Machine Learning Techniques for Malaria Incidence and Tuberculosis Prediction

A DISSERTATION

Submitted to the Department of Computer Science
African University of Science and Technology (AUST), Abuja, Nigeria
In Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy in the Department of Computer Science.

Odu Nkiruka Bridget
70152

Advisors:
Dr. Rajesh Prasad
Dr. Clement Onime

July, 2021

CERTIFICATION

This is to certify that the thesis titled “A Machine Learning Techniques for Malaria Incidence and Tuberculosis Prediction” submitted to the school of postgraduate studies, African university of Science and Technology (AUST), Abuja, Nigeria, for the award of the Doctor of Philosophy is a record of original research carried out by Odu Nkiruka Bridget in the Department of Computer Science.

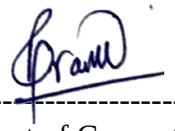
MACHINE LEARNING TECHNIQUES FOR MALARIA INCIDENCE AND TUBERCULOSIS PREDICTION

By
Odu Nkiruka Bridget
A Thesis Approved by the Computer Science Department

Recommended:

Dr. Rajesh Prasad

Dr. Clement Onime



Head, Department of Computer Science

APPROVED:

Chief Academic Officer

Date

ABSTRACT

This research proposes machine learning techniques to develop models that would facilitate decision-making in health informatics. It focuses on using efficient machine learning techniques to solve the pressing need in the two main disease burdens of Africa, which are Malaria and Tuberculosis. In 2019, there were an estimated 229 million malaria cases and 409,000 deaths worldwide, with Africa having the 94% of these cases and deaths. In 2019, Nigeria, Niger Republic, DRC, Burkina Faso, Tanzania, and Mozambique accounted for approximately 50% of the malaria deaths worldwide. Climate variability is one of the leading factors that influence malaria prevalence and transmission, especially in Africa. However, the effect of climate variability on malaria varies across geographical locations. Implementation of a surveillance system that could predict possible malaria outbreak is one of the efforts to eradicate malaria. A surveillance system is domain-specific since what works for one location may not work for another. This research employed an eXtreme Gradient Boosting (XGBoost) algorithm, a machine learning-based model, to predict the incidence of malaria in the six malaria-endemic countries of sub-Saharan Africa. XGBoost is scalable and efficient in memory usage and drives fast learning through parallel and distributed computing. It is used here to develop a malaria incidence classification system that enables early detection of malaria outbreak or epidemics and typically helps policymakers to take pre-informed decisions on malaria intervention.

Tuberculosis has an estimated 10 million cases and about 1.4 million deaths in 2019 while multidrug-resistant Tuberculosis remains a public health crisis and a threat to health security. Methods of diagnosing Tuberculosis is sometimes invasive, takes much time and demands the presence of an expert. Therefore, this research focuses on applying the Frequent Pattern growth algorithm to discover hidden reoccurring patterns on Drug-Resistant Tuberculosis symptoms and generates relevant association rules used to fit a logistic regression model and classify the patient into two target classes. The system is a knowledge capturing one that assists in the quick diagnosis of Drug-Resistance Tuberculosis and leads to breakthroughs in treatments based on correlations found from the collection and integration of Drug-Resistance Tuberculosis data.

Accuracy of diagnosis is crucial in the medical field because wrong diagnosis or prediction might lead to severe consequences. The performance of the proposed models was evaluated using some performance metrics such as Area under the curve (AUC) of Receiver operating characteristic (ROC), classification accuracy, precision, recall and F1-score. The model was also compared with other models using the same metrics; an Akaike information criterion was used to select the best model. The comparisons showed that the proposed models performed better than other models for the intended applications; this proved the efficiency of the models. This research work presents a unique knowledge-based decision support system that can aid physicians, governments, and other health policy makers in making the informed clinical decisions.

Keywords: Machine Learning, Health Informatics, Malaria, Drug-resistance Tuberculosis, XGBoost, Logistic regression, FP-Growth algorithm.

ACKNOWLEDGEMENT

Firstly, I give all thanks to God almighty for his constant love, favour and wisdom, even toward the success of this program.

