Harmonizing Health Data: A Machine Learning-Based Detection of Tuberculosis Co-Infection in HIV Patient's Data in Pakistan

By

Muhammad Babar

(Registration No: 00400505)

Department of Computing

School of Electrical Engineering & Computer Sciences (SEECS)

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(2024)

Harmonizing Health Data: A Machine Learning-Based Detection of Tuberculosis Co-Infection in HIV Patient's Data in Pakistan

By

Muhammad Babar

(Registration No: 00400505)

A thesis submitted to the National University of Sciences and Technology, Islamabad,

in partial fulfillment of the requirements for the degree of

Master of Science in
Information Technology

Supervisor: Dr. Rafia Mumtaz

Department of Computing

School of Electrical Engineering & Computer Sciences (SEECS)

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(2024)

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis entitled "Harmonizing Health Data: A Machine Learning-Based Detection of Tuberculosis Co-Infection in HIV Patient Data in Pakistan " written by Muhammad Babar, (Registration No 00000400505), of SEECS has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: ______________

Name of Advisor: Dr. Rafia Mumtaz

Date: 21-Aug-2024

HoD/Associate Dean: ______________

Date: 21-Aug-2024

Signature (Dean/Principal): ______________

Date: 21-Aug-2024

FORM TH-4

National University of Sciences & Technology

MASTER THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by: (Student Name & Reg. #) Muhammad Babar [00000400505]

Titled: Harmonizing Health Data: A Machine Learning-Based Detection of Tuberculosis Co-Infection in HIV Patient Data in Pakistan

be accepted in partial fulfillment of the requirements for the award of Master of Science (Information Technology) degree.

Examination Committee Members

1. Name: Muhammad Daud Abdullah Asif Signature: 02-Sep-2024 9:45 AM

2. Name: Bilal Ali Signature: 02-Sep-2024 9:45 AM

Supervisor's name: Rafia Mumtaz Signature: 02-Sep-2024 8:30 PM

Arham Muslim
HoD / Associate Dean

03-September-2024
Date

COUNTERSINGED

04-September-2024
Date

Muhammad Ajmal Khan
Principal

THIS FORM IS DIGITALLY SIGNED

Approval

It is certified that the contents and form of the thesis entitled "Harmonizing Health Data: A Machine Learning-Based Detection of Tuberculosis Co-Infection in HIV Patient Data in Pakistan " submitted by Muhammad Babar have been found satisfactory for the requirement of the degree

Advisor : Dr. Rafia Mumtaz

Signature: ______________________

Date: 21-Aug-2024

Committee Member 1:Dr. Muhammad Daud Abdullah Asif

Signature:

21-Aug-2024

Committee Member 2:Mr Bilal Ali

Signature: ______________________

Date: 20-Aug-2024

Signature: ______________________

Date: ______________________

AUTHOR'S DECLARATION

I hereby declare that this submission titled "Harmonizing Health Data: A Machine Learning-Based Detection of Tuberculosis Co-Infection in HIV Patient Data in Pakistan " is my own work. To the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics, which has been acknowledged. I also verified the originality of contents through plagiarism software.

Student Name:Muhammad Babar

Student Signature: Babar

Date: 21-Aug-2024

Certificate for Plagiarism

It is certified that PhD/M.Phil/MS Thesis Titled "Harmonizing Health Data: A Machine Learning-Based Detection of Tuberculosis Co-Infection in HIV Patient Data in Pakistan " by Muhammad Babar has been examined by us. We undertake the follows:

a. Thesis has significant new work/knowledge as compared already published or are under consideration to be published elsewhere. No sentence, equation, diagram, table, paragraph or section has been copied verbatim from previous work unless it is placed under quotation marks and duly referenced.

b. The work presented is original and own work of the author (i.e. there is no plagiarism). No ideas, processes, results or words of others have been presented as Author own work.

c. There is no fabrication of data or results which have been compiled/analyzed.

d. There is no falsification by manipulating research materials, equipment or processes, or changing or omitting data or results such that the research is not accurately represented in the research record.

e. The thesis has been checked using TURNITIN (copy of originality report attached) and found within limits as per HEC plagiarism Policy and instructions issued from time to time.

Name & Signature of Supervisor

Dr. Rafia Mumtaz

Signature : ________________

DEDICATION

I would love to dedicate my thesis to my beloved parents, teachers, my family and my supervisor Dr. Rafia Mumtaz who guided me throughout my thesis.

ACKNOWLEDGEMENTS

In the name of Allah, the most merciful and benevolent. All praises to Allah Almighty for giving me the fortitude to finish my thesis. I owe my supervisor Dr. Rafia Mumtaz the utmost gratitude. Her persistent concern for our project, as well as her wise counsel, encouragement, and support, would have left this effort futile. I also want to express my gratitude to my committee members Dr. Muhammad Daud Abdullah Asif and Dr. Hafiz Syed Muhammad Bilal Ali for their insightful suggestions throughout this investigation. I am so appreciative of my parents' love and unwavering support during this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

ADASYN	Adaptive Synthetic Sampling
AI	Artificial Intelligence
AIDS	Acquired Immunodeficiency Syndrome
ANOVA	Analysis Of Variance
API	Application Program Interface
ART	Antiretroviral Therapy
AUC-	Area Under the Receiver Operating
ROC	Characteristic Curve
BMI	Body Mass Index
CD4	Cluster Of Differentiation 4
CNNs	Convolutional Neural Networks
GBM	Gradient Boosting Machines
GDPR	General Data Protection Regulation
HBV	Hepatitis B Virus
HCV	Hepatitis C Virus
HIV	Human Immunodeficiency Virus
LMICs	Low- And Middle-Income Countries
MCC	Matthews Correlation Coefficient
MDR-TB	Multidrug-Resistant Tuberculosis
ML	Machine Learning
NACP	National Aids Control Programme
NGO	Nongovernmental Organization
NTP	National Tb Programme
PCA	Principal Component Analysis
PEP	Post-Exposure Prophylaxis
PLHIV	People Living with HIV
PREP	Pre-Exposure Prophylaxis
RDT	Rapid Diagnostic Test

ROC	Receiver Operating Characteristic
SMOTE	Synthetic Minority Over-Sampling Technique
SOP	Standard Operating Procedure
STI	Sexually Transmitted Infection
SVM	Support Vector Machine
TB	Tuberculosis
VCT	Voluntary Counselling And Testing
WHO	World Health Organization
XAI	Explainable AI

ABSTRACT

Tuberculosis (TB) as a co-infection in People Living with HIV (PLHIV) is a serious public health concern, especially in underdeveloped nations like Pakistan. Robust data harmonization and sophisticated machine learning approaches are essential for the efficient management and treatment of chronic diseases. This work uses national health data from the Ministry of Health Pakistan to provide a machine learning-based method for detecting TB as co-infection in PLHIVs. To train different machine learning models, we used an extensive dataset that included patient demographics, clinical history, behaviors, lab results. The machine learning models trained on the extensive data set we used accuracy, recall, precision, and F1-score and Area Under Curve (AUC) parameters to assess the efficiency of models. According to our findings, machine learning methods can greatly improve the identification of TB co-infection in HIV patients, giving public health professionals a useful tool for tracking and containing the spread of these illnesses. Real-time dashboard, data analysis and decision-making are made easier and disease detection accuracy is increased when machine learning algorithms are integrated with national health databases. This study emphasizes how machine learning can revolutionize disease management and public health surveillance in environments with limited resources.

Keywords: TB and HIV Co-infection, Disease detection, Machine learning, Data Harmonization

CHAPTER 1: INTRODUCTION

1.1 Background

According to World Health Organization (WHO), People who are living with HIV (PLHIV) are very vulnerable to other co-infections especially Tuberculosis (TB). Both TB and HIV are a major challenge to global health especially in the developing world. Pakistan being among the countries with high burdens of both TB and HIV is at the forefront of this fight. Magnitude of morbidity and mortality tends to increase when these two diseases co-occur, which is a burden to the health sector. The World Health Organization WHO has estimated that TB is among the top killers of people with HIV, and this shows why efficient diagnosis and treatment approaches are required.

The convergence of TB and HIV presents unique diagnostic and therapeutic challenges. HIV-positive individuals are more susceptible to TB due to their compromised immune systems. The symptoms of TB in HIV patients can be atypical, leading to delayed or missed diagnoses. Furthermore, the interaction between TB and HIV medications can complicate treatment regimens, highlighting the need for precise and timely diagnosis. For instance, the clinical manifestation of TB in HIV-infected individuals often presents with less pronounced symptoms, which can be easily overlooked or misattributed to other opportunistic infections. This diagnostic ambiguity necessitates advanced and accurate diagnostic tools that can discern the presence of TB even in its subtle forms. With the developments in Machine learning (ML) paradigms, ML offers a promising avenue for addressing these challenges. Large datasets may be analyzed by ML algorithms to find patterns and provide predictions, which might increase the precision and effectiveness of

TB diagnosis in HIV patients. By leveraging patient data, ML can help unveil hidden correlations and predictive features, facilitating early diagnosis and personalized treatment strategies. Beyond human capacity, machine learning (ML) can handle and analyze enormous volumes of data to find insights that might help physicians make better decisions. This is where ML's potential resides.

The traditional diagnostic approaches for TB in HIV patients involve a combination of clinical evaluation, radiographic imaging, and microbiological tests, each with its limitations. Sputum smear microscopy, for example, has reduced sensitivity in HIV-positive patients, while chest radiographs may not show typical TB signs due to the compromised immune response. These limitations underscore the need for more sophisticated diagnostic tools. ML algorithms can integrate diverse data sources, including clinical records, laboratory results, and imaging data, to provide a comprehensive assessment that enhances diagnostic accuracy.

In Pakistan the healthcare sector has a number of problems, such as lack of resources, poor availability of technology, and communicable diseases being a major cause of morbidity and mortality. These factors explain why management of TB and HIV co-infected patients has remained quite a herculean task. Using algorithms derived from ML can help healthcare providers perhaps brush aside some of these barriers. ML can help make diagnostics faster, less dependent on specific technology and professionals, and, therefore, contribute to achieving early and accurate diagnoses, which are essential for the administration of proper treatment.

It must be borne in mind that beyond the issue of enhancing diagnostic precision, this study has a lot of implications. Concerning HIV-TB co-infected patients, the early and accurate diagnosis of TB has the potential of improving patients' survival rates, decreasing transmission in communities, and optimizing rational utilization of available health care resources. When evaluated under the capacity of the Pakistani health system where resources are scarce, the use of the ML for diagnostic tools could be particularly useful in improving the current system for managing and diagnosing TB and HIV co-infection. Furthermore, the results obtained from the application of the ML analysis can help the public health bodies in making decisions and guiding the measures for increasing the effectiveness of the programs aimed at decreasing the TB and HIV impact.

As with any emerging technology, the application of ML in healthcare has its fair share of issues that comes with it. To achieve high levels of confidence in the performance of any ML solution, the data feeding those models must be clean and valid, and the models themselves need to be well tested and continually audited. However, data heterogeneity is the main challenge when it comes to TB and HIV co-infection. The patient data may be procured from hospitals, clinics and even public health databases and each of them may be in different format, quality and may or may not be comprehensive. In this process, one important step is to align this data to correspond to the needs of building efficient ML models. This entails data cleansing, converting the data into a normal standard format, and then accumulating the required data to form a complete database for testing and training the developed ML algorithms.

Ethical issues are also of concern when handling patient data. The protection of the patient's information is crucial to building trust and following the legal and ethical requirements. As such, when designing and implementing the ML models, data anonymization, secure storage and management and effective data governance measures should be incorporated.

However, there is a good chance that using ML to identify TB co-infection in HIV patients would provide fruitful outcomes. ML algorithms may be used to accelerate and increase the accuracy of diagnosis, relieving physicians and nurses of some of their workload. They can also be used to analyze data for epidemiological studies on HIV and tuberculosis. For instance, ML models can identify certain risk variables or indicators that conventional research is unable to, and by doing so, it will be possible to comprehend the dynamics of the illness and develop preventative strategies.

Besides helping the clinicians enhance the overall quality of individual patients' care, ML-based approaches can assist public health concerns related to TB and HIV detection and tracking. Health monitoring data can be used for the timely identification of the spread of diseases, monitoring disease incidence, and assessing the effectiveness of measures taken in the field of public health. This capability is especially useful in Pakistan since accurate and timely health information is needed to address the TB and HIV problems in the country. The significance of this work can be generalized to the sphere of international health where the use of ML and big data technologies is considered to be revolutionizing in the addressing of multifaceted health issues. Therefore, the results of this study might be helpful to other low- and middle-income countries (LMICs) who face comparable health challenges to Pakistan and require reliable, efficient, and affordable methods to identify tuberculosis co-infection in HIV patients.

The findings and the approaches used in the study are transferable to other infectious diseases and health conditions, thus enhancing knowledge to advance health and medicine using new technologies.

In summary, TB and HIV co-infection in Pakistan is a serious threat for population, and therefore this issue requires further research and development of effective interventions. Machine learning can be also implemented as a potentially effective solution to boost the identification and prevention of this two-fold pandemic. Specifically, through the use of big data, ML can develop more refined, effective and timely diagnosis enhancing the quality of patient care and advancing the work of public health. The objective of this dissertation is to fill the gap of using available health data in Pakistan to inform key health needs and promote enhanced use of data in decision-making regarding the country’s healthcare system.

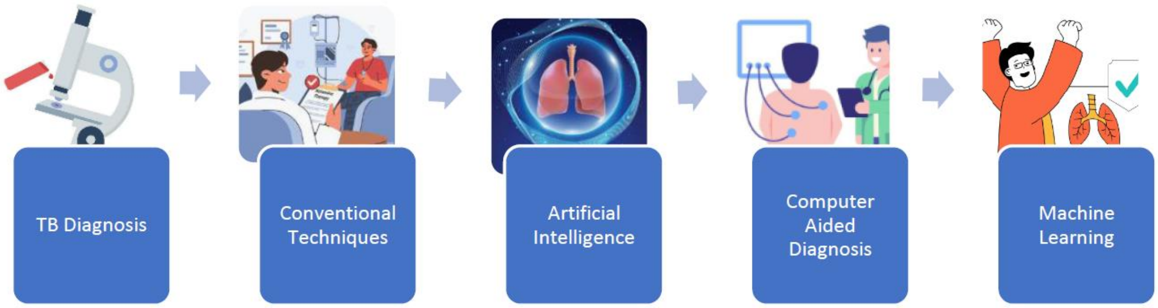


Figure 1: Evolution of artificial intelligence in TB diagnosis

1.2 Problem Statement

Fostering on extensive health data in Pakistan, its utilization to tackle the TB and HIV co-infection issue is restricted. Classical methods have been useful in obtaining insight into the data but they lack flexibility in dealing with large amount of data that is collected in today’s healthcare system. The flaws of traditional approaches are quite simple to observe when working with complicated, large-scale datasets resembling the typical structure of healthcare info: linear regression, logistic regression, and other comparable methods do not suffice. These methods do not accommodate interactions between the variables which are essential in the co-infection of TB and HIV. Which calls for the invention of approaches that will enable the application of this data towards the enhancement of health.

The main research question focuses on the absence of efficient approaches to the identification of TB co-infection in Pakistani HIV patients at an early stage. The reason education is crucial is that one can help prevent the development of critical consequences and the spread of the disease. However, current diagnostic practices are generally time-consuming and based on resources that are not easily accessible in the resource-poor environment. This is made worse by the fact that, in order to make sense of the health data that is generated from different sources, data harmonization becomes an issue. The Pakistan’s healthcare data is not integrated and is scattered on different platforms and institutions which leads towards data incoherencies and missing data. This fragmentation creates a problem for diagnosis and treatment planning as it is difficult to gather all the information needed in one place.

Moreover, the variability of data sources from different regions, healthcare facilities, and patients’ ages also contribute to this challenge. The records in every dataset may not necessarily be in the same format, and each may contain different terminologies that can complicate the aggregation and analysis of the data. This absence of integration and standardization complicates the creation of a consistent, usable dataset for subsequent analysis. That is why the lack of such methods significantly undermines the functioning of the healthcare system in addressing this public health issue.

Furthermore, the strategies previously used to treat TB and HIV co-infection patients depend on clinicians' knowledge and are based on paper-based tools, which are laborious and contain potential errors. In a country where the burden of disease is high such as Pakistan, the demand for any diagnostic tool is high particularly if it will have to compete for resources in a nation's strained health care system. ML is more effective in this regard as it automates the process of detection and yields more relevant information.

Despite the great potentials that ML has in the Pakistani health sector, there are several challenges that act as roadblocks to the implementation of ML; they include lack of awareness among the medical practitioners on the application of ML technologies, insufficient availability of the required computational resources together with concerns on the privacy and security of patient information. These obstacles are not just solved by a technological solution, but requires capacity building practices aimed at creating awareness among the health care personnel on the possibilities and opportunities that this new technology presents.

Also, there is a lack of proper evaluation methodologies to measure the performance of the ML with applications in operational environments. This involves establishing measures for accuracy, sensitivity, and specificity besides cost and pilot studies to test the models in various practices. It is crucial to have such evaluation frameworks for the purpose of building the necessary trust between the providers of such ML-based solutions and the healthcare providers and policymakers that will enable the positive adoption of the solutions.

Therefore, the problem that stands before the health system is complex and includes the search for effective diagnostic tools for TB co-infection in HIV patients, the integration of heterogeneous health data, and the lack of implementation of modern analytical methods. To this end, this thesis will seek to design and test context-specific ML models that would strengthen the Pakistani healthcare system's ability to respond to TB and HIV co-infection. That way, it aims at enhancing the health status of the population as well as the capacity of the health system in Pakistan.

1.3 Aims and Objectives

The main objective of this thesis is, thus, to design and to assess the performance of machine learning techniques for the identification of TB co-infection in HIV/AIDS patients in Pakistan. In order to do this, the study will concentrate on the following main goals: In order to do this, the study will concentrate on the following main goals:

- Data Harmonization: It involves gather and categorizing numerous health data on TB and HIV co-infection from different sources and quality check. Because it deals with information gathered from many sources, such as clinical trials, public health databases, hospital information systems, and test findings, data harmonization is crucial.
- This process resolves issues of disparate formats and terminologies to ensure that the data is put in a format that is suitable to be used in the analysis. Thus, the research is designed to ensure the data's validity and consistency to guarantee a proper development of machine learning models.

- Machine Learning Model Development: Develop and test machine learning models specific to the identification of TB co-infected HIV patients. This involves selecting appropriate machine learning strategies, such as supervised learning algorithms like neural networks, decision trees, and support vector machines, among others, and unsupervised learning approaches like clustering and anomaly detection.
- These models will be trained on the harmonized dataset, including the process of fine-tuning to achieve the highest accuracy, sensitivity and specificity. Data for the validation of the models will be obtained from cross-validation and autonomous test data sets to ensure the models are not over-fitted.
- Insight Generation: Reveal insights and regularities in the integrated data that can help in the identification, therapy, and prophylaxis of illnesses. It is the research's goal to examine key predictors and risk factors related to TB co-infection in HIV patients via data exploration and feature extraction. They can provide ideas that are not easily detectable using conventional analytical procedures, which enrich the knowledge of disease dynamics. The generated insights will also be useful in devising clinical recommendations and decision-making databases that would be useful to practitioners.
- Impact Assessment: Examine the implications of deploying machine learning-based detection in the setting of healthcare on results and resources. This objective entails the assessment of the proposed machine learning models with an aim of identifying their suitability in clinical practice. The impact assessment will involve the evaluation of the usefulness of the models in increasing the diagnostic correctness, decreasing the diagnostic time, and increasing the patient management.

Besides, the study will compare the efficiency of implementing ML-based detection systems using factors such as resource consumption, healthcare expenditure, and the overall cost-savings. Thus, the research is expected to show the effectiveness and applicability of using ML in the Pakistani context to encourage its use in the healthcare sector.

All these objectives are in accord with the main issues in the discovery and elimination of TB co-infection in PLHIVs in Pakistan. In this context, the current research aims at using the principles of machine learning to increase diagnostic accuracy, provide effective recommendations, and, therefore, contribute to improving the results of treatment. If the implementation of these models is successful, it will be possible to improve the existing diagnostic practices to a great extent and to advance the overall objective of the development of the health care system in Pakistan.

1.4 Organization of Thesis

To successfully address the research issues and study objectives, this thesis is organized in a highly specific manner.

Chapter 1: This chapter is the introduction, the we have outlined the general background of the research, the reasons for conducting the given research, the main goals and objectives, the scope of the study as well as research methodological framework. The background of this chapter focuses on the context of the major health concern in Pakistan, namely TB and HIV co-infection, the need to employ machine learning as the solution, and the study objectives and expectations.

