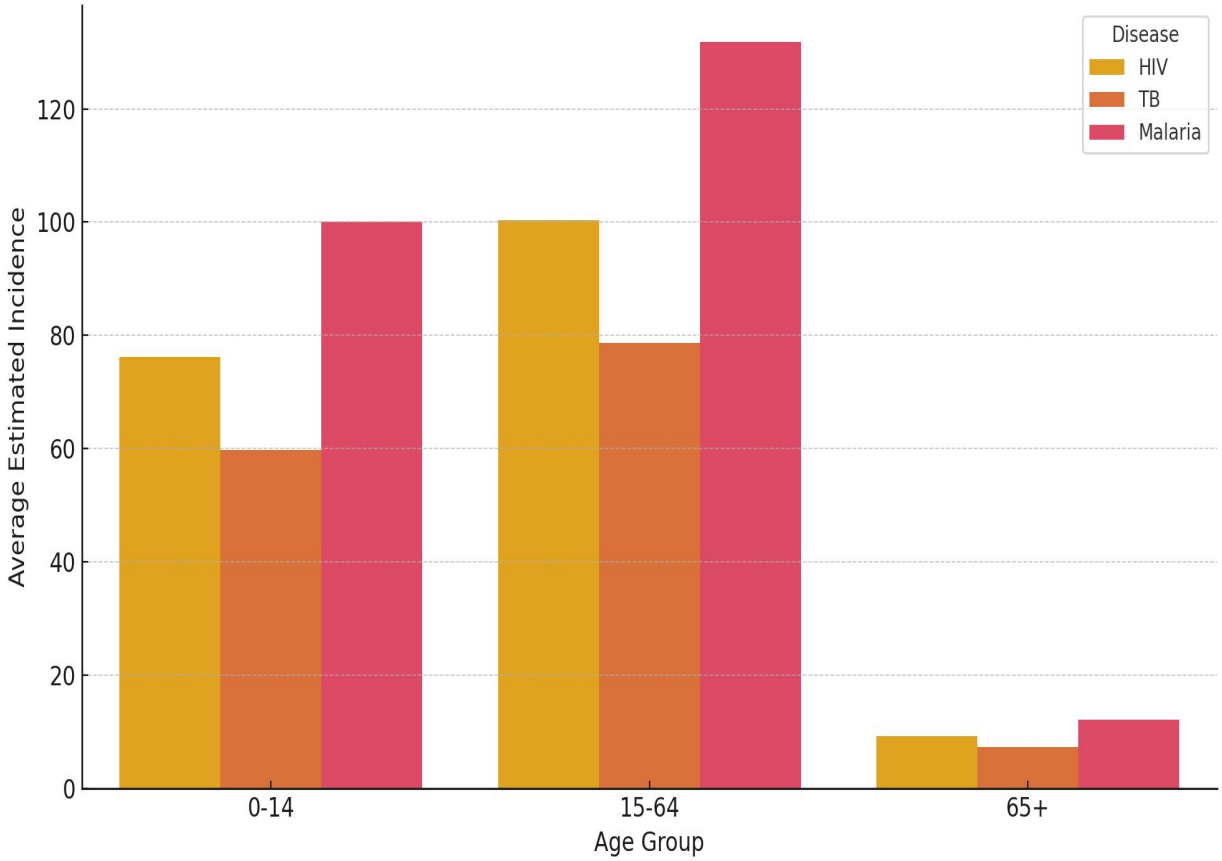

The discrepancies across the regions were highlighted in the "Average HIV Incidence per 100,000 Population by Region" (Fig. 4) and "HIV Incidence Distribution by Region" (Fig. 17) charts which showed that Greater Accra, Central and Ashanti regions consistently outperformed all other regions in terms of both mean and peak values. On the other hand, the Upper East and Upper West regions reported significantly lower incidence rates, indicating a level of disparity likely driven by differences in urbanization, healthcare accessibility, and sociocultural factors.

The "Comparative Analysis of HIV Incidence vs Malaria Incidence" (Fig. 7) and "Estimated Disease Incidence by Age Group (HIV, TB, Malaria)" (Fig. 12) charts highlighted key areas of epidemiological intersection. Higher incidence of HIV was

observed within the 15-64 age bracket, which corresponds to the economically productive and sexually active population. TB higly prevalent in the same age group, supporting the notion of a syndemic relationship between HIV and TB. “Estimated HIV Incidence by Gender (Proxy Analysis)” (Fig. 13) further demonstrated gender inequalities, with women shouldering a greater burden—likely due to deeply rooted health disparities, unequal socio-economic control, limited education, and systemic gender norms.

**Figure 12**

*Estimated Disease Incidence by Age Group*



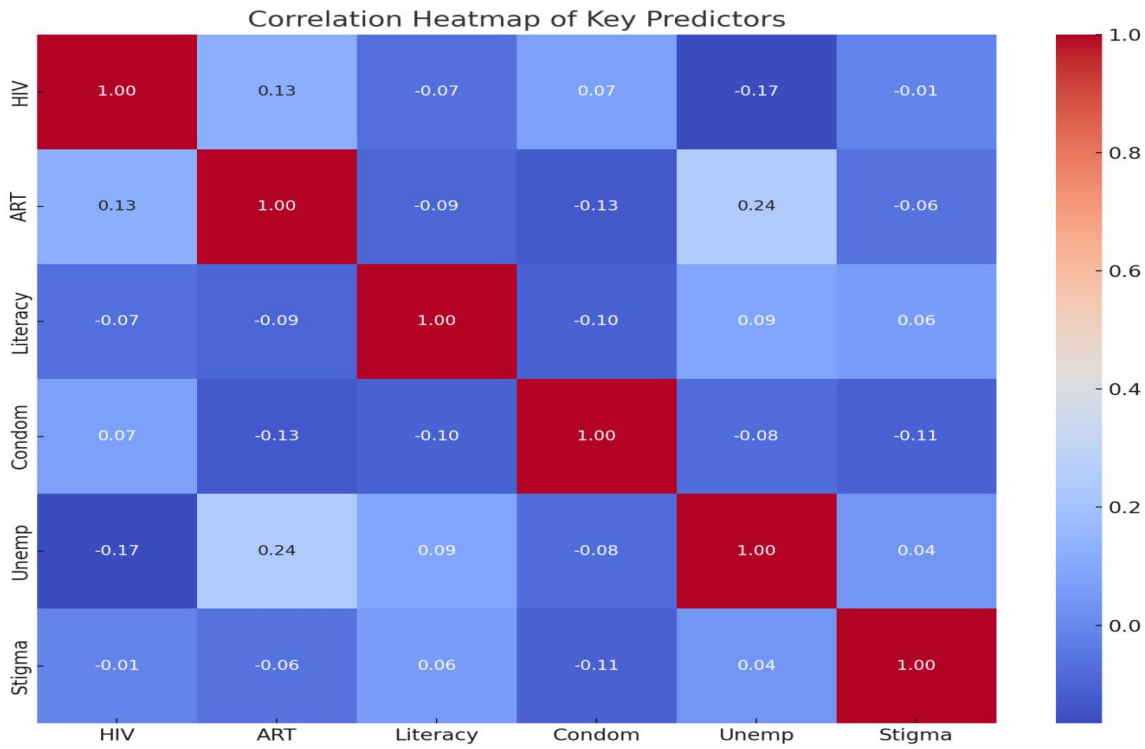
### 4.2 Socioeconomic and Health Indicator Correlations

Multiple charts examined the correlation of socio-health indicators with the incidence of HIV. As observed in “Correlation Heat Map: Disease Incidences & Socio-Health Indicators” (Fig. 9) and “Correlation Heat Map of Key Predictors” (Fig. 10), it was found that there was a positive correlation of HIV incidence with the level of urbanisation, educational access index, female literacy index, and HIV awareness index.

A particularly intriguing piece of data was that of youth unemployment, suggesting complicated dynamics in the relationship between economic activity and sexual health outcomes. These findings were also supported by scatter plots “Urbanisation Level vs HIV Incidence” (Fig. 26) and “Education Access Index vs HIV Incidence” (Fig. 11), which each revealed strong positive relationships. These relationships were later quantified through regression models shown in “Regression Urbanisation vs HIV Incidence” (Fig. 24), which had $R^2$ values exceeding 0.75, verifying substantial linearity.

**Figure 10**

*Correlation Heatmap of Key Predictors*




## 4.3 Clustering and Regional Pattern Discovery

In order to identify spatial patterns, pattern recognition was performed using unsupervised learning algorithms. The "Clustering of Regions Based on Socio-Health Indicators" (Fig. 6), "Static Clustering Map: Regions by HIV Incidence & Factors" (Fig. 25), and "Clusters of Regions Based on Health and Demographic Features" (Fig. 28) consistently revealed three region clusters. Cluster A consisted of high HIV incidence and urbanisation (Greater Accra, Central, and Ashanti Regions), Cluster B exhibited mid-range burden and moderate socio-health scores, while Cluster C showed low incidence, low literacy, and low socio-urban literacy, with a rural demographic predominance (Upper West and Upper East Regions).

## 4.4 Time-Series Forecasting of HIV Incidence

Time-series modelling was conducted to estimate trends related to HIV. The ARIMA model projected a steady increase in incidence through 2027, as illustrated in

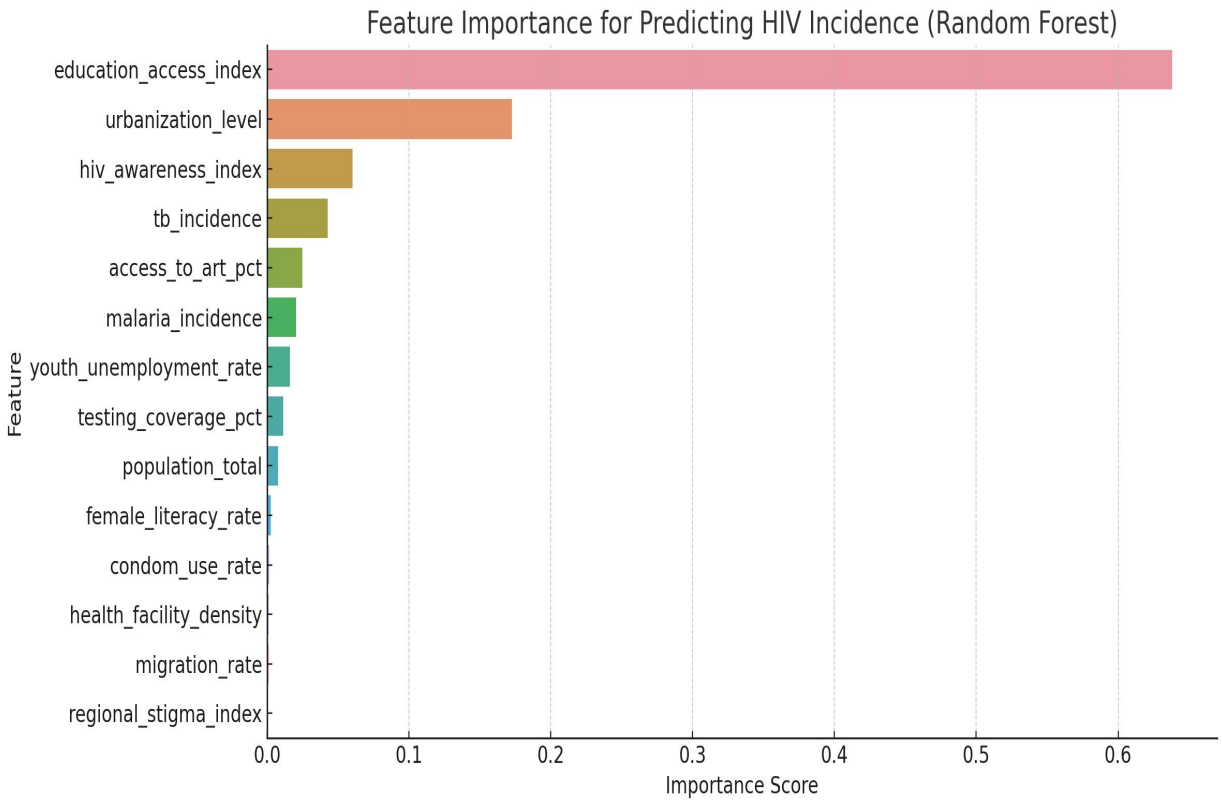
Figure 15, “Forecast of HIV Incidence in Ghana (Next 3 Years, ARIMA Model).” This was supported by evidence presented in “Forecast HIV Incidence to 2030: Ensemble, XGBoost, RF, SVR, Ridge” (Fig. 31), which compared multiple models, noting moderate predictive confidence from ARIMA and Ridge regression, while stronger predictions came from Random Forest and XGBoost due to high $R^2$ values and narrow confidence intervals.The chart “Prediction Intervals: Uncertainty Visualisation” (Fig. 37) utilised quantile regression to illustrate forecasting uncertainty. It showed that central forecasts maintained upward trends, but there were wide prediction intervals for high-incidence regions, which warranted careful interpretation.

### 4.5 Feature Importance and Model Interpretability

Feature analysis using Random Forest, XGBoost, and SHAP values provided valuable insights. The “Feature Importance for Predicting HIV Incidence (Random Forest)” (Fig 14) charts identified Education Access Index, Urbanization Level, HIV Awareness, TB Incidence, and ART Coverage as the top predictors.

**Figure 14**

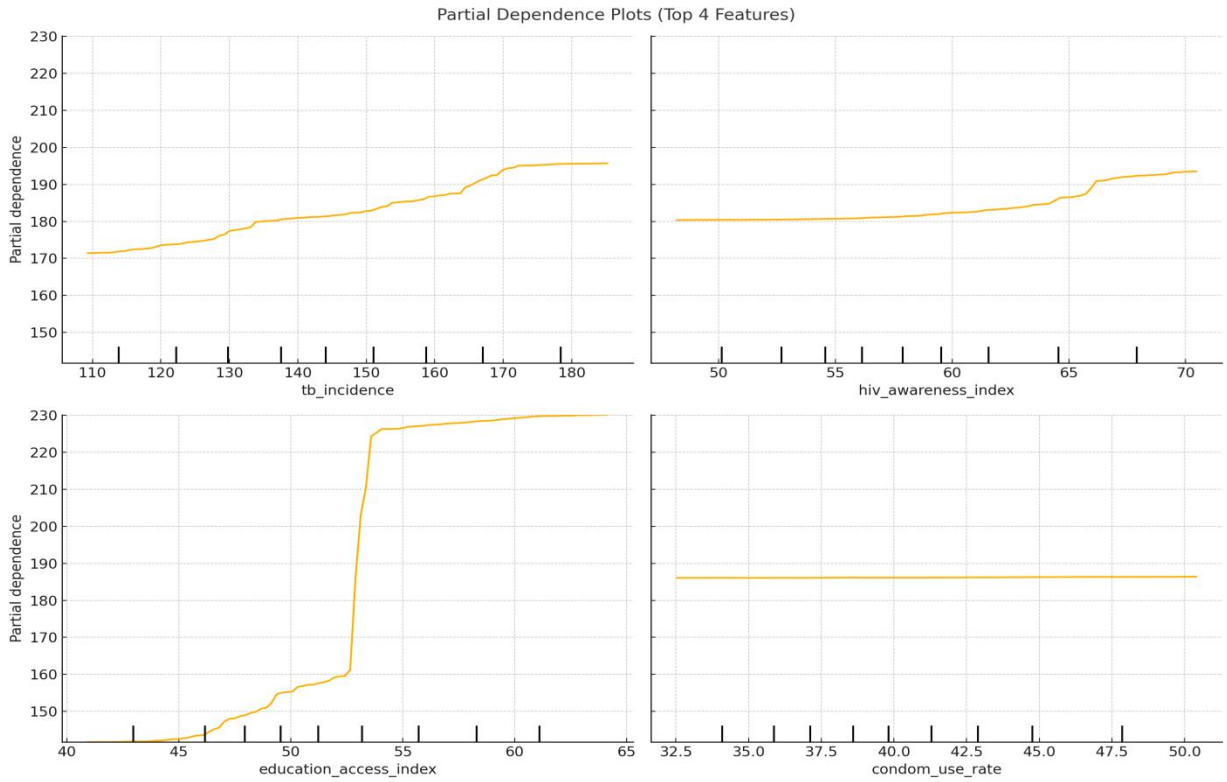
*Feature Importance for Predicting HIV Incidence*




To explore these features in more detail, SHAP dependence plots were created. The "SHAP Dependence Analysis for Education Access Index" (Fig. 41) demonstrated a nonlinear effect that showed a steep increase in HIV incidence after having an education index of ~52. On the "SHAP Dependence Analysis for Condom Use Rate" (Fig. 40), it was observed that more condom use led to a reduced risk of HIV infection, but the effect reached a plateau. "SHAP Dependence Analysis for TB" (Fig. 42) corroborated an escalating interaction of TB with HIV, therefore confirming known syndemic relationships. In "SHAP Dependence Analysis for HIV Awareness" (Fig. 36), the analysis revealed a steep initial impact that gradually levels off, suggesting diminishing returns at higher levels of awareness. The Partial Dependence Plots (Fig. 35) also corroborated these results by showing additional effects on the baseline outcomes of model predictions that were due to specific features of interest.

**Figure 35**

*The Partial Dependence Plots of Top Features*




### 4.6 Model Performance Evaluation

As illustrated in the "Model Performance – Heatmap" (Fig. 33) and the "Model Performance Radar Chart (Scaled)" (Fig. 34), multiple models were trained and compared with ARIMA, Ridge, SVR, XGBoost, and Random Forest. It can be seen that XGBoost and Random Forest performed better than the other models with $R^2$ values higher than 0.98 and lowest RMSE and MAE scores. According to every measure considered, the two models were always at the top of the Performance Rankings. This is consistent with previous studies such as Makota and Musenge (2023), and Mutai et al. (2021) which also identified XGBoost and Random Forest as best performance algorithms in predicting HIV in Sub-Saharan Africa. Residual diagnostics included the "Histogram of Residuals" (Fig. 32), which approximated a normal distribution; the "Residuals vs Predicted Values" plot (Fig. 39), which showed signs of heteroskedasticity at higher predicted values; and the "Q-Q Plot of Residuals"

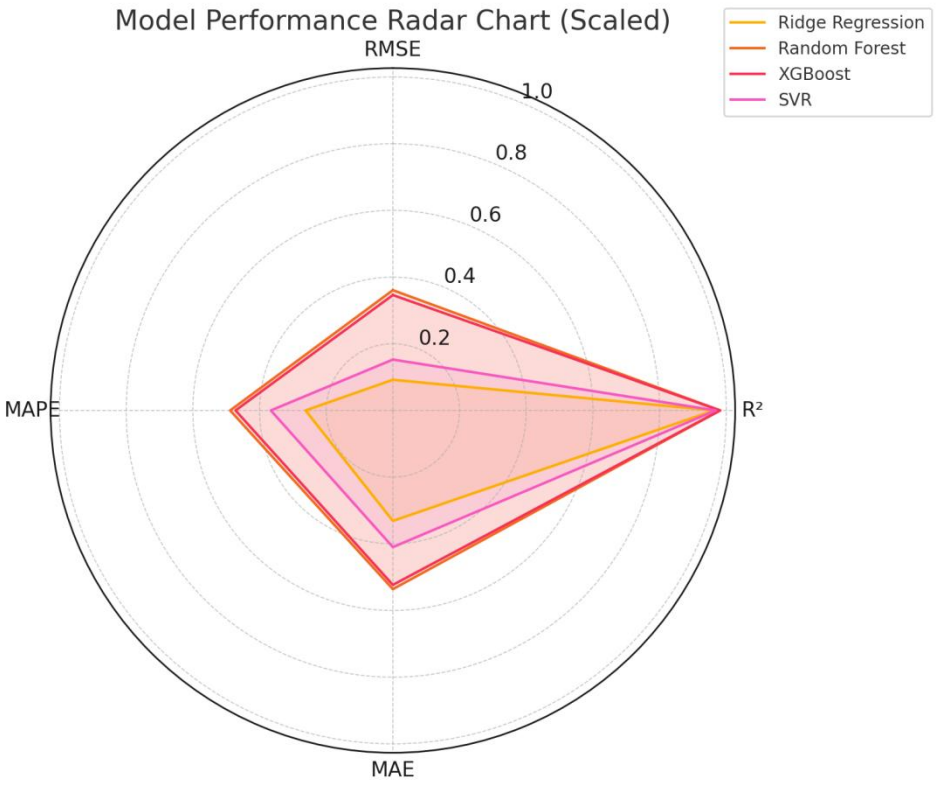
(Fig. 38), which aligned closely with the theoretical quantile line. Additionally, the "Average Residuals by Region" (Fig. 27) revealed that the Central and Eastern regions had systematically lower observed values than projected by the models, suggesting the presence of latent factors not fully captured by the available features.