It is my genuine pleasure to appreciate my supervisors for the privilege to work with them; they took out time from their busy schedules to always suggests methods for the analysis, verify the results obtained in this research and proofread many drafts. I sincerely thank Dr. Rajesh Prasad for his untiring efforts and advice to ensure that I successfully graduate timely. Your constant check-up on the progress of this research is second to none. I learnt a lot from every bit of your corrections and suggestions. You never cease to encourage me to work harder and smarter; I am indeed grateful to you, sir. To Dr. Clement Onime, for his immeasurable supports throughout this program, his constant encouragement and academic advices has helped me in various ways, coupled with the remarkable opportunity he granted me to visit ICTP, Italy to attend some programs that enhanced my research experience, thank you for such exposure, sir.

I wish to thank the African Development Bank (AfDB) for giving me the award of a full scholarship to start and complete this program.

I appreciate the efforts of the dissertation committee for taking out ample amount of time to read this dissertation carefully and for their suggestions on improving the work. My thanks also go to the external examiner for reading through and suggesting ways of making this work better.

Also, I wish to thank Prof. Charles Chidume, the Acting President, for his guidance and discipline during my stay in AUST.

I am immensely thankful to the following faculty members: Prof. Amos, Dr Ekpe Okoroafor, Prof. Hamada, Prof. Cohen, Prof. Csato, and Prof. Samuel Ajila.

I thank the entire staff and students of AUST who have helped me in different ways, especially Mr. Ben, Ms. Amaka, Mrs. Bolade, Mrs. Paulina, Mr. Saheed, Mrs. Buchi Ekpoloma, Mr. Bidemi and Mr. Emeka.

I also wish to thank some of my colleagues for their help and support and for making my stay in AUST fun and memorable, Dr Kehinde Samuel, Hajara Abdulwahab, David Clement, Latifat Abdulsalam, Bunmi Alabi, Chukwudalu Nwazojie, Ifeyinwa Obianyo, Ayeni Gbenga and Killian Onwudiwe.

I want to specially thank Dr Charlene for her professional advice and guidance on how to mine climatic data, Mr Barnabas Ifebude for helping me come through this limelight, Roseline Uneke for her sisterly love, Mrs. Ifeoma Eze for always praying for me, Chioma Okpala for always inspiring me and Very Rev. Professor Ulu Ogbonnaya for his constant fatherly advice and prayers.

I owe a deep sense of gratitude to my loving parents Mr and Mrs David Odu, for setting this pace of academic excellence for me to ride upon; God bless and keep you both. I sincerely appreciate my God-given siblings Uche, Onyedika, Chinedu and Tochukwu, for their constant supports, encouragements, love and prayers.

Finally, but very important, my heartfelt appreciation goes to my loving husband, Mr. James Eneh, for standing by me throughout those busy times, his constant supports, encouragements and above all, he never stops praying for me every morning. I love you, my sweetness!

DEDICATION

To God almighty for his wisdom, favour and strength to complete this research,
To my parents and siblings for their unfailing love, supports and encouragements,
To my sweet hubby for his immeasurable supports, patience and dedication.

TABLE OF CONTENTS

CERTIFICATION i
ABSTRACT iv
ACKNOWLEDGEMENT v
DEDICATION vi
TABLE OF CONTENTS vii
LIST OF TABLES x
LIST OF FIGURES xi
LIST OF ABBREVIATIONS AND SYMBOLS xiii
LIST OF PEER-REVIEWED PUBLICATIONS xvii
CHAPTER 1 1
Introduction 1
1.1 Background of the Study: 1
1.2 Research Question and Motivation 11
1.3 Research Contribution 12
1.4 Aim and Objectives 13
1.5 Scope and Organization of the Research 14
CHAPTER 2 16
Literature Review 16
2.1 Introduction 16
2.2 Overview of the Study Sites and High Incidences of Malaria 17
2.3 The Need for a Malaria Surveillance System 21
2.3.1 Mathematical Models for Malaria Surveillance System 24
2.3.2 Statistical Model for Predicting Malaria Incidences 31
2.3.3 Machine Learning Models for Predicting Malaria Incidence Outbreak 35
2.4 Application of Machine Learning Techniques in TB Diagnosis 39
2.4.1 Association Rule Mining (ARM) 40
2.4.2 Applications of ARM in the Health Care System 41
2.4.3 Classification and Prediction Techniques for TB Diagnosis 43
2.5 State of the Art of Machine Learning Techniques in Healthcare 45
2.6 Summary of the Related Works and our Major Contributions 46
CHAPTER 3 48
Prediction of Malaria Incidence using Climate Variability and Machine Learning 48