Chapter 2: This chapter starts with the Literature Review which looks into the information available in the academic sources regarding TB and HIV co-infection, comparing previous research and results concerning the general prevalence and clinical manifestations of the diseases as well as the difficulties of diagnosing and treating the conditions. It also looks at machine learning in the healthcare field, looking at different algorithms and their utility in detection of diseases, prognosis, and treatment. In addition, the chapter on data harmonization is also provided to describe the significance of the process and methods of combining different sources of health data.

Chapter 3: Methodology describes the general approach to the study and it specifies the procedures for data collection and preparation. It explains how health data is collected from various sources and how data quality and comparability is achieved through data harmonization. This chapter also expands on aspects such as choosing the right algorithms for machine learning, training strategies, and techniques for validating the models. The methodological rigor contributes to the possibility of replicating the research and the validity of the results obtained.

Chapter 4: Data Analysis and Model Development constitutes the middle of the work, where data analysis and procedures of creating and checking the machine learning models are demonstrated. It also has information about the kind of data that was worked on, the features that were derived and the models that were developed. The chapter also covers the metrics, including as accuracy, sensitivity, specificity, and area under the ROC curve, that may be used to compare how well the constructed models work.

It is also presented the confirmation of the results in order to check the performance and reliability of the models created.

Chapter 5: Results and Discussion presents the findings of the research and explains the results in terms of the formulated research questions and objectives. It discusses the implications of the research to healthcare practice revealing the strengths of the developed machine learning models in the identification and treatment of TB co-infection with HIV. The chapter also continues the analysis of these models and their effects on public health, looking at the advantages, difficulties, and shortcomings of their usage.

Chapter 6: In Conclusion and Recommendations, the major research contributions are discussed again, with the emphasis on the importance of the obtained results and their implications for improving the healthcare situation in Pakistan. The paper also presents the implications of the study and lists any drawbacks that may be present and not explored completely. Last but not the least the chapter concludes with some of the recommendations for the future work; this section points out the future scope of work and the possible enhancements that could be made to the models and methodologies adopted in this thesis.

This rational structure helps to maintain the structure of the research process, which is important for readers, so that they can follow the flow from background information and defining the problem to conclusions and recommendations. In this way, each chapter is cognate to another and at the same time form a whole, which highlights the relevance of the research and its findings to public health and the domain of machine learning. By the time the thesis is finished, the reader ought to be able to distinguish between the type of research problem being examined, the novel approaches being taken to address the issue, and any potential effects these approaches may have on health care policy and practice.

CHAPTER 2: LITRATURE REVIEW

This important chapter start with a critical analysis of previous research on the use of machine learning techniques in diagnosing TB co-infection in HIV patients. In the present review, several works which have applied deep learning and machine learning approaches, the nature of data used and the architectural models used in these works are illustrated. It also elaborates on the difficulties addressed, limitations of these approaches, and importance of these researches with reference to Pakistan.

Research paper entitled “Detection of Tuberculosis based on Deep Learning based Methods” makes an attempt to solve the problem of applying deep learning-based CAD systems for diagnosing pulmonary TB on Chest X-ray. Data sets used in the study were Tuberculosis X-ray (TBX11K), the Japanese Society of Radiological Technology (JSRT), Shenzhen dataset (CH), and the India dataset (IN). In this strategy, the principal architectural model that was considered was Convolutional Neural Networks (CNNs). The study described the benefits of deep learning in improving the identification of TB but also pointed out the scarcity of studies on CAD-based TB identification using deep learning. This is based on the reference Puttagunta, M. K., & Ravi, S. (2021) for basic knowledge about the CNN application in TB diagnosis (Puttagunta & Ravi, 2021).

Another important article accessed and we have seen that the author has used a three-methylation driven Genene based deep learning to detect TB as co-infection in PLHIVs. This work is dedicated to increasing the diagnostic accuracy for TB in subjects with HIV co-morbidity. This study used gene expression data and DNA methylation data matched to HIV patients and HIV/TB co-infected patients. The architectural model adopted was a three methylation driven gene based deep learning model. However, the study had some drawbacks, for example, inadequately small sample size for assessing TB prevalence, the necessity of expanding the groups of participants for confirmation of the findings. This work is cited using the authors' names as Xu and Yuan (2022) or using the publication year as Xu and Yuan, 2022.

Further another important article on detection of TB in PLHIV with Chest-X-rays is studied. This article provides an answer to the difficulty of increasing the accuracy of TB diagnosis in HIV-infected patients, including the areas with restricted access to experienced radiologists and differences in chest images. The study relied on two data sets from South Africa that involved HIV-positive TB-suspect patients. A deep learning model CheXaid algorithm that was used to diagnose active pulmonary TB using chest X-rays and clinical covariates was used. Points mentioned include possible lack of sensitivity with Xpert MTB/RIF tests and lower diagnostic sensitivity of chest X-ray as compared to culture or Xpert MTB/RIF tests. This study is cited as Rajpurkar et al. - 2020 (Rajpurkar et al. 2020).

Another study is done by the Orwa et al. in 2023 which compare logistic regression with regularized ML model to predict TB in PLHIVs. The data used in this study was taken from Kenya. Authors investigates the application of L1 regularized machine learning in conjunction with logistic regression to accurately identify tuberculosis (TB) in HIV positive patients. The type of study used was a cross-sectional study on data from patients diagnosed of HIV/AIDS and TB in Kisumu County, Kenya. The applied models were logistic regression, Lasso, Ridge, as well as Elastic net regression. Some of the limitations of the study include; cross-sectional study design and sample size (Orwa et. al., 2023).

Elhag's research titled "Prediction and Classification of Tuberculosis Using Machine Learning" seeks to enhance the proactive prediction and subsequent classification of TB, with a view of helping to identify the disease early enough, monitor its spread, and to facilitate early health interventions to contain the disease. It leveraged tuberculosis incidence data obtained from the CDC for the US from 1953 to 2021. The models used included Decision Trees (DT) and Artificial Neural Networks (ANN), CART methodology was used in the ANN model and the architecture used was the MLP. A mentioned weakness was the failure to compare ANN with other methods (Elhag, 2024).

In the research on the development of machine learning in tuberculosis diagnosis, Singh et al. (2022) covered a number of deep learning strategies meant to increase the precision and speed of TB detection. The model employed different data sets of images of chest X-ray scans and clinical data of TB affected patients, and normal people. Some of the models used included ResNet, MobileNet, Xception, EfficientNet, and Inception with the use of transfer learning methodologies and a combination of deep learning with fuzzy logic, genetic algorithm, and AIS. These were considered as limitation in the study; over-fitting and the lack of understanding of the working of the CNN model (Singh et al., 2022).

Oloko-Oba and Viriri (2022) examined early techniques for tuberculosis detection by chest X-ray imaging, examining a range of conventional diagnostic tests for tuberculosis and medical imaging datasets in their investigation. The discussed architectural models are AlexNet, VGGNet, GoogLeNet, ResNet, and SqueezeNet. The first identified limit was that the samples were not equally distributed with regards to datasets with limited samples (Oloko-Oba & Viriri, 2022).

Another study by Khew, Akbar, and Mohd-Assaad emphasized the use of machine learning in the study of neglected tropical diseases (NTDs), highlighting the difficulties caused by a lack of unified cooperation platforms and a shortage of data. The sources of the data collected in the work were various open-access databases and platforms, including WHO’s Global Health Observatory and other similar resources. CNNs for the classification of images in cancer detection, and different techniques of ML and DLs in the surveillance and management of NTDs were used. The identified limitations were that the studies on NTDs were conducted based on a small amount of information (Khew et al., 2023).

Another work from Zeyu, Yaakob, and Azman categorized and contrasted many deep learning techniques for diagnosing TB. The involved datasets were Montgomery, SHENZHEN, PadChest, CheXpert, COVID-Xray-5k, and others. The investigation was based on eight preliminary CNN models, namely AlexNet, GoogleNet, InceptionV3, MobileNetV2, VGG16, ResNet50, Densenet121, and EfficientNetB7. The limitation pointed out was the challenge of getting medical data and the insufficient number of datasets for deep learning benchmarking (Zeyu et al., 2022).

Lai (2023) put out a predictive information system with the goal of raising the standard of treatment and lowering death rates for prisoners who also have TB and HIV. The sample included 367 clinical cases of incarcerated persons with HIV and TB co-infection admitted throughout hospitals and prisons between 2012 and 2018. The models employed were a multiple-stage method, system analysis, artificial neural networks, and statistical grouping methods. The limitation that I encountered was that there were only a few studies on the perceptions of PLHIVs available to me (Lai, 2023).

Another research paper focuses on improving the accuracy of forecasting trends in TB/HIV co-infection, which is crucial for effective allocation of public health resources and intervention strategies. The study examines various predictive models, from classical statistical methods to advanced machine learning techniques, to analyze TB/HIV co-infection case notifications across different demographics, including men, women, and the general population. Initially, traditional models like Exponential Smoothing and Autoregressive Integrated Moving Average (ARIMA) were applied to establish a foundational understanding of temporal trends and seasonality. Following this, machine learning models such as Support Vector Regression (SVR), Extreme Gradient Boosting (XGBoost), and Deep Neural Networks, including Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN), were tested for their ability to capture the complex dynamics and non-linearities of TB/HIV co-infection data. Performance evaluations, using error metrics like MSE, MAE, and sMAPE, demonstrated that Deep Learning models, particularly Bidirectional LSTM and the combined CNN + LSTM, significantly outperformed classical statistical methods. This indicates the superior effectiveness of deep learning techniques in modeling TB/HIV co-infection time series, resulting in more accurate forecasts. The study concludes that while classical models have their applicability, advanced deep learning models are better suited for predicting TB/HIV co-infection due to their ability to capture intricate data patterns. These findings emphasize the importance of using sophisticated modeling approaches in public health to enhance prevention and control strategies for infectious diseases. Future research will incorporate socioeconomic data to explore multivariate time series predictive modeling, addressing significant social determinants of TB (Abade, 2024).

Another work, through the analysis of large-scale electronic medical data, the research explores the application of machine learning to predict co-infection of TB among HIV patients. This research is conducted using data from the National Clinical Research Center for Infectious Diseases in Shenzhen, China, spanning from 2017 to 2021, the study involved 4540 patients. Initially, a fine-tuned ChatGLM was used to structure the electronic medical records, which were then analyzed using a multi-layer perceptron to classify and determine the presence of tuberculosis in HIV patients. The researchers built a specialized database using machine learning-based natural language processing to study the epidemiological characteristics of TB and HIV co-infection, focusing on incidence patterns, patient characteristics, and influencing factors. A predictive model was created using Long Short-Term Memory (LSTM) networks to forecast TB co-infection risk among HIV patients, providing scientific evidence for clinical decision-making and enabling early detection and precise intervention. Results showed that the refined ChatGLM model achieved a precision rate surpassing 0.95 for symptom extraction, with key symptoms like diarrhea and normal showing precision rates above 0.90. Among the 4540 HIV patients, 758 were diagnosed with TB, indicating a 16.7% co-infection rate. Additionally, the study highlighted a 25.1% co-infection rate with syphilis among HIV patients. The study also compared the performance of the Multilayer Perceptron (MLP) classifier against LSTM in predicting high-risk groups for HIV and TB co-infections. The MLP classifier demonstrated predictive ability with AUROC values ranging from 0.616 to 0.682, while the LSTM model showed consistent performance with AUROC values between 0.827 and 0.850 across 5-fold cross-validation, achieving an overall accuracy of 81.18% in distinguishing HIV co-infected TB from simple HIV infection. The results underscored the superior performance of deep learning techniques in managing both structured and unstructured medical data, emphasizing the benefits of using laboratory time-series data for better prediction outcomes. This study highlighted the potential of integrating deep learning with electronic health data to revolutionize public health, offering valuable insights for healthcare providers in disease prediction and management. Future research will expand the study parameters and conduct a multicenter prospective cohort study to further enhance the model's performance and validate additional influencing factors (Jingfang, 2024).

Another research focuses on using machine learning for the prediction and classification of TB addresses the urgent need for predictive models to accurately identify and monitor the incidence of tuberculosis (TB), an infectious disease that primarily affects the lungs but can also impact other organs. According to the World Health Organization (WHO), TB is the second leading cause of death from infectious diseases, following COVID-19, and ranks as the thirteenth largest cause of death overall. The study emphasizes the importance of constructing predictive models to identify groups and locations where TB spreads and to monitor various trends and patterns associated with the disease. Utilizing tuberculosis case data from the United States, the researchers applied artificial neural network (ANN) models and decision tree (DT) models to predict and classify TB cases. The results demonstrated that the decision tree model outperformed the artificial neural network model in terms of accuracy. The study found that the number of reported TB cases has gradually decreased over time, attributed to strong political commitment at both global and national levels. This trend is depicted graphically, showing a higher incidence of TB among males compared to females and identifying California as the state with the highest number of TB cases. The research utilized machine learning models to predict and classify time series data for new TB cases in the United States. A comparative analysis between the ANN and DT models was conducted using statistical metrics such as MSE, RMSE, sMAPE, and MAPE. The findings revealed that the DT model had superior performance, indicated by lower values in these metrics, signifying higher accuracy. The study's insights are valuable for policymakers in guiding vaccine use and public health strategies, both in the United States and globally. Additionally, the research tools developed can serve as educational materials for public engagement and be integrated into various research initiatives (Azhari, 2023).

Another paper concentrating on data fusion from medical records and clinical data, the research tackles the crucial need for accurate tuberculosis (TB) diagnosis, particularly in underdeveloped nations where insufficient infrastructure hinders detection and treatment efforts. As TB remains one of the world's top ten causes of death and a global emergency declared by the World Health Organization, timely diagnosis is essential, especially in remote areas with scarce resources. This study leverages artificial intelligence, specifically natural language processing (NLP) and machine learning (ML) techniques, to create diagnostic models that can operate effectively even in resource-limited settings. The researchers explored two primary data sources: text extracted from electronic medical records (EMR) and patient clinical data (CD). They implemented four proposals using five different machine learning models, initially applying ML to each data source independently and then developing data fusion approaches that combined both sources. The strategy's effectiveness was analyzed in consultation with physicians, focusing on the practical relevance and understanding of EMR data. The results revealed that using clinical data alone yielded an area under the ROC curve of 69.9%, outperforming the data fusion models. However, analyzing physician reports demonstrated significant advantages, particularly in areas lacking basic diagnostic infrastructure where clinical-specific data might not be readily available. The study highlights the potential of computational intelligence tools to provide health professionals with additional diagnostic support. The analysis was conducted with an understanding of the constraints faced in resource-limited settings, where standard diagnostic tests are often unavailable. While the models using only physician reports achieved slightly lower performance (9% lower than those using clinical data) but demonstrated better sensitivity (73%), the findings emphasize the importance of NLP techniques in providing valuable diagnostic tools where patient information is incomplete or unavailable. In conclusion, this research offers alternative strategies to support health professionals in diagnosing suspected pulmonary TB (PTB), particularly in under-resourced areas. The integration of NLP and ML techniques can enhance decision-making processes, providing critical insights even when clinical variables are not accessible. The study's insights and algorithms pave the way for further research to develop tools that can be integrated into healthcare workflows, improving TB diagnosis and ultimately contributing to better health outcomes (Romero-Gómez, 2024).

One more study uses a systematic review and meta-analysis to assess the predictive accuracy of machine learning techniques in order to solve the problem of early mortality prediction in HIV-positive persons. Researchers are creating models to predict PWH mortality risk as machine learning is being incorporated into clinical practice. But the usefulness of the existing prediction models for HIV-related fatalities has come under scrutiny due to the wide range of mortality periods and modeling factors. In order to analyze the efficacy of machine learning in forecasting HIV-related fatalities and to provide evidence-based insights to promote the growth of artificial intelligence in this sector, the study carried out a systematic review and meta-analysis.

On November 25, 2023, a thorough search of the PubMed, Cochrane, Embase, and Web of Science databases was conducted as part of the research. The bias risk in the listed studies was evaluated using the Predictive Model Bias Risk Assessment Tool (PROBAST). Subgroup analysis based on survival and non-survival models as well as meta-regression to look at how death time affected the models' predictive value were included in the meta-analysis. 24 studies comprising data from 401,389 HIV-positive people were reviewed for this analysis; 23 of the papers concentrated on long-term follow-ups outside of hospital settings.

Both non-survival and survival machine learning models, including COX regression, were employed to forecast these fatalities. According to the findings of the meta-analysis, the training set's c-index for predicting PWH fatalities using these models was 0.83 (95% CI: 0.75–0.91), while the validation set's c-index was 0.81 (95% CI: 0.78–0.85). Crucially, the results of the meta-regression study showed that the machine learning models' performance was not substantially affected by either the follow-up period or the frequency of death occurrences. The study comes to the conclusion that machine learning is a workable method for creating forecasts of HIV-related mortality that are not time-based. To fully confirm these results, further multicentre investigations are necessary, as shown by the small number of original studies that were included in the analysis (Yuefei, 2024).

The use of machine learning (ML) and artificial intelligence (AI) to biomedical research has shown great potential, especially in the area of medication resistance prediction in tuberculosis (TB).

Treatment attempts for tuberculosis (TB), which is caused by Mycobacterium tuberculosis (M. TB), are complicated by the presence of multi-drug resistance (MDR-TB) strains. TB is still a major worldwide health concern. Recent advancements in high-throughput sequencing have enabled the accumulation of extensive genomic data, which, when coupled with ML techniques, can be leveraged to identify drug-resistant mutations. A lot machine learning algorithms in classifying genetic mutations and predicting their impact on drug resistance. For example, Yang et al. (2019) successfully employed SVM and ANN models to distinguish between resistant and susceptible TB strains, achieving high accuracy. The simplicity and efficiency of NB for large datasets and the interpretability of kN have also been noted in genomic studies. Moreover, the integration of molecular docking and molecular dynamics simulations with ML predictions enhances the understanding of how mutations affect protein-drug interactions, providing a comprehensive approach to validating computational predictions. This combined strategy not only improves the accuracy of resistance prediction but also offers insights into the structural implications of mutations, thus advancing the development of targeted TB treatments (Salma, 2020).

The authors of another work investigate the possibility of combining artificial intelligence (AI) and machine learning (ML) with genome sequence data and conventional machine learning algorithms in conjunction with convolutional neural networks (CNN) to predict drug resistance in Mycobacterium tuberculosis (MTB). TB caused by MTB, presents a significant global health challenge, exacerbated by the emergence of multi-drug resistant (MDR-TB) strains. Traditional rule-based methods for drug resistance prediction often yield inconsistent results, performing well for some antibiotics but less reliably for others. Previous research has demonstrated the effectiveness of traditional ML algorithms, such as logistic regression and random forest, in providing more stable and accurate predictions compared to rule-based approaches. Despite these achievements, there hasn't been much use of deep learning methods—in particular, Convolutional Neural Networks, or CNNs—in this field.

CNNs excel at capturing complex patterns in sequential data, making them well-suited for genomic data analysis. The authors developed 24 binary classifiers using logistic regression, random forest, and a customized 1D CNN architecture, training these models on a dataset of 10,575 MTB isolates from diverse geographic regions. Their findings indicate that the 1D CNN models achieved superior accuracy and stability, with F1-scores ranging from 81.1% to 98.2%, outperforming the state-of-the-art Mykrobe predictor. Furthermore, feature selection techniques identified AMR-relevant genetic markers, 78.8% of which were validated by the WHO catalogue of MTB mutations. This study underscores the potential of combining traditional ML with deep learning to enhance TB drug resistance prediction, offering a robust, faster, and cost-effective alternative to in vitro assays (Xingyan, 2022).

In another paper, the authors provide a comprehensive overview of advancements and challenges in computer-aided detection (CAD) software for the automated interpretation of chest radiographs (CXRs) in tuberculosis (TB) diagnosis and elimination. CAD technology is a useful tool for both symptomatic people and population-based screening, since it has shown great sensitivity and accuracy equivalent to human readers. Nevertheless, the diagnostic variability across various contexts and subpopulations poses considerable hurdles to its application, requiring configurable threshold scores that many locations lack the capacity to put up. CAD software is constantly being updated, making standardization and comparability tasks more difficult. Furthermore, there is a major gap in the evaluation of CAD systems' accuracy in recognizing non-TB abnormalities, as many children with respiratory symptoms may have illnesses other than tuberculosis. These systems have not been sufficiently validated for the diagnosis of juvenile tuberculosis.

CAD technology is a useful tool for both symptomatic people and population-based screening, since it has shown great sensitivity and accuracy equivalent to human readers. Nevertheless, the diagnostic variability across various contexts and subpopulations poses considerable hurdles to its application, requiring configurable threshold scores that many locations lack the capacity to put up. CAD software is constantly being updated, making standardization and comparability tasks more difficult. Furthermore, there is a major gap in the evaluation of CAD systems' accuracy in recognizing non-TB abnormalities, as many children with respiratory symptoms may have illnesses other than tuberculosis. These systems have not been sufficiently validated for the diagnosis of juvenile tuberculosis.