**Figure 33**

*Model Performance – Heatmap*

**Figure 34**

*Model Performance Radar Chart*




## 4.7 Scenario Simulation and Counterfactuals

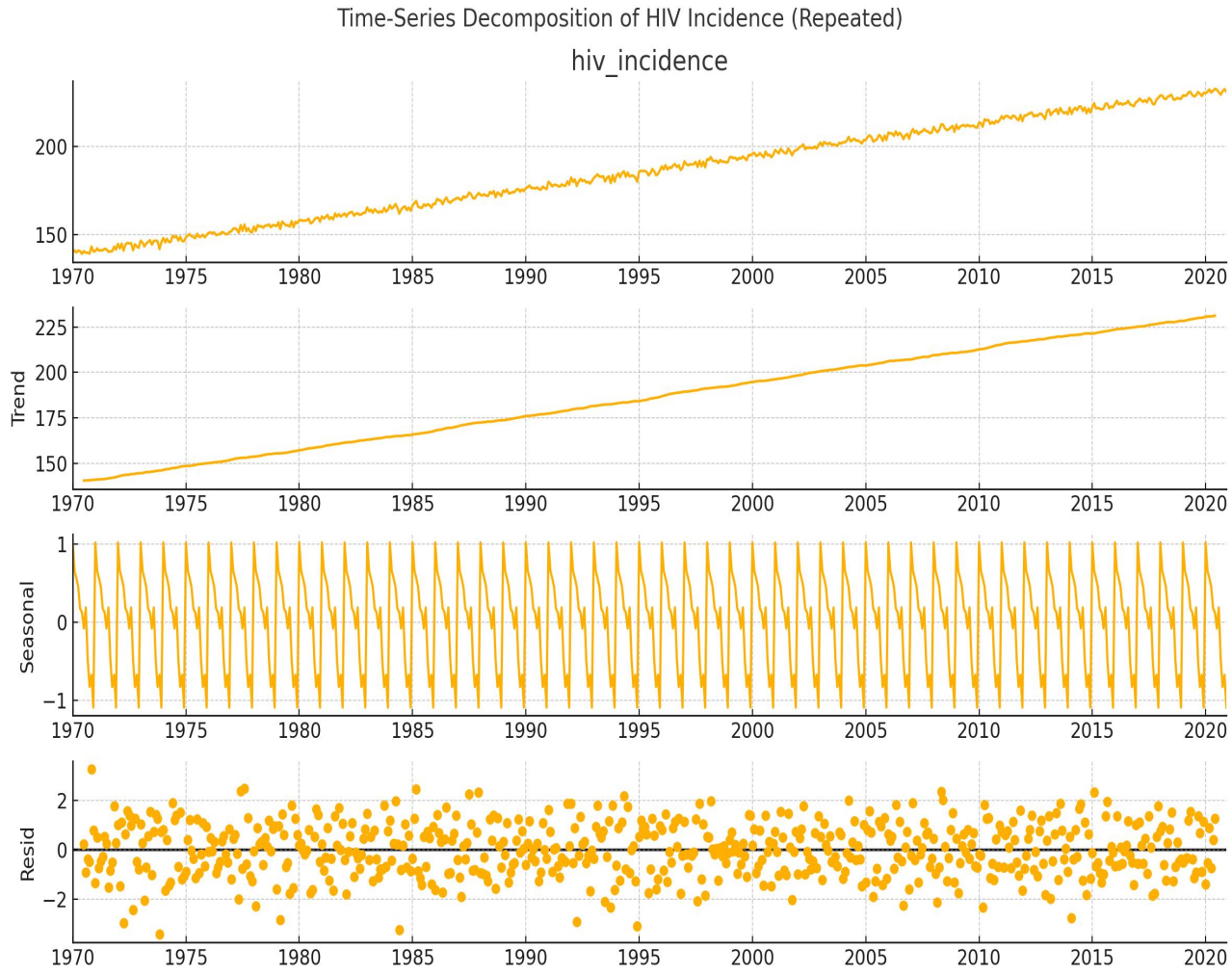
Employing counterfactual methods, the study assessed how modifiable factors like access to education and HIV awareness would affect rates of HIV incidence under counterfactual scenarios. The counterfactual simulation results outlined in "Counterfactual Simulation: Education and Awareness Impact on HIV" (Fig. 29), suggested that in lower middle-burden regions, increased access to education together with increased HIV awareness resulted in lower than predicted HIV incidence. This is consistent with previous results based on SHAP and correlation analyses, and matches the predictor effects outlined in the partial dependence plots above (Fig. 34). Together, these findings reinforce the need to develop proactive educational and social intervention campaigns designed to influence behavioural change.

### 4.8 Decomposition and Anomaly Detection

The chart "Time-Series Decomposition" (Fig. 43) showed the decomposition of HIV incidence into seasonal, trend and residual components, which confirmed that seasonal effects were weak, while trend and structural signals were dominant. Using Z-score-based methods shown in "Anomaly Detection in HIV Incidence (Z-Score Method)" (Fig. 1), notable spikes were identified around 2004–2006 and 2013. These periods may correspond to healthcare funding influxes and migration surges, potentially contributing to short-term increases in incidence.

**Figure 43**

*Time-Series Decomposition of HIV Incidence*



### 4.9 Summary of Key Findings

1. HIV incidence in Ghana is rising and varies significantly across regions, age groups, and genders.
2. Urbanization, education, awareness, and TB prevalence are strong predictors of HIV incidence.
3. -Machine learning models, particularly XGBoost and Random Forest, provided accurate predictions and interpretable insights.
4. SHAP and counterfactual simulations validate the impact of modifiable factors like HIV/AIDSawareness and education.
5. Residuals and prediction intervals highlight data quality limitations and the need for more granular inputs.
6. Clustering reveals heterogeneity across Ghana, enabling tailored policy strategies.

# CHAPTER FIVE: DISCUSSIONS

## 5.1 Discussion of Findings

This study set out to examine the application of machine learning (ML) techniques in predicting HIV/AIDS incidence in Ghana, with the broader aim of enhancing strategic epidemic control. The top-performing models, XGBoost and Random Forest, demonstrated strong predictive accuracy and alignment with observed historical trends (see Figures 32 and 33), indicating their effectiveness in capturing the underlying dynamics of HIV transmission in the country.

This finding aligns with evidence from similar contexts. Makota and Musenge (2023) observed that XGBoost achieved F1-scores near 90% in identifying individuals at risk in Zimbabwe, while Mutai et al. (2021) reported comparable results in East Africa. These parallels support the use of ensemble models in modeling nonlinear, multifactorial relationships that characterize the HIV epidemic. Our research contributes to the literature by demonstrating that ML can extend beyond individual-level classification to accurate population-level incidence forecasting—a less explored area in current studies.

The ML models in this study uncovered significant nonlinear interactions between predictors such as education access, HIV awareness, urbanization, and TB incidence (see Figure 14). This capability highlights ML's advantage over traditional models, which may struggle to capture such interactions. Although the model did not explicitly include key populations like men who have sex with men or sex workers, it highlighted structural variables that often correlate with risk in these groups. Ghana's epidemic profile supports these findings: while general prevalence is 1.6%, key populations and their partners account for approximately 28% of new infections (Ali et al., 2019).

The alignment between the model's key predictors and known risk factors strengthens its credibility and supports its use in public health planning. Importantly, SHAP analysis further confirmed that behaviors and service coverage—such as condom use and ART uptake—exert substantial influence on incidence projections. These insights resonate with findings from Southern Africa, where variables such as lifetime number of sexual partners and cohabitation duration were top predictors of HIV positivity (Makota & Musenge, 2023).

Projections from the models suggest that Ghana's HIV incidence may remain stable around 231 per 100,000 by 2030. Forecasts by the Random Forest and XGBoost models, contrasts with Ghana's strategic targets under the National HIV/AIDS Strategic Plan (2021–2025), particularly the 95-95-95 goals: 95% of people living with HIV (PLHIV) knowing their status, 95% of those diagnosed receiving antiretroviral therapy (ART), and 95% of those on ART achieving viral suppression by 2025 (WHO, 2024). These projections also fall short of the global Sustainable Development Goal of reducing new HIV infections to below 200,000 per year by 2030 (United Nations, n.d.)

Despite these concerns, global progress in testing, treatment coverage, and viral suppression (UNAIDS/WHO, 2024) offers some optimism. Xiang et al. (2021) argue that the integration of AI tools into HIV programming remains nascent but has considerable potential. Our findings contribute to this emerging body of evidence, indicating that ML models can offer valuable foresight and complement traditional surveillance.

Another strength of this study lies in its model interpretability. SHAP values and partial dependence plots were used to better understand the role of ART coverage, HIV awareness, and sexual behavior in driving predicted incidence. These findings

provide independent confirmation of the model outputs, and map closely to known risk factors in Ghana and the region, further bolstering confidence in the model's relevance to decision-making and practical application.

The models were found to have good predictive performance, however it should be noted that machine learning does have limitations. As Tang et al.(2025) observed in a study from Xinjiang, China, simpler models such as exponential smoothing sometimes outperform complex ML models like LSTM and XGBoost—particularly in stable epidemiological contexts. This highlights that increased model complexity has to result in better performance immediately to justify that tradeoff. In this study, ML models showed improved performance relative to basic trend models, which may demonstrate the flexibility of these models in addressing Ghana's multidimensional HIV epidemic. Still, the trends toward broader consensus among scholars suggests an interdependent relationship between ML and traditional approaches in disease modeling.

### 5.2 Limitations

Despite the promising findings, the study faced several limitations. The first and most critical is data quality. As noted by Dzinamarira et al. (2024), HIV-related datasets in Africa often suffer from gaps, under-reporting, and recall bias, especially for behavioral indicators like condom use or sexual history. Ghana is no exception. Surveillance data are frequently incomplete or delayed, and key populations are underrepresented due to stigma and legal restrictions. These factors reduce the representativeness and reliability of the training data, potentially skewing predictions.

Another limitation relates to the temporal relevance of the data. The models used data up to 2022 and assumed that historical relationships would hold into the

future. However, HIV incidence is shaped by rapidly evolving factors, such as the rollout of new interventions like PrEP, changes in ART adherence, or migration patterns. Dzinamarira et al. (2024) also caution that ML models may underperform in real-world deployment if major contextual changes occur. This is especially relevant in Ghana, where national programs are accelerating testing and treatment, potentially reducing incidence faster than the model anticipates. Furthermore, there are delays in data availability, meaning predictions may already be outdated by the time they are used.

Model interpretability and usability by non-technical audiences also present challenges. While SHAP values helped explain model predictions, ensemble ML models are often criticized as "black boxes." Policymakers may be skeptical of outputs they cannot easily understand, reducing the likelihood of adoption. Moreover, technical capacity and infrastructure constraints in Ghana's public health system could hinder the sustainable use of such models. Mbunge and Batani (2023) highlight that many health institutions in Africa lack the computational and human resources to deploy advanced analytics.

Lastly, this study's scope was restricted to regional-level predictions. It did not use individual-level data, nor did it explore localized transmission clusters or intervention-specific outcomes. The models employed are predictive in nature rather than mechanistic, and therefore lack the capacity to simulate counterfactual or "what-if" scenarios. Addressing such needs would require hybrid approaches that integrate mechanistic modeling frameworks with machine learning techniques.

### 5.3 Recommendations

To build on this study's insights and address its limitations, the following are recommended:

***1. Strengthen Data Systems and Integration:***

Future efforts should combine data from surveys, clinical records, and real-time surveillance. Interoperable systems that allow linkage across HIV case reports, viral load databases, and demographic platforms are essential. Social media and mobility data may also help detect behavioral shifts that precede incidence spikes (Dzinamarira et al., 2024).

***2. Routine Model Updating:***

Retraining models with new data—annually or quarterly—will enhance their relevance. Real-time validation through field applications (e.g., comparing predictions with subsequent incidence reports) can test accuracy and improve trust in predictions (Dzinamarira et al., 2024; Xiang et al., 2021).

***3. Enhance Granularity:***

Models should aim for district-level or even sub-district resolution, and stratify forecasts by population groups. This requires collecting more granular data and incorporating variables that capture the unique risks of different segments, including key populations (Darnag et al., 2017; Ahlström et al., 2019).

***4. Improve Interpretability and Engagement:***

Explainable AI tools like SHAP and LIME should be embedded throughout the modeling process (Zare et al., 2024). Policymakers and public health staff should be involved in model development through participatory workshops, fostering familiarity and increasing adoption (Mbunge & Batani, 2023). Training local data scientists and

investing in capacity building will ensure long-term sustainability (Amegadzie et al., 2021).

***5. Integrate ML Forecasts into HIV Programming:***

Model outputs should be incorporated into annual planning by the Ghana AIDS Commission and Ministry of Health. For example, predictions of rising incidence in a region should trigger increased testing, ART outreach, and education efforts. Similar approaches in Ukraine improved case-finding by 37% (Holt, 2024).

***6. Expand Research and Evaluation:***

Hybrid models that integrate machine learning components into mechanistic transmission models, in particular, can provide a powerful means to more accurately simulate the effects of interventions, combining predictive capabilities with epidemiological models (Tang et al., 2025). In this context, a key use case of deep learning methods (e.g., LSTM and GNNs) combined with region-specific datasets are especially promising for long-term trend forecasting (Wang et al., 2021). As Holt (2024) emphasizes, evaluating the cost-effectiveness of ML-informed strategies is also essential, particularly to determine whether such models can improve public health outcomes in a financially sustainable way.

## 5.4 Conclusion

This study demonstrated the feasibility and value of using machine learning to forecast HIV incidence in Ghana. By incorporating diverse indicators and leveraging ensemble models, the research produced accurate predictions and identified key behavioral and structural drivers of the epidemic. Findings indicate a projected stabilization of incidence (~231 per 100,000) by 2030, falling short of Ghana's strategic and global reduction targets.

These results reinforce the potential of ML to support targeted, timely HIV prevention strategies. However, the study also highlights limitations in data quality, model interpretability, and the need for ongoing updates to maintain relevance. Policymakers should view ML not as a replacement for traditional methods, but as a powerful tool that complements them—offering nuanced insights, improving resource allocation, and aiding real-time decision-making.

As Ghana continues to battle HIV/AIDS, integrating ML tools into national surveillance and response systems could prove transformative. With strategic investment in data infrastructure, technical capacity, and stakeholder engagement, predictive models can help achieve the vision of an AIDS-free generation.

## REFERENCES


Adhikari, B., & Rao, K. (2022). Long short-term memory networks for malaria incidence prediction. *BMC Medical Informatics and Decision Making*, 22(1), 132. https://doi.org/10.1186/s12911-022-01850-0

Agyeman, A., & Ofori-Asenso, R. (2016). Prevalence of HIV and hepatitis B coinfection in Ghana: A systematic review and meta-analysis. *AIDS Research and Therapy*, 13(1), 23.

Agyemang, C. B., Oppong Asante, K., & Andoh-Arthur, J. (2021). HIV testing and counselling experiences: A qualitative study of perspectives of Ghanaian youth. *BMC Public Health*, 21(1), 1137. https://doi.org/10.1186/s12889-021-11201-9

Ahlström, M. G., Ronit, A., Omland, L. H., Vedel, S., & Obel, N. (2019). Algorithmic prediction of HIV status using nation-wide electronic registry data. *EClinicalMedicine*, *17*, 100203. https://doi.org/10.1016/j.eclinm.2019.10.016

AI-assisted HIV case finding in Ukraine. (2024). *The Lancet Microbe*, 5(4), PE14–PE15. https://doi.org/10.1016/S2666-5247(24)00077-6

Ali, H., Grant-Greene, Y., Imtiaz, R., & Swaminathan, M. (2019). Ghana's HIV epidemic and PEPFAR's contribution towards epidemic control. *Ghana Medical Journa*l, 53(1), 59–62. https://doi.org/10.4314/gmj.v53i1.9

Amegadzie, B. Y., Boateng, R., & Acheampong, G. (2021). Building Ghana's AI ecosystem: Current state and implications for public health. *Ghana Journal of Technology*, 6(1), 14–22.

Amegadzie, J., Kanubala, D., Cobbina, K. A., & Acquaye, C. (2021). State and future prospects of artificial intelligence (AI) in Ghana. *Proceedings of the 27th iSTEAMS Multidisciplinary & Inter-tertiary Research Conference*. https://doi.org/10.22624/aims/isteams-2021/v27p1.

Ameyaw, E. K., Budu, E., Sambah, F., & Baatiema, L. (2018). Prevalence and correlates of HIV testing among young women in Ghana. *PLOS ONE*, 13(10), e0205281. https://doi.org/10.1371/journal.pone.0205281

Application of artificial intelligence and machine learning for HIV prevention interventions. (2021). *The Lancet HIV*, 8(11), 661. https://doi.org/10.1016/S2352-3018(21)00247-2

Ba, D., Ssentongo, P., & Sznajder, K. (2019). Prevalence, behavioral and socioeconomic factors associated with human immunodeficiency virus in Ghana: A population-based cross-sectional study. *Journal of Global Health Reports*, 3. https://doi.org/10.29392/joghr.3.e2019092

Babanejaddehaki, G., An, A., & Davoudi, H. (2024). Ontology-based data collection for a hybrid outbreak detection method using social media. *IEEE Transactions on NanoBioscience*, 23(1), 591–602. https://doi.org/10.1109/TNB.2024.3442912

Bauermeister, J., Pingel, E., Zimmerman, M., Couper, M., Carballo-Diéguez, A., & Strecher, V. (2012). Data quality in HIV/AIDS web-based surveys. *Field Methods*, 24(3), 272–291. https://doi.org/10.1177/1525822X12443097

Beltrami, J., Wang, G., Usman, H., & Lin, L. S. (2017). Quality of HIV testing data before and after the implementation of a national data quality assessment and feedback system. *Journal of Public Health Management and Practice*, 23(3), 269–275. https://doi.org/10.1097/PHH.0000000000000376

Boutayeb, A. (2009). The impact of HIV/AIDS on human development in African countries. *BMC Public Health*, 9(S1), S3. https://doi.org/10.1186/1471-2458-9-S1-S3

Bulstra, C., Hontelez, J., Giardina, F., Steen, R., Nagelkerke, N., Bärnighausen, T., & de Vlas, S. D. (2020). Mapping and characterising areas with high levels of HIV transmission in sub-Saharan Africa: A geospatial analysis. *PLoS Medicine*, 17. https://doi.org/10.1371/journal.pmed.1003042

Clauset, A. (2021). Improved disease tracking with AI. *Science*, 372(6548), 1300. https://doi.org/10.1126/science.372.6548.1300-d

Darnag, R., Minaoui, B., & Fakir, M. (2017). QSAR models for prediction study of HIV protease inhibitors using support vector machines, neural networks and multiple linear regression. *Arabian Journal of Chemistry*, 10, S600–S605. https://doi.org/10.1016/j.arabjc.2012.10.021

Bosheng, D., Linlin, L., Lidong, B., Kruengkrai, C., Nguyen, T. H., Joty, S., Si, L., & Miao, C. (2020). DAGA: *Data augmentation with a generation approach for low-resource tagging tasks*. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6045–6057. https://doi.org/10.18653/v1/2020.emnlp-main.488

Fairclough, D. L., Thijs, H., Huang, I., Finnern, H., & Wu, A. (2008). Handling missing quality of life data in HIV clinical trials: What is practical? *Quality of Life Research*, 17, 61–73. https://doi.org/10.1007/s11136-007-9284-3

Dzinamarira, T., Mbunge, E., Chingombe, I., Cuadros, D. F., Moyo, E., Chitungo, I., & Musuka, G. (2024). Using machine learning models to plan HIV services: Emerging opportunities in design, implementation and evaluation. *South African Medical Journal*, 114(6b), e1439. https://doi.org/10.7196/SAMJ.2024.v114i6b.1439

Ebulue, C. C., Ekkeh, O. V., Ebulue, O. R., & Ekesiobi, C. S. (2024). Machine learning insights into HIV outbreak predictions in Sub-Saharan Africa. *International Medical Science Research Journal*. https://doi.org/10.51594/imsrj.v4i5.1121

Ebulue, C. C., Ekkeh, O. V., Ebulue, O. R., & Ekesiobi, C. S. (2024). Machine learning insights into HIV outbreak predictions in Sub-Saharan Africa. *International Medical Science Research Journal,* 2(1), 23–34. [preprint]

Fieggen, J., Smith, E., Arora, L., & Segal, B. M. (2022). The role of machine learning in HIV risk prediction. *Frontiers in Reproductive Health*, 4, 847382. https://doi.org/10.3389/frph.2022.847382

Gareta, D., Lauren, E., Shumba, K., Herbst, K., Onoya, D., & Bor, J. (2024). Developing an integrated HIV clinical and population database in rural South Africa. International *Journal of Population Data Science*, 9(5). https://doi.org/10.23889/ijpds.v9i5.2830

Ghanem, V. G. (2025). Analysis and Prediction of Infectious Disease Trends in Ghana: A Machine Learning Approach, with Emphasis on HIV/AIDS [Data and software repository]. Zenodo. https://doi.org/10.5281/zenodo.15292209.

Gyaase, P., Aanienang, E. K., Mensah, P., Adzordor, P., Tweneboah, S. K., Amoakoaa, E. A., ... & Sampson, D. (2024). Determinants of human immuno deficiency virus (HIV) prevalence in the Central Region of Ghana. *Asian Journal of Research in Infectious Diseases*. https://doi.org/10.9734/ajrid/2024/v15i11386

Hamilton, A., Strauss, A. T., Martínez, D., Hinson, J., Levin, S., Lin, G., & Klein, E. (2021). Machine learning and artificial intelligence: Applications in healthcare epidemiology. *Antimicrobial Stewardship & Healthcare Epidemiology (ASHE)*, 1. https://doi.org/10.1017/ash.2021.192

Holt, E. (2024). AI-assisted HIV case finding in Ukraine. *The Lancet Microbe*, 5(7), 631. https://doi.org/10.1016/s2666-5247(24)00077-6

Lu, X., Wang, L., & Jiang, Z. (2018). The application of deep learning in the prediction of HIV-1 protease cleavage site. *5th International Conference on Systems and Informatics (ICSAI)*, 1299–1304. IEEE. https://doi.org/10.1109/ICSAI.2018.8599496

Mahto, R. V., & Sood, K. (2024). HIV progression and outcome prediction to enhance patient matching for clinical trials. *IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 278–284). IEEE. https://doi.org/10.1109/CCWC57342.2024.10382198

Makota, R. B. B., & Musenge, E. (2023). Predicting HIV infection in the decade (2005–2015) pre-COVID-19 in Zimbabwe: A supervised classification-based machine learning approach. *PLOS Digital Health*, 2(6), e0000260. https://doi.org/10.1371/journal.pdig.0000260

Mazrouee, S., Little, S., & Wertheim, J. (2021). Incorporating metadata in HIV transmission network reconstruction: A machine learning feasibility assessment. *PLoS Computational Biology*, 17. https://doi.org/10.1371/journal.pcbi.1009336

Mbunge, E., & Batani, J. (2023). Application of deep learning and machine learning models to improve healthcare in sub-Saharan Africa: Emerging opportunities, trends and implications. *Telematics and Informatics Reports*, 11, 100097. https://doi.org/10.1016/j.teler.2023.100097

Mumuni, A., & Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array*, 16, 100258. https://doi.org/10.1016/j.array.2022.100258

Mutai, C. K., McSharry, P. E., Ngaruye, I., & Musabanganji, E. (2021). Use of machine learning techniques to identify HIV predictors for screening in sub-Saharan Africa. *BMC Medical Research Methodology*, 21(1), 159. https://doi.org/10.1186/s12874-021-01346-2

Ndayishimiyepas, P., Wilson, C., & Kimwele, M. W. (2022). A hybrid model for predicting missing records in data using XGBoost. *2022 IEEE International Symposium on Product Compliance Engineering - Asia (ISPCE-ASIA)*, 1–5. https://doi.org/10.1109/ISPCE-ASIA57917.2022.9971092

Nketiah-Amponsah, E., Codjoe, E., & Ampaw, S. (2018). HIV/AIDS awareness and knowledge among Ghanaian women of reproductive age: What are the correlates? *Journal of Asian and African Studies*, 54(2), 267–281. https://doi.org/10.1177/0021909618810037

Parums, D. (2023). Editorial: Infectious disease surveillance using artificial intelligence (AI) and its role in epidemic and pandemic preparedness. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, 29, e941209. https://doi.org/10.12659/MSM.941209

Rezaei, I., Amirshahi, S. H., & Mahbadi, A. A. (2023). Utilizing support vector and kernel ridge regression methods in spectral reconstruction. *Results in Optics*, 12, 100405. https://doi.org/10.1016/j.rio.2023.100405