3.1 Introduction 48
3.2 Study Site 48
3.2.1 Clinical Data 49
3.2.2 Climate Data 50
3.3 Experimental Dataset 51
3.3.1 Data Pre-processing 52
3.3.2 Feature Engineering using Statistical Significance 54
3.4 Data Visualization 55
3.5 Implementation Frameworks 62
3.5.1 K-means Clustering 62
3.5.2 Extreme Gradient Boosting 63
3.5.3 Performance Metrics and Model Selection 65
3.6 Malaria Incidence Classification Model 68
3.7 Results and Discussion 72
3.7.1 Statistical Significance of Predictor Variables 72
3.7.2 Result of MIC Model 74
3.7.3 Results Comparisons 78
3.8 Discussion 80
3.9 Summary and Concluding Remarks 81
CHAPTER 4 83
Drug-Resistant Tuberculosis Classification using Logistic Regression and FP-Growth Algorithm 83
4.1 Introduction 83
4.2 Experimental Dataset 83
4.2.1 Data Pre-processing and Feature Selection 84
4.3. FP-grwoth Algorithm for Mining Frequent Patterns and Association Rules 86
4.3.1 FP-tree Construction 86
4.3.2 Mining Frequent Pattern on FP-tree 89
4.3.3 Tree Prunning 90
4.4. TB Pattern Discovery and Association Rule Model 90
4.5 Logistic Regression-based Classification Model 91
4.6 Implementing Drug-resistant TB Classification Model 92
4.6.1 Choice of the Programming Tool Kit 93
4.6.2 Evaluation Metrics: 93
4.7 Results 96
4.7.1 Statistical Significance of Climate Variables 96

4.7.2 Frequent Patterns and Association Rules 98
4.7.3 Classification Report 104
4.7.4 Comparing Other Classifiers 106
4.8. Discussion 108
4.9 Summary and Concluding Remark 109
CHAPTER 5 110
Conclusion and Suggestion for Future Work 110
5.1 Conclusion 110
5.2 Suggestion for Future Work 113
REFERENCES 115

LIST OF TABLES

Table 3.1: Sample of the dataset before preprocessing

Table 3.2: Sample of the dataset after preprocessing

Table 3.3: Significance table

Table 3.4: Input variables (predictors) for malaria incidence classification.

Table 3.5: Accuracy value of original dataset without feature engineering.

Table 3.6: Accuracy values for dataset modelled with feature engineered dataset + K-means clustering.

Table 3.7: Akaike weight for the four fitted models.

Table 4.1: Sample of phase1 dataset

Table 4.2: Sample of Phase 2 dataset

Table 4.3: Transaction DB

Table 4.4: Pattern-mining through the creation of conditional-pattern bases

Table 4.5: Relationship between predictors and class variable for Phase1 classification

Table 4.6: Relationship between predictors and class variable for Phase2 classification

Table 4.7: Frequent pattern for pulmonary TB

Table 4.8. Frequent patterns for DR-TB datasets

Table 4.9. The seven most relevant association rules for first-case TB

Table 4.10 The first ten most relevant association rules for DR-TB

Table 4.11 AIC score and Akaike weight

Table 4.12 Comparing other classifiers

LIST OF FIGURES

Figure 1.1: Global Malaria death rate (Data source: WHO global malaria report 2019).