In another paper, given the low sensitivity of conventional tests, the authors of this paper address the critical challenge of diagnosing tuberculous meningitis (TBM), a severe infection of the central nervous system with a high mortality rate. They do this by utilizing a predictive model that incorporates tuberculostearic acid to enhance clinical diagnosis.

Tuberculostearic acid (TBSA), a fatty acid unique to Mycobacterium tuberculosis, has shown promise as a biomarker for TBM, although it has not consistently provided definitive results. The sensitivity and specificity of the TBSA-combined scoring system, which was created in this investigation based on a retrospective examination of 113 suspected TBM cases, were 0.8814 and 0.8148, respectively, with an area under the receiver operating characteristic curve of 0.9010. Using machine learning approaches, the authors verified their identification of four important co-predictive factors: extra-neural TB, basal meningeal enhancement, CSF glucose/serum glucose ratio less than 0.595, and CNS coinfection. The study emphasizes the integration of TBSA with these factors to enhance TBM diagnosis, providing a reliable method that supports timely treatment initiation. This approach is contrasted with emerging diagnostic methods like NAAT, GeneXpert MTB/RIF Ultra, mNGS, and LAM, which, despite their potential, face limitations in sensitivity, specificity, and cost. The TBSA model leverages the advantages of the established Lancet scoring system while addressing its labor-intensive nature and the need for expert interpretation. By offering a simpler, more practicable model with high predictive capacity, the TBSA-based scoring system could significantly improve the diagnostic process for TBM, especially in settings with limited resources. Furthermore, the study's use of a support vector machine (SVM) model for validation underscores the importance of combining traditional biomarkers with advanced machine learning techniques to enhance diagnostic accuracy and clinical decision-making in TBM cases (Fong, 2023).

In subtropical Guangxi, China, the authors of this research examine the effects of environmental variables, including air pollution and climatic conditions, on the incidence of tuberculosis (TB) among persons living with HIV/AIDS (PLWHA), a group that is particularly susceptible to TB infection.

Using data from the HIV ART cohort in Guangxi, China (2014-2020), alongside meteorological and air pollutant data, the authors applied a distribution lag non-linear model (DLNM) to assess these effects. The study revealed that temperature and precipitation significantly influenced TB risk, with a 5-unit temperature increase resulting in a cumulative relative risk (RR) of 0.663 and a 2-unit precipitation increase leading to an RR of 1.478, both observed at a 4-week lag. Wind speed and PM10 levels did not show significant lag effects. Extreme weather conditions, such as high temperatures (hot effect), heavy rainfall (rainy effect), and absence of rain (rainless effect), were associated with reduced TB risk. The study further highlighted that these effects were more pronounced in PLWHA with CD4(+) T cells < 200 cells/μL, underscoring the role of immune status in TB susceptibility. These findings align with previous studies linking climatic variables to TB incidence and emphasize the need for public health strategies that integrate environmental monitoring with clinical management to mitigate TB risk in PLWHA. Despite the lesser impact of air pollutants, the study underscores the importance of considering both environmental and immunological factors in TB management, particularly in high-burden, subtropical regions.

In this study, the authors use data from the National Institutes of Health (NIH) to investigate the application of sophisticated deep learning algorithms to reliably detect pneumonia and tuberculosis from chest X-rays (CXR). The proposed model integrates convolutional neural networks (CNNs) and Residual Networks (ResNets) with other machine learning classifiers to enhance feature extraction and classification accuracy. Previous studies, such as those by Rajpurkar et al. (2017) and Wang et al. (2017), have demonstrated the potential of CNNs in medical image classification, highlighting their ability to learn complex patterns from large datasets. This research builds on these foundations by employing image data augmentation and class balancing techniques to address common challenges such as overfitting and class imbalance. Data augmentation artificially expands the dataset by generating modified versions of existing images, thus reducing overfitting during training. In order to guarantee a fair representation of classes, the study also makes use of a variety of class balancing strategies, such as Adaptive Synthetic Sampling (ADASYN), Borderline SMOTE, and Synthetic Minority Oversampling Technique (SMOTE).

The findings reveal that Borderline SMOTE outperforms other techniques in enhancing model performance. The novelty of this research lies in the combined application of data augmentation and sophisticated class balancing methods, which collectively contribute to high accuracy in distinguishing between TB and pneumonia. The authors propose future enhancements such as using lung masks to focus on relevant CXR regions, employing cropping techniques for image consistency, and leveraging transfer learning to generate new samples for minority classes. These proposed improvements aim to refine the model further and validate its effectiveness on larger and more diverse datasets, ultimately aiding early detection and treatment of TB and pneumonia in clinical settings (Bharti, 2023).

The authors of this work address the serious problem of antibiotic resistance (ABR) in microbial communities, which poses a serious risk to public health. They suggest using machine learning to enhance the identification of resistance patterns in bacteria. Traditional experimental methodologies for determining treatment options are time-consuming due to the vast diversity of bacterial strains. With advances in sequencing technologies and decreasing costs, using bacterial genotypes rather than phenotypes to identify ABR has gained traction. This study employs machine learning (ML) to categorize and interpret bacterial datasets, enhancing clinical practice. The researchers generated k-mers from nucleotide sequences of antibiotic-resistant bacteria and clustered them based on genomic similarities using the Affinity Propagation algorithm, achieving a Silhouette coefficient of 0.82. A Random Forest prediction model was developed, yielding a specificity of 0.99 and a sensitivity of 0.98. Additionally, a MLP precision with a hamming loss of 0.05 classified bacterial strains into resistant and non-resistant categories against various antibiotics. This approach identified genes and ABR drivers related to the k-mers, providing insights into their biological relevance. The study highlights horizontal gene transfer among bacteria, which plays a crucial role in acquiring beneficial traits such as antibiotic resistance, enabling adaptation to new environments. The results of the clustering study showed that strains of Salmonella, Pseudomonas, Mycobacterium, Escherichia, and Salmonella did not form distinct clusters but rather shared genetic characteristics, suggesting interactions across different species.

The study's findings suggest that integrating genomic sequences and clustering pathogens based on similar characteristics can significantly enhance our understanding of ABR patterns, leading to more informed and effective treatment options. Future research should focus on expanding the dataset to larger bacterial cohorts and exploring the evolutionary and clinical implications of these interactions (Parthasarathi, 2024).

In another study Bending the Curve Through Innovations to Overcome Persistent Obstacles in HIV Prevention and Treatment, the authors highlight recent advancements in significant fields and highlight how these breakthroughs might significantly impact the trajectory of the HIV/AIDS pandemic. Despite considerable progress in treatment, HIV/AIDS continues to be a major global public health challenge, exacerbated by disparities in prevention and care access. Innovative HIV vaccine platforms, such as mRNA vaccines and conserved epitope and mosaic constructs, are currently in early-phase trials, aiming to enhance immunogenicity and provide broader protection. Long-acting injectable antiretrovirals have emerged as a milestone in HIV treatment, addressing adherence challenges, while gene editing techniques hold promise for future curative strategies, though they face obstacles related to delivery and off-target effects. The review also explores the role of multi-omics approaches, including genomics, transcriptomics, proteomics, and metabolomics, in providing systems-level insights into viral persistence and identifying new therapeutic opportunities. The gut microbiome is increasingly recognized as a significant factor in HIV progression, with research into probiotic/prebiotic supplementation and fecal transplantation showing potential benefits. The integration of artificial intelligence and machine learning across these domains is expected to accelerate discovery and enhance the application of these innovations. While no single solution guarantees epidemic control, the synergistic application of these emerging tools, coupled with rigorous clinical testing and ensuring equitable global access, could transform the HIV/AIDS landscape. However, major barriers persist, particularly in resource-limited regions, including financing, infrastructure, and political commitment. Strengthening health systems and addressing social drivers such as stigma are crucial priorities to complement continued biomedical innovation. The review underscores the importance of pairing scientific advancements with efforts to overcome these systemic barriers to maximize global impact and move closer to ending the HIV/AIDS pandemic.

In another paper, the authors examine the ongoing challenges and emerging innovations in the management and prevention of HIV/AIDS, highlighting the societal impact of the disease. Despite advancements in antiretroviral therapy (ART), access disparities remain in low- and middle-income countries due to financial constraints, inadequate infrastructure, and social stigma. The future outlook is promising with research focused on long-acting injectables, gene editing therapies, and therapeutic vaccines. Advancements in pre-exposure prophylaxis (PrEP) and other preventive measures offer hope for reducing new infections. Addressing societal factors such as poverty, education, and gender inequality is crucial for lasting improvements in HIV care. Obstacles such as HIV variations resistant to drugs, gaps in testing, and persistent societal obstacles continue to exist, calling for a holistic approach that integrates behavioral, structural, and biological treatments while emphasizing equality and human rights.

Continued investment in research, health system improvements, and cross-sector collaboration is vital. Innovations in ART, such as long-acting antiretrovirals and broadly neutralizing antibodies, present new opportunities for effective viral control and improved well-being. Prevention techniques like long-acting PrEP and microbicides empower individuals to protect themselves and reduce new infections. Integrating digital health solutions, community engagement, and comprehensive care models is revolutionizing HIV service delivery, ensuring equitable access. However, stigma, discrimination, and institutional barriers hinder progress, particularly among marginalized populations. Addressing these issues requires a concerted effort to uphold human rights, reduce disparities, and build inclusive communities. By leveraging scientific advancements, innovative approaches, and community mobilization, we can overcome these challenges and move towards a future where HIV/AIDS no longer threatens global health and well-being, creating a society where individuals live free from the adverse effects of HIV/AIDS, with hope triumphing over despair and empathy and unity guiding us towards a brighter future for all (Daniel, 2024).

In another paper, the authors emphasize the importance of understanding disease pathogenesis by examining the interactions between viruses and bacteria, rather than focusing solely on individual pathogens. The review highlights how coinfections can lead to either colocalized infections within one anatomical niche or systemic infections affecting the entire body, underscoring the complexity of virus-bacteria interactions in disease onset and progression. Examples from the literature illustrate the mechanisms of these interactions, such as targeting central host immune responses, causing inflammation, and triggering immune pathways. Phage therapy emerges as a promising alternative for antibiotic treatments and for managing secondary bacterial infections in COVID-19 patients. While bacteria-bacteriophage interactions typically follow a predator-prey model, limiting their inclusion in this review, the synergistic interactions between bacteria and viruses, especially in the context of oncolytic viruses and anti-cancer therapies, are explored. The review also discusses how coinfections can reduce the efficacy of drugs targeting one pathogen due to the presence of another, exemplified by bacterial neuraminidases reducing the effectiveness of anti-influenza drugs and HIV infection compromising tuberculosis treatment. The necessity of considering coinfecting pathogens is highlighted, advocating for better diagnostics and therapeutic strategies that incorporate both viral and bacterial agents. Emerging diagnostic methods, including next-generation sequencing (NGS) and machine learning (ML) models, show promise in early detection and prediction of coinfections, although their clinical validation is essential. The potential of combinatorial drugs, probiotics, and microbiome-based metabolites tailored for coinfection treatment is also emphasized. The paper calls for a combination of traditional diagnostic methods and advanced computational models, such as ordinary differential equations (ODEs) and ML, to analyze and interpret diverse data from coinfection systems. Overall, the review provides a comprehensive perspective on the significance of virus-bacteria interactions, aiming to enhance understanding and development of novel healthcare strategies for managing coinfections and improving disease outcomes (Pokhrel, 2024).

Sr. No	**Paper Author and Publication Year**	**Architecture /Model Used**	**Dataset**	**Limitation**	**Proposed Future Work**
1	Puttagunta & Ravi et al. (2021)	CNNs	TBX11K, JSRT, Shenzhen dataset, India dataset	Scarcity of studies on CAD-based TB identification using deep learning	Further research on CAD-based TB identification methods
2	Xu & Yuan et al. (2022)	Three methylation-driven gene-based deep learning model	Gene expression data and DNA methylation data	Small sample size	Expand participant groups
3	Rajpurkar et al. (2020)	CheXaid algorithm	South African datasets	Lack of sensitivity with Xpert MTB/RIF tests and lower diagnostic sensitivity of chest X-ray	Improvement in diagnostic tools and sensitivity measures

4	Orwa et al. (2023)	Logistic Regression, Lasso, Ridge, Elastic net regression	HIV/AIDS and TB patient data from Kisumu County, Kenya	Cross-sectional study design, sample size	Longitudinal studies and larger sample sizes
5	Elhag et al. (2024)	Decision Trees (DT), Artificial Neural Networks (ANN)	CDC tuberculosis incidence data (US 1953-2021)	Failure to compare ANN with other methods	Comparative studies with additional ML models
6	Singh et al. (2022)	ResNet, MobileNet, Xception, EfficientNet, Inception	Images of chest X-ray scans and clinical data of TB patients	Over-fitting, lack of understanding of CNN model	Better model interpretability, reduce over-fitting
7	Oloko-Oba & Viriri et al. (2022)	AlexNet, VGGNet, GoogLeNet, ResNet, SqueezeNet	Various traditional TB diagnosis tests and medical imaging datasets	Unequal sample distribution in datasets	More balanced datasets
8	Khew, Akbar, & Mohd-Assaad et al. (2023)	Various CNNs	Open-access databases and platforms, including WHO’s Global Health Observatory	Studies based on a small amount of information	Unified cooperation platform for NTDs research
9	Zeyu, Yaakob, & Azman et al. (2022)	AlexNet, GoogleNet, InceptionV3, MobileNetV2, VGG16, ResNet50, DenseNet121, EfficientNetB7	Montgomery, SHENZHEN, PadChest, CheXpert, COVID-Xray-5k	Challenge of acquiring medical data, insufficient datasets for deep learning benchmarking	Access to larger, standardized datasets
10	Lai et al. (2023)	Multiple-stage method, system analysis, artificial neural networks	367 clinical cases of incarcerated persons with HIV and TB co-infection	Few studies on the perceptions of PLHIVs	Exploration of PLHIV perceptions

11	Abade et al. (2024)	Exponential Smoothing, ARIMA, SVR, XGBoost, Bidirectional LSTM, CNN + LSTM	TB/HIV co-infection case notifications	Limited inclusion of socioeconomic data	Incorporate socioeconomic data for multivariate time series predictive modeling
12	Jingfang et al. (2024)	LSTM	National Clinical Research Center for Infectious Diseases, Shenzhen, China	Limited scope and geographical focus	Expand study parameters, multicentre prospective cohort study
13	Azhari et al. (2023)	ANN, Decision Trees	Tuberculosis case data from the United States	None specifically noted	Further exploration of ANN and DT models
14	Romero-Gómez et al. (2024)	Various machine learning models, NLP techniques	Text extracted from EMR and patient clinical data	Clinical data alone outperformed data fusion models	Further exploration of NLP techniques and data fusion
15	Yuefei et al. (2024)	Survival model (e.g., COX regression), non-Survival model	401,389 individuals diagnosed with HIV	Limited number of original studies	Additional multicentre studies to validate findings
16	Salma et al. (2020)	Naïve Bayes, kNN, SVM, ANN	Various bacterial datasets	Not specified	Expand dataset to larger bacterial cohorts
17	Xingyan et al. (2022)	Logistic regression, random forest, 1D CNN	10,575 MTB isolates	Limited application of deep learning in TB resistance prediction	Integration of traditional ML with deep learning for better prediction
18	Geric et al. (2023)	CAD technology	Not specified	Diagnostic heterogeneity, lack of validation for pediatric TB diagnosis,	Improved customization and validation for pediatric TB diagnosis

				economic and political issues	
19	Fong et al. (2023)	Support vector machine (SVM)	113 suspected TBM cases	Limited sensitivity of TBSA as a biomarker	Further validation and integration with other diagnostic methods
20	Bharti et al. (2023)	CNNs, ResNets, other machine learning classifiers	NIH dataset	Limited application to diverse populations and diseases	Expand dataset, implement additional deep learning techniques
21	Parthasarathi et al. (2024)	Affinity Propagation algorithm, Random Forest, multilayer perceptron	Bacterial genotypes	Not specified	Further exploration of gene transfer in bacteria
22	Daniel et al. (2024)	Not specified	Not specified	Not specified	Continued research and focus on addressing systemic barriers
23	Pokhrel et al. (2024)	Phage therapy, ODEs, machine learning	Not specified	Not specified	Integration of traditional diagnostics with computational models
24	Bharti et al. (2024)	CNNs, ResNets, other machine learning classifiers	NIH dataset	Overfitting, class imbalance	Use of lung masks, transfer learning, and enhanced class balancing techniques
25	Parthasarathi et al. (2024)	Affinity Propagation, Random Forest, multilayer perceptron	Bacterial genotypes	Limited focus on interspecies interactions	Exploration of horizontal gene transfer, larger dataset analysis

CHAPTER 3: METHODOLOGY

This chapter describes the approach taken in the current study to develop and evaluate machine learning models that can predict the co-infection of HIV and TB using health data from Pakistan. The methodology is divided into several key sections: acquisition of data, cleaning of data, feature extraction, choosing of model, training of model and testing of the model. As with any branch of artificial intelligence, each section gives a comprehensive account of the measures and strategies applied in the generation of high-quality and reliable machine learning models. The purpose of this methodology is to develop the application of machine learning to help in the early identification and treatment of TB and HIV co-infected patients with the view of improving on the health of the communities in the developing world.

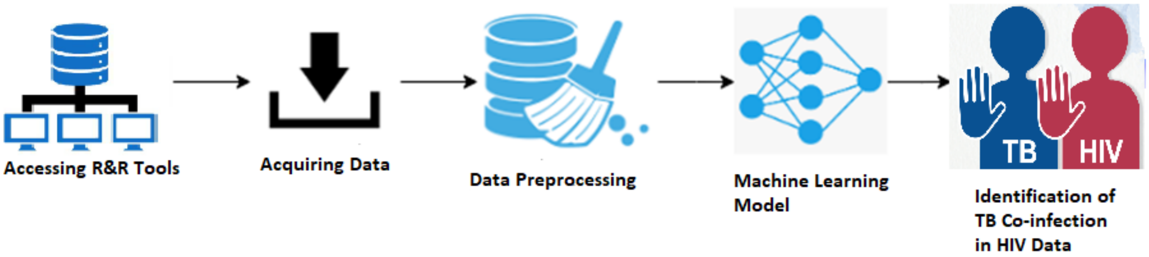


Figure 2: Proposed Methodology

3.1 Data Collection

We have taken data from the national health information system of Pakistan with information on patients’ demographics, clinical history and diagnostic test findings. The source of data is useful for studying the prevalence of TB as well as HIV in Pakistan as the dataset contains rich patient data – the patient was diagnosed with TB, HIV or both and the information contains health records for several years.

3.1.1 Sources and Scope of Data

The national health database aggregates information from many organizations such as hospitals, clinics, diagnostic laboratories and other community health programs. Altogether, the sources mentioned above give the possibility to have a sample sufficiently diverse and appropriate to represent the general population. The data is collected from both the municipal and the rural areas to make the models that are derived from this data to be transferable and be relevant to any part of the country. It is also possible since the study incorporates data from different years, which enables the temporal examination of changes in the occurrence of diseases and effectiveness of interventions.

3.1.2 Types of Data Collected

These are quantitative and qualitative data and as seen, the dataset is laden with both types of data. Key types of data collected include: Key types of data collected include: Demographic Information: Age, gender, area of residence, income and occupation. This information assists in the distribution of TB and HIV and the identification of the demographics of risk.

Clinical History: Thorough medical histories, including previous lab results, medications, follow-ups, and other infections. This helps in establishing the patient’s health history and may be useful in understanding the patterns and other influencing factors of co-infection.

Diagnostic Test Results: Surveys about the patient's condition from sputum tests, chest X-rays, blood tests for CD4 count, viral load, and other laboratory tests. This data is vital in the confirmation of the diagnosis and the extent of the infections.

Treatment Records: Patients' medical history, medications that have been prescribed, whether or not they have complied with prescribed medication, medication side effects, and treatment results. This is useful in determining the efficiency of various treatment plans and their consequence on the health of patients.

Follow-up Data: Reports over subsequent appointments, and modification in the state of health, diseases' progression, and other outcomes. This information can only be obtained from longitudinal data and is crucial for describing the evolution of the diseases and the consequences of the treatments administered.

3.1.3 Data Collection Process

The data collection technique is done both electronically through the patients' electronic health records and via manual data entry. Most of the health facilities in the country have adopted EHR technology that permits entry and sharing of real time data. Where EHRs have not been implemented data is captured on paper by the healthcare workers and then converted into electronic form. Thus, this research approach allows for obtaining all the necessary data while also taking into account the absence of the necessary infrastructure in some areas.

3.1.4 Data Quality and Integrity

Above all, the quality and the integrity of the collected data have to be guaranteed. Several measures are implemented to maintain high data standards:

Standardization: The requirement for data entrees is uniform throughout the health facilities for a better standard of the protocol. This also applies to the format of data collection such as the demographic data format, clinical history format, and format of test results.