Saldana, C. S., Burkhardt, E., Pennisi, A., Oliver, K., Olmstead, J., Holland, D. P., Gettings, J., Mauck, D., Austin, D., Wortley, P., & Ochoa, K. V. S. (2024). Development of a machine learning modeling tool for predicting human immunodeficiency virus incidence using public health data from a county in the southern United States. *Clinical Infectious Diseases*, *79*(3), 717–726. https://doi.org/10.1093/cid/ciae100

Tang, D., Jin, Y., Hu, X., & Lin, D. (2025). Study on the prediction performance of AIDS monthly incidence in Xinjiang based on time series and deep learning models. *BMC Public Health*, 25(1), 780. https://doi.org/10.1186/s12889-025-21982-3

Tetteh, J., Ntumy, M. Y., Asamoah, I. E., Ekem-Ferguson, G., Swaray, S. M., Yamson, P., Addo, S. A., Torpey, K., & Yawson, A. E. (2024). Risky sexual behavior and HIV infection among Ghanaians aged 15–49: evidence from the Ghana Demographic Health Survey 2003 and 2014. *Journal of Psychosexual Health*, *6*(3), 265–273. https://doi.org/10.1177/26318318241276989

United Nations Development Programme. (2015). *Sustainable Development Goals*. https://www.undp.org/sustainable-development-goals

United Nations. (n.d.). *Countries commit to action to end AIDS by 2030*. https://www.un.org/en/academic-impact/countries-commit-action-end-aids-2030

United Nations. (n.d.). *Sustainable Development Goal 3*. Sustainable Development Goals Knowledge Platform. https://sdgs.un.org/goals/goal3

World Health Organization. (2023). *Health Topics (Ghana)*. WHO Regional Office for Africa. https://www.afro.who.int/countries/ghana

World Health Organization. (2024). *HIV data and statistics*. https://www.who.int/teams/global-hiv-hepatitis-and-stis-programmes/hiv/strategic-information/hiv-data-and-statistics

World Health Organization. (2024). *Proportion of people living with HIV with suppressed viral load among those on antiretroviral therapy (ART)*. Global Health Observatory. https://www.who.int/data/gho/indicator-metadata-registry/imr-details/4752

Xiang, Y., Du, J., Fujimoto, K., Li, F., Schneider, J., & Tao, C. (2021). Application of artificial intelligence and machine learning for HIV prevention interventions. *The Lancet HIV*, 9(1), e54–e62. https://doi.org/10.1016/s2352-3018(21)00247-2

Yang, X., Cai, R., Ma, Y., Zhang, H. H., Sun, X., Olatosi, B., et al. (2024). Using machine learning techniques to predict viral suppression among people with HIV. *Journal of Acquired Immune Deficiency Syndromes*, 95(2), 187–195. https://doi.org/10.1097/QAI.0000000000003226

Zare, A., Bajestani, N. S., & Khandehroo, M. (2024). Machine learning in public health. Journal of Research and Health, 14(03), 207–208. https://doi.org/10.32598/jrh.14.3.2417.1

# APPENDIX 1

Tables

**Table 12**

*Detailed Data Quality Assessment Summary*

| Fetures | Data Type | Missing Values | Missing % | Unique Values | Outliers Detcted | Post Winsorisation |
|---|---|---|---|---|---|---|
| region | object | 0 | 0 | 10 | 64 | 0 |
| date | object | 0 | 0 | 612 | 9 | 0 |
| tb_outlier | bool | 0 | 0 | 1 | 0 | 0 |
| malaria_outlier | bool | 0 | 0 | 2 | 379 | 0 |
| urban_rural_sum_pct | float64 | 0 | 0 | 1 | 0 | 0 |
| age_sum_pct | float64 | 0 | 0 | 3 | 57 | 0 |
| migration_rate | float64 | 0 | 0 | 9261 | N/A | 0 |
| regional_stigma_index | float64 | 0 | 0 | 6579 | 0 | 0 |
| urbanization_level | float64 | 0 | 0 | 9792 | 112 | 0 |
| health_facility_density | float64 | 0 | 0 | 9761 | N/A | 0 |
| testing_coverage_pct | float64 | 0 | 0 | 9792 | N/A | 0 |
| access_to_art_pct | float64 | 0 | 0 | 9792 | 0 | 0 |
| hiv_awareness_index | float64 | 0 | 0 | 9792 | 0 | 0 |
| youth_unemployment_rate | float64 | 0 | 0 | 9792 | 0 | 0 |
| female_literacy_rate | float64 | 0 | 0 | 9792 | 52 | 0 |
| condom_use_rate | float64 | 0 | 0 | 9792 | 69 | 0 |
| education_access_index | float64 | 0 | 0 | 9792 | 45 | 0 |
| region_fixed | object | 0 | 0 | 10 | 0 | 0 |
| population_rural_pct | float64 | 0 | 0 | 9531 | 52 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **population_urban_pct** | float64 | 0 | 0 | 9531 | N/A | 0 |
| **population_15_64_pct** | float64 | 0 | 0 | 9792 | 334 | 0 |
| **population_65_plus_pct** | float64 | 0 | 0 | 9781 | N/A | 0 |
| **population_0_14_pct** | float64 | 0 | 0 | 9751 | 0 | 0 |
| **population_total** | float64 | 0 | 0 | 9631 | 44 | 0 |
| **hiv_incidence** | float64 | 0 | 0 | 9598 | 53 | 0 |
| **tb_incidence** | float64 | 0 | 0 | 9598 | 0 | 0 |
| **malaria_incidence** | float64 | 0 | 0 | 9598 | 59 | 0 |
| **month** | int64 | 0 | 0 | 12 | N/A | 0 |
| **year** | int64 | 0 | 0 | 51 | 92 | 0 |
| **hiv_outlier** | bool | 0 | 0 | 1 | 2 | 0 |

**Table 13**

*Key Feature Importance and SHAP Summary for hiv_incidence*

| Feature | SHAP Score | Key Interaction Highlighted |
|---|---|---|
| SHAP for **condom_use_rate** | 0.127 | Positive effect increases with literacy |
| SHAP for **education_access_index** | 0.241 | Nonlinear jump effect around score ~55 |
| SHAP for **hiv_awareness_index** | 0.198 | Steep increase above awareness ~65 |
| SHAP for **tb_incidence** | 0.113 | Moderate rise, esp. with higher urbanization |

**Table 14 A.**

*Spatial Analysis: HIV & TB Summary*

| Region | HIV Incidence (Mean) | Std_ Dev | Min | Max | TB Incidence (Mean) |
|---|---|---|---|---|---|
| Ashanti | 224.1 | 33.71 | 154.35 | 288.78 | 159.89 |
| Brong-Ahafo | 199.97 | 29.76 | 140.46 | 260.1 | 147.87 |
| Central | 231.78 | 35.44 | 163.72 | 298.04 | 152 |
| Eastern | 207.73 | 31.01 | 148.43 | 267.18 | 144.06 |
| Greater Accra | 254.44 | 35.92 | 184.7 | 300.98 | 167.26 |
| Northern | 144 | 21.41 | 97.13 | 188.14 | 136.07 |
| Upper East | 120.46 | 17.44 | 94.77 | 155.66 | 128.23 |
| Upper West | 120.27 | 17.53 | 94.77 | 154.8 | 128.29 |
| Volta | 191.92 | 28.69 | 132.32 | 251.86 | 148.01 |

**Table 14 B.**

*Spatial Analysis: Malaria, Education, Urbanization Summary*

| Region | Malaria Incidence (Mean) | Education Access Index (Mean) | Urbanization Level (Mean) |
|---|---|---|---|
| Ashanti | 199.64 | 56.84 | 62.94 |
| Brong-Ahafo | 216.1 | 53.38 | 50.76 |
| Central | 207.88 | 54.21 | 48.15 |
| Eastern | 199.9 | 58.66 | 56.87 |
| Greater Accra | 129.12 | 65.6 | 78.71 |
| Northern | 304.14 | 43.76 | 39.34 |
| Upper East | 359.51 | 48.11 | 34.98 |
| Upper West | 367.21 | 45.52 | 36.72 |
| Volta | 216.06 | 52.49 | 45.47 |
| Western | 224.07 | 55.16 | 52.52 |