Figure 2.1 Process model for outbreak detection

Figure 3.1. Geographical map of the six selected regions and their endemicity to malaria

Figure 3.2. Annual malaria incidence case per 1000 population for the six selected regions

Figure 3.3: Trend in Annual climatic variability and malaria incidence in Nigeria

Figure 3.4: Trend in Annual climatic variability and malaria incidence in Mali

Figure 3.5: Trend in Annual climate variability and malaria incidence in DRC

Figure 3.6: Trend in Annual climatic variability and malaria incidence in Burkina Faso

Figure 3.7: Trend in Annual climatic variability and malaria incidence in Cameroon

Figure 3.8. Trends in climatic variability and malaria incidence in Niger Republic

Figure 3.9. Flow diagram of the Malaria incidence classification (MIC) model

Figure 3.10a: ROC and AUC score of MIC model in Burkina Faso

Figure 3.10b: ROC and AUC score of MIC model in Cameroon

Figure 3.10c: ROC and AUC score of MIC model in DRC

Figure 3.10d: ROC and AUC score of MIC model in Mali

Figure 3.10e: ROC and AUC score of MIC model in Niger Republic

Figure 3.10f ROC and AUC score of MIC model in Nigeria

Figure 4.1a: Summary of the proposed dataset after preprocessing

Figure 4.1b: Summary of the DR-TB Dataset after preprocessing

Table 4.2: Transaction DB

Figure 4.3: DR-TB pattern and association rule discovery model.

Figure 4.4. LRDR-TB classification system framework

Figure 4.5. Network graph for the Association rule for pulmonary TB.

Figure 4.6. Network graph for the Association Rule for DR-TB

Figure 4.7. AUC score and ROC plot for Phase 1 classification

Figure 4.8. AUC score and ROC plot for Phase 2 classification

Figure 4.9a. Classification report of Phase1 classification

Figure 4.9b. Classification report of Phase2 classification

LIST OF ABBREVIATIONS AND SYMBOLS

ACT	Artemisinin combination therapy
AIC	Akaike Information Criteria
ANFIS	Adaptive Neuro-fuzzy Inference System
ANN	Artificial Neural Network
AR	Auto-regressive
ARIMA	Auto-regressive Integrated Moving Average
ARM	Association Rule Mining
AUC	Area Under Curve
CA	Classification Accuracy
CART	Classification and Regression Tree
CDC	Centre for disease control
CO	Carbon(ii) Oxide
CSCP	Clinical State Correlation Prediction
CV	Cross-Validation
DB	Databases
DDT	dichlorodiphenyltrichloroethane
DNA	Deoxyribonucleic Acid
DRC	Democratic Republic of Congo
DR-TB	Drug Resistance Tuberculosis
DSS	Decision Support System
DST	Drug susceptibility test
DS-TB	Drug Sensitivity Tuberculosis

ECLAT	Equivalence Class Transformation
FiS	Frequent Item-Set
FP-Growth	Frequent Pattern Mining
FPR	False Positive Rate
FP-Tree	Frequent Pattern Tree
HI	Health Informatics
HIV	Human Immune Virus
IPT_p	Intermittent Preventive Treatment for Pregnant Women
ITN	Insecticide-Treated Bed Nets
LR	Logistic Regression
LRDR	Logistic Regression Drug-resistance
LSTM	Long Short-Term Memory
MDG	Millennium Development Goals
MDR	Multi-drug Resistant
MEP	Malaria Epidemic Prediction
MEWS	Malaria Early Warning System
MIC	Malaria Incidence Classification
minConf	Minimum Confidence
minSup	Minimum Support
ML	Machine Learning
MLP	Multi-layer Perceptron
MLP	Multi-layer Perceptron
MSS	Malaria Surveillance System

NB	Naïve Bayes
NCAR	national Center for Atmospheric Research
NDVI	Normalized difference vegetation index
ODE	Ordinary differential equation
OLTP	Online Transaction Processing
PACF	Partial autocorrelation function
PDL	Polynomial distributed lag models
PM10	Particulate matter
RDT	Rapid diagnostic test
ROC	Receiver Operating Characteristics
SARIMA	Seasonal Auto-regressive Integrated Moving Average
SEI	Susceptible-Exposed-Infectious
SEIRS	Susceptible-Exposed-Infectious-Recovered-Susceptible
SL	Supervised Learning
SNS	Social Network Systems
SVM	Support Vector Machine
TB	Tuberculosis
TPR	True Positive Rate
UL	Unsupervised Learning
UNDP	United Nations Development Programme
UNICEF	United Nations Children's Fund
VDCP	Vector-borne Disease Control Programme
VIF	Variance Inflation Factor

VSEIRS	Vectors-susceptible-exposed-infected-recovered-susceptible
WEKA	Waikato Environment for Knowledge Analysis
WHO	World Health Organization
XGBoost	Extreme Gradient Descent Boosting

LIST OF PEER-REVIEWED PUBLICATIONS

1. Bridget, O. N., Prasad, R., Onime, C., & Ali, A. A. (2021). Drug resistant tuberculosis classification using logistic regression. *International Journal of Information Technology*, Vol. 13, 741–749, Springer(Scopus indxed).