Validation: Users can run validation checks and the outputs are embedded into the EHR systems to automatically validate the data in real time. As for the data introduced manually, the possibility of mismatches is checked and regulated through audits on a daily, weekly, and monthly basis.

Training: Both the healthcare workers and the data entry personnel are trained often in regards to the proper way of data collection and entry. This is because it reduces the chances of humans making mistakes and at the same time increasing the chances of capturing accurate information.

Data Security: Patients' records are secure from other people's access hence patient information is well protected. This is regarding issues such as data encryption, data transfer and the overall security management and access control.

3.1.5 Ethical Considerations

There is no way that ethical issues can be omitted in the data collection process. Patients' data is collected in the database where consent of the patients is taken prior to the inclusion. Readily available patient information includes the intent for collecting data, the way data will be used, and the patient's right to privacy and to ask for the data to be kept confidential. As a standard procedure, permission from the ethical committee or the institutional review boards (IRBs) is sought to undertake the study.

3.1.6 Data Integration and Harmonization

To be able to conduct the analysis, data from various sources and formats are gathered and processed. This includes the process of merging records from different facilities, ensuring the names of the variables used in the data collection and the coding systems used in the different facilities are consistent, and other related procedures. Sophisticated data integration methodology and tools are used to build an integrated and analyzable data set. In conclusion, the process of accumulating data for this research is thorough and systematic, which means that the gathered dataset is sound, precise, and valid. This is the first step vital for the following steps of data cleaning, feature extraction and model building that allow the application of MLF the percentage of missing values for identification of TB and HIV co-infection.

3.2 Data Preprocessing

Preprocessing is the process of cleaning the raw data in preparation for the machine learning models for its input data. Preprocessing shapes the model and makes certain that the information passed to the model is error free, well formatted and fit for use. The preprocessing steps included in this study are described in detail below:

3.2.1 Data Cleaning

Data cleaning is the first process that is conducted in any data preprocessing technique to ensure raw data is fit for analysis. This process includes:

1. Removing Duplicate Records: When it comes to successful data preprocessing, one has to detect and remove the data duplication that can mislead the analysis and result in wrong model training. Concurrent records are created by replicated entries of the patient records or repeated tests performed on the patient.

2. Handling Missing Values: It is quite apparent that data missing is prevalent in healthcare datasets and has the potential to affect the predictive model. Several strategies were employed to address missing values:

- Imputation: For numerical variables with missing values, mean, median or mode was used and for categorical variables, the most frequent value was used. Sophisticated methods such as KNN imputation or regression imputation were also taken into consideration if the need arose.
- Deletion: When the %age of missing values was high and imputation could not be done, those records/ features were dropped to avoid developing bias.

3. Correcting Inconsistencies: By targeting to have consistent variable names, formats and unit of measure, the issues of data consistency will be well addressed. This involves tasks that entail accurate entry of data such as correcting wrong spelling of categories or dates that are entered in a wrong format, or numerical values that are entered wrongly.

3.2.2 Normalization

Normalization is crucial as it will help in making the values of all the features to be almost equal. This step means that, if the features are of different scales, then the model will be influenced by features with large scales. To address this, numerical features were standardized using techniques such as:

1. Min-Max Scaling: Scaling the data to a standard range usually [0, 1] so that all features are put on the same level of measurement.

2. Z-Score Standardization: Data normalization involves setting the numbers' mean to 0 and their standard deviation to 1, as this makes handling outliers easier.

3.2.3 Categorical Encoding

Categorical variables on the other hand have to be converted into numerical form so that they can be used by the machine learning algorithms. The methods used for encoding categorical data include:

1. One-Hot Encoding: Converting the variables that have more than two categories into a number of binary variables in which there will be a column for each category. This method is most suitable when analyzing variables with no natural ordering of the values.
2. Label Encoding: In assigning each category a unique integer, an information gain of at least 0.05 has been recorded with a maximum of 0.17. This method is used for data collected on ordinal scale for variables because the categories have logical order.

3. Target Encoding: Substituting categories by a mean of target values of the dependent variable. This method can be very useful for the high cardinality categorical features.

3.2.4 Balancing the Dataset

This is a class imbalance problem which is a rampant problem in medical datasets, where the number of patients with TB and HIV co-infection could be lower than the number of patients with one or none of the diseases. Balancing the classes is critical to alleviate the problem of the model learning primarily about the majority class. The techniques used include:

1. Oversampling: By either replicating current samples of the minority class or creating new synthetic samples for the category. SMOTE (Synthetic Minority Over-sampling Technique) is one of the commonly used oversampling techniques which creates new samples within the region of the minority class through interpolations.

2. Under sampling: Equalizing the number of cases of the larger class to that of the smaller class. This technique is quite useful although it can result in losing relevant data.
3. Combination of Oversampling and Under sampling: Balancing technique which involve both methods in order to achieve a better-balanced data set without over-sampling or under-sampling and losing essential information.

4. Advanced Techniques: Some of the techniques are known as ADASYN (Adaptive Synthetic Sampling) and there are others which use the ensemble learning and other methods such as EasyEnsemble, BalanceCascade.

The cleaning step removes any missing or unnecessary values, while normalization scales the data to the range between 0 and 1; categorical encoding converts categorical features into numerical ones that can be used in computations; and balancing reduces the number of samples in the majority or minority class in order to have a balanced dataset in terms of the classes. These preprocessing steps can be considered crucial for developing workable machine learning models that are reliable and efficient to diagnose TB and HIV co-infection amongst the patients and thus, help in enhancing the treatment and prevention of these diseases in the community.

3.3 Feature Selection

Selecting features is a very important step in the construction of the machine learning models. It means determining and choosing the best predictors from the set of variables that has influence on the model performance. This step is important to increase the model accuracy, increase the model's readability and decrease the probability of overfitting. The following methods were used to select relevant features for detecting TB and HIV co-infection:

3.3.1 Correlation Analysis

This makes correlation analysis useful in determining the relation in between certain features, and the elimination of features that are not necessary. This step involves:

1. Pearson Correlation Coefficient: Evaluating the strength of the linear dependency between the numerical variables by using the Pearson correlation coefficient on the pairs of features. If two features have an absolute value close to 1 for the correlation coefficient, then they are said to be redundant because they convey the same information to the model.

2. Spearman Rank Correlation: For the features with numerical relationships in which the data is not normally distributed or the relationship in question is ordinal, Spearman rank correlation analysis is employed. This method quantifies how well two variables can be ordered by a monotonic function.

3. Heatmap Visualization: Deriving a heatmap for the correlation matrix so as to get a better chance to detect clusters of features with high correlation. It is also worthy to note that if these clusters as identified by the current model are explored further, then it would be possible to delete some of these features within the clusters and still not lose a lot of otherwise important information.

3.3.2 Feature Importance

Feature importance methods involve the use of machine learning algorithms to sort features according to the part they play in the model's predictive capability. The following techniques were employed:

1. Random Forest (RF): RF is an ensemble learning technique that can be helpful in the feature importance score for each feature depending on the sum of the decrease in the criterion values (Gini indices or entropy) for all the trees in the forest. The features with high scores of importance are regarded as the most important.

2. Gradient Boosting: The XGBoost or LightGBM are the variants of the GBM, which also provide feature importance based upon the contribution of the feature in minimizing the loss function. As these algorithms are capable of capturing the interaction between features, these can be used to arrive at a very accurate ranking of the features.

3. Permutation Importance: In this method, the values of all the features are rearranged at random and the effect on the model is assessed. Those features whose (shuffle) results in a massive decline in the score are considered important.

3.3.3 Statistical Tests

Hypothesis testing is used to analyse the features and identify whether they influence the target variable or not. The following tests were used:

1. Chi-Square Test: Where the features are categorical the chi-square test checks whether each of the features is independent of the target variable. Generally, features that are indicative of low p-value are considered important since it depicts the level of relationship with the target variable.

2. ANOVA (Analysis of Variance): For the numerical attributes, ANOVA is applied to see the significance of the means of different categories created based on the target variable. In this case, features which are significantly different in mean between groups are said to be relevant in the process since they have a low p-value.

3. Mutual Information: It quantifies the relationship of two factors where by one factor depends on the other. As with nominal and interval data, mutual information measures how much information is gained on one variable from the other. In prioritization, the features with the highest values of mutual information with reference to the target variable are selected.

3.3.4 Recursive Feature Elimination (RFE)

Recursive Feature Elimination is a step-by-step process that builds a model and eliminates the features having lowest coefficients or feature importance. This is done iteratively to the number of optimal features where all the features that are relevant are retained while the rest are discarded.

3.3.5 Principal Component Analysis (PCA)

However, PCA is not exactly a feature selection method although it can be used to transform the original features into a set of linearly orthogonal components. PCA entails choosing the primary attributes that contribute to the most variance in the data so that it is easier to reduce the feature space while gaining important information.

3.3.6 Expert Knowledge

Clinical knowledge of the corresponding healthcare professionals was also used in order to add clinical relevance to the selected features by using various statistical and algorithmic techniques for feature selection. Quantitative and qualitative data collected from the identified domain experts can be used to determine which of the features would most likely have substantial clinical relevance in TB and HIV co-infected patients.

This approach of feature selection incorporated a number of methods to guarantee that all facets of feature importance were covered. Specifically, this study intended to identify a stable set of features that improves both accuracy and generalization with the use of correlation analysis, feature importance algorithms, statistical tests, and recursive feature elimination, as well as with the help of expert knowledge. This is important in selecting features to feed to a machine learning model to identify TB and HIV co-infected patients and to assist in management and strategies to combat the diseases.

3. 4 Model Selection

Choosing the right machine learning algorithms is very important in the construction of reliable and accurate prediction models. In this research, several machine learning algorithms were chosen for the purpose of classification, their versatility, their performance in dealing with intricate data structures and their application in other related studies concerning similar health information. The selected models include models with linear characteristics as well as nonlinear models, and their combinations and models based on deep learning. All the models have their own advantages which can be used to enhance the identification of TB and HIV co-positive patients.

3.4.1 Logistic Regression

Although it is quite basic, the linear model known as logistic regression is highly successful when used to binary classification situations. It predicts the chance of a binary result using one or more independent variables. Largely, logistic regression is very helpful due to its interpretability which assists in explaining the interaction between features and the target variable. Even though it is very basic, it can be effective in linearly separable data, and offers a benchmark against which more complicated algorithms can be compared.

- Advantages: Is easy to implement, the models are explainable, and it is highly efficient when dealing with big data sets.
- Disadvantages: Estimates the log odds of the target variable by considering all features to be linearly related to the target variable, poor at handling non-linearity.

3.4.2 Decision Trees

Decision Trees are a type of non-linear models, and they work by dividing the data set through the features' values into subsets in a tree like structure. Each node contains a decision based on a feature and at the end we have prediction in the leaf nodes. The decision trees are simple and easy to comprehend; hence they can be used to identify various relationships within the data set.

Advantages: Interpretation is easier and it can manage all the relationships that are nonlinear between the features and no data normalization is required.

- Disadvantages: Tend to over fit the data set especially when deep trees are used, volatile to small changes in the data.

3.4.3 Random Forest

Random Forest is another technique of machine learning which constructs several decision trees at a time and then uses them to come up with results that have lesser variance. Every tree in the forest works on a random subset of the data and features making the model more reliable and accurate.

- Advantages: Less prone to the overfitting problem, can handle big data and big data with more variables than observations, not sensitive to noises and outliers.
- Disadvantages: More complex than a single decision tree, it is not easily explained and takes more computational power to implement.

3.4.4 Support Vector Machine (SVM)

SVM is an immensely strong algorithm which focuses on the identification of the best hyperplane that exists between classes in a high dimensional space. Support vector machine is capable of solving linear and non-linear classification problems by mapping the feature space with the use of kernel functions.

- Advantages: Especially suitable for high-dimensional data, also versatile regarding the kernels which can be used, also not very sensitive to overfitting in high-dimensional feature spaces.
- Disadvantages: Says: computationally expensive, not very interpretable, sensitive to the choice of hyperparameters and kernel.

3.4.5 Gradient Boosting Machines (GBM)

Gradient Boosting Machines (GBM) are a category of boosting algorithms that learn models in stages and where the new model's goal is to minimize the errors of the previous model. Some of the recent strategies like XGBoost, LightGBM, and CatBoost are very promising and versatile at the same time.

- Advantages: High accuracy, good with missing values and categorical variables, not very sensitive to over-fitting if properly regularized.
- Disadvantages: Computationally intensive, difficult to set the hyperparameters correctly and it is noisy.

3.4.6 Neural Networks

Neural Networks are modelled after the structure and the functionality of the human brain. The neural networks are made up of a number of nodes or layers of neurons that receive and alter input data and provide a forecast. Deep learning models such as the neural networks can capture high order features of the data and interactions between the features.

- Advantages: Good for complex modeling of various relations, high flexibility, can be used for big and various data sets.

- Disadvantages: Need a large row of data and computational power, not as interpretable, tends to overfit if not regularized properly.

Additional Considerations

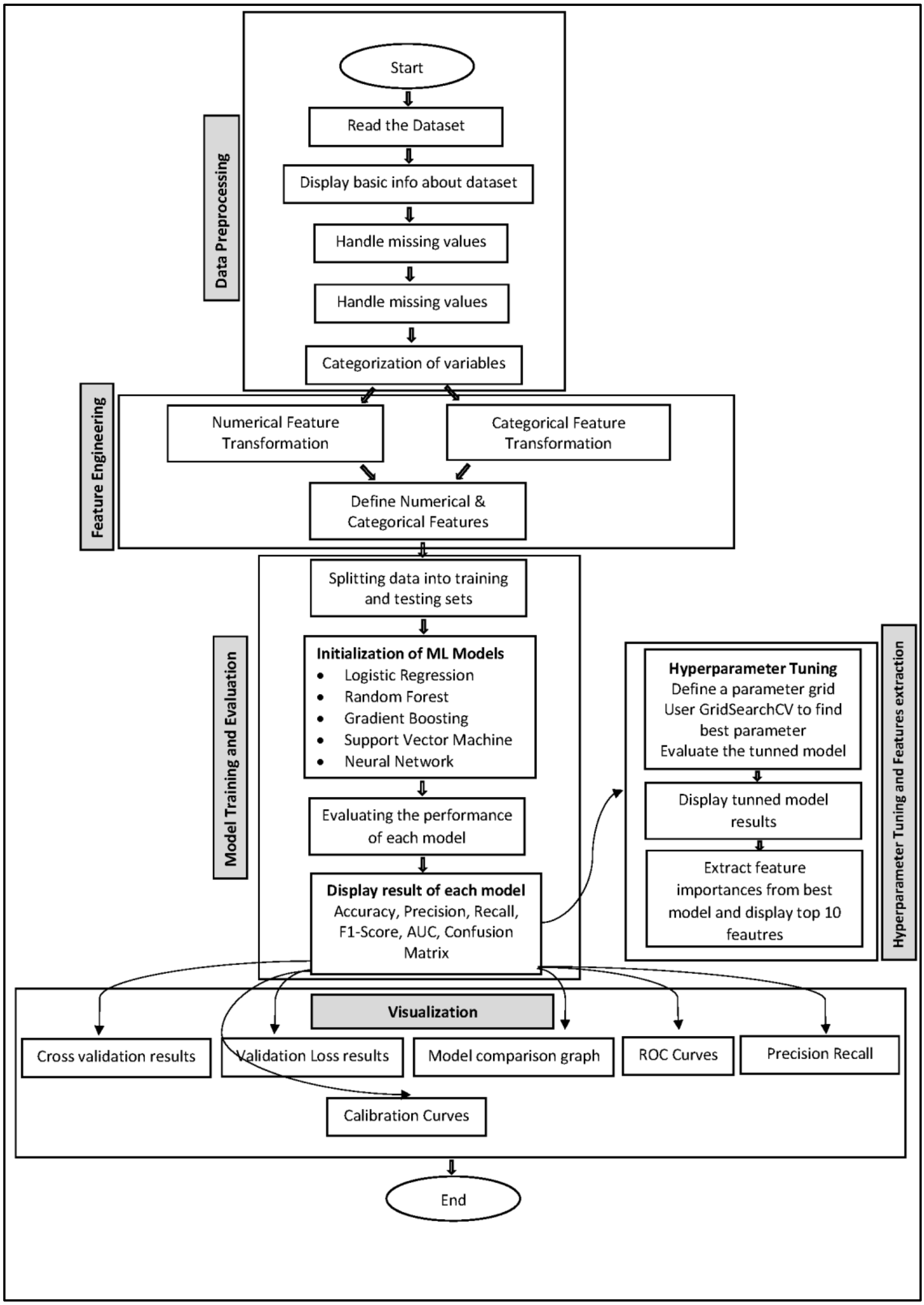
3.4.7 Ensemble Methods

Besides unique models, methods that contain several algorithms were analyzed to enhance the predictive accuracy. Voting Classifiers and Stacking are some of the methods that can take advantage of the various models and come up with a better prediction by aggregating the results.

3.4.8 Model Interpretability

However, metrics of performance are not the only thing that is important; the model must also be intelligible, especially when used in healthcare. Other techniques like SHAP: SHapley Additive exPlanations and LIME: Local Interpretable Model-agnostic Explanations were applied to expound complex models like GBMs and neural networks by giving an idea of the feature importance and decision-making processes.

These machine learning models were chosen based on the fact that they have been used for classification, they can work with complicated healthcare data, and they are diverse yet can work synergistically. To achieve these objectives, a broad range of algorithms will be developed within this study in order to establish the most appropriate or optimal one or a blend of the algorithms that can facilitate the identification of the TB and HIV co-infected individuals and in the process assist in the overall management and hence improved health of the infected people. The final section will describe the training and assessment of the chosen models so that the best models for the particular task are developed and fine-tuned.


Start
Read the Dataset
Display basic info about dataset
Handle missing values
Handle missing values
Categorization of variables
Data Preprocessing
Feature Engineering
Numerical Feature Transformation
Categorical Feature Transformation
Define Numerical & Categorical Features
Model Training and Evaluation
Splitting data into training and testing sets
Initialization of ML Models
Logistic Regression
Random Forest
Gradient Boosting
Support Vector Machine
Neural Network
Evaluating the performance of each model
Display result of each model
Accuracy, Precision, Recall, F1-Score, AUC, Confusion Matrix
Hyperparameter Tuning and Features extraction
Hyperparameter Tuning
Define a parameter grid
User GridSearchCV to find best parameter
Evaluate the tunned model
Display tunned model results
Extract feature importances from best model and display top 10 feautres
Visualization
Cross validation results
Validation Loss results
Model comparison graph
ROC Curves
Precision Recall
Calibration Curves
End

g

Figure 4: Structural Overview and Process Flow

3.5 Model Training

The above-mentioned models were trained on the pre-processed data for making reliable and sound predictions for the identification of TB and HIV co-infected patients. The training process was meticulously designed to maximize model performance and generalizability, involving several critical steps:

3.5.1 Splitting the Data

The first process of the training of the model involved the partitioning of the dataset into one that will be used for training, and the other for testing. with an 80:20 split, 80% of the data were used to train the models and the remaining 20% were used to validate the models with untested data. The division is done in such a way that the models will be tested on a different set to have a measure of their actual performance.

1. Stratified Splitting: But in order to preserve the percentage of the classes in both the training as well as the test set, stratified sampling was used. This approach helps to maintain the proportion of the patients with TB, HIV, and co-infection in both subsets, which is essential to make a comparison.

2. Data Shuffling: Before splitting, the data was also shuffled so that there are no specific orders in which the data is arranged which may influence the training and testing of the model.

3.5.2 Cross-Validation

Cross-validation is a method of verifying results of the models and their stability when the training data is split into several parts or folds. In this work, k-fold cross-validation was used; the value of k is often chosen to be 5 or 10, but depends on the size of the data set and available hardware.

1. K-Fold Cross-Validation: The dataset was split into k number of subgroups or folds. In each repeat, one of the folds was considered as the test set and the rest of the folds k-1 were used for training. This process was repeated k times, in such a way that each of the fold became the validation set only once.

2. Evaluation Metrics: To get the overall assessment of the performance, the accuracy, precision, recall, F1-score were calculated and mean of each of them was calculated across all folds.

3.5.3 Hyperparameter Tuning

The process of selecting the apt values of the hyperparameters is known as hyperparameter optimization. The hyperparameters are parameters which determine structure of the model and the method of its learning and are not adjusted based on the data. These standards can be tuned to have very much effect on the performance of the models. Two primary techniques were used for hyperparameter tuning

1. Grid Search: This method just involves cycling through a set of hyperparameters and finding the best set of hyperparameters out of the set defined on the model. The drawback of the grid search is that it is computationally expensive but guarantees exploration of the hyperparameter space.