## APPENDIX 2

Charts

**Figure 1:**

*Anomaly Detection in HIV Incidence (Z-Score Method)*

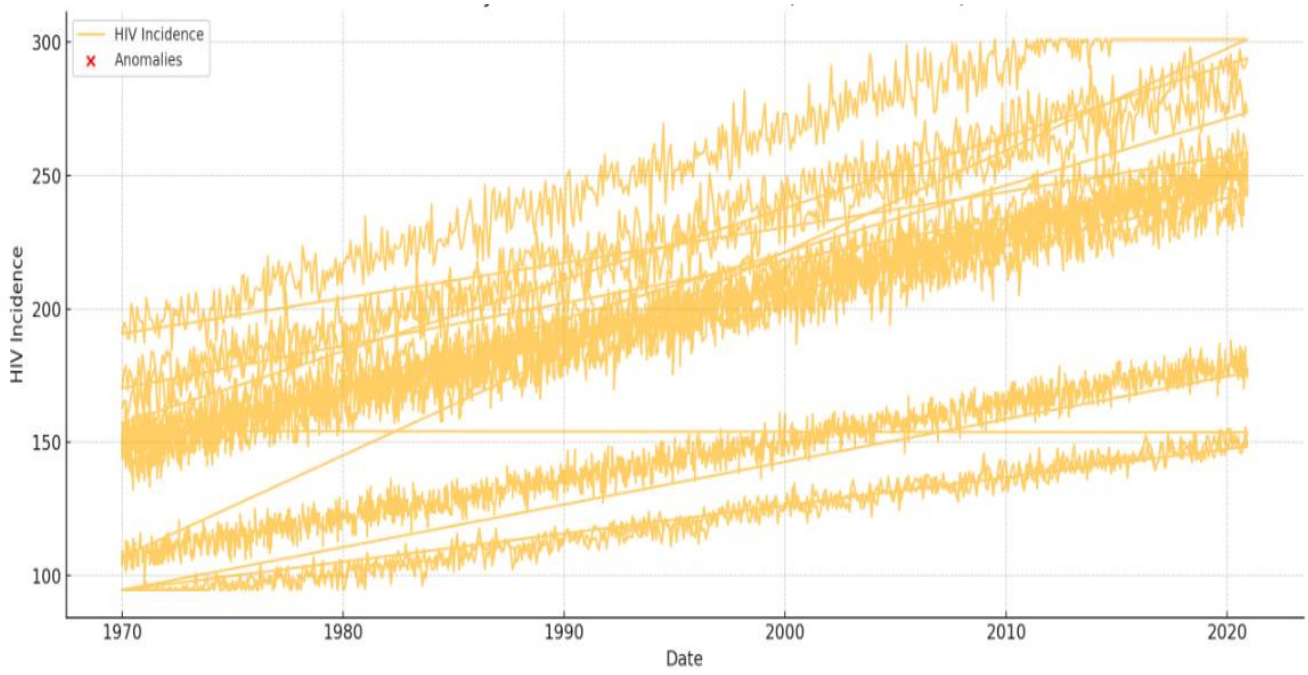



**Figure 2:**

*ART Coverage Vs HIV Incidence By Region*

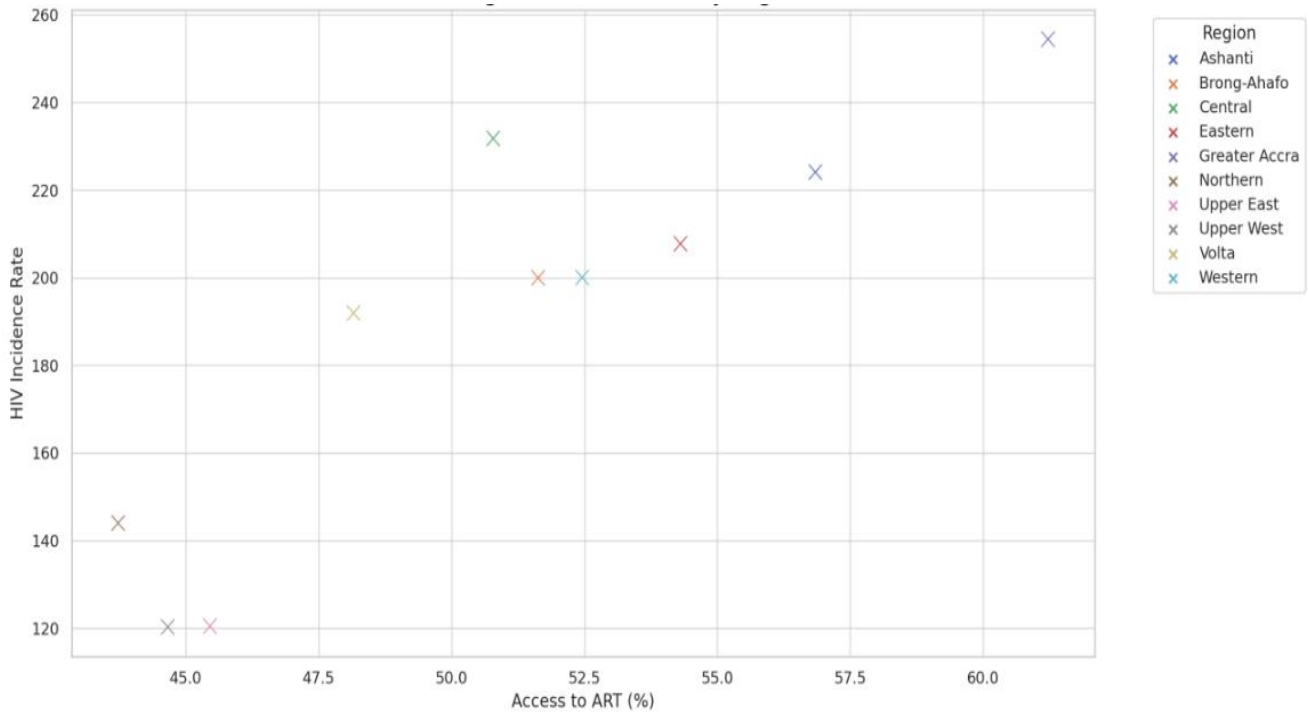


**Figure 3:**

*Average Disease Incidence By Region (HIV, Malaria, TB)*

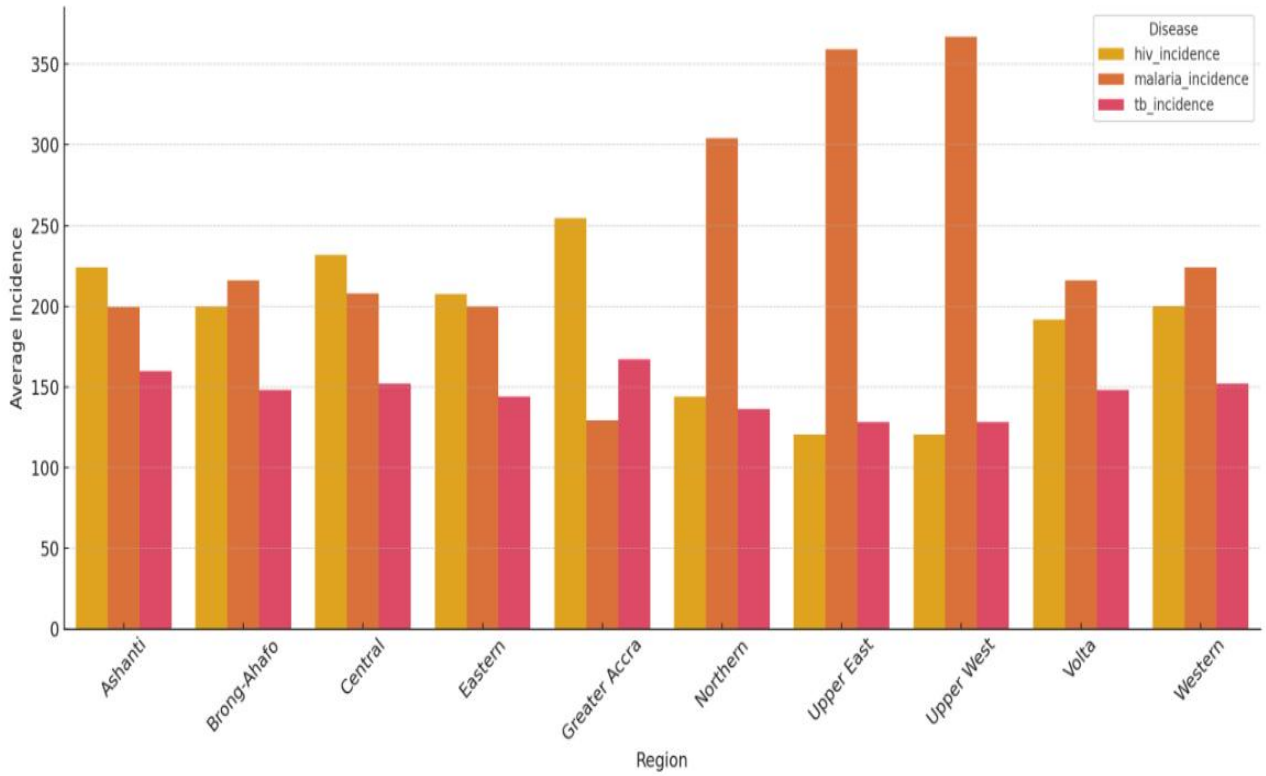



**Figure 4:**

*Average HIV Incidence Per 100,000 Population By Region*

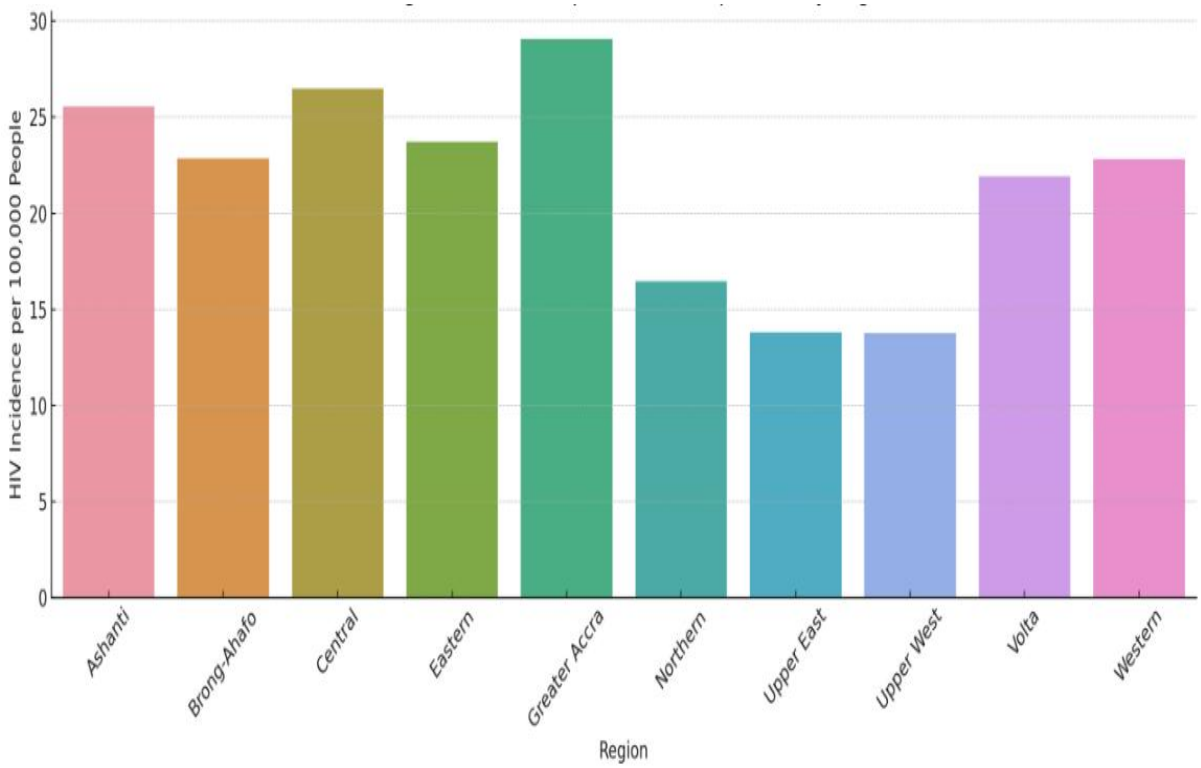


**Figure 5:**

*HIV Incidence Across Ghanaian Regions*

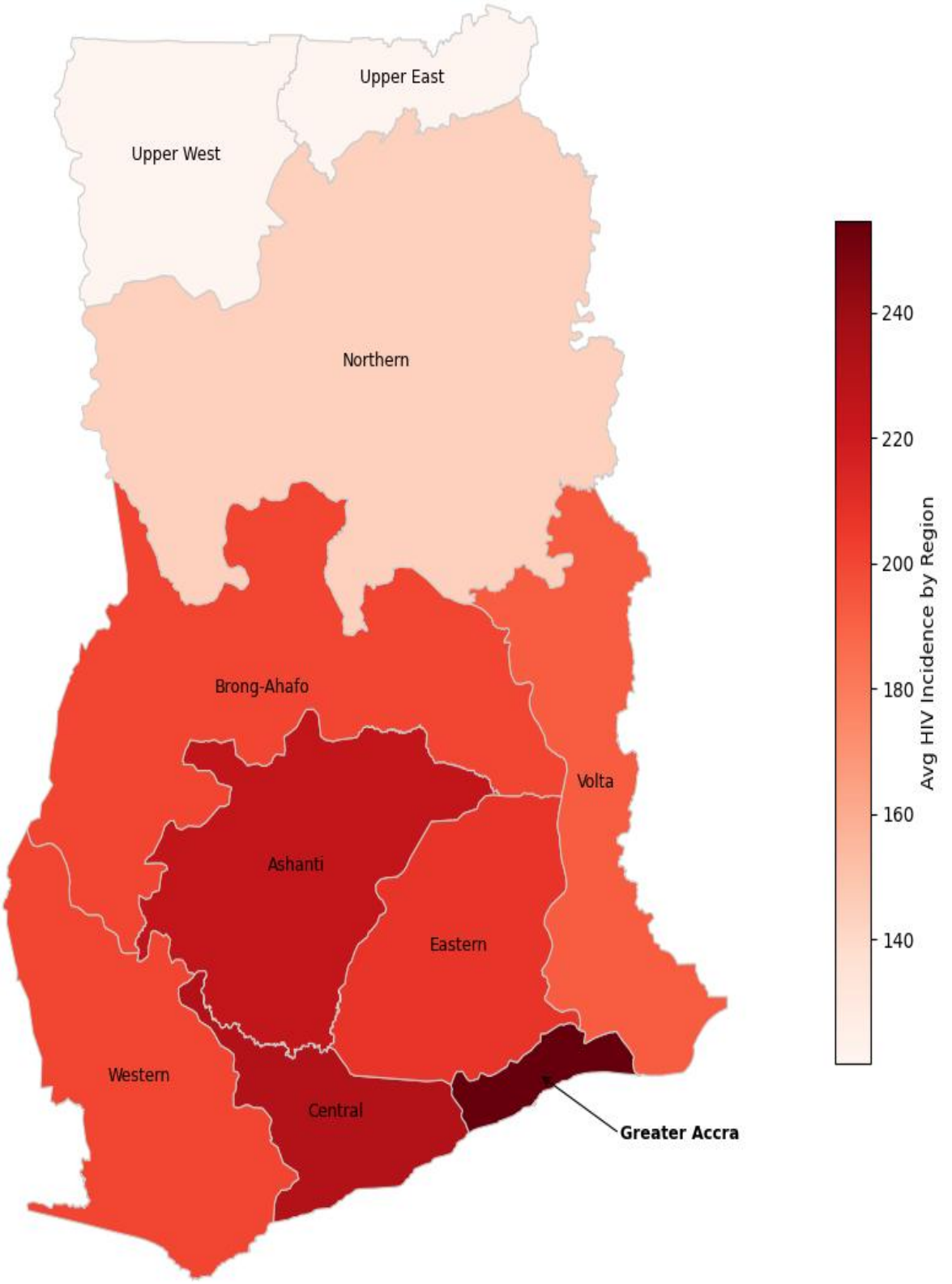


**Figure 6:**

*Clustering Of Regions Based On Socio-Health Indicators*

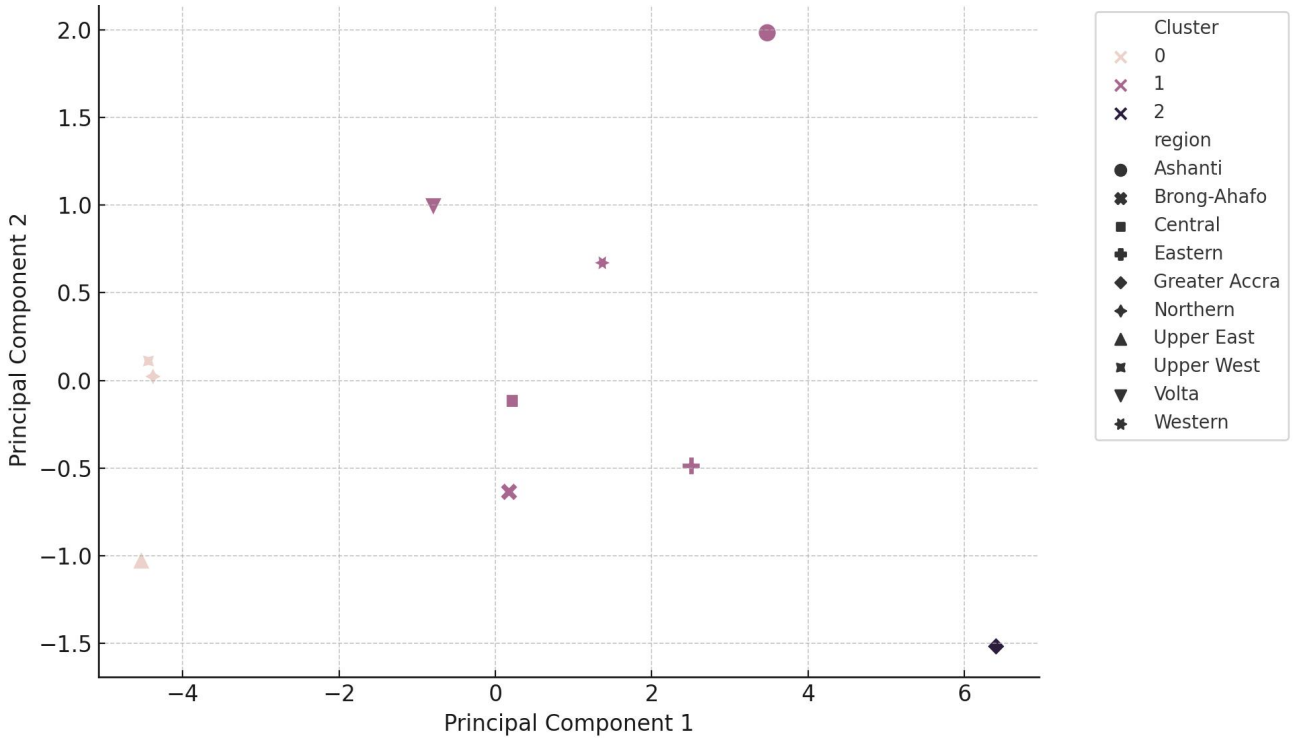



**Figure 7:**

*Comparative Analysis HIV Incidence Vs Malaria Incidence*

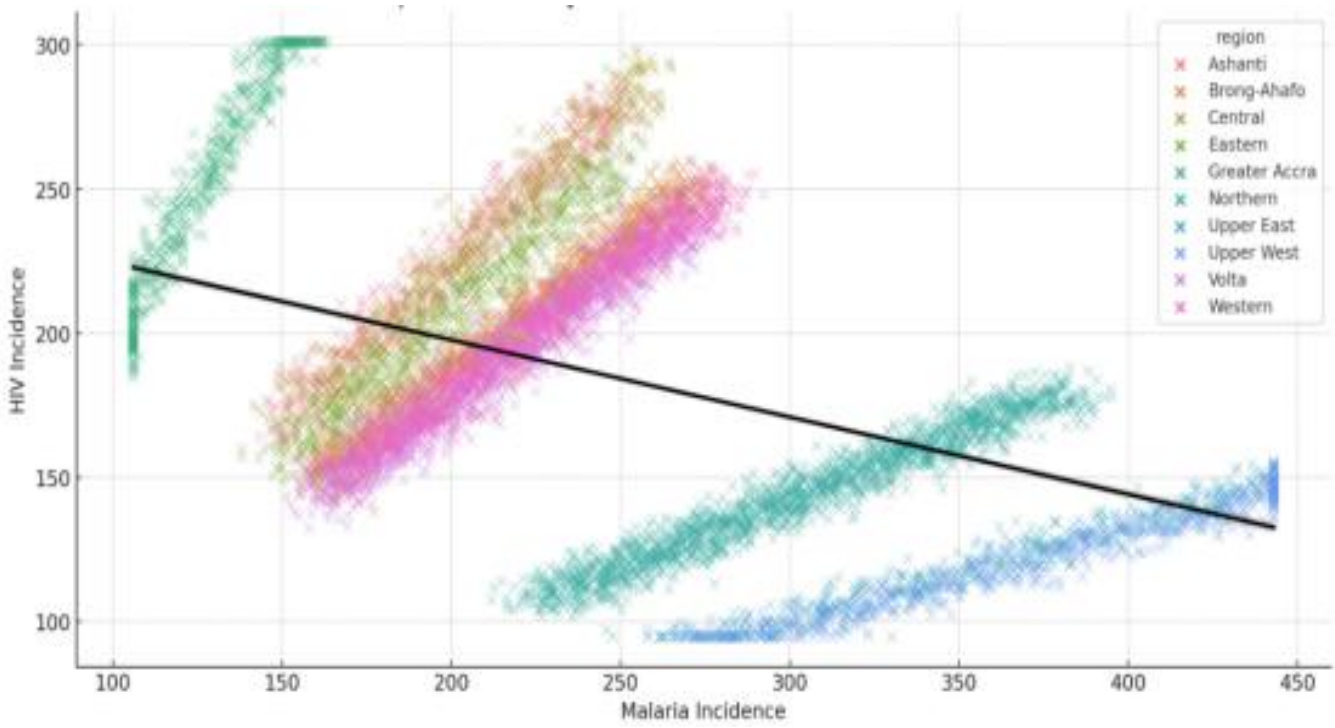


**Figure 8:**

*Comparative Analysis HIV Incidence Vs TB Incidence*

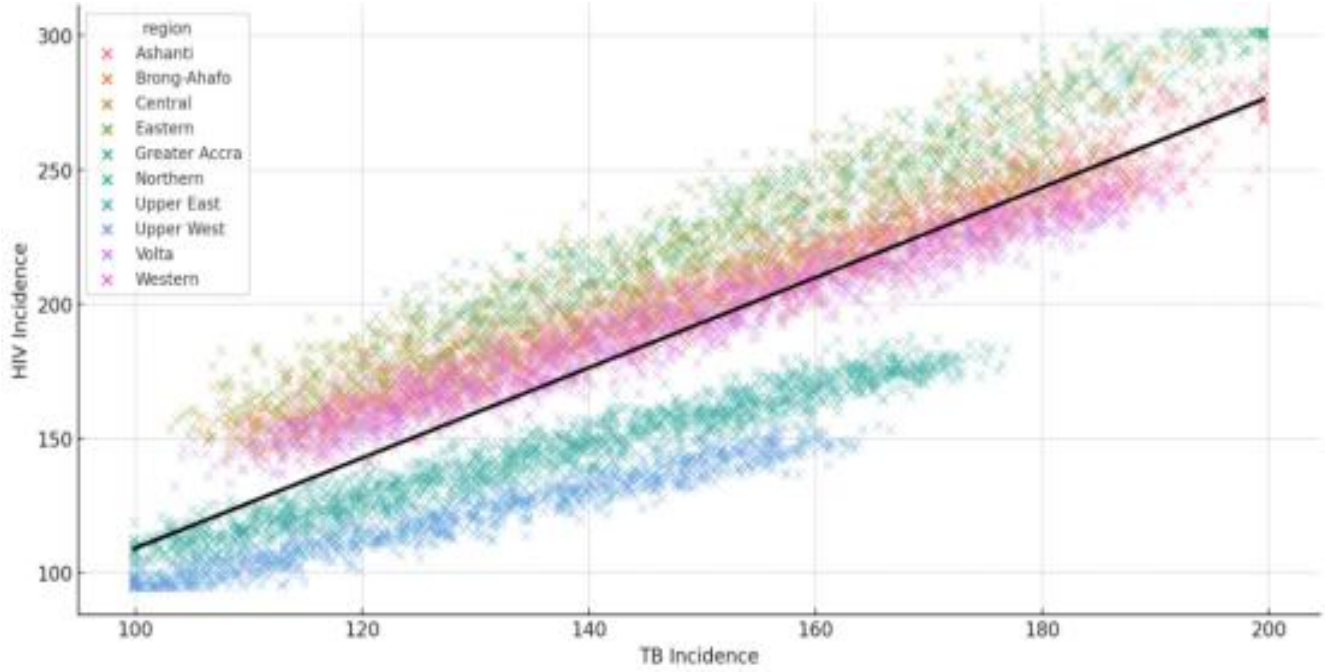



**Figure 9:**

*Correlation Heatmap Disease Incidences & Socio-Health Indicators*