2. Nkiruka, O., Prasad, R., & Clement, O. (2021). Prediction of malaria incidence using climate variability and machine learning. *Informatics in Medicine Unlocked*, *22*, 100508, Sciencedirect(Scopus indexed).

3. Nkiruka, O., Prasad, R., & Clement, O. (2021), Drug-Resistance Tuberculosis Pattern Discovery System: A Decision Support System. Submitted to the journal of Data Science and Engineering, communicated

CHAPTER 1

Introduction

1.1 Background of the Study:

Machine learning is a term coined in the late 1950s by Arthur Samuel with the ultimate objective to develop algorithms that are capable of learning, improving over time and useful for predictions (Samuel, 1959). It is a field of computer science that gives computer systems the ability to learn progressively with data, without being explicitly programmed. Healthcare Informatics (HI) is a modern and rapidly developing arena that deals with medical and health data by integrating computer science and information technology. Machine Learning advancements help health care move a step ahead of challenges (Baranauskas & Macedo, 2012). Two of the most common machine learning applications in healthcare are decision support systems (DSS) and knowledge discovery (Wojtusiak, 2014). A decision support system (DSS) can be constructed and maintained using machine learning models, and it aids decision-makers in various situations. Also, knowledge discovery, which is primarily derived from medical datasets, is applied to study healthcare delivery systems, management, and billing patterns (Wojtusiak, 2014). ML is divided into supervised, unsupervised, and reinforcement learning (Kulkarni, 2017)

i. Supervised learning (SL): is a type of Machine Learning that learns from labelled training data to predict unforeseen data outcomes. SL solves both classification and regression tasks by finding specific relationships or structures in the input data that allow effective output data. Some of the SL algorithms include Linear regression, Logistic regression, Support Vector Machine (SVM), Artificial Neural Network, Naïve Bayes.

ii. Unsupervised learning (UL): Uses unlabeled data to find patterns in data. This algorithm is very efficient for detecting latent characteristics that may not be immediately obvious and for exploratory analysis because it can automatically identify structures in data. Clustering and association pattern mining is the major form of UL. Clustering involves dividing a given number of data points into similar groups referred to as clusters. Examples of clustering algorithm includes, k-means algorithm, k-medians algorithm, k-medoids algorithm, hierarchical clustering algorithm, and graph-based algorithm. Association pattern mining is a technique used mostly in market-basket analysis to find the associations between group of items or attributes. The following are the most frequently used algorithm for Association pattern mining, Apriori Algorithm, Equivalence Class Transformation (ECLAT) algorithm, and Frequent-pattern growth algorithm.

iii. Reinforcement Learning (RL): RL is a type of Machine Learning technique that teaches a computer to learn in a collaborative environment via trial and error using feedback from its actions and experiences. It is mostly applied in domains where simulated data is readily available, like gameplay and robotics.

Healthcare service providers generate significant volumes of structured and unstructured data daily that are difficult to analyze and process using traditional methods (Chowriappa et al., 2014). The four major healthcare applications that can benefit from Machine Learning techniques are prognosis, diagnosis, treatment, and clinical workflow (Ahmad et al., 2018). Prognosis involves predicting the expected development of a disease, the likelihood of survival, identifying symptoms and signs related to a specific disease, and finding out if they will worsen, improve, or remain stable over time (Maity & Das, 2017). Diagnosis is a systematic way of identifying a disease by its symptoms and signs (Sutabri et al., 2019). Machine Learning is used

in medical treatment to detect the effects of drugs on diseases, creating room for further diagnosis (Manjiri et al., 2019). Then clinical workflow is defined as directed series of steps comprising a clinical process that is: (1) performed by people or equipment/computers and (2) consumes, transforms, and produces information. Some already claim that machine learning and Artificial Intelligence diagnose disease and treat illness earlier and better (Seneviratne & Shah, 2019). Healthcare delivery concerns are most predominant in Africa, and it is imperative that the system of medical diagnosis in Africa must be automated(Araújo et al., 2016). One of the main objectives of Machine Learning in healthcare is to help automate its operation and predict diseases and gain insight from the hidden patterns, facts, or trends that may have been hidden in the data (Skorburg, 2020).