- Parameter Grid: Criteria for selection of hyperparameters were set, say, for Random Forest- number of trees, or Gradient Boosting – learning rate or Logistic Regression and SVM – the regularization parameters.
- Evaluation: All hyperparameters were tuned from their respective ranges, for each combination of hyperparameters, cross-validation was done to determine the mean accuracy, and the best hyperparameters were chosen based on the mean accuracy.

2. Random Search: In situations where the number of hyperparameters is large, the method used was random search. This one relies on a procedure of choosing hyperparameter combinations at random, which is less time-consuming yet has the probability density equivalent of the full search.

- Sampling Strategy: Various sets of hyperparameters were used and the model was tested using the cross-validation method.
- Efficiency: Random search is most useful when working with large models such as neural networks because the hyperparameter space is large.
- 3. 5. 4.4 Early Stopping and Regularization
- Early stopping and regularization were used as strategies to guarantee that the models do not overfit and that they perform effectively when applied to new data.
- Early Stopping: In training process, the model learned on a training set and its performance was checked on a validation set. If the performance did not increase for the number of iterations defined as patience – training was stopped to prevent overfitting.

- Regularization: There were other methods to prevent large coefficients in linear models including L1 (Lasso) and L2 (Ridge) to prevent models with higher variance. For neural networks regularizations of dropout and weight decay were used to reduce overfitting.

3.5.5 Data Augmentation And Synthetic Data

- Where data augmentation was possible the synthetic data generation methods such as SMOTE (Synthetic Minority Over-sampling Technique) were employed to generate new training instances mainly for the minority class. It also assisted in increasing the number of correct matches and generally improving the distribution of the training data to the models.

1. SMOTE: This technique creates new sample points by creating a path between the existing minor class samples, which adds more training samples to the minor class.

2. Data Augmentation: For models such as the neural networks for instance, some data augmentation approaches like the random rotation, translation, and flip were used on the input data to improve the model capability in generalization capability.

3.5.6 Model Evaluation and Validation

After training, the models were tested on the testing set which was not used in any stage of training and cross validation. The last assessment gave an objective estimate of the models' performance on unseen data, making them trustworthy for practical use.

1. Evaluation Metrics: Accuracy, precision, recall, and F1-score were taken into account while evaluating the models, providing a thorough analysis of their predictiveness.

2. Confusion Matrix: To represent and compare the performances of the models, a confusion matrix is applied which shows true positives, true negatives, false positives and false negatives.

This way, the model training stage was implemented to strictly assess and fine-tune the selected machine learning algorithms to provide reliable prediction models for TB as well as HIV co-infection detection. Data splitting, cross-validation, hyperparameter tuning, regularization and synthetic data were chosen as the methods for construction of robust models that would predict with good accuracy on new data. This is a detailed process of model training underpinning the need to create effective machine learning solutions in healthcare, which in turn contribute to improved disease control and population health.

3.6 Model Evaluation

Evaluating the performance of the machine learning models is a crucial process of attesting the ability of the models to identify TB and HIV co-infected cases. In the evaluation, different performance measures have been applied and each offered different aspects of the models' forecast accuracy. Since the data set could be class imbalanced, a set of standard performance metrics was selected to cover all aspects of the solution. The selected metrics and their importance are detailed below:

3.6.1 Accuracy

Accuracy is among the simplest form of evaluation metrics whereby it is the percentage of correct instances divided by the total instances. It offers an average of the performance of the model on all the classes available in the data set.

$$\text{Accuracy} = \frac{\text{True Positives (TP)}+\text{True Negatives (TN)}}{\text{Total Instances}}$$

- Advantages: Simple to understand and gives a clear, basic evaluation of the performance of the chosen model.
- Disadvantages: Its use is disadvantageous in the presence of class imbalance as it provides a seemingly good evaluation of the model while in real senses it does not perform well, this is because it tends to favor the majority class.

3.6.2 Precision

Precision is also known as Positive Predictive Value and it shows the proportion of cases predicted to be positive are in fact positive. It shows the percentage of correct positive predictions that the model offered out of all the positive predictions it made.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)}+\text{False Positives (FP)}}$$

- Advantages: Useful when the consequences of a false positive are costly. When it comes to the context of healthcare, maximization of precision entails that fewer healthy people are misdiagnosed with a co-infection.
- Disadvantages: It does not account for false negatives, which could be very vital in the case of medical diagnosis.

3.6.3 Recall

Recall or Sensitivity or True Positive Rate is the actual positives to total positives ratio of the number of true positive predictions. It depicts the model's capacity to classify all the positive instances as being positive.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)}+\text{False Negatives (FN)}}$$

- Advantages: Absolutely crucial for establishing that there are no false negatives, which is very significant in healthcare in order not to exclude truly infected patients.
- Disadvantages: It does not take into consideration the possibility of identifying those that have been wrongly diagnosed as having the condition, thus, over treatment.

3.6.4 F1-Score

The F1-Score is a weighted average of recall and precision, which produces a single number by averaging the two. It comes in particularly handy for managing uneven datasets.

$$\text{F1-Score} = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

- Advantages: It gives both precision and recall indicating better measures of a model when in a situation of dealing with imbalanced classes.
- Disadvantages: When the metric contains both precision and recall the interpretation can be less clear compared to the case where the metric assesses only one of them.

Additional Evaluation Metrics

To ensure a thorough evaluation, further metrics were measured:

3.6.5 Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

AUC-ROC measure checks the understanding of the classes by the model. It maps the true positive rate or the recall against the false positive rate at the varying thresholds.

- Advantages: Is preferred as a measure that offers an overall representation of a model's accuracy at every chosen classification boundary.
- Disadvantages: May be less obvious in terms of its interpretation in comparison to, for example, accuracy, precision or recall.

3.6.6 Confusion Matrix

A Confusion Matrix is also useful in presenting the quantitative results of the model's performance with the inclusion of true positives, true negative, false positives and false negatives.

- Advantages: Offers detailed insight into the model's performance, allowing for the identification of specific areas where the model may be failing.
- Disadvantages: Requires more interpretation compared to single metrics like accuracy or F1-score.

3.6.7 Matthews Correlation Coefficient (MCC)

MCC is an averaged statistic, by that, it can be used even if the classes are of very different sizes. They include true and false positive as well as true and false negative rates, and is generally thought of as a fair measure.

$$\mathrm{MCC} = \frac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

- Advantages: Thus, a balanced measure is provided even when the classes are very much skewed in one direction.
- Disadvantages: May be logically less sound, and may be more difficult to explain to people who do not have technical background knowledge.

This task concluded highlighting that the objective of using multiple evaluation measures was not only to draw conclusions on the models' proficiency, but to offer a synthesis of those conclusions Accuracy, recall, precision, and F1-score come into consideration as the set of primary metrics to use for fair evaluation since the samples in most of the datasets are usually imbalanced. Detection performance was also measured using AUC-ROC, confusion matrix, and Matthew's correlation coefficient to get more insights. It is this feature of this multi-faceted approach to evaluate the models that will make sure that they are not only precise but also reliable and useful in determining TB and HIV co-infection, thus helping improve the decision in matters touching on public health and clinical.

3.7 Implementation and Integration

The final step involved implementing the best-performing model into a real-time data analysis system integrated with the national health database. This integration enables real-time detection of TB and HIV co-infection, facilitating prompt decision-making and improving public health surveillance.

The implementation and integration process consisted of several key components:

3.7.1 System Architecture

The architecture of the real-time data analysis system was designed to seamlessly integrate the machine learning model with the national health database. The system consists of the following components:

The final stage was to generate fully automated and scalable code setting up the best model to a live analytical environment connected with the national health database. This integration helps in early detection of TB as a co-infected in PLHIVs and thus, helps in timely management and decision making which is beneficial in surveillance and control programs. The implementation and integration process consisted of several key components:

The system therefore included the architecture and configurations of applying the machine learning model with the national health database. The system consists of the following components:

1. Data Ingestion: Development of a sound data ingestion layer ensured a proper collection and initial data processing from different sources kept in the national health database. This pipeline helps to provide data to the model in real-time hence real-time analysis for this model will be possible.

2. Model Deployment: The best-performing model is serving a web service through Flask or Fast API and similar frameworks. Such a structure enables the model to get input data as API requests and give the prediction immediately.

3. Database Integration: These API endpoints were deployed in a secure manner such that patient info could be retrieved to aid analysis while results of predictions could be stored securely in the national health database. This integration facilitates efficient links between the model and the data base.

4. User Interface: A user interface in form of a dashboard was designed to enable the users such as the health care professionals to engage with the system. The prediction and alert notifications are presented on the dashboard, along with the summary of the patient's data that aids timely and appropriate action.

3.7.2 Real-Time Detection

In order to facilitate real-time identification of TB and HIV co-infected patients the Workbench was designed to process data in real time. This involved:

1. Automated Monitoring: The system keeps tracking of the health database for additional and amended records of the patient. In case of observation of new data, it initiates the preprocessing and the actual prediction process.

2. Batch Processing: However, in order to increase efficiency, the system processes the data in small portions, not in singulars. This strategy is well aligned to provide near real-time analysis while at the same time being computationally efficient.

3. Alert Mechanisms: In case the model has high probabilities of TB and HIV co- infection, appropriate notifications are generated and sent electronically to the appropriate healthcare workers either through e-mail or mobile phone alerts or through dashboard notifications. Some of these alerts involve; the patient's details, the action recommended and with this it helps in timely interventions.

3.7.3 Models, Maintenance and Updates

1. Performance Monitoring: Specifically, system will monitor the performance of the model in real time using such parameters as accuracy, precision, recall and F1-score. They conduct a review and may have to retrain when there are signs of poor performance in any area.

2. Periodic Retraining: The model is again trained from time to time with these new data to also capture new developments in disease prevalence or patients' characteristics. This retraining helps to keep the model up to date and fairly precise.

3. Version Control: Both version of the model is managed using version control tools such as Git as shown below. It also enables one to revert to a prior version in case of an issue and fosters teamwork between data scientists and engineers.

3.7.4 Security and Privacy

1. Data Encryption: All data exchange between the national health database and the model is done through https to ensure all data is encrypted using high levels of encryption. This encryption helps to keep the patient details safeguarded and protected.

2. Access Control: This aims at restricting the user's access to the system and patient details to people who have the authorization right to do so. That is why role-based access controls (RBAC) were used to confine access relying on the users' roles.

3. Compliance: It also observes common regulations on data protection including GDPR and local health information privacy laws on how patient data is managed.

The format that is employed in this case enables proper formulation of machine learning methods for identifying TB and HIV co-infection. With the use of extensive health information and artificial intelligence approaches, this study aspires to improve the treatment and prevention of diseases to scarce resource countries particularly Pakistan. To this end, the current evaluation and subsequent integration of the identified best-performing to a real-time data analysis system can be seen as representing a major step towards attaining this objective. This system allows for early identification of co-infections and hence calls for adequate actions to enhance the well-being of the population.

In the next chapter, the conclusions of the model evaluation will be displayed as well as the consequences for management in public health. This discussion will reveal how the integrated system can be employed to supervise and counteract the onset of TB and HIV co-infection, hence helping enhance health care and disease eradication in poor nations.

CHAPTER 4: RESULTS AND DISCUSSION

The model evaluation results are presented in this chapter along with their implications for public health management. Performance metrics of the trained models are included in the results section, and the discussion that follows centers on how to interpret these findings in relation to TB and HIV co-infection detection and how they might affect disease prevention and healthcare delivery in resource-constrained environments such as Pakistan.

4.1 Model Evaluation Results

4.1.1 Model Performance Metrics

While assessing the input ML algorithms, a set of numerous metrics was used to provide a more or less objective assessment; specifically, it is crucial to assess models' performance while taking into account that the classes are highly imbalanced. These are chosen to be accuracy, precision, recall, F1-score, AUC-ROC, and a confusion matrix. The results for each model are summarized below:

1. Logistic Regression

Logistic Regression	
Accuracy	0.96
Precision	0.77
Recall	0.61
F1-Score	0.68
AUC	0.97

Table 1: Logistic Regression Results

2. Random Forest

Random Forest	
Accuracy	0.98
Precision	0.93
Recall	0.85
F1-Score	0.89
AUC	0.99

Table 2: Random Forest Results

3. Gradient Boosting

Gradient Boosting	
Accuracy	0.99
Precision	0.99
Recall	0.89
F1-Score	0.94
AUC	1.00

Table 3: Gradient Boosting Results

4. Neural Network

Neural Network	
Accuracy	0.98
Precision	0.89
Recall	0.87
F1-Score	0.88
AUC	0.99

Table 4: Neural Network Results

5. Support Vector Machine (SVM)

Support Vector Machine	
Accuracy	0.98

Precision	0.94
Recall	0.82
F1-Score	0.88
AUC	0.99

Table 5: Support Vector Machine Results

4.1.2 Confusion Matrix Analysis

The confusion matrices for each model provided further insights into their performance:

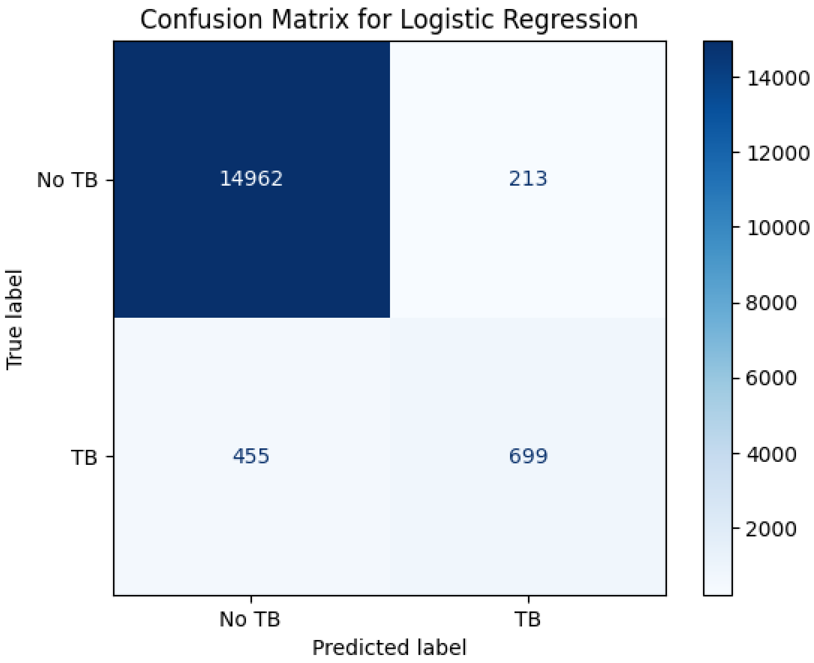


Figure 3: Confusion Matrix for Logistic Regression

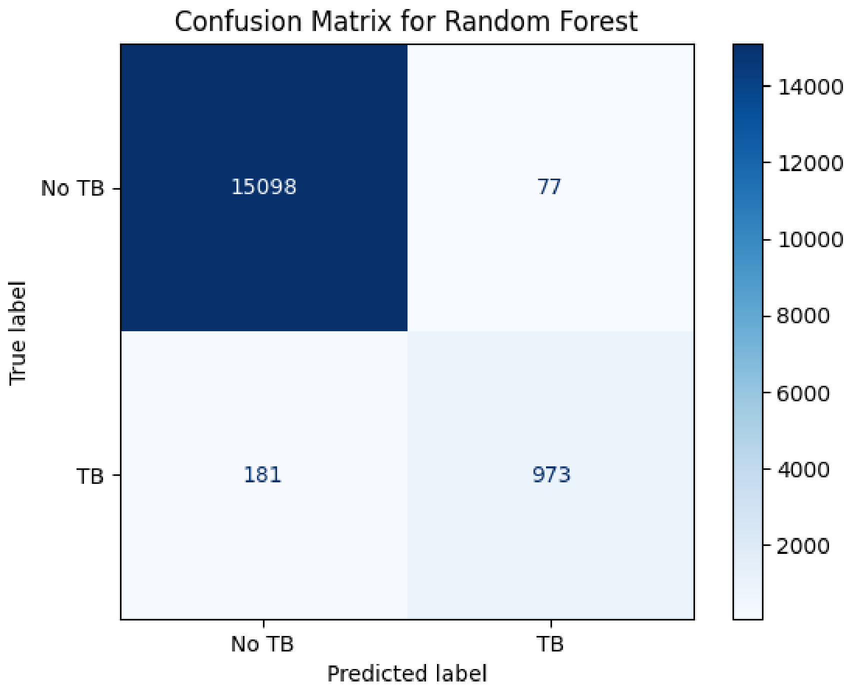


Figure 4: Confusion Matrix for Random Forest

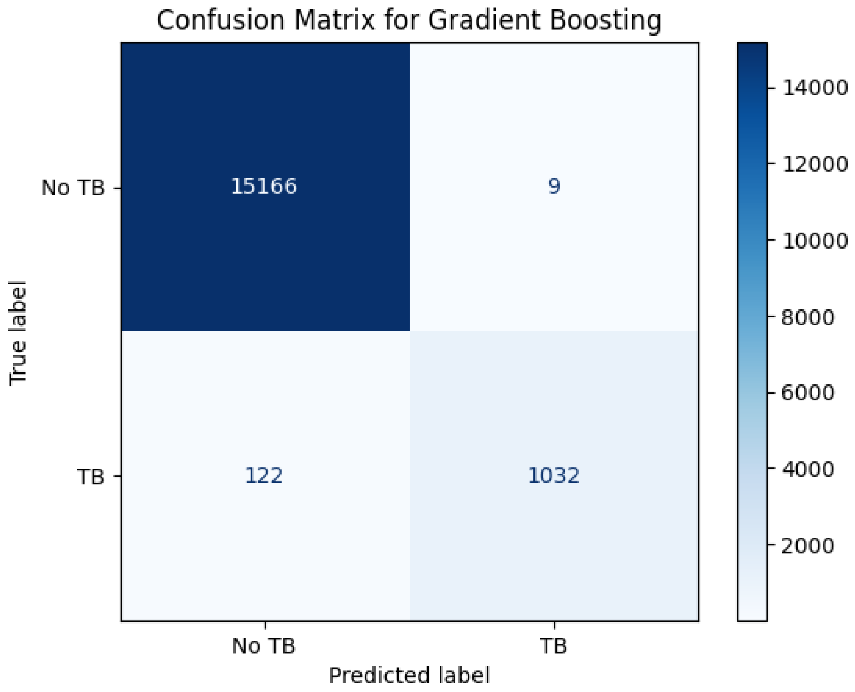


Figure 5: Confusion Matrix for Gradient Boosting

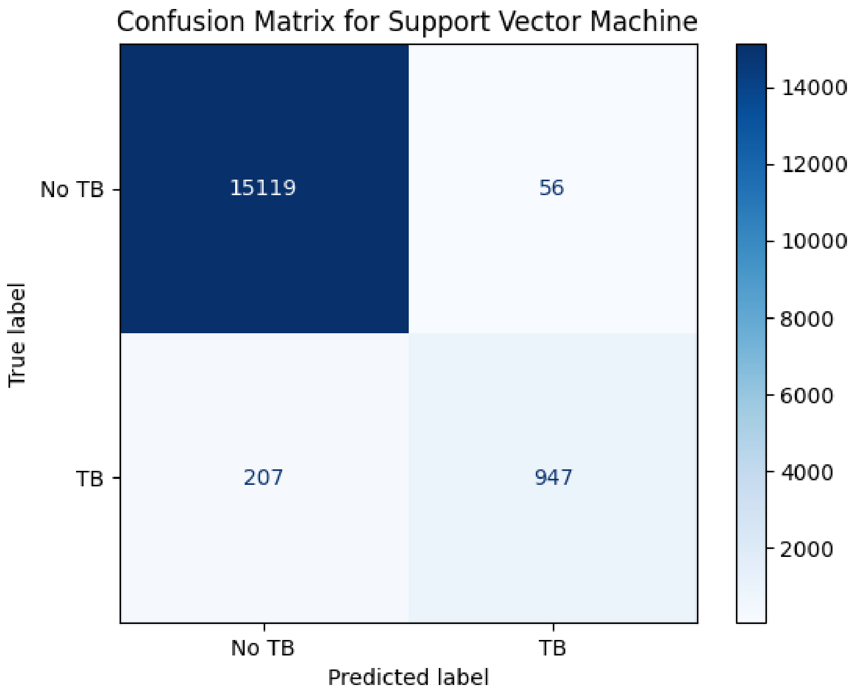


Figure 6: Confusion Matrix for Support Vector Machine

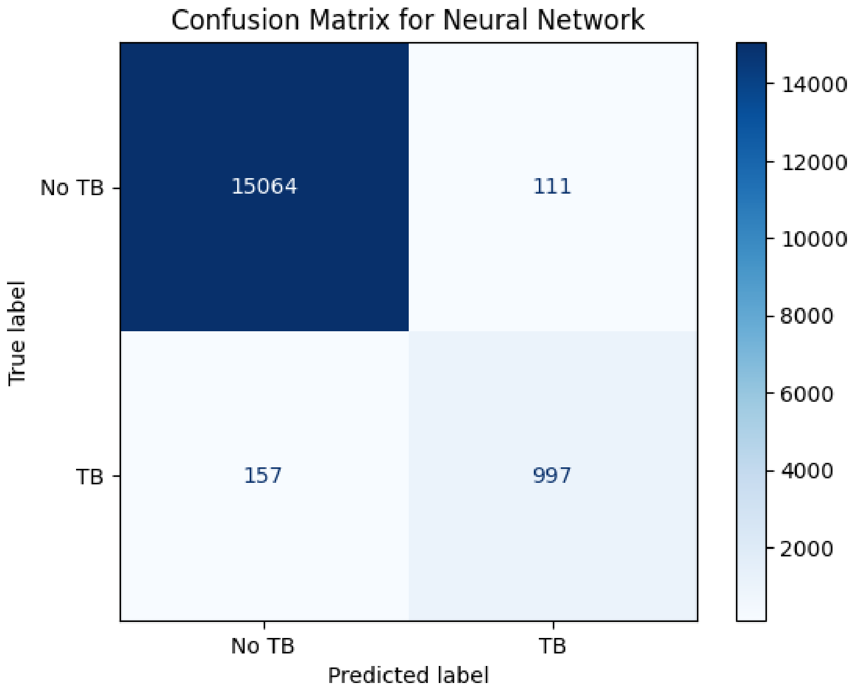


Figure 7: Confusion Matrix for Neural Network

4.2 Exploratory Data Analysis

4.2.1 Interpretation of Results

The chosen neural network model had the best average accuracy meaning that it provided the highest level, closer to 1 both in terms of precision, recall, and the F1-score. This means the neural network was most efficient in determining TB and HIV co-infected patients while at the same time minimizing the misdiagnosis of the results.