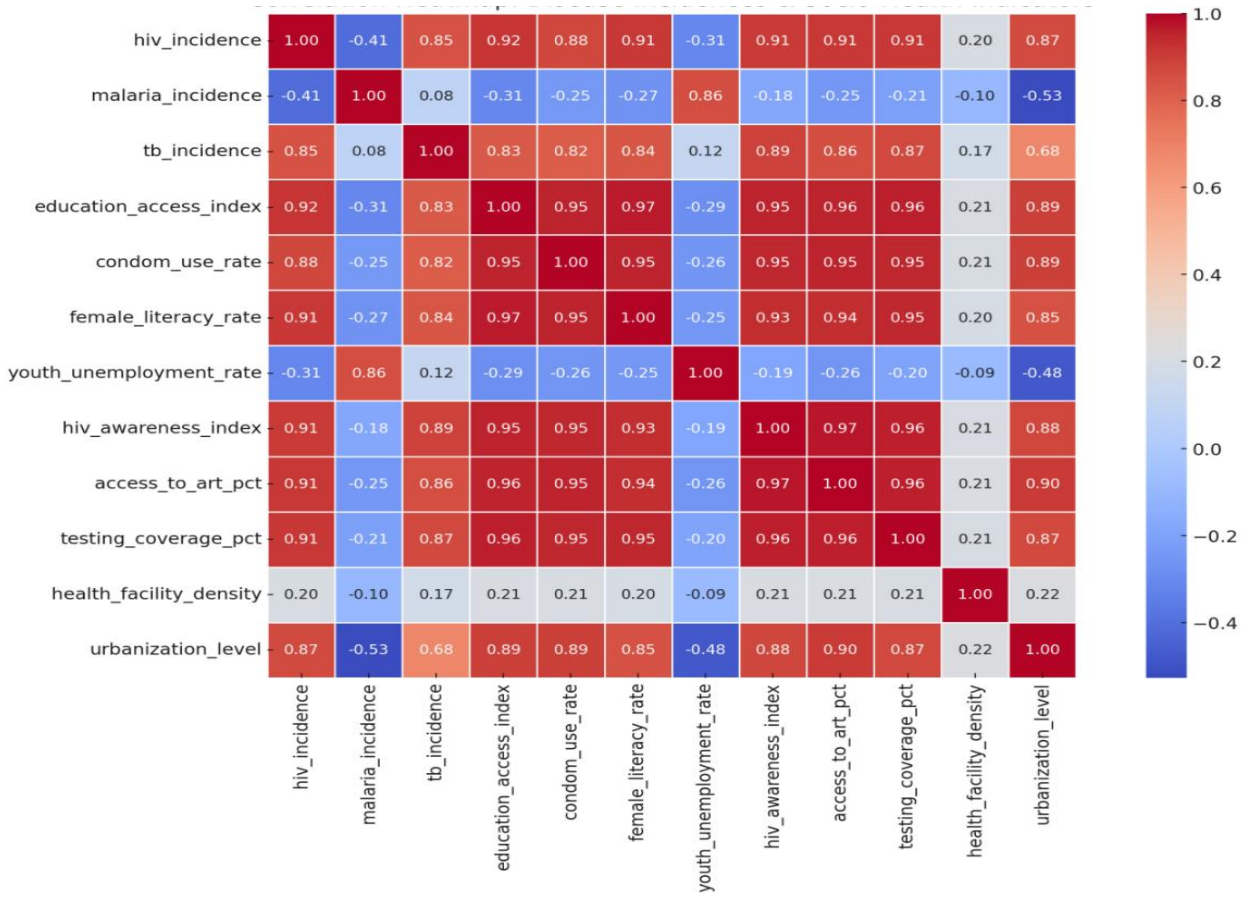


**Figure 11:**

*Education Access Index Vs HIV Incidence*

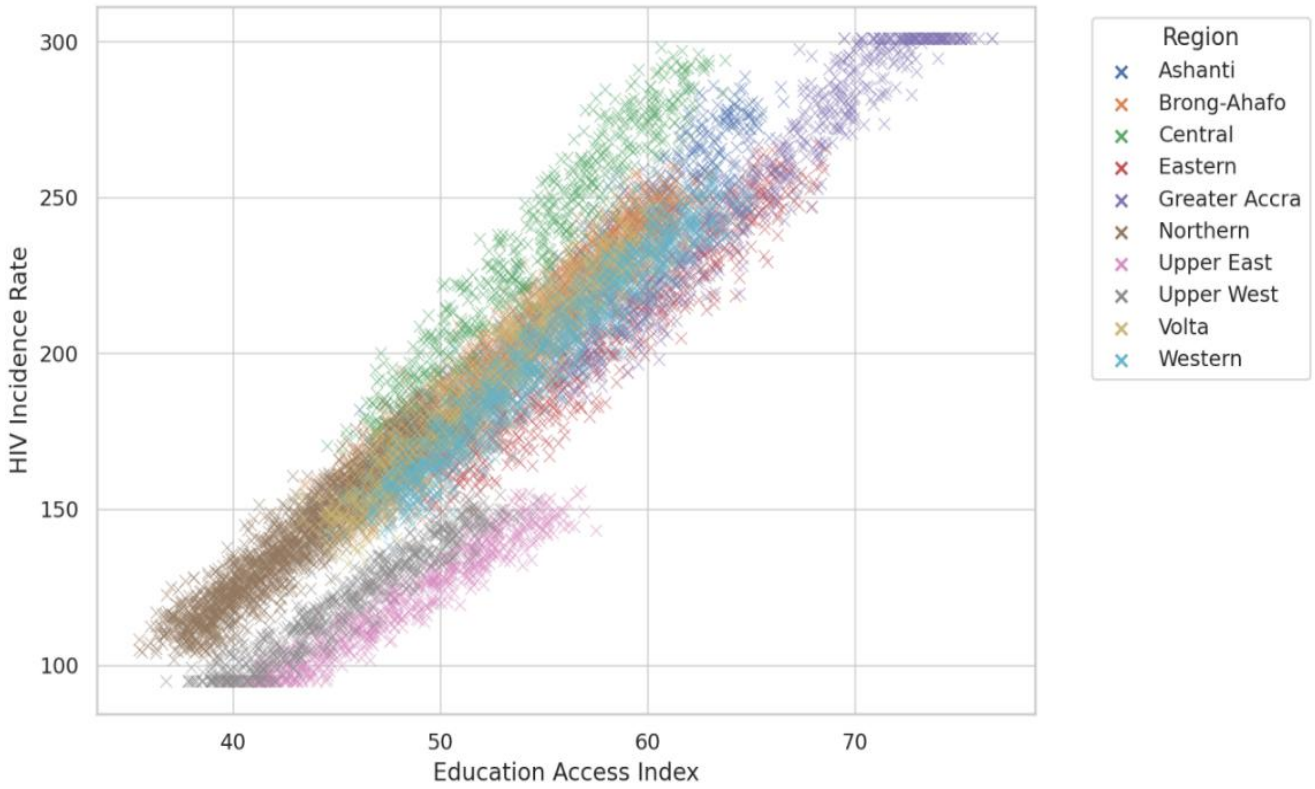


**Figure 13:**

*Estimated HIV Incidence By Gender (Proxy Analysis)*

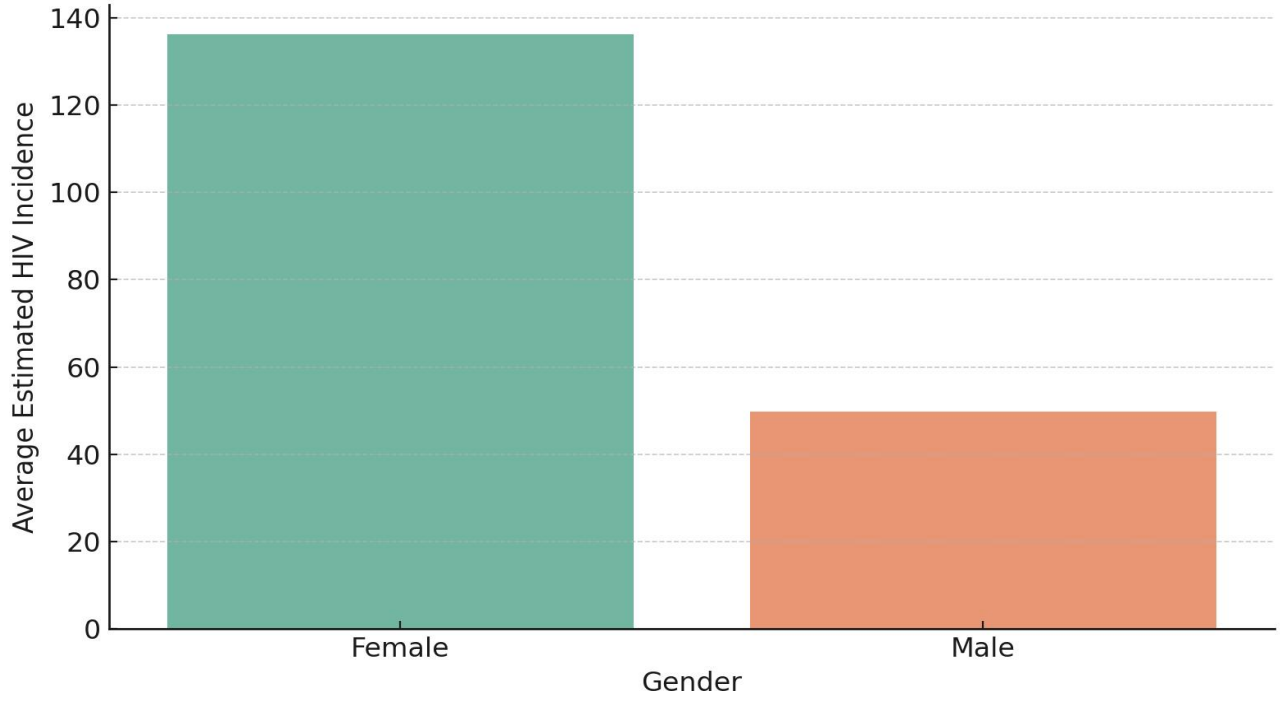



**Figure 15:**

*Forecast Of HIV Incidence In Ghana (Next 3 Years, ARIMA Model)*

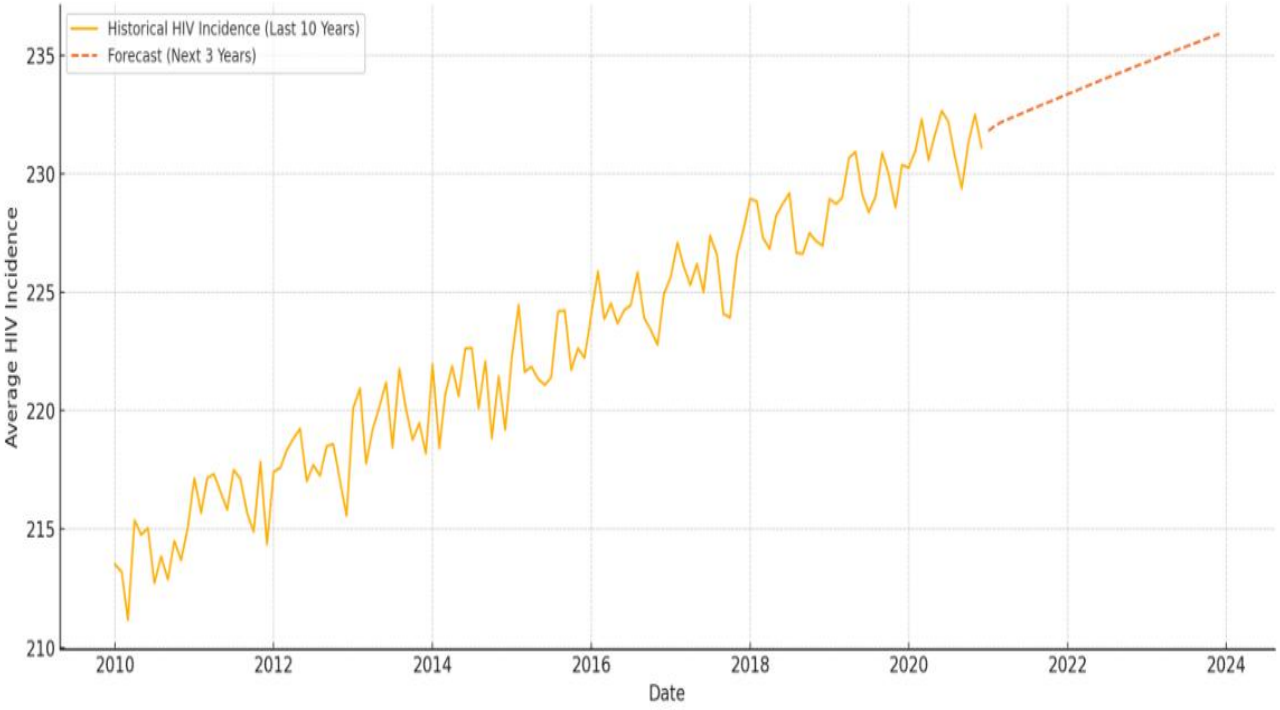


**Figure 17:**

*HIV Incidence By Age Group And Region*

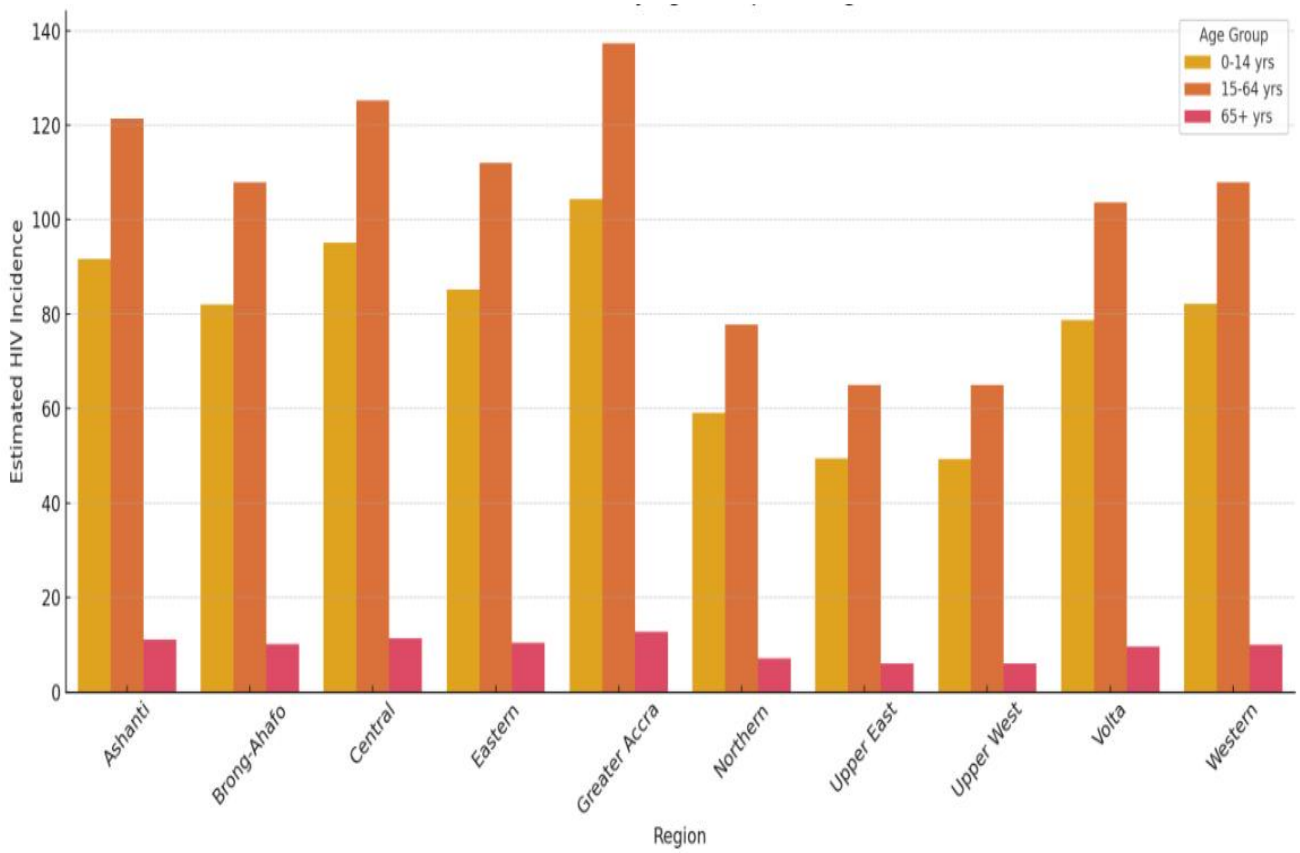



**Figure 18:**

HIV Incidence Distribution By Region

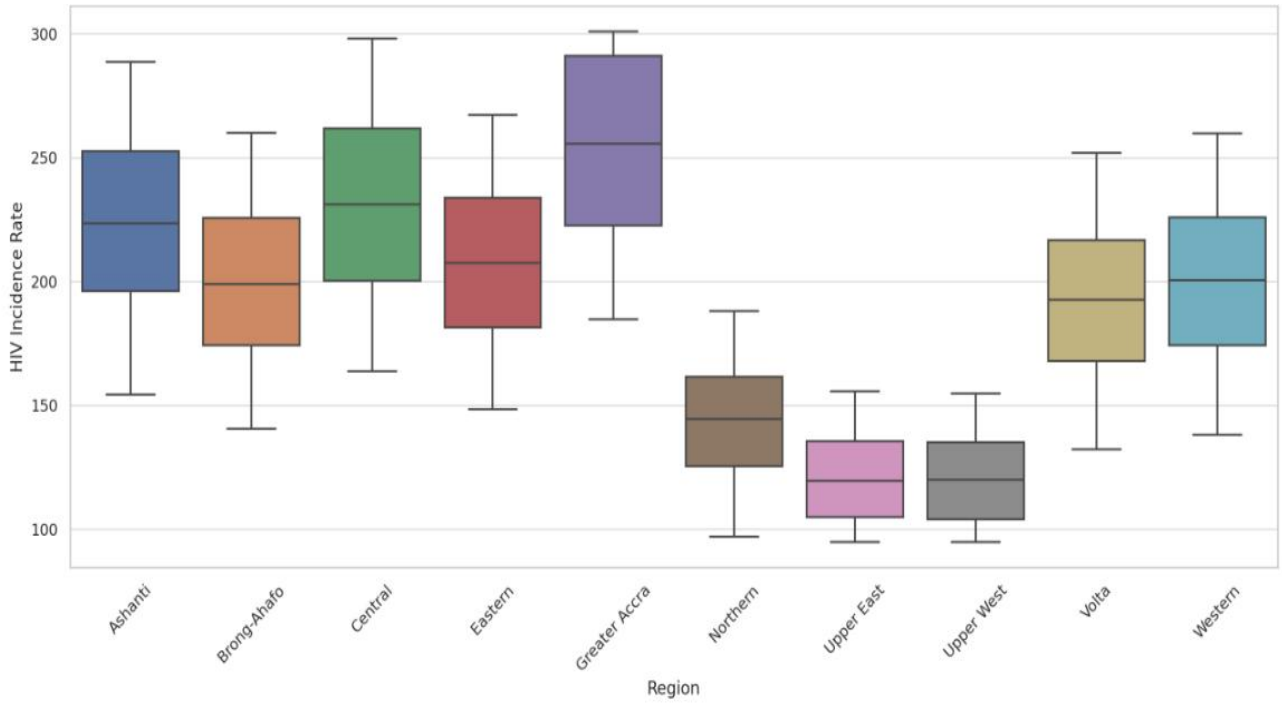


**Figure 19:**

*HIV Incidence Trends (Grouped Regions 1)*

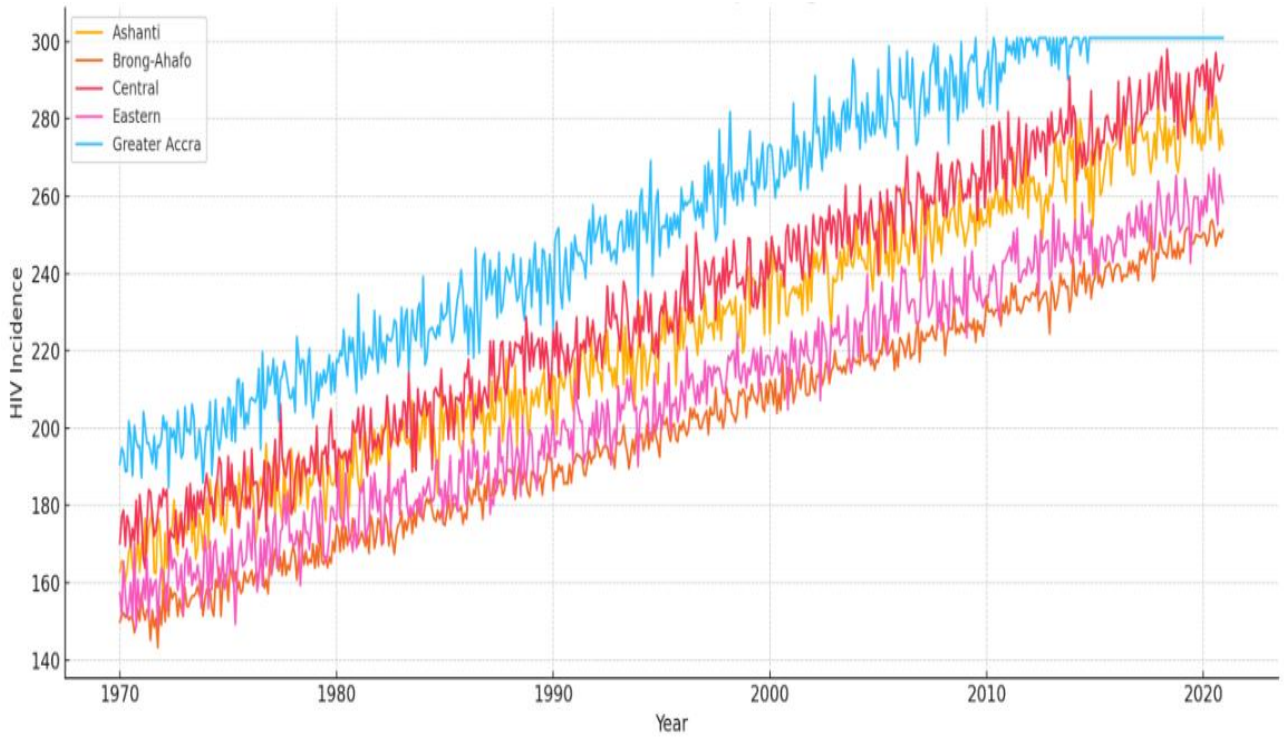



**Figure 20:**

*HIV Incidence Trends (Grouped Regions 2)*

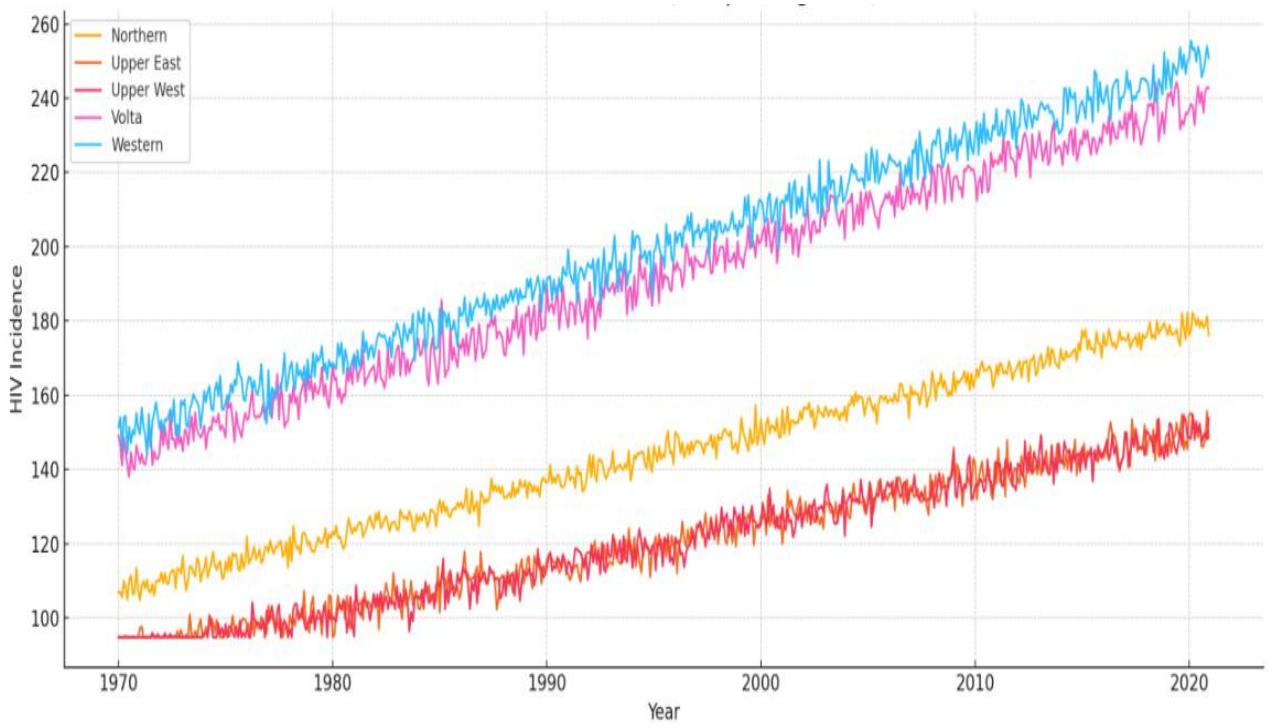


**Figure 21:**

*Infectious Disease Outliers (Malaria, TB, HIV)*

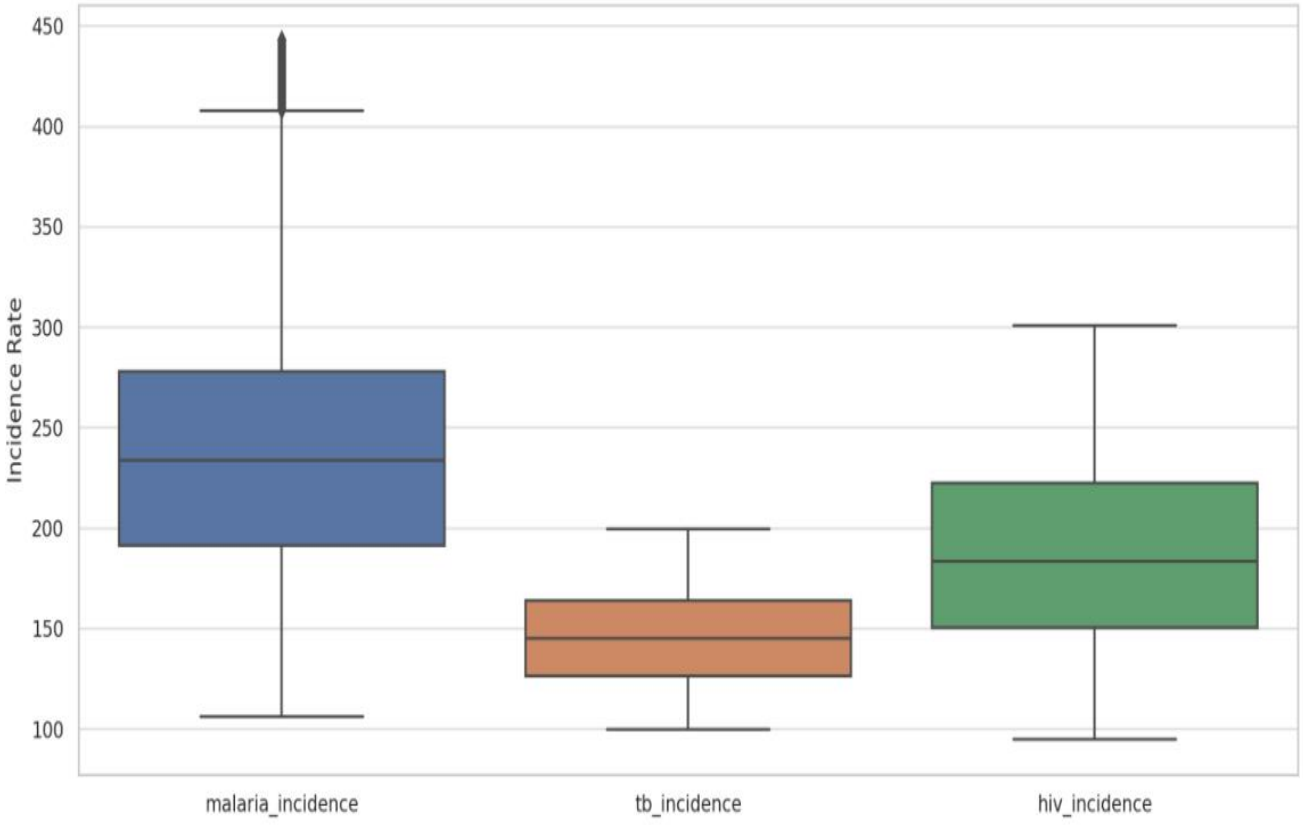


**Figure 23:**

*Predicted HIV Prevalence Across Ghanaian Regions (2030)*

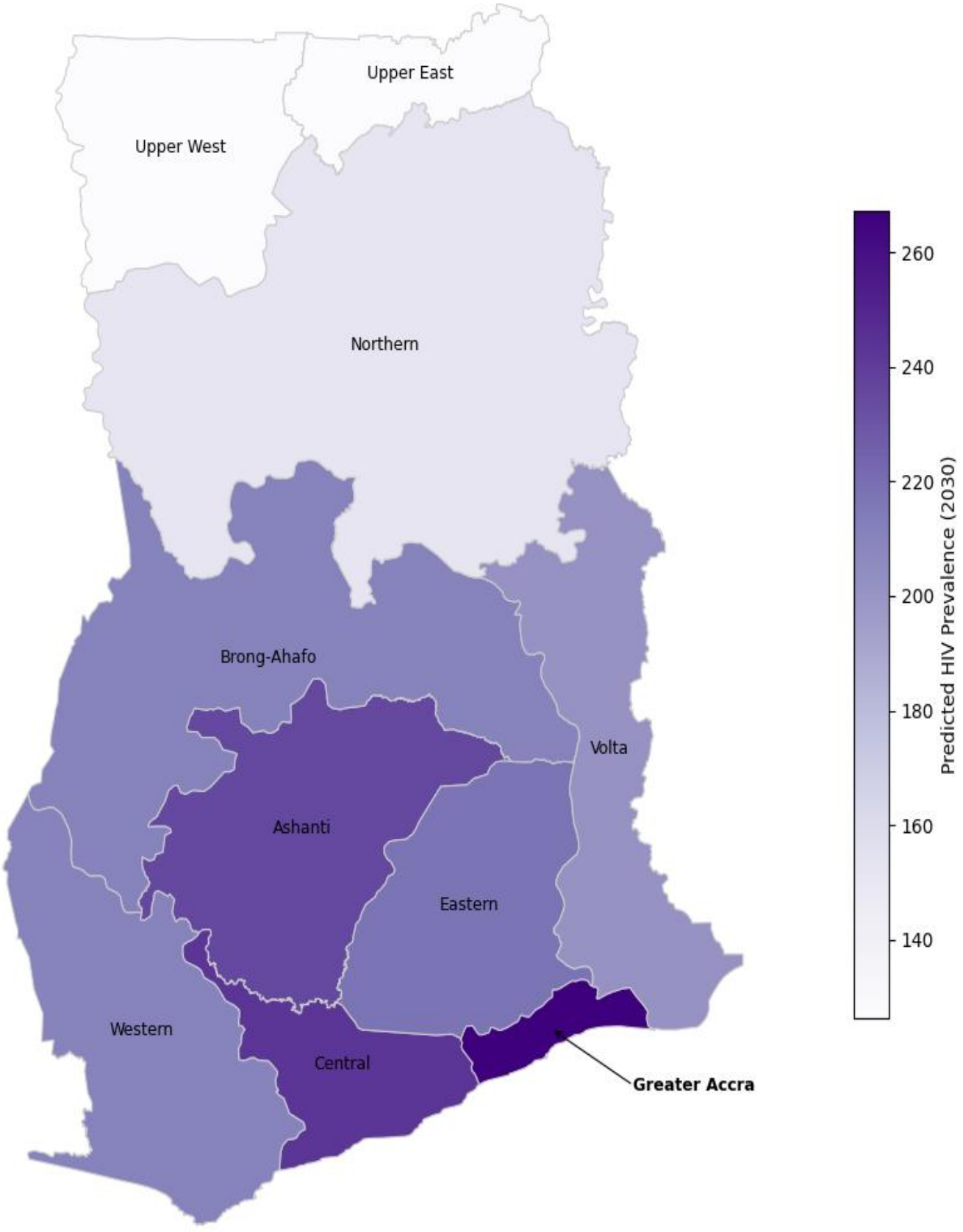


**Figure 24:**

*Regression Urbanization Vs HIV Incidence*

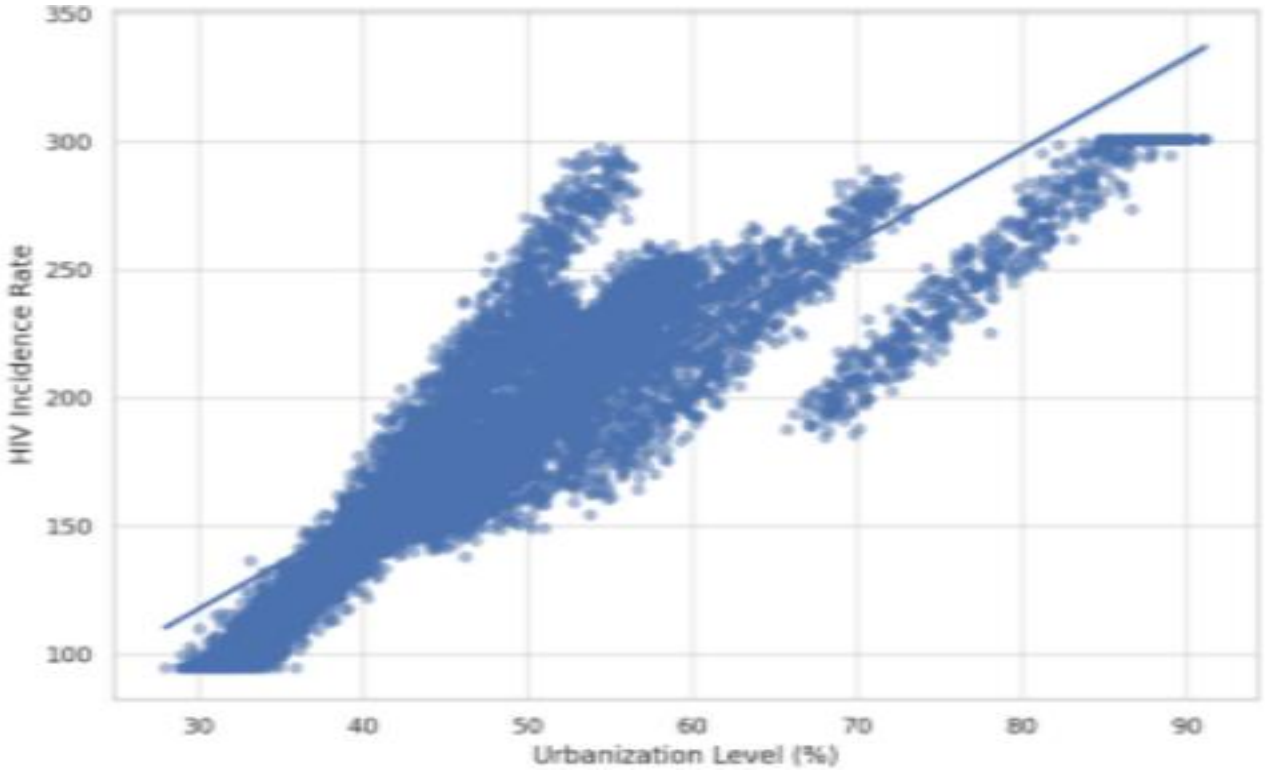



**Figure 25:**

*Static Clustering Map: Regions By HIV Incidence & Factors*

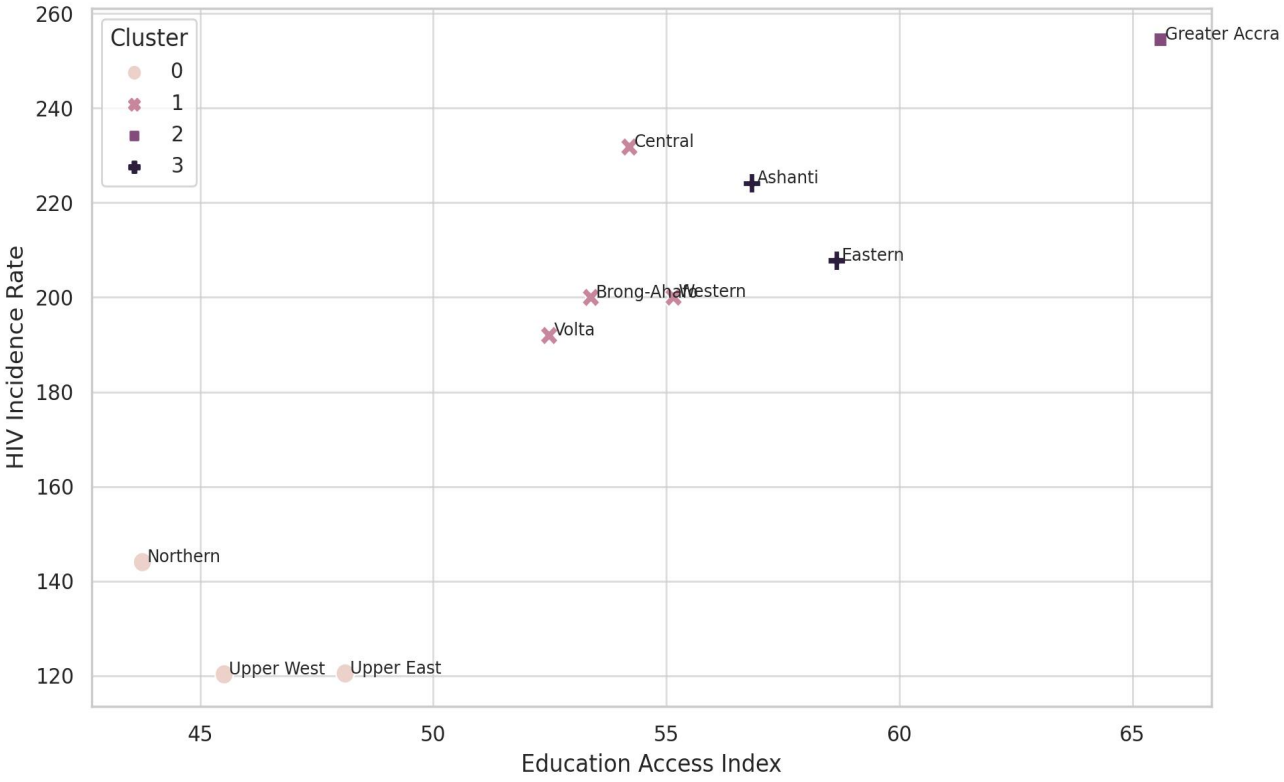


**Figure 26:**

*Urbanization Vs HIV Incidence*

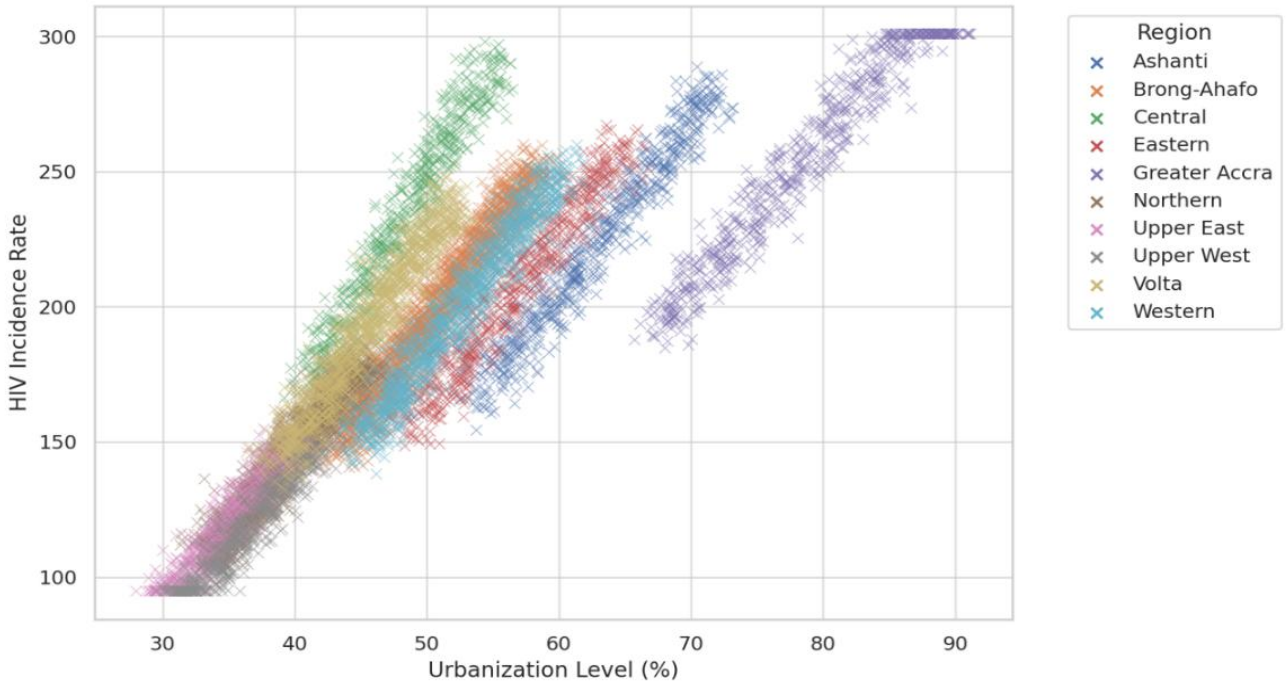



**Figure 27:**

*Average Residuals By Region*

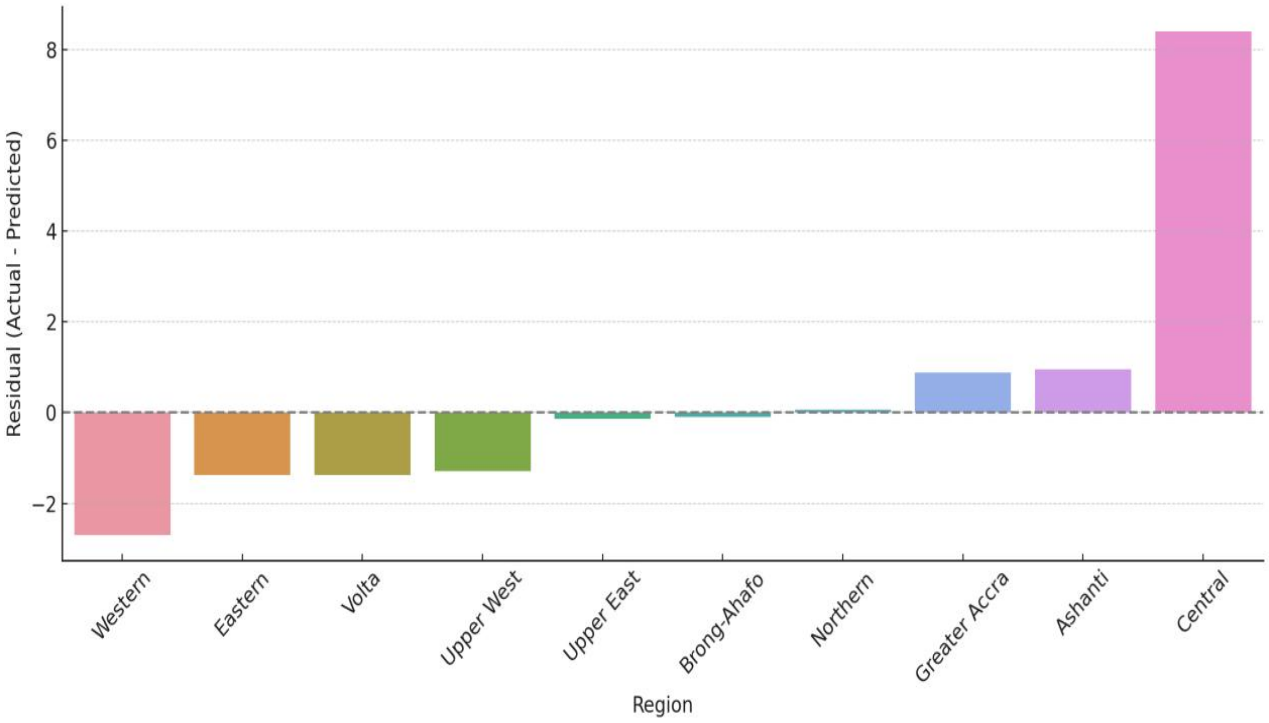


**Figure 28:**

*Clustering Of Regions Based On Health And Demographic Features*

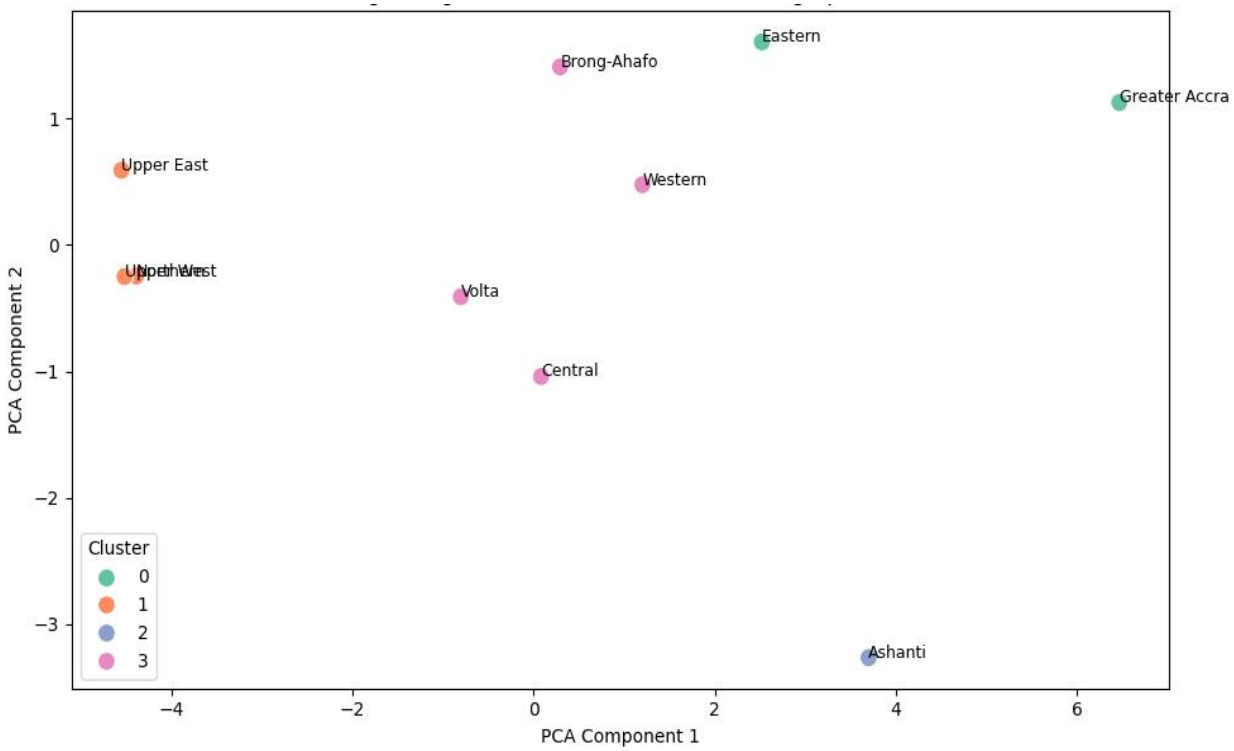



**Figure 29:**

*Counterfactual Simulation Education and Awareness Impact On HIV*

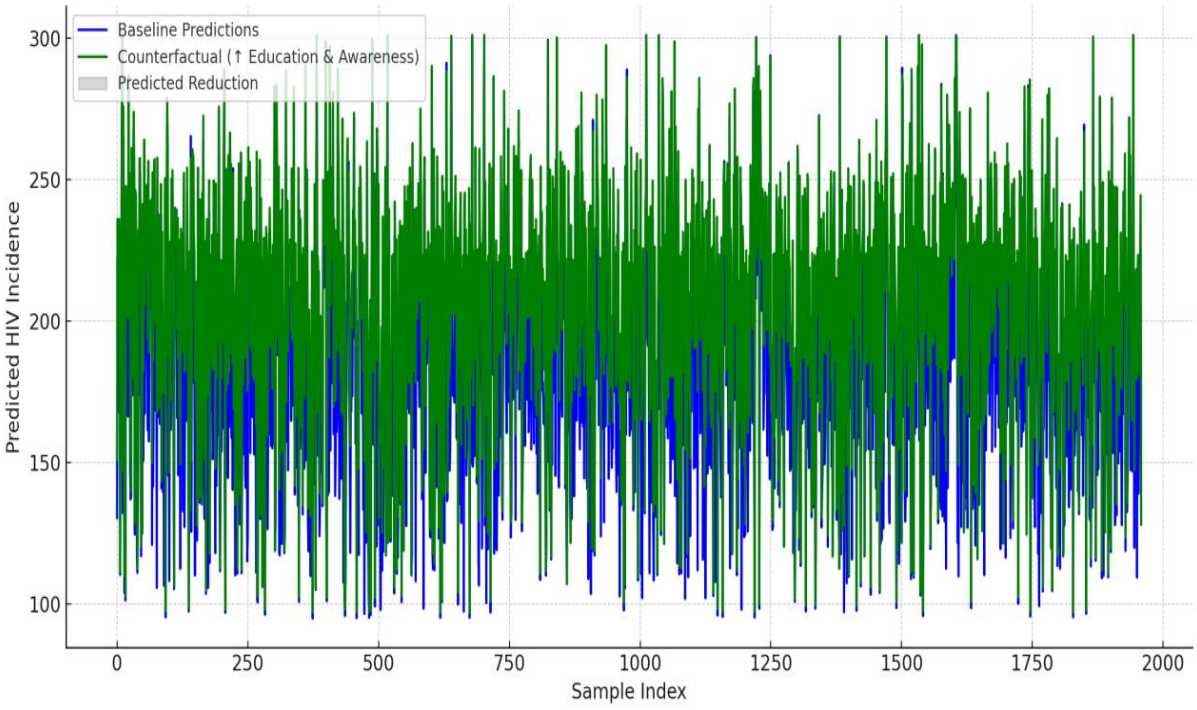


**Figure 30:**

*Feature Importance – Random Forest*

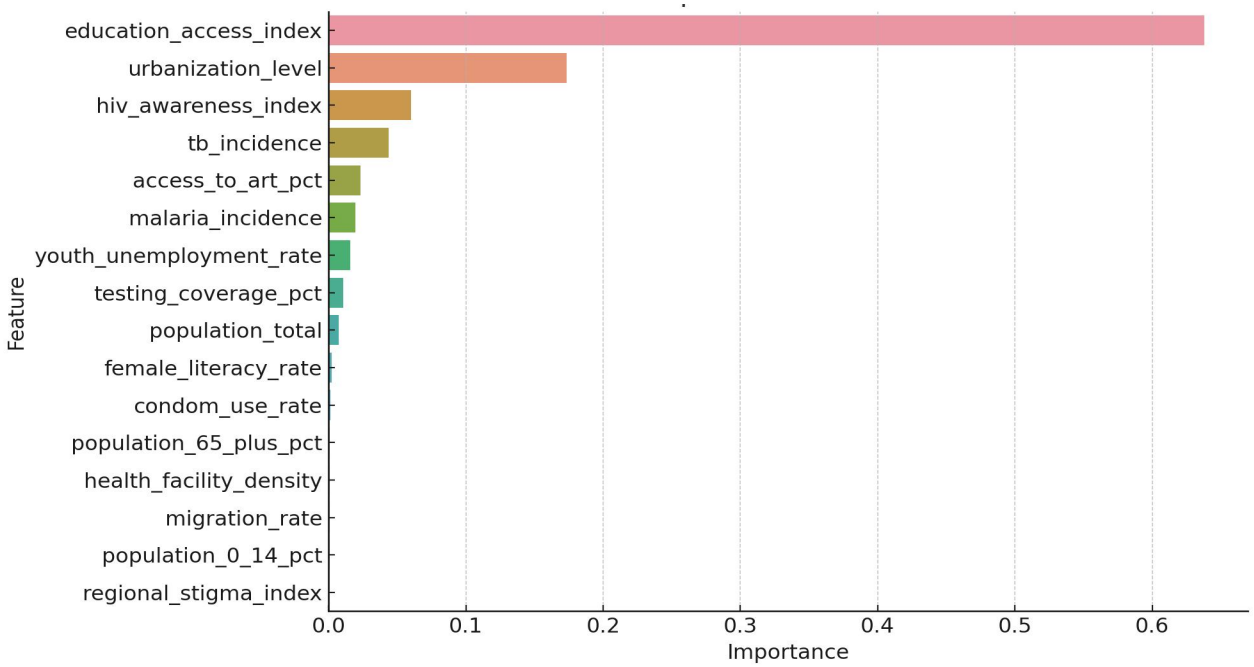



**Figure 31:**

*Forecast HIV Incidence to 2030: Ensemble, XGBoost, RF, SVR, Ridge*

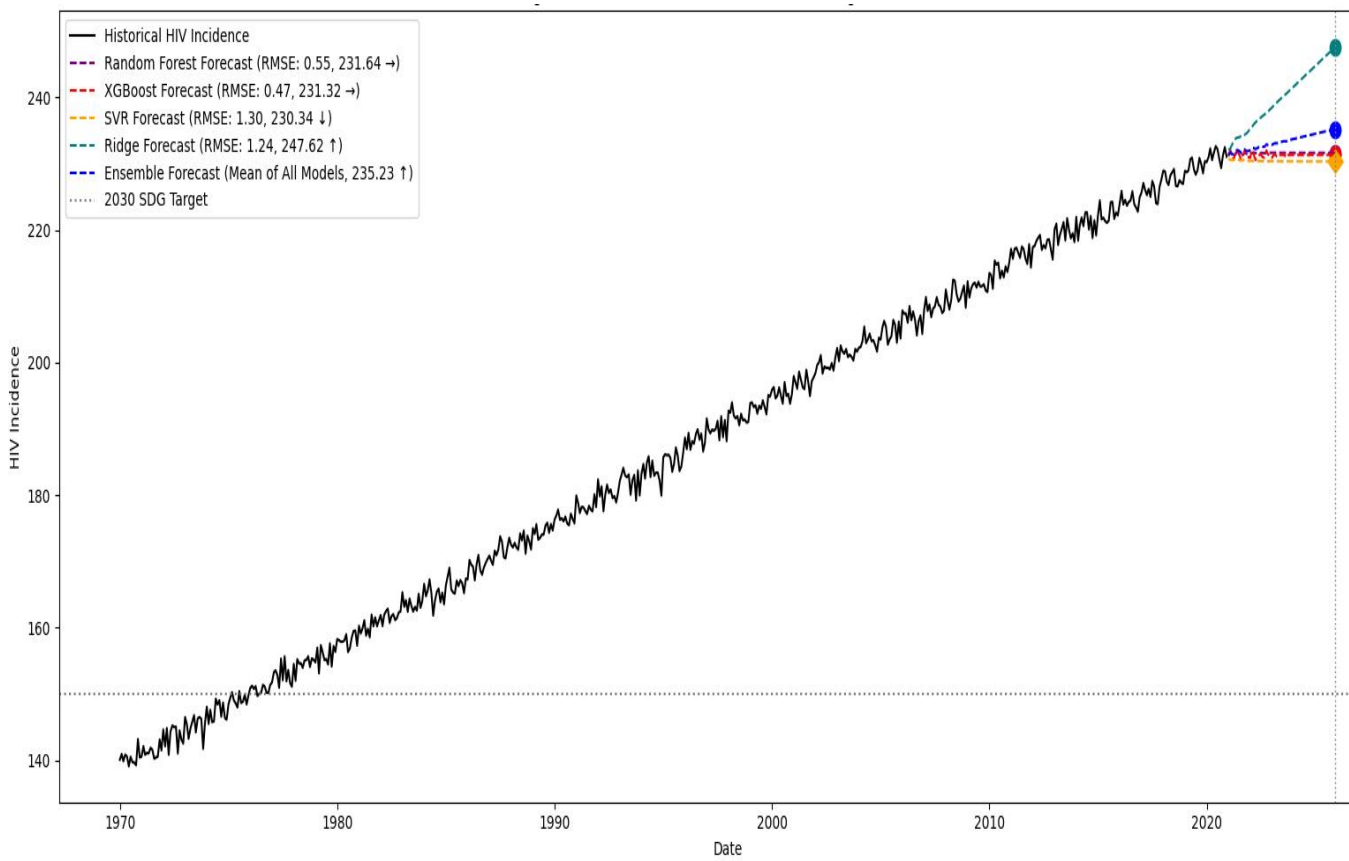


**Figure 32:**

*Histogram Of Residuals*