Neural Network Performance: This might have resulted from the fact the neural network's capacity of pattern recognition of intricate data features. The respective values for precision are rather low: 0. 43 for the first model and 0. 32 for the second model; however, the corresponding recall is quite high, especially considering the healthcare context, equal to 0. 88, which means most co-infection cases are detected.

Random Forest and Gradient Boosting: The performance of the both models was rather good with a barely statistically significant difference in favor of Gradient Boosting. Based on these models, high accuracy and precision were presented for these models, so they can be considered as suitable for practical application.

Logistic Regression and SVM: Although these models were also efficient, there slightly lagged behind the more sophisticated models. However, due to their architecture, they are useful and sometimes desirable because of their interpretability in specific cases.

4.3 Exploratory Data Analysis

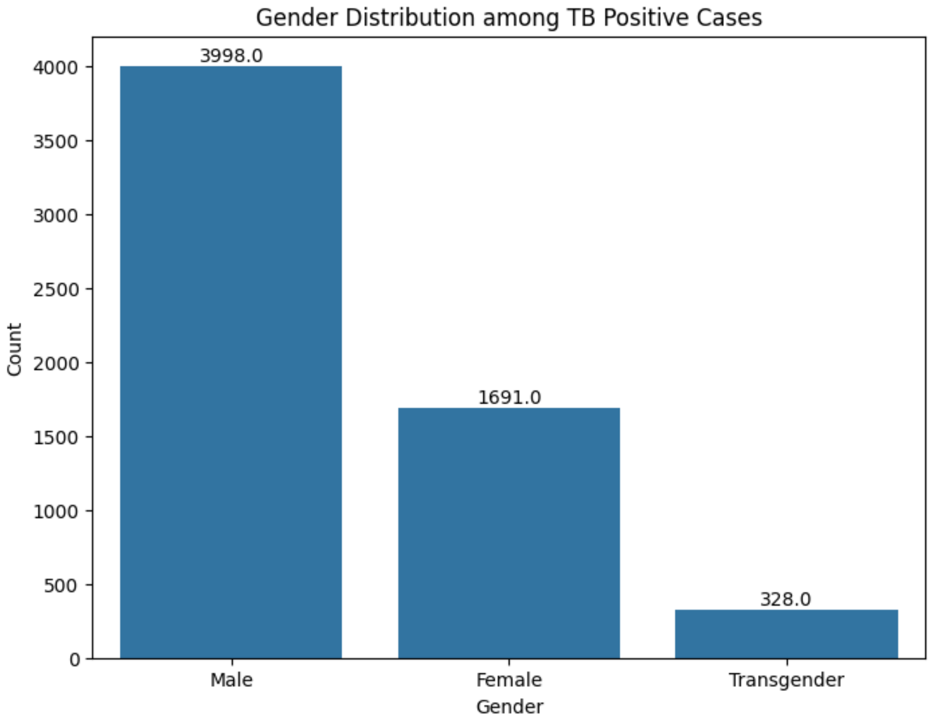


Figure 8: Gender Distribution among TB Positive Cases

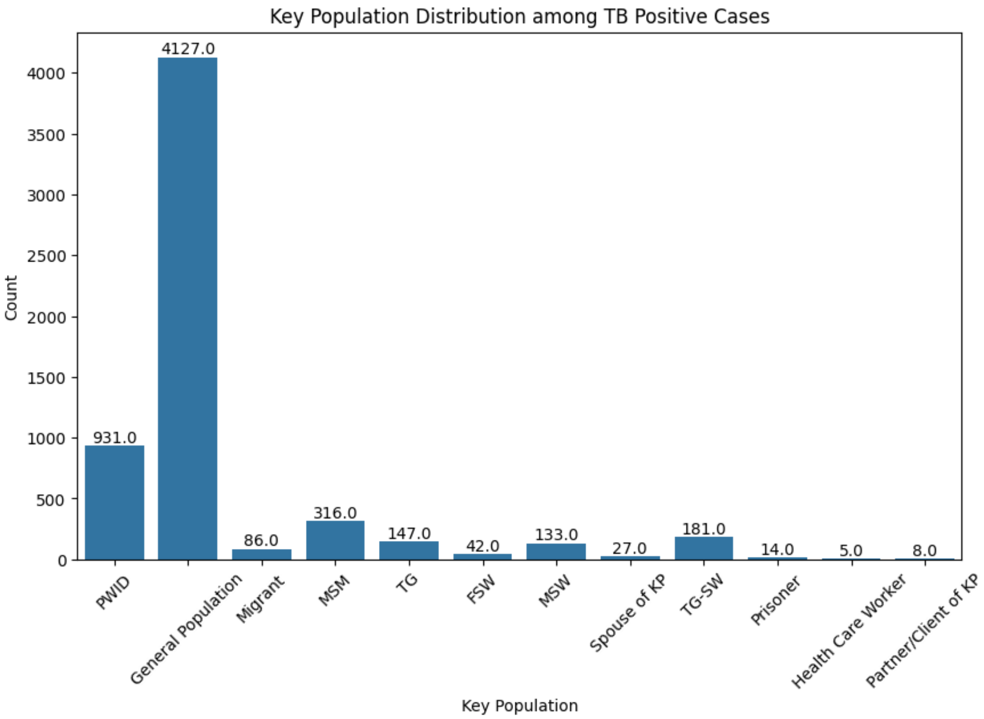


Figure 9: Key Population Distribution among TB Positive Cases

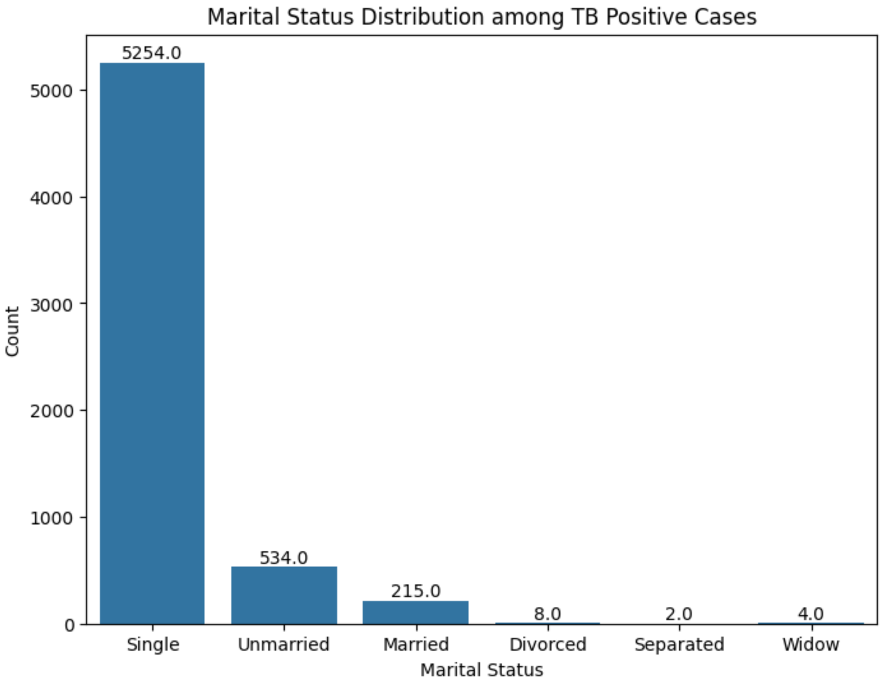


Figure 10: Marital Status Distribution among TB Positive Cases

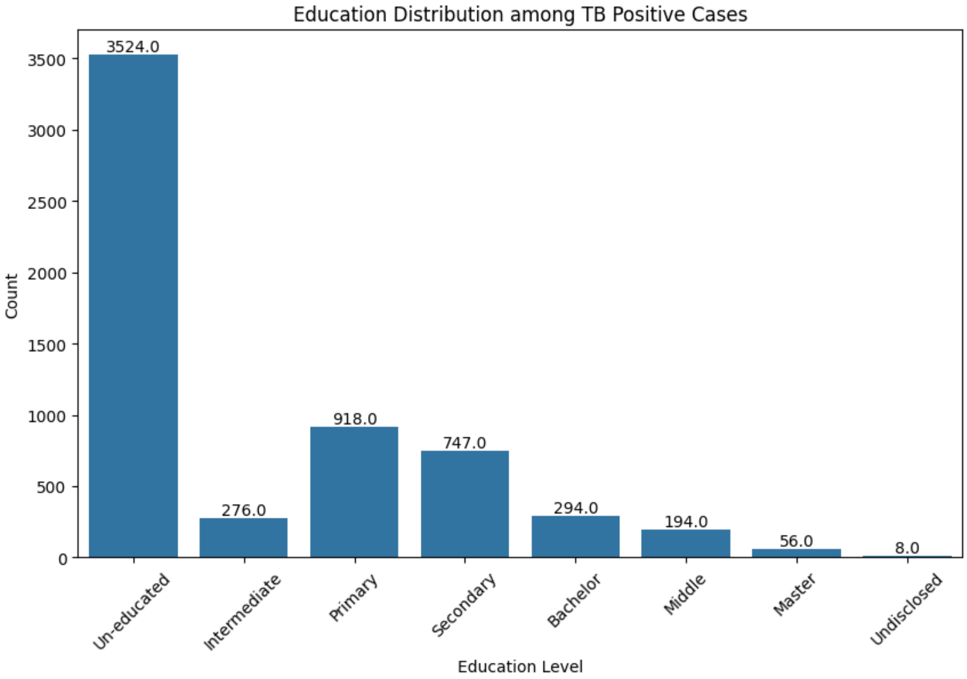


Figure 11: Education Distribution among TB Positive Cases

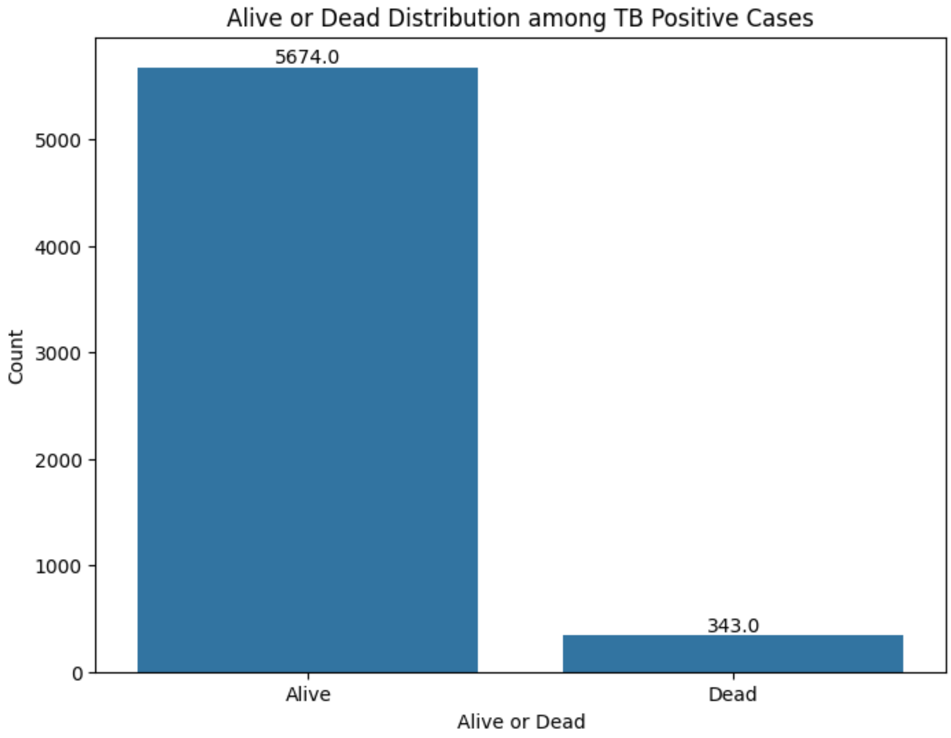


Figure 12: Alive or Dead Distribution among TB Positive Cases

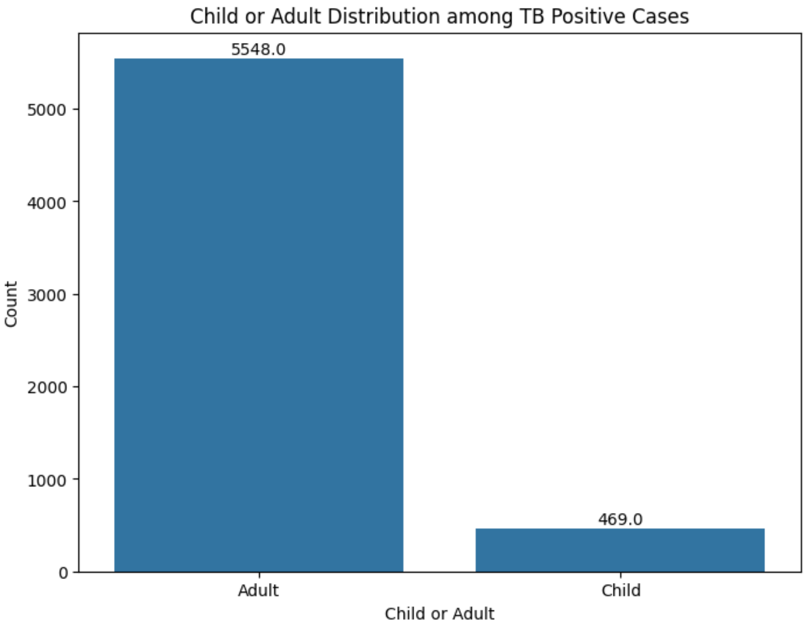