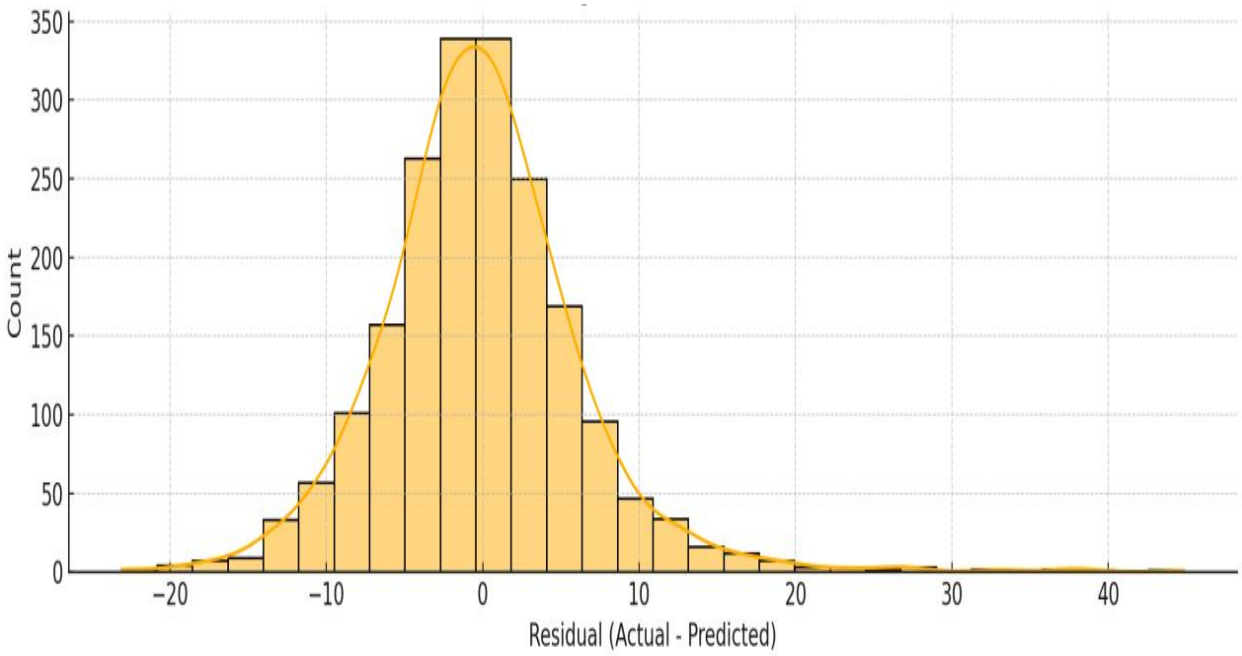



**Figure 36:**

*SHAP Dependence Analysis For HIV Awareness*

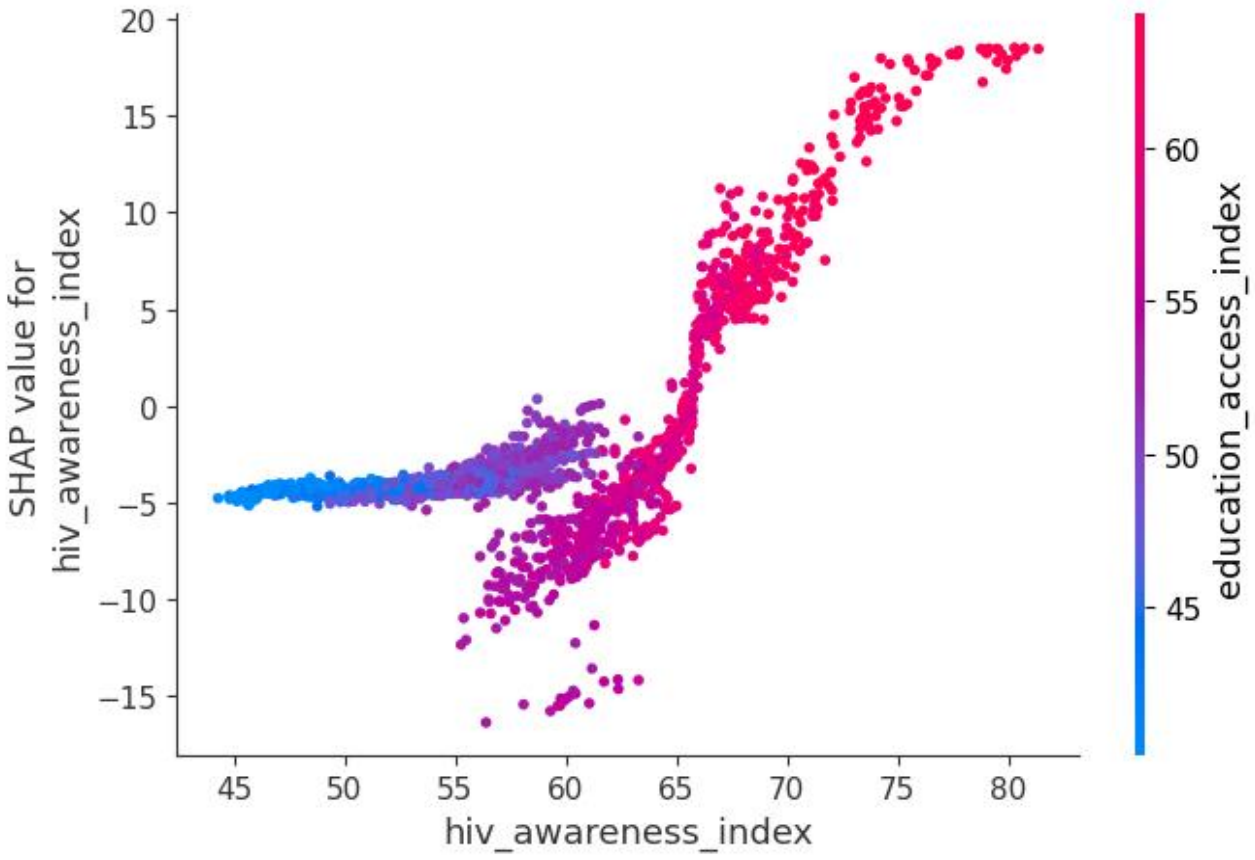


**Figure 37:**

*Prediction Intervals. Uncertainty Visualization*

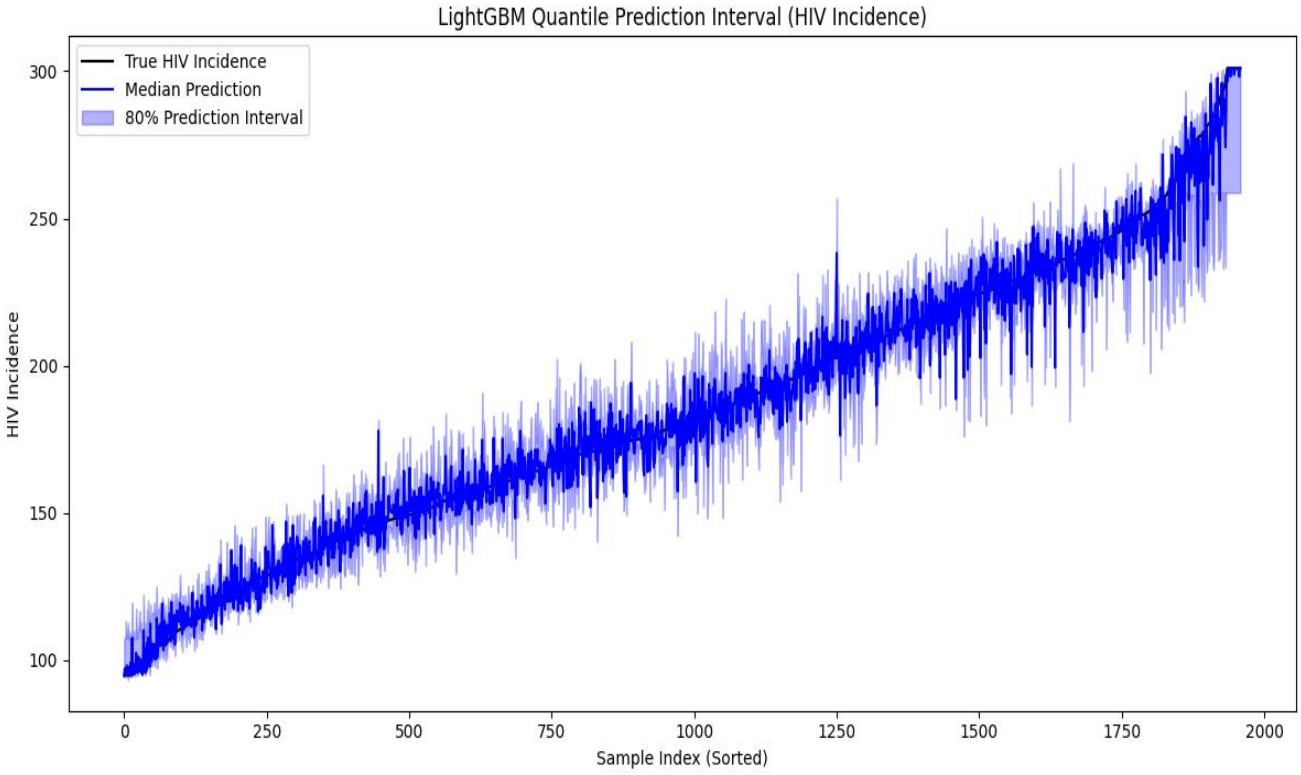



**Figure 38:**

*Q-Q Plot Of Residuals*

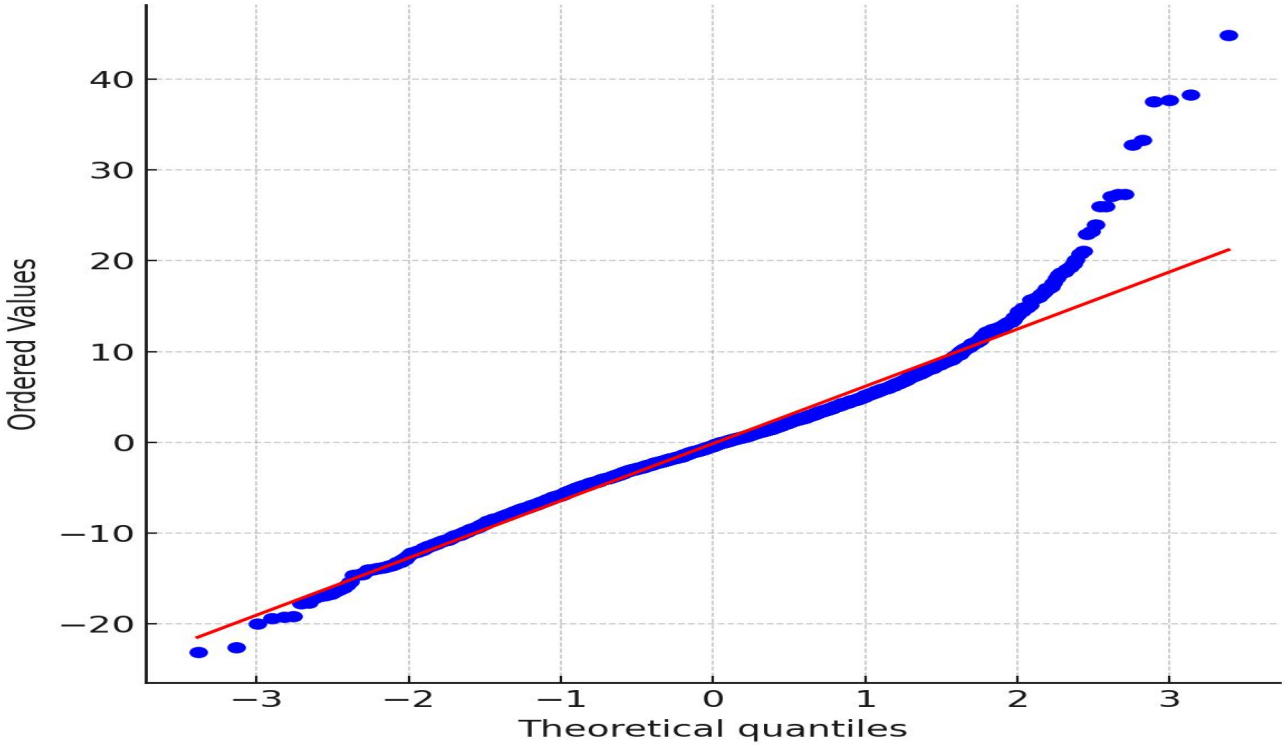


**Figure 39:**

*Residuals Vs Predicted Values*

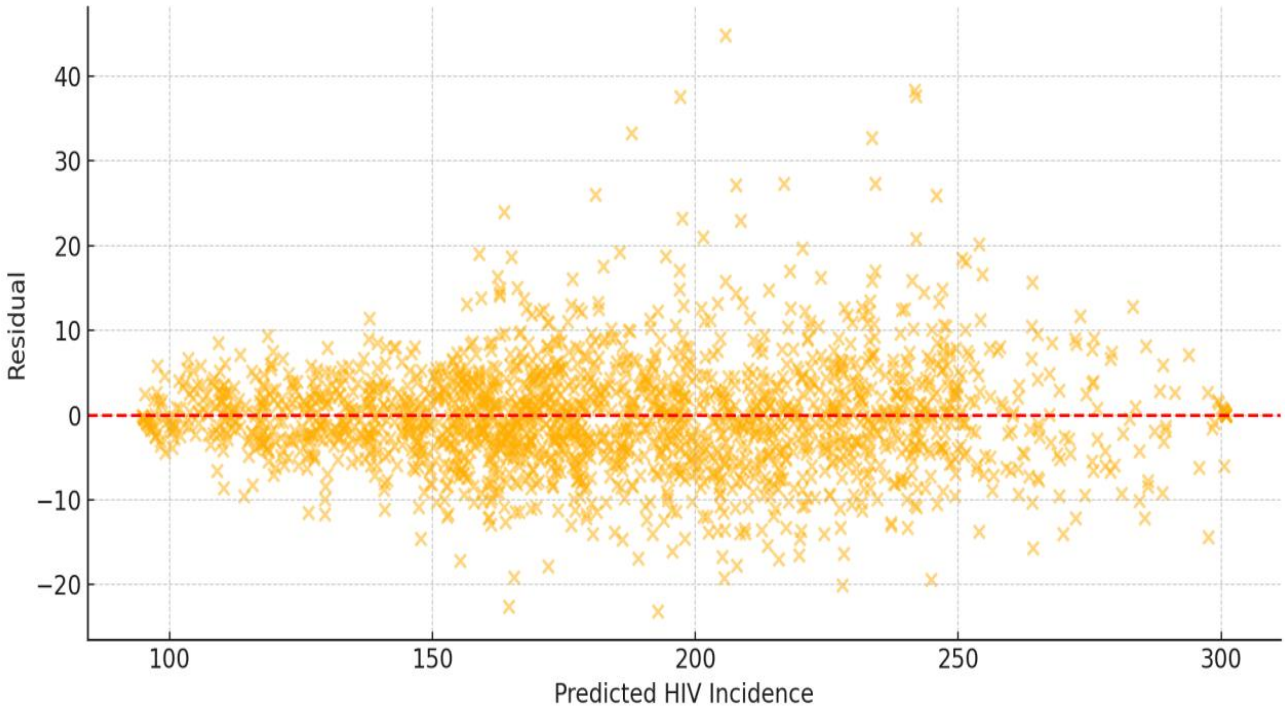



**Figure 40:**

*SHAP Dependence Analysis For Condom Use Rate*

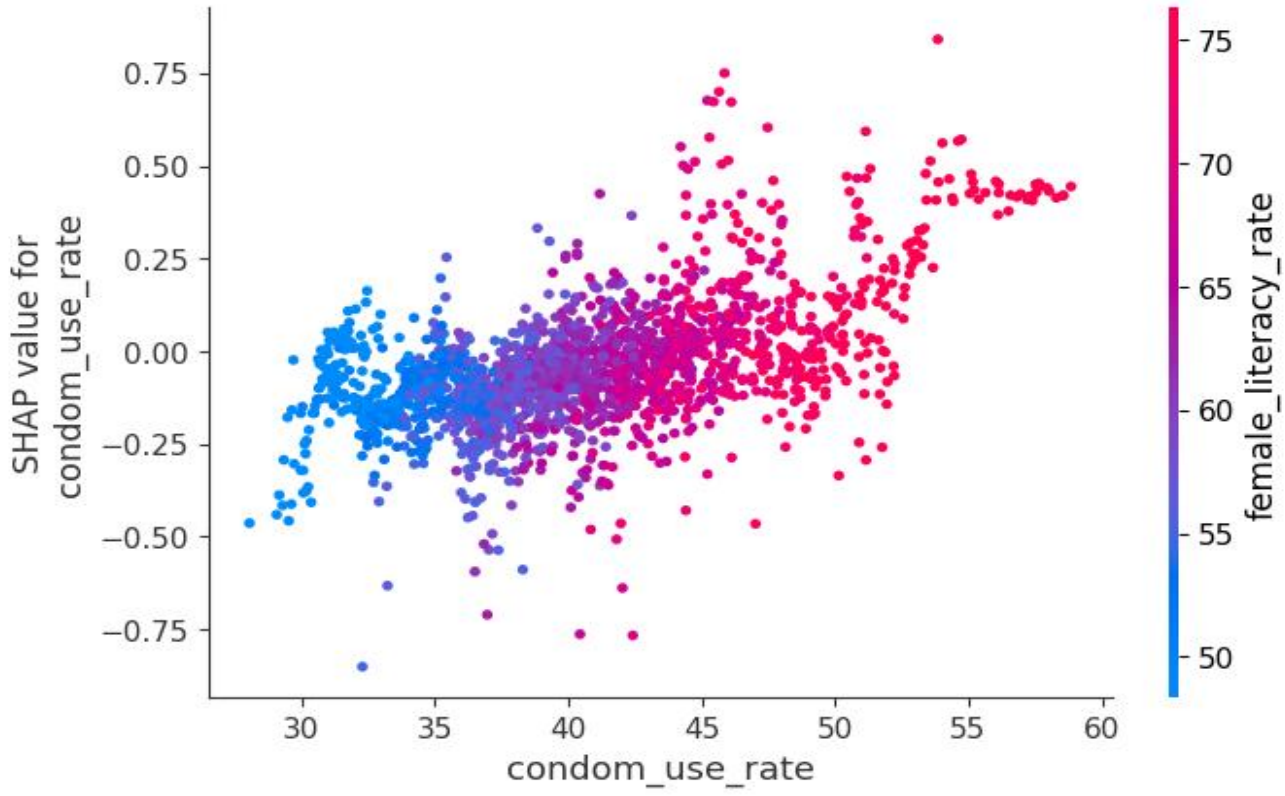


**Figure 41:**

*SHAP Dependence Analysis For Education Access Index*

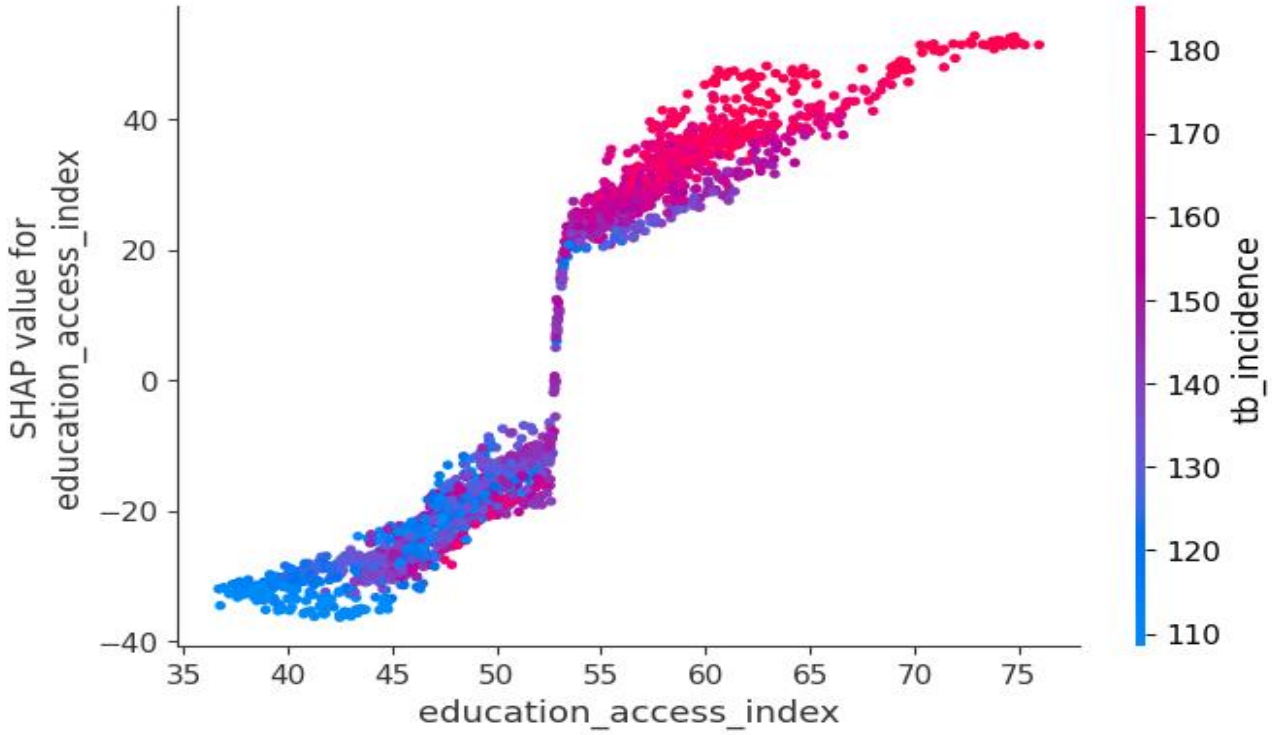
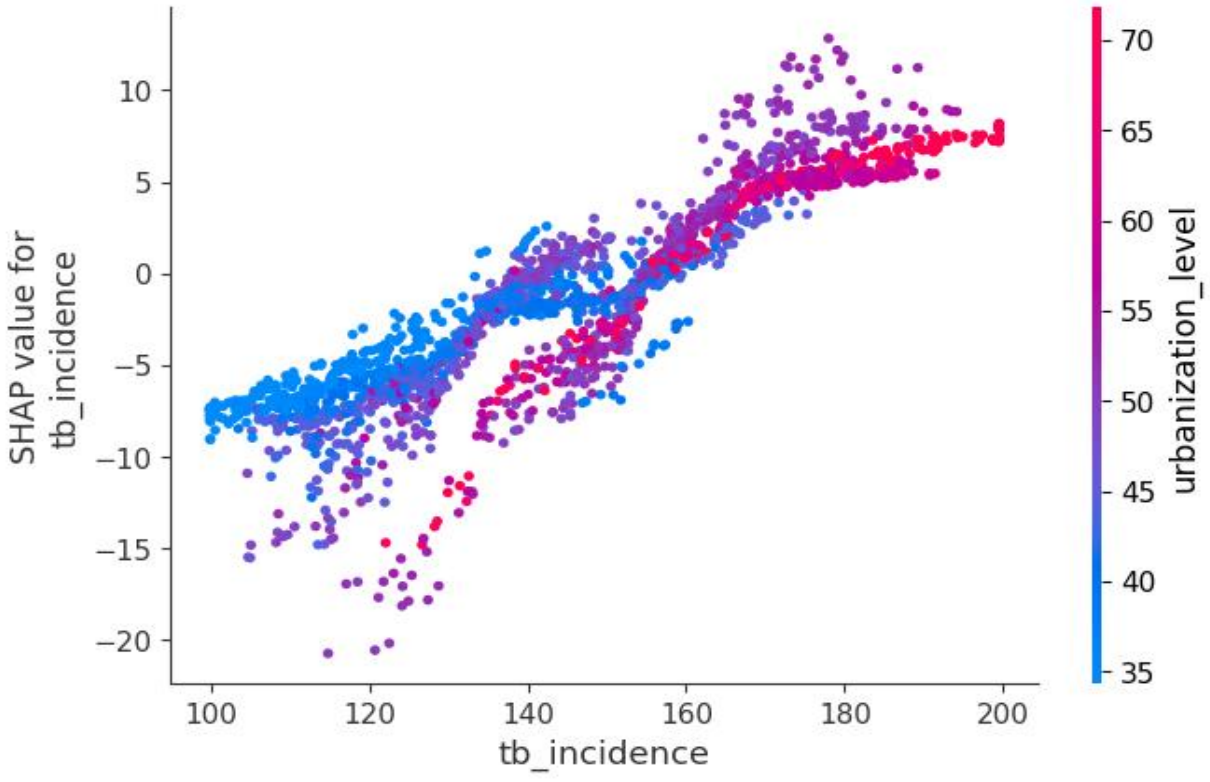



**Figure 42:**

SHAP Dependence Analysis For TB



## APPENDIX 3

Dashboard Deployment

### Interactive Dasboard Deployment

As an extension of the analysis and modeling efforts, a dynamic dashboard was developed in python using Streamlit to showcase key findings interactively.