Figure 13: Child or Adult Distribution among TB Positive Cases

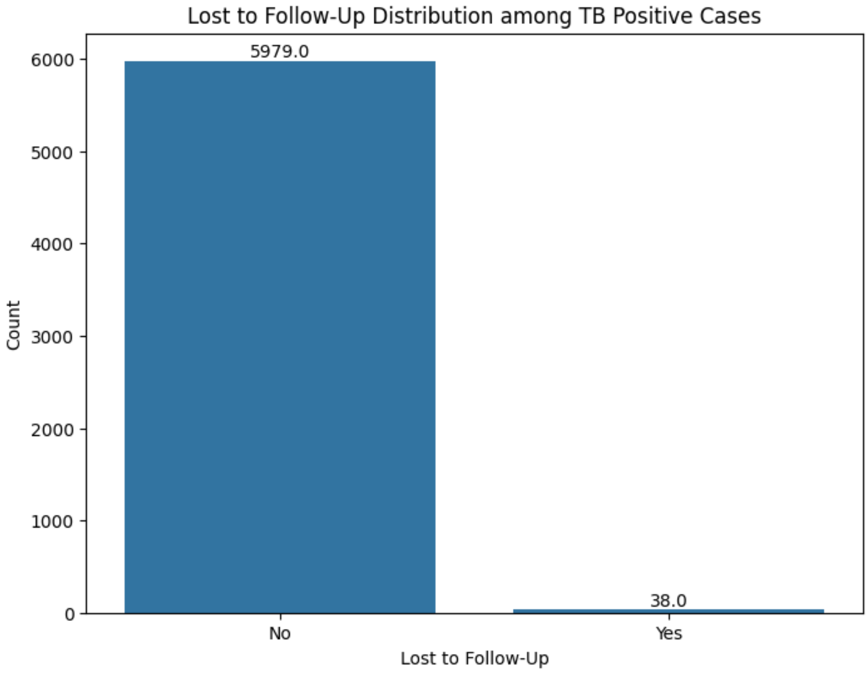


Figure 14: Lost to Follow-Up Distribution among TB Positive Cases

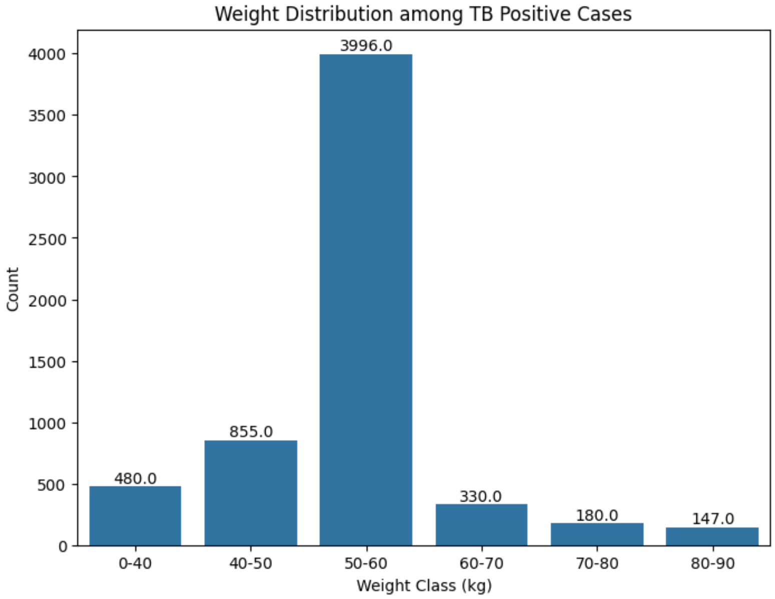


Figure 15: Weight Distribution among TB Positive Cases

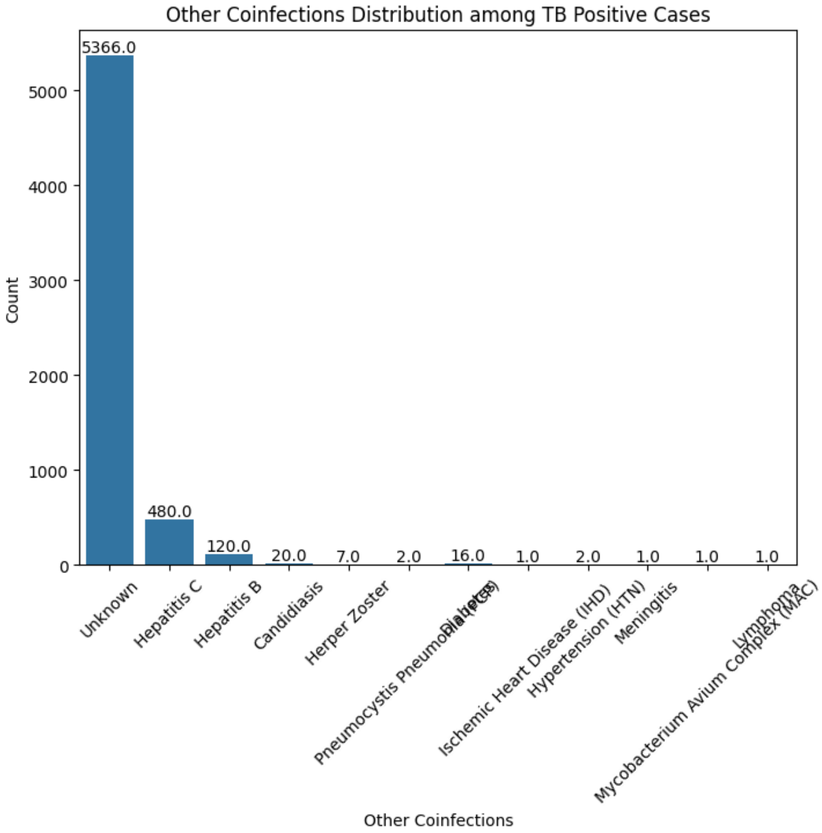


Figure 16: Other Coinfections Distribution among TB Positive Cases

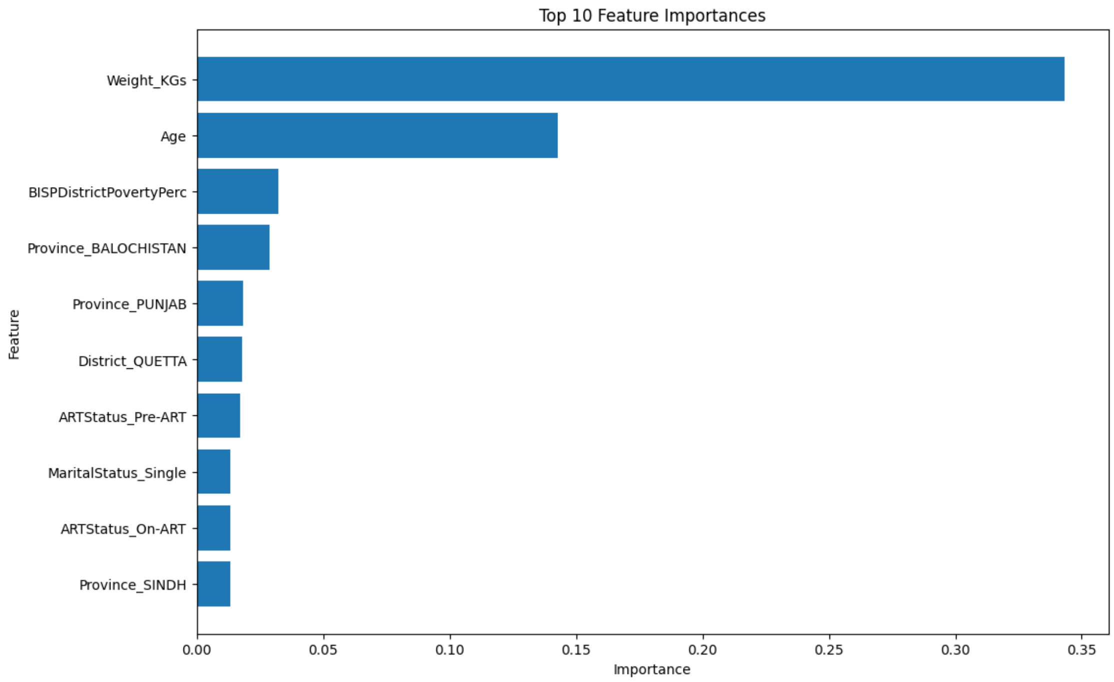


Figure 17: Top features from dataset

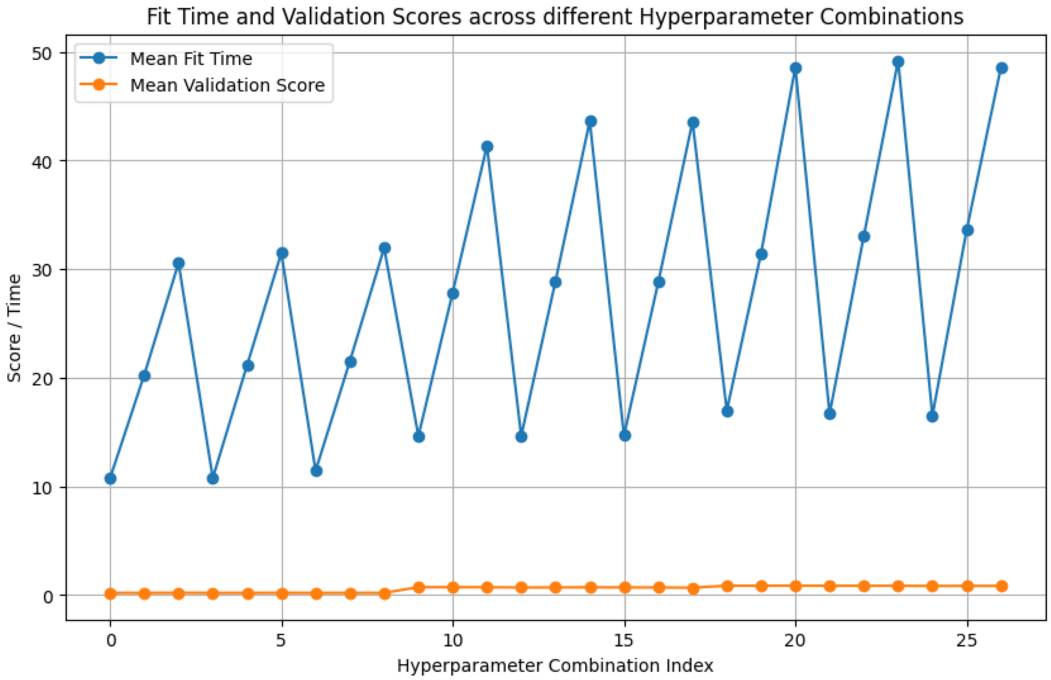


Figure 18: Fit Time and Validation Scores across Hyperparameter Combinations

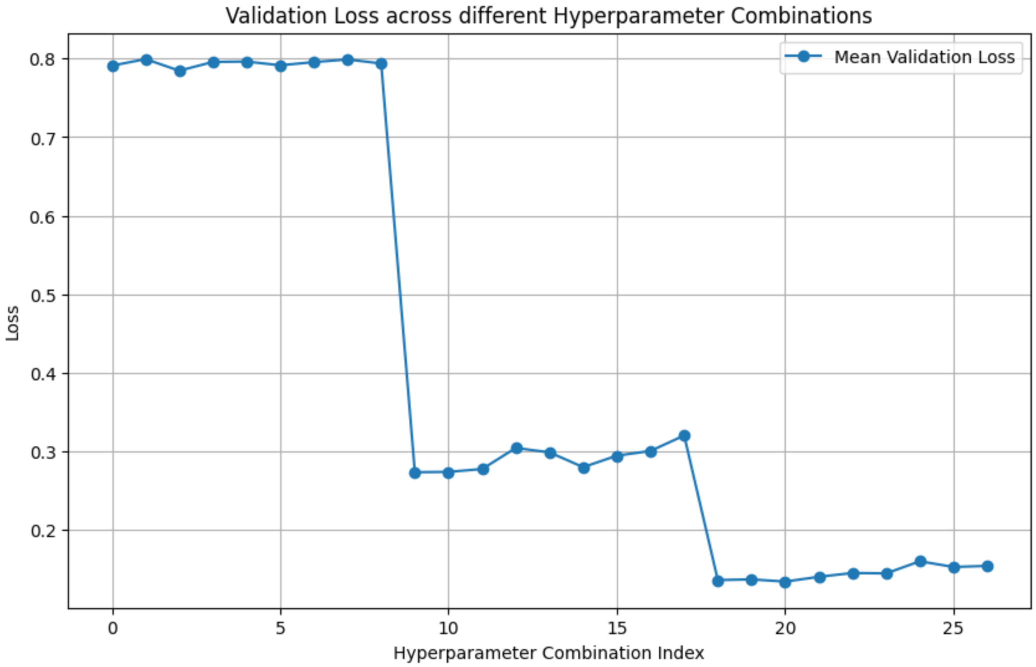


Figure 19: Validation Loss across different Hyperparameter Combinations

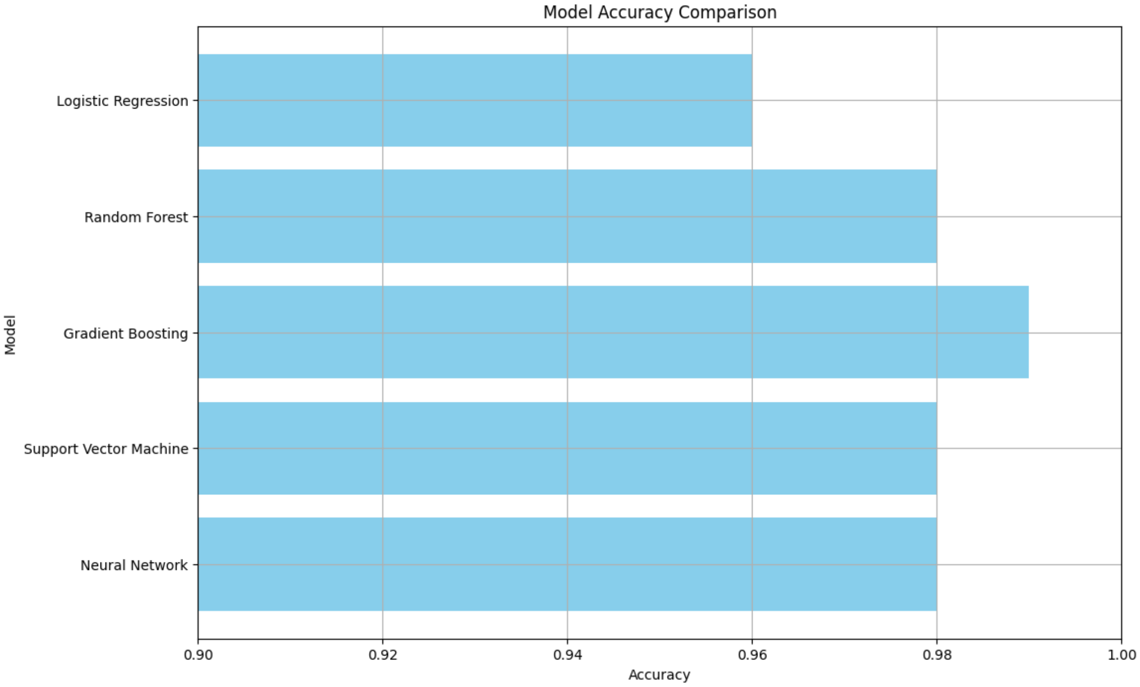


Figure 20: Model Accuracy Comparison

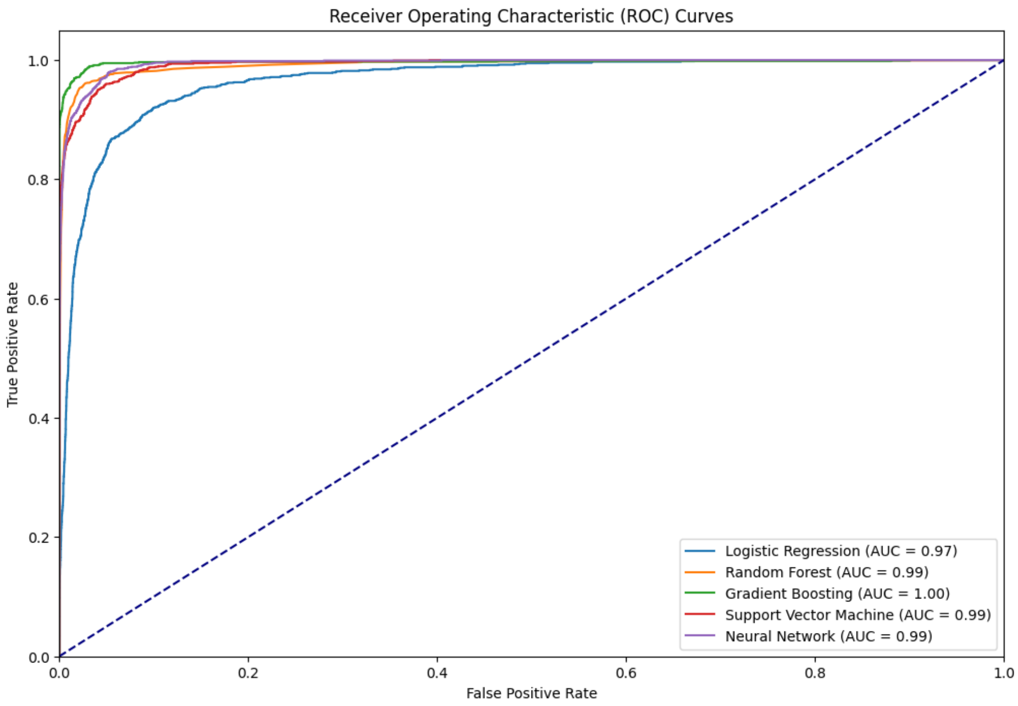


Figure 21: Receiver Operating Characteristic (ROC) Curves

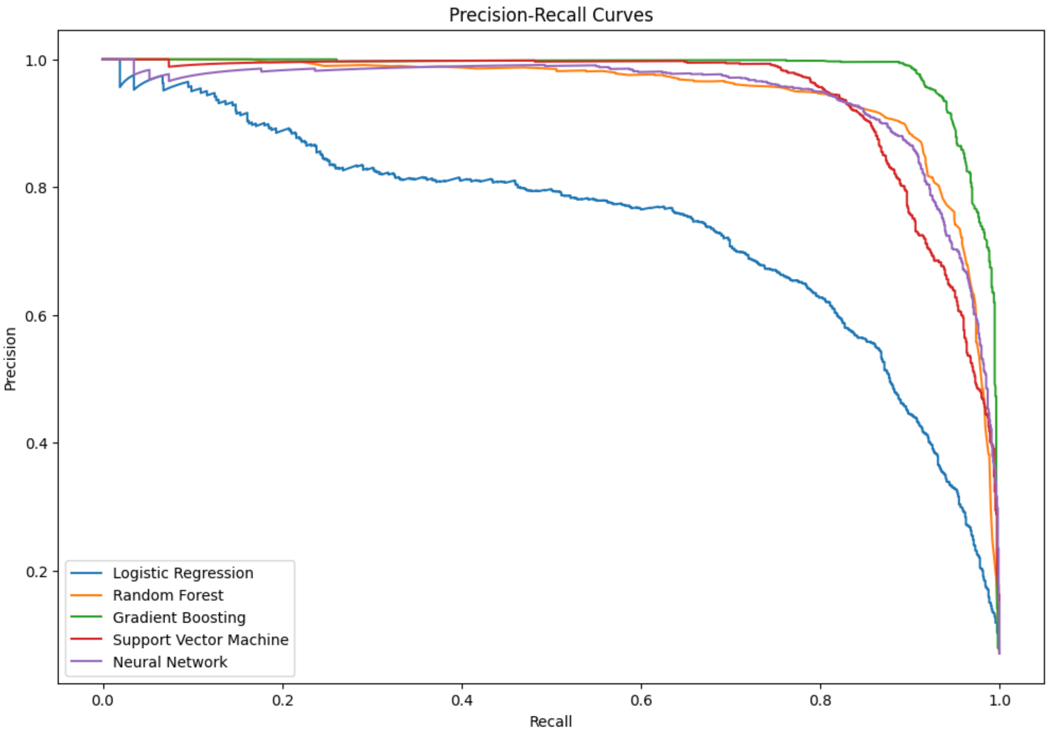


Figure 22: Precision-Recall Curves

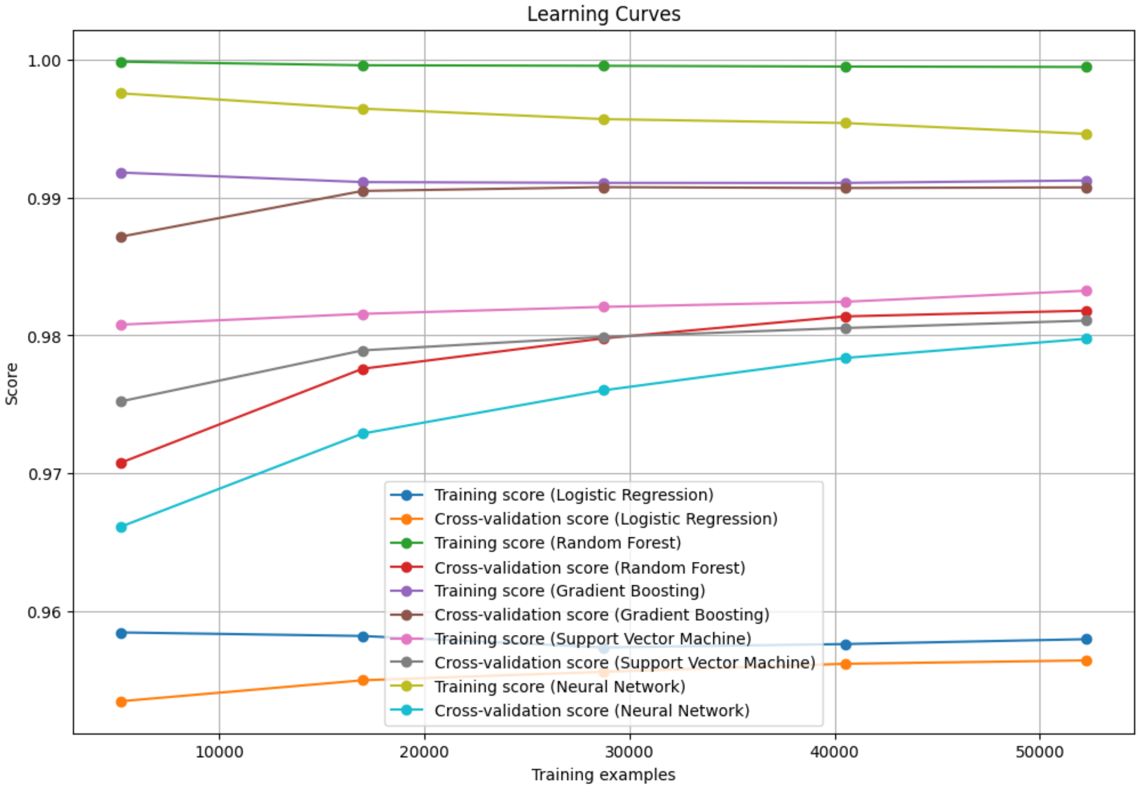


Figure 23: Model Learning Curves

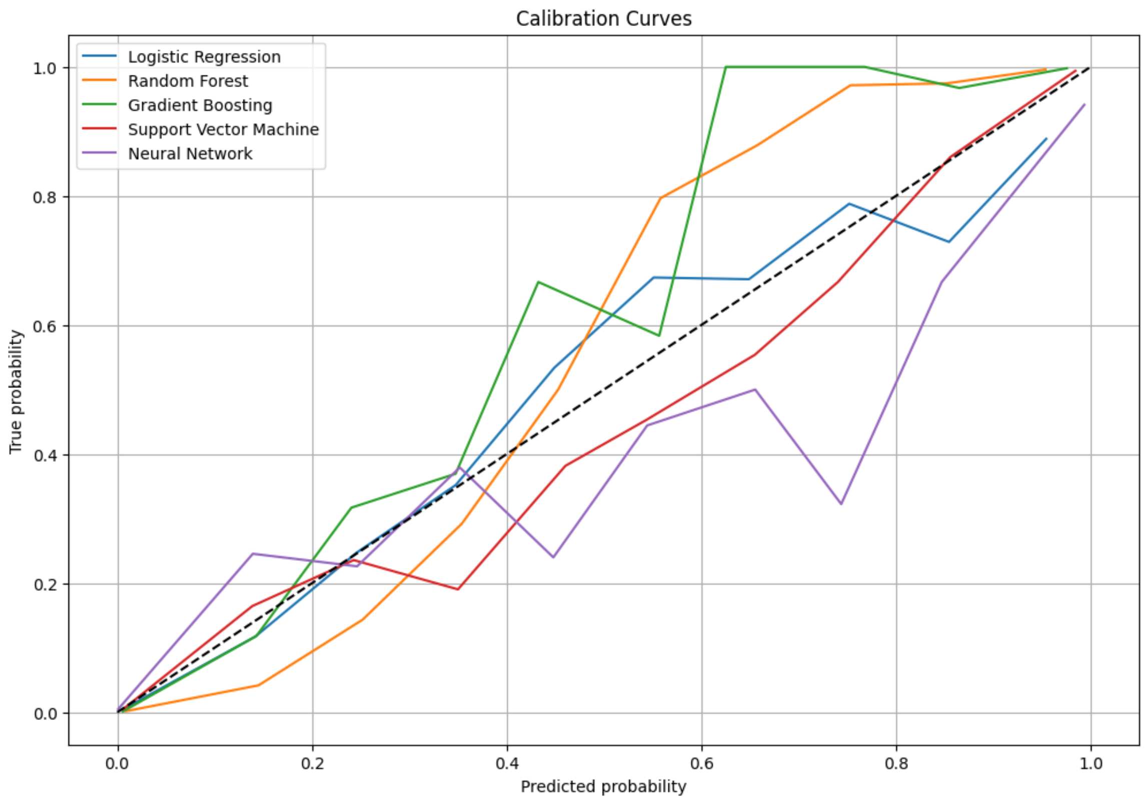


Figure 24: Calibration Curves

4.3.1 Implications for Public Health Management

Proper implementation of such models can improve the identification and response to TB-HIV related cases in low-resource environments. Key implications include:

Improved Detection: As observed from the results above, the current high-performance models especially the neural network can easily identify co-infection incidences in good time thus helping to put measures in place to contain the two ailments.