The dashboard includes:

"Regional choropleth maps of HIV incidence",

"Behavioral and socio-health variable correlations",

"Correlation Heatmap of Key Predictors"

"Disease incidence forecasts up to 2030",

"Granular HIV Trends by Region Over time"

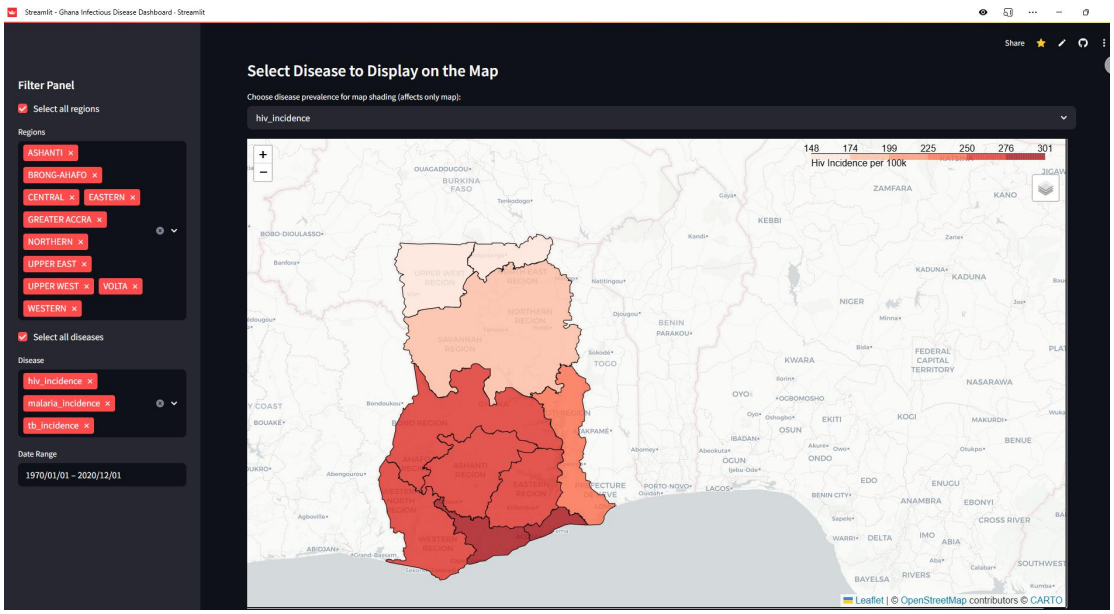
"Model performance comparisons".

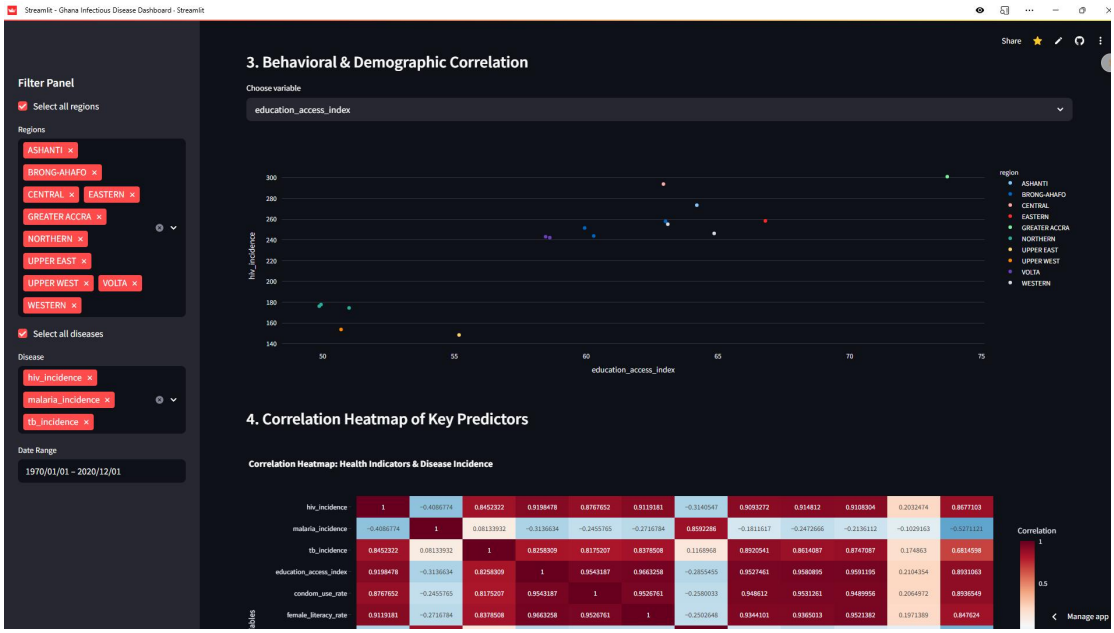
It is intended as a tool for public health researchers, planners, and policymakers to better understand and interact with Ghana's HIV/AIDS epidemic patterns.

### Access Link:

https://dashboardapppy-3ryuuaxnlmrsrrkqoxpcfw.streamlit.app

## Screenshots:



## APPENDIX 4

Codes

```python
# 1. RIDGE REGRESSOR TRAINING
import pandas as pd
import numpy as np
from sklearn.linear_model import Ridge
from sklearn.model_selection import train_test_split, GridSearchCV, KFold
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error, mean_absolute_percentage_error

# Load dataset
df = pd.read_csv('ghana_infectious_disease_model_dataset_cleaned.csv')  # Replace with your uploaded file

# Features & target
features = ['malaria_incidence', 'tb_incidence', 'population_total',
            'population_0_14_pct', 'population_65_plus_pct', 'education_access_index',
            'condom_use_rate', 'female_literacy_rate', 'youth_unemployment_rate',
            'hiv_awareness_index', 'access_to_art_pct', 'testing_coverage_pct',
            'health_facility_density', 'urbanization_level', 'regional_stigma_index',
            'migration_rate']
X = df[features]
y = df['hiv_incidence']

# Train-test split and scaling
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Ridge model with hyperparameter tuning
ridge_params = {'alpha': [0.01, 0.1, 1, 10, 100]}
cv = KFold(n_splits=5, shuffle=True, random_state=42)
ridge = GridSearchCV(Ridge(), ridge_params, cv=cv, scoring='neg_mean_squared_error')
ridge.fit(X_train_scaled, y_train)

# Predict and evaluate
y_pred = ridge.predict(X_test_scaled)
print("Best Hyperparameters:", ridge.best_params_)
print("RMSE:", np.sqrt(mean_squared_error(y_test, y_pred)))
print("R²:", r2_score(y_test, y_pred))
print("MAE:", mean_absolute_error(y_test, y_pred))
print("MAPE:", mean_absolute_percentage_error(y_test, y_pred))
```

```python
# 2. RANDOM FOREST REGRESSOR TRAINING

from sklearn.ensemble import RandomForestRegressor

rf_params = {
    'n_estimators': [100, 200],
    'max_depth': [10, 20, None],
    'min_samples_split': [2, 5]
}
rf = GridSearchCV(RandomForestRegressor(random_state=42), rf_params, cv=cv, scoring='neg_mean_squared_error', n_jobs=-1)
rf.fit(X_train, y_train)

y_pred = rf.predict(X_test)
print("Best Hyperparameters:", rf.best_params_)
print("RMSE:", np.sqrt(mean_squared_error(y_test, y_pred)))
print("R²:", r2_score(y_test, y_pred))
print("MAE:", mean_absolute_error(y_test, y_pred))
print("MAPE:", mean_absolute_percentage_error(y_test, y_pred))
```

```python
# 4. SUPPORT VECTOR REGRESSOR TRAINING

from sklearn.svm import SVR

svr_params = {
    'C': [1, 10],
    'epsilon': [0.1, 0.2],
    'kernel': ['rbf']
}
svr = GridSearchCV(SVR(), svr_params, cv=cv, scoring='neg_mean_squared_error', n_jobs=-1)
svr.fit(X_train_scaled, y_train)

y_pred = svr.predict(X_test_scaled)
print("Best Hyperparameters:", svr.best_params_)
print("RMSE:", np.sqrt(mean_squared_error(y_test, y_pred)))
print("R²:", r2_score(y_test, y_pred))
print("MAE:", mean_absolute_error(y_test, y_pred))
print("MAPE:", mean_absolute_percentage_error(y_test, y_pred))
```

```python
# Model Performance Heatmap

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from matplotlib.colors import Normalize

# Sample model performance table (replace with your real data if different)
table2 = pd.DataFrame({
    "Model": ["Ridge Regression", "Random Forest Regressor", "XGBoost Regressor", "Support Vector Regressor"],
    "RMSE": [9.0802, 6.3937, 6.5393, 8.4746],
    "R2": [0.9640, 0.9821, 0.9813, 0.9686],
    "MAE": [6.6865, 4.6401, 4.7674, 5.9004],
    "MAPE": [0.0369, 0.0256, 0.0264, 0.0317]
})

# Normalize values to 0-1 for color scaling
normalized_df = table2.set_index("Model").copy()
norm = Normalize()
normalized_df = normalized_df.apply(norm)

# Create heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(normalized_df, annot=table2.set_index("Model"), fmt=".4f", cmap="coolwarm", cbar=True)
plt.title("Model Performance Heatmap")
plt.tight_layout()
plt.show()
```

```
# Clusters of Regions Based on Health and Demographic Features (1)
# Install required libraries
!pip install seaborn matplotlib scikit-learn

# Import libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA

# Load dataset
df = pd.read_csv("ghana_infectious_disease_model_dataset_cleaned.csv")  # Replace with actual file

# Features to use
features = ['malaria_incidence', 'tb_incidence', 'population_total',
            'population_0_14_pct', 'population_65_plus_pct', 'education_access_index',
            'condom_use_rate', 'female_literacy_rate', 'youth_unemployment_rate',
            'hiv_awareness_index', 'access_to_art_pct', 'testing_coverage_pct',
            'health_facility_density', 'urbanization_level', 'regional_stigma_index',
            'migration_rate']
```

```
# Clusters of Regions Based on Health and Demographic Features (2)
# Group by region and average to remove time series
region_grouped = df.groupby("region")[features].mean()

# Standardize
scaler = StandardScaler()
scaled = scaler.fit_transform(region_grouped)

# PCA for 2D visualization
pca = PCA(n_components=2)
pca_result = pca.fit_transform(scaled)

# KMeans Clustering
kmeans = KMeans(n_clusters=4, random_state=42)
region_grouped["Cluster"] = kmeans.fit_predict(scaled)

# Plot
plt.figure(figsize=(10, 6))
sns.scatterplot(x=pca_result[:, 0], y=pca_result[:, 1],
                hue=region_grouped["Cluster"], palette="Set2", s=100)
for i, txt in enumerate(region_grouped.index):
    plt.text(pca_result[i, 0], pca_result[i, 1], txt, fontsize=9)
plt.title("Clustering of Regions Based on Health and Demographic Features")
plt.xlabel("PCA Component 1")
plt.ylabel("PCA Component 2")
plt.legend(title="Cluster")
plt.tight_layout()
```

```
# Quantile Prediction Intervals / Uncertainty Visualization (1)

!pip install lightgbm

import lightgbm as lgb
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split

# Load your dataset
df = pd.read_csv("ghana_infectious_disease_model_dataset_cleaned.csv")  # Upload your file and use its name here

# Define features
features = ['malaria_incidence', 'tb_incidence', 'population_total',
            'population_0_14_pct', 'population_65_plus_pct', 'education_access_index',
            'condom_use_rate', 'female_literacy_rate', 'youth_unemployment_rate',
            'hiv_awareness_index', 'access_to_art_pct', 'testing_coverage_pct',
            'health_facility_density', 'urbanization_level', 'regional_stigma_index',
            'migration_rate']
X = df[features]
y = df['hiv_incidence']
```

```
# Quantile Prediction Intervals / Uncertainty Visualization (2)
# Split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train Quantile Models
params = {
    'objective': 'quantile',
    'learning_rate': 0.1,
    'min_data_in_leaf': 5,
    'verbose': -1
}

quantiles = [0.1, 0.5, 0.9]
models = {}
train_data = lgb.Dataset(X_train, label=y_train)

for q in quantiles:
    params['alpha'] = q
    models[q] = lgb.train(params, train_data, num_boost_round=100)

# Predictions
pred_10 = models[0.1].predict(X_test)
pred_50 = models[0.5].predict(X_test)
pred_90 = models[0.9].predict(X_test)

# Plot
sorted_idx = np.argsort(y_test.values)
plt.figure(figsize=(12, 6))
plt.plot(y_test.values[sorted_idx], label="True HIV Incidence", color="black")
plt.plot(pred_50[sorted_idx], label="Median Prediction", color="blue")
plt.fill_between(range(len(y_test)), pred_10[sorted_idx], pred_90[sorted_idx],
                 color="blue", alpha=0.3, label="80% Prediction Interval")
plt.title("LightGBM Quantile Prediction Interval (HIV Incidence)")
plt.xlabel("Sample Index (Sorted)")
plt.ylabel("HIV Incidence")
plt.legend()
plt.tight_layout()
plt.show()
```

```python
# 2030 Forcast, All Models: ARIMA, XGBoost, RF, SVR, Ridge, Ensemble (1)


# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestRegressor
from sklearn.linear_model import Ridge
from sklearn.svm import SVR
from sklearn.metrics import mean_squared_error
from xgboost import XGBRegressor

# Load and preprocess dataset
df = pd.read_csv("ghana_infectious_disease_model_dataset_cleaned.csv")
df['date'] = pd.to_datetime(df['date'])
df = df.sort_values('date')

# Aggregate HIV incidence to monthly level
hiv_monthly = df.groupby('date')['hiv_incidence'].mean()

# Fill missing dates and interpolate values
last_actual_date = hiv_monthly.index.max()
historical_dates = pd.date_range(start=hiv_monthly.index.min(), end=last_actual_date, freq='MS')
hiv_full = hiv_monthly.reindex(historical_dates)
hiv_full = hiv_full.interpolate()
hiv_full.name = "hiv_incidence"

# Create lag features for forecasting
df_forecast = hiv_full.to_frame().copy()
for lag in range(1, 13):
    df_forecast[f'lag_{lag}'] = df_forecast['hiv_incidence'].shift(lag)
df_forecast.dropna(inplace=True)

# Define features and target variable
X = df_forecast[[f'lag_{i}' for i in range(1, 13)]]
y = df_forecast['hiv_incidence']
X_train = X.copy()
y_train = y.copy()

# Prepare latest lag features for forecasting
last_known = df_forecast.iloc[-1][[f'lag_{i}' for i in range(1, 13)]].values
```

```python
# 2030 Forcast, All Models: ARIMA, XGBoost, RF, SVR, Ridge, Ensemble (2)

# Initialize forecasting models
models = {
    'Random Forest': RandomForestRegressor(n_estimators=200, max_depth=20, random_state=42),
    'XGBoost': XGBRegressor(n_estimators=100, learning_rate=0.1, max_depth=5, random_state=42),
    'SVR': SVR(C=10, epsilon=0.2),
    'Ridge': Ridge(alpha=1.0)
}

# Perform forecasts and compute RMSE
forecast_results = {}
rmse_values = {}
forecast_horizon = 60
forecast_index = pd.date_range(start=hiv_full.index[-1] + pd.offsets.MonthBegin(1), periods=forecast_horizon, freq='MS')

for name, model in models.items():
    model.fit(X_train, y_train)
    preds = []
    current_input = last_known.copy()
    for _ in range(forecast_horizon):
        input_df = pd.DataFrame(current_input.reshape(1, -1), columns=X_train.columns)
        pred = model.predict(input_df)[0]
        preds.append(pred)
        current_input = np.roll(current_input, -1)
        current_input[-1] = pred
    forecast_results[name] = pd.Series(preds, index=forecast_index)
    y_pred_train = model.predict(X_train)
    rmse = mean_squared_error(y_train, y_pred_train) ** 0.5
    rmse_values[name] = rmse
    print(f"{name} Training RMSE: {rmse:.2f}")
```

```python
# 2030 Forcast, All Models: ARIMA, XGBoost, RF, SVR, Ridge, Ensemble (3)

# Generate ensemble forecast as mean of all models
ensemble_forecast = pd.DataFrame(forecast_results).mean(axis=1)
forecast_results['Ensemble'] = ensemble_forecast

# Combine forecasts and export
all_forecasts = pd.DataFrame(forecast_results)
all_forecasts.to_csv("ghana_hiv_forecasts_2025_2030.csv")

# Determine model closest to SDG proxy target
sdg_target_value = 150
un_sdg_target_2030 = 0.005  # New official SDG incidence target (per 1,000 uninfected population)
forecast_2030 = all_forecasts.iloc[-1]
closest_model = forecast_2030.sub(sdg_target_value).abs().idxmin()
closest_value = forecast_2030[closest_model]
print(f"Model closest to SDG target: {closest_model} with value {closest_value:.2f}")

# Plot forecasts and targets
plt.figure(figsize=(16, 7))
plt.plot(hiv_full.index, hiv_full.values, label="Historical HIV Incidence", color='black')

colors = {
    "Random Forest": "purple",
    "XGBoost": "red",
    "SVR": "orange",
    "Ridge": "teal",
    "Ensemble": "blue"
}
```

```
# 2030 Forcast, All Models: ARIMA, XGBoost, RF, SVR, Ridge, Ensemble (4)
# Plot forecast lines and markers
for name, forecast in forecast_results.items():
    final_value = forecast.iloc[-1]
    previous_value = forecast.iloc[-2]
    trend = "↑" if final_value > previous_value else ("↓" if final_value < previous_value else "→")
    color = colors.get(name, 'black')
    if name == 'Ensemble':
        label = f"Ensemble Forecast (Mean of All Models, {final_value:.2f} {trend})"
    else:
        rmse_display = rmse_values.get(name, 'N/A')
        label = f"{name} Forecast (RMSE: {rmse_display:.2f}, {final_value:.2f} {trend})"
    plt.plot(forecast.index, forecast.values, label=label, linestyle='--', color=color)
    marker_style = 'D' if name == closest_model else 'o'
    plt.scatter(forecast.index[-1], forecast.iloc[-1], color=color, s=80, marker=marker_style)

# Add SDG target lines to plot
plt.axhline(y=sdg_target_value, color='gray', linestyle=':', linewidth=1.5, label="2030 National SDG Proxy Target")
plt.axhline(y=un_sdg_target_2030, color='green', linestyle=':', linewidth=2, label="2030 UN Global SDG Target (0.005)")
plt.axvline(x=forecast_index[-1], color='gray', linestyle=':', linewidth=1.0)

# Final plot formatting
plt.title("Forecasting HIV Incidence in Ghana to 2030 (Including Ensemble Method)")
plt.xlabel("Date")
plt.ylabel("HIV Incidence (Per 1,000 Uninfected Population)")
plt.legend()
plt.tight_layout()
plt.show()

# Display forecast summary for 2030
summary_df = pd.DataFrame(forecast_2030).reset_index()
summary_df.columns = ['Model', '2030 HIV Incidence']
print("\nForecasted HIV Incidence by Dec 2030 (Including Ensemble):\n")
print(summary_df.sort_values(by='2030 HIV Incidence'))
```

For access to additional codes used in this work, including those developed for the dashboard, kindly refer to my online repository at:

[ https://github.com/vghanem/ghana_disease_dashboard/blob/main]

## APPENDIX 5

File Integrity Verification

### SHA-256 Checksums

To maintain transparency and enable independent validation, the following SHA-256 checksums are provided for key files used in this study:

| File | SHA-256 Checksum |
|---|---|
| `ghana_infectious_disease_model_dataset_cleaned.csv` | `F0CA3B11EAB890A22AEB910AEC2870C2C3E09523349AE1DB3BE2832ADC2DC71` |
| `dashboard_app.py` | `652817CDD1BE012630D86B2071980563D3F0381E1F5BFAA6CE1CD59632AF828F` |

**Verification Instructions:**

1. Download the file from the project repository (Github).
2. Use a SHA-256 checksum tool appropriate for your operating system:

   Windows: PowerShell command Get-FileHash -Algorithm SHA256 [path\to\file]

   Linux/macOS: Terminal command sha256sum [path/to/file]
3. Compare the computed checksum to the corresponding value in the table above.
4. A matching hash confirms that the file has remained unaltered.

This process reinforces data integrity, research transparency, and supports reproducibility standards.