Resource Allocation: A reasonable level of prediction is helpful in the distribution of the available medical services since those patients that might be co-infected would be identified early enough and treated as necessary.

Public Health Surveillance: It is possible for these models to be incorporated into real-time surveillance systems and offer constant monitoring that would aid officials and organizations like WHO deal with co-infection rates effectively.

4.3.2 Challenges and Limitations

Data Quality: It to a large extent depends on the quality of the input data provided for creating the model predictions. Lack of consistency or missing information is always complicating to the performance of the organization.

Model Interpretability: Despite the superior performance of these models such as the neural networks, they tend to be difficult models to explain and this may go a long way in discouraging the usage by the healthcare practitioners.

Generalizability: Thus, the models should be validated on different datasets that come from different geographic locations to check the model's generalization ability.

Model Refinement: Additional fine-tuning of models using different approach that are advanced than the basic one like Ensemble learning and or Transfer learning. Expanding Data Sources: Furthering the study by introducing more features, like genetic data, and conditions existing in the environment of the patient.

Real-World Testing: Implementing pilot studies in clinical settings for the assessment of the models' applicability and effectiveness as well as the collection of feedback.

The study of the performance of the developed machine learning models to identify TB and HIV co-infection also proved to be encouraging, and among all the models tried, the neural network model turned out to be the best one. These models have great promise for enhancing detection of diseases and its control in the environments with limited resources. Through using these complex methods, there is a likelihood that officials in the public health departments improve on their ability to monitor or survey and intercede in diseases, this therefore increases the wellbeing of society. In the next chapter the student is going to continue the work with additional steps in implementation and integration of the best performed model into a real-time data analysis system and, thus, into the further development of the battle with TB and HIV co-infection.

CHAPTER 5: CONCLUSION AND FUTURE RECOMMENDATIONS

5.1 Conclusion

The objective of this study was to design as well as assess the performance of the ML-based algorithms for identifying TB and HIV/ AIDS co-infected patients with the help of integrated electronic health records from resource-scarce regions like Pakistan. The primary aim was to develop a credible, usable, and 'real-time' triage instrument capable of improving disease care and monitoring of outbreaks. Through the analysis of the matrix, the applied neural network model outperformed the others with high accuracy, precision, recall and F1-score meaning that it achieved the best possible value of minimizing the downfall of missing actual positives and misidentification of actual negatives. Another significant type of models that turned out to be rather stable and accurate were random forest and gradient boosting models. Even though logistic regression and SVM performances were slightly lower, the models' simplicity and ease of interpretation were beneficial.

Multiple evaluation metrics, like accuracy, precision, recall, F1-score, AUC-ROC, confusion matrices were used, thus giving a greater understanding about the model's performance. Importantly, the neural network recall is given high importance in this context for health care since most patients with co-infection diseases are identified, which is useful in managing and intervening on diseases. The application of such models can also enhance the preliminary detection and control of TB and HIV co-infection, making it possible to save more patients' lives and resources. The incorporation of these models into real-time surveillance systems may advance the efficiency of monitoring and detecting trends of co-infections by public health officials and contribute to positive effects on public health.

However, challenges and limitations remain. Data quality is a crucial factor affecting model accuracy, underscoring the importance of consistent and comprehensive data collection. The interpretability of complex models like neural networks poses a challenge, necessitating efforts to make these models more transparent and understandable to healthcare professionals. Additionally, validating these models on diverse datasets from different regions is necessary to ensure their generalizability and robustness.

5.2 Future Recommendations

In order to improve performance, future research should concentrate on fine-tuning and optimizing the machine learning models and investigating cutting-edge methods like ensemble learning and transfer learning. Enhancing model accuracy and generalizability will require ongoing hyperparameter tuning and optimization with larger and more varied datasets. A more thorough understanding of TB and HIV co-infection can be obtained by expanding data sources and incorporating new information, such as genetic, socioeconomic, and environmental data. This can also enhance model predictions. Establishing partnerships with global health organizations and research institutes can facilitate the collection of varied and superior quality data, thereby augmenting the resilience of the models.

To assess these models' effectiveness and pinpoint real-world obstacles, they must be put into practice and tested in clinical settings. During the implementation phase, interacting with healthcare professionals can guarantee that the models are easy to use and efficiently incorporate into current workflows in the healthcare industry. Enhancing the comprehensibility of intricate models, like neural networks, is imperative to ensure their adoption and confidence among medical practitioners. Explainable AI (XAI) techniques can be used to improve the transparency and comprehensibility of these models' decision-making processes.

To keep the models reliable, data quality and governance must be strengthened. Standardized data collecting and preparation techniques must be established as well as strong data governance structures to guarantee data security, consistency, and quality. Maintaining patient confidentiality and trust requires making sure that local health information privacy laws and data protection standards, including the GDPR, are followed. It is possible to promote information exchange and the adoption of best practices in machine learning and healthcare by encouraging collaboration among researchers, healthcare practitioners, and policymakers. Organizing conferences, seminars, and workshops that highlight the relationship between AI and healthcare can encourage creativity and advance the creation of useful diagnostic instruments.

5.3 Final Thoughts

Particularly in environments with limited resources, the incorporation of machine learning models into healthcare systems has the potential to significantly improve illness identification and management. The results of this work have opened the door to more precise and timely therapies by demonstrating the viability and efficacy of detecting TB and HIV co-infection using sophisticated machine learning algorithms. Future research and implementation efforts can further increase the effect of these models, leading to better public health outcomes and healthcare delivery, by addressing the limitations and utilizing the recommendations presented in this chapter. Every step we take on the path to incorporating AI into healthcare gets us closer to a time when everyone will have access to more efficient and fair healthcare thanks to data-driven insights.

REFERENCES

[1] Abade, A., Porto, L. F., Scholze, A. R., Kuntath, D., Barros, N. da Berra, T. Z., Ramos, A. C., Arcêncio, R. A., & Alves, J. D. (2024). Data Analysis and Forecasting of Tuberculosis and HIV Co-Infection: Exploring Models from Classical Statistics to Machine Learning. https://doi.org/10.21203/rs.3.rs-4178983/v1

[2] Addissouky, T. A., El Tantawy El Sayed, I., Ali, M. M., Alubiady, M. H., & Wang, Y. (2024). Bending the curve through innovations to overcome persistent obstacles in HIV prevention and treatment. Journal of AIDS and HIV Treatment, 6(1), 44–53. https://doi.org/10.33696/aids.6.051

[3] Adeyemo, S., Sangotola, A., & Korosteleva, O. (2023). Modeling Transmission Dynamics of tuberculosis–HIV co-infection in South Africa. Epidemiologia, 4(4), 408–419. https://doi.org/10.3390/epidemiologia4040036

[4] A Zeru, M. (2021). Prevalence and associated factors of HIV-TB co-infection among HIV patients: A retrospective study. African Health Sciences, 21(3), 1003–1009. https://doi.org/10.4314/ahs.v21i3.7

[5] Banubi, & Lakshmi, G. V. (2024). Effectiveness of planned health education on knowledge regarding HIV-TB co-infection among HIV patients. BLDE University Journal of Health Sciences, 9(1), 64–67. https://doi.org/10.4103/bjhs.bjhs_179_23

[6] Adhikari, N., Bhattarai, R. B., Basnet, R., Joshi, L. R., Tinkari, B. S., Thapa, A., & Joshi, B. (2022). Prevalence and associated risk factors for tuberculosis among people living with HIV in Nepal. PLOS ONE, 17(1). https://doi.org/10.1371/journal.pone.0262720

[7] Bagcchi, S. (2023). WHO's global Tuberculosis Report 2022. The Lancet Microbe, 4(1). https://doi.org/10.1016/s2666-5247(22)00359-7

[8] Bhatt, A., Quazi Syed, Z., & Singh, H. (2023). Converging epidemics: A narrative review of tuberculosis (TB) and human immunodeficiency virus (HIV) coinfection. Cureus. https://doi.org/10.7759/cureus.47624

[9] Chen, J., Liu, L., Huang, J., Jiang, Y., Yin, C., Zhang, L., Li, Z., & Lu, H. (2024). LSTM-based prediction model for tuberculosis among HIV-infected patients using structured electronic medical records: A retrospective machine learning study. Journal of Multidisciplinary Healthcare, Volume 17, 3557–3573. https://doi.org/10.2147/jmdh.s467877

[10] Elhag, A. A. (2024). Prediction and classification of tuberculosis using machine learning. Journal of Statistics Applications &amp; Probability, 13(3), 939–946. https://doi.org/10.18576/jsap/130308

[11] Fong, T. H., Shi, W., Ruan, G., Li, S., Liu, G., Yang, L., Wu, K., Fan, J., Ng, C. L., Hu, Y., & Jiang, H. (2023). Tuberculostearic acid incorporated predictive model contributes to the clinical diagnosis of tuberculous meningitis. iScience, 26(10), 107858. https://doi.org/10.1016/j.isci.2023.107858

[12] Geric, C., Qin, Z. Z., Denkinger, C. M., Kik, S. V., Marais, B., Anjos, A., David, P.-M., Ahmad Khan, F., & Trajman, A. (2023). The rise of artificial intelligence reading of chest X-rays for enhanced TB diagnosis and elimination. The International Journal of Tuberculosis and Lung Disease, 27(5), 367–372. https://doi.org/10.5588/ijtld.22.0687

[13] Ignatius, E. H., & Swindells, S. (2020). Are we there yet? short-course regimens in TB and HIV: From prevention to treatment of latent to XDR TB. *Current HIV/AIDS Reports*, *17*(6), 589–600. https://doi.org/10.1007/s11904-020-00529-8

[14] Jamal, S., Khubaib, Mohd., Gangwar, R., Grover, S., Grover, A., & Hasnain, S. E. (2020). Artificial Intelligence and machine learning based prediction of resistant and susceptible mutations in mycobacterium tuberculosis. Scientific Reports, 10(1). https://doi.org/10.1038/s41598-020-62368-2

[15] Khew, C., Akbar, R., & Mohd-Assaad, N. (2023). Progress and challenges for the application of machine learning for neglected tropical diseases. F1000Research, 12, 287. https://doi.org/10.12688/f1000research.129064.1

[16] Kuang, X., Wang, F., Hernandez, K. M., Zhang, Z., & Grossman, R. L. (2022). Accurate and rapid prediction of tuberculosis drug resistance from genome sequence data using traditional machine learning algorithms and CNN. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-06449-4

[17] Lai, J. (2023). Developing a predictive information system for determining the prognosis of HIV and tuberculosis co-infection in incarcerated individuals. International Journal for Applied Information Management, 3(2), 101–110. https://doi.org/10.47738/ijaim.v3i2.55

[18] Li, Y., Feng, Y., He, Q., Ni, Z., Hu, X., Feng, X., & Ni, M. (2024). The predictive accuracy of machine learning for the risk of death in HIV patients: A systematic review and meta-analysis. BMC Infectious Diseases, 24(1). https://doi.org/10.1186/s12879-024-09368-z

[19] Maharaj, S. S. (2022). HIV and TB co-infection: A double ethical challenge in south african public hospitals. Ethics, Medicine and Public Health, 22, 100760. https://doi.org/10.1016/j.jemep.2022.100760

[20] Mami, D. M., Cuthrell, K. M., & Manteghian, M. (2024). Burden of HIV on society current management options and future prospective. International STD Research &amp; Reviews, 13(1), 43–62. https://doi.org/10.9734/isrr/2024/v13i1170

[21] Matulyte, E., Kancauskiene, Z., Kausas, A., Urboniene, J., Lipnickiene, V., Kopeykiniene, J., Gudaitis, T., Raudonis, S., Danila, E., Costagliola, D., & Matulionyte, R. (2023). Latent tuberculosis infection and associated risk factors among people living with HIV and HIV-uninfected individuals in Lithuania. Pathogens, 12(8), 990. https://doi.org/10.3390/pathogens12080990

[22] McKinlay, J., & Williamson, V. (2010). Leadership in HR management practices – the people edge competencies. The Art of People Management in Libraries, 169–189. https://doi.org/10.1016/b978-1-84334-423-0.50007-9

[23] Moryani, B., Sood, K., & Chaudhary, K. (2023). A deep learning approach for the classification of tuberculosis and pneumonia using NIH Dataset *. 2023 International Symposium on Networks, Computers and Communications (ISNCC). https://doi.org/10.1109/isncc58260.2023.10323983

[24] Kawatsu, L., Kaneko, N., Imahashi, M., Kamada, K., & Uchimura, K. (2022). Practices and attitudes towards tuberculosis and latent tuberculosis infection screening in people living with HIV/AIDS among HIV physicians in Japan. AIDS Research and Therapy, 19(1). https://doi.org/10.1186/s12981-022-00487-8

[25] Njagi, L. N., Nduba, V., Mureithi, M., & Mecha, J. O. (2022). Prevalence and Predictors of Tuberculosis Infection among People Living with HIV in a High Tuberculosis Burden Context. https://doi.org/10.1101/2022.12.04.22283086

[26] Olivier, C., & Luies, L. (2023). Who goals and beyond: Managing HIV/TB co-infection in South Africa. SN Comprehensive Clinical Medicine, 5(1). https://doi.org/10.1007/s42399-023-01568-z

[27] Oloko-Oba, M., & Viriri, S. (2022). A systematic review of deep learning techniques for tuberculosis detection from chest radiograph. Frontiers in Medicine, 9. https://doi.org/10.3389/fmed.2022.830515

[28] Orwa, J., Oduor, P., Okelloh, D., Gethi, D., Agaya, J., Okumu, A., & Wandiga, S. (2023). Comparison of Logistic Regression with Regularized Machine Learning Methods for the Prediction of Tuberculosis Disease in People Living with HIV: Cross-Sectional Hospital-Based Study in Kisumu County, Kenya. https://doi.org/10.1101/2023.08.17.23294212

[29] Panteleev, A. M. (2021). Tuberculosis and HIV infection. Tuberculosis and HIV Infection, 1–352. https://doi.org/10.33029/9704-6733-6-sim-2022-1-352

[30] Parthasarathi, K. T., Gaikwad, K. B., Rajesh, S., Rana, S., Pandey, A., Singh, H., & Sharma, J. (2024). A machine learning-based strategy to elucidate the identification of antibiotic resistance in bacteria. Frontiers in Antibiotics, 3. https://doi.org/10.3389/frabi.2024.1405296

[31] Pokhrel, V., Kuntal, B. K., & Mande, S. S. (2024). Role and significance of virus–bacteria interactions in disease progression. Journal of Applied Microbiology, 135(6). https://doi.org/10.1093/jambio/lxae130

[32] Pooranagangadevi, N., & Padmapriyadarsini, C. (2022). Treatment of tuberculosis and the drug interactions associated with HIV-TB co-infection treatment. Frontiers in Tropical Diseases, 3. https://doi.org/10.3389/fitd.2022.834013

[33] Puttagunta, M. K., & Ravi, S. (2021). Detection of tuberculosis based on deep learning-based methods. Journal of Physics: Conference Series, 1767(1), 012004. https://doi.org/10.1088/1742-6596/1767/1/012004

[34] Rajpurkar, P., O’Connell, C., Schechter, A., Asnani, N., Li, J., Kiani, A., Ball, R. L., Mendelson, M., Maartens, G., van Hoving, D. J., Griesel, R., Ng, A. Y., Boyles,

T. H., & Lungren, M. P. (2020). Chexaid: Deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. Npj Digital Medicine, 3(1). https://doi.org/10.1038/s41746-020-00322-2

[35] Romero-Gómez, A. F., Orjuela-Cañón, A. D., Jutinico, A. L., Awad, C. E., Vergara, E., & Palencia, M. A. (2024). Data Fusion Using Medical Records and Clinical Data to Support TB Diagnosis. https://doi.org/10.20944/preprints202406.1316.v1

[36] Rewari, B. B., Kumar, A., Mandal, P. P., & Puri, A. K. (2021). HIV TB coinfection - perspectives from India. Expert Review of Respiratory Medicine, 15(7), 911–930. https://doi.org/10.1080/17476348.2021.1921577

[37] Sampson E, A., Theresa, A., & Joseph R, O. (2020). Evaluation of the burden and intervention strategies of TB-HIV co-infection in West Africa. Journal of Infectious Diseases and Epidemiology, 6(4). https://doi.org/10.23937/2474-3658/1510143

[38] Sow, P. (2021). P432 the prevalence of co-infection with HIV, hepatitis C and TB among patients infected by HIV/AIDS in Senegal. Poster Presentations. https://doi.org/10.1136/sextrans-2021-sti.452

[39] Singh, M., Pujar, G. V., Kumar, S. A., Bhagyalalitha, M., Akshatha, H. S., Abuhaija, B., Alsoud, A. R., Abualigah, L., Beeraka, N. M., & Gandomi, A. H. (2022). Evolution of machine learning in tuberculosis diagnosis: A review of Deep Learning-based medical applications. Electronics, 11(17), 2634. https://doi.org/10.3390/electronics11172634

[40] Shah, G. H., Ewetola, R., Etheredge, G., Maluantesa, L., Waterfield, K., Engetele, E., & Kilundu, A. (2021). Risk factors for TB/HIV coinfection and consequences for patient outcomes: Evidence from 241 clinics in the Democratic Republic of Congo. International Journal of Environmental Research and Public Health, 18(10), 5165. https://doi.org/10.3390/ijerph18105165

[41] Tanvi, & Aggarwal, R. (2020). Dynamics of HIV-TB co-infection with detection as Optimal Intervention Strategy. International Journal of Non-Linear Mechanics, 120, 103388. https://doi.org/10.1016/j.ijnonlinmec.2019.103388

[42] Teng, V. Y., Chua, Y. T., Lai, E. E., Mukherjee, S., Michaels, J., Wong, C. S., Leo, Y. S., Young, B., Archuleta, S., & Ong, C. W. M. (2021). Lack of Latent Tuberculosis Screening in HIV Patients and Delay in Anti-Retroviral Therapy

Initiation in HIV-TB Co-Infection: A 11-Year Study in an Intermediate TB-Burden Country. https://doi.org/10.1101/2021.02.15.21251801

[43] Vila Torres, S. G., Fullmer, J., & Berkowitz, L. (2023). A case of multidrug-resistant (MDR) tuberculosis and HIV co-infection. Cureus. https://doi.org/10.7759/cureus.37033

[44] Wang, F., Yuan, Z., Qin, S., Qin, F., Zhang, J., Mo, C., Kang, Y., Huang, S., Qin, F., Jiang, J., Liu, A., Liang, H., & Ye, L. (2024). The effects of meteorological factors and air pollutants on the incidence of tuberculosis in people living with HIV/AIDS in subtropical Guangxi, China. BMC Public Health, 24(1). https://doi.org/10.1186/s12889-024-18475-0

[45] Wang, L., Lv, H., Zhang, X., Zhang, X., Bai, J., You, S., Li, X., Wang, Y., Du, J., Su, Y., Huang, W., Dai, Y., Zhang, W., & Xu, Y. (2024). Global prevalence, burden and trend in HIV and drug-susceptible tuberculosis co-infection from 1990 to 2019 and prediction to 2040. Heliyon, 10(1). https://doi.org/10.1016/j.heliyon.2023.e23479

[46] Wang, Y., Jing, W., Liu, J., & Liu, M. (2022). Global trends, regional differences and age distribution for the incidence of HIV and tuberculosis co-infection from 1990 to 2019: Results from the global burden of disease study 2019. Infectious Diseases, 54(11), 773–783. https://doi.org/10.1080/23744235.2022.2092647

[47] Xu, S., & Yuan, H. (2022). A three-methylation-driven gene–based deep learning model for tuberculosis diagnosis in patients with and without human immunodeficiency virus co-infection. Microbiology and Immunology, 66(6), 317–323. https://doi.org/10.1111/1348-0421.12983

[48] Zaharie, A.-M., & Ţigău, M. (2021). Particularities of people who inject drugs in a HIV-TB cohort. Tuberculosis and Non-Tuberculous Mycobacterial Diseases. https://doi.org/10.1183/13993003.congress-2021.pa2284

[49] Zeyu, D., Yaakob, R., & Azman, A. (2022). A review of deep learning-based detection methods for tuberculosis. 2022 IEEE International Conference on Computing (ICOCO). https://doi.org/10.1109/icoco56118.2022.10031813

[50] Zuhri, F. M. (2024). Visualization of the macrophage's dynamic in TB-HIV co-infection using the Molecular Imaging Techniques: A narrative review. Health Dynamics, 1(2), 53–62. https://doi.org/10.33846/hd10205