Malaria disease and Tuberculosis continue to be the major disease burden of the world, especially in Africa (National Academies of Sciences and Medicine, 2017), (Gouda et al., 2019) and (Abebe et al., 2019). The World Health Organization (WHO) estimate in 2019 showed that about 229 million clinical malaria cases occurred, and 409,000 people died (World malaria report, 2019). Malaria has profoundly affected human lives for thousands of years and remains one of the most serious life-threatening diseases in many developing countries (Molineaux, 1988). Sub-Saharan Africa is a hotspot for its transmission, as the region recently has about 90% estimated incidence and death rates in the world. *Plasmodium* parasites of four main species: *Plasmodium malariae, Plasmodium falciparum, Plasmodium ovale, and Plasmodium vivax* cause malaria but among these, *P. falciparum* and *P. vivax* is the most common cause of malaria infection and is mostly transmitted to human through the bite of infected female anopheles mosquito known as malaria vector (White & Ho, 1992). The intensity of malaria transmission is

dependent on the type of parasite, the vector, the human host, and the environment that supports the life of the vector. Malaria diagnosis involves identifying malaria parasites or antigens in the patient blood. The most common methods of diagnosing malaria include clinical diagnosis where the physician diagnoses the patient based on physical signs and symptoms and other physical examination. It is the least expensive and most widely practiced method. However, this method poses a lot of challenge as it lacks specificity in signs and symptoms of malaria which may result in mis-diagnosis and mistaken treatment of malaria. Another method of diagnosing malaria is a microscopic diagnosis by staining thin and thick blood smears using Giemsa, Wrights or Field's stains. It is referred to as a gold standard for laboratory diagnosis and has been in use for more than one century(Tangpukdee et al., 2009). However, this method is slow, requires an expert's presence, and cannot function without electricity. The WHO recognizes rapid diagnostic test (RDT) as a simple, cost-effective and accurate diagnostic test for malaria that could overcome the expertise limitation linked with microscopy. It is fast, easy to perform; it is easy to interpret and does not require electricity or specific equipment. Nevertheless, unpredictable sensitivity might occur due to environmental factors. Polymerase Chain Reaction is another method of diagnosing malaria that is reliable and fast, and can even diagnose drug-resistance threads in a patient (Zimmerman & Howes, 2015). However, it is costly, mostly beyond the reach of many developing countries, it also requires highly skilled experts to operate. The World Health Organization (WHO) advises presumptive diagnosis as the basis for first-line treatment of uncomplicated malaria in places where a parasitological test is not possible. Symptoms associated with malaria infection are fever, headache, chills, fatigue, severe anaemia, dizziness, vomiting, anorexia, and pruritus, myalgia, and respiratory distress; although people in malaria endemic areas may develop some immunity that allows the occurrence of asymptomatic malaria.

Anti-malaria drugs like chloroquine and sulfadoxine-pyrimethamine (SP) are administered for treating malaria disease. In African settings, controlling the spread of malaria and mosquitoes (vectors) include use of anti-malaria drugs, insecticides, larvicides, the destruction of breeding grounds, and insecticide-treated nets (Aikins et al., 1993).

The effects of weather and climatic factors on diseases' epidemiology have received increasing attention in recent years; consequently, malaria, although a vector-borne disease, is thought to be particularly influenced by climate change (Lindsay & Birley, 1996). Meteorological variables such as rainfall, relative humidity, and temperature influence malaria transmission through their effects on parasite-vector development and survival rates (Paaijmans et al., 2009). The development of parasites in the mosquitoes, which is part of the malaria transmission cycle, is very sensitive to external temperatures. The rate of larval development (and subsequent increase in the size of mosquito populations) is also dependent upon water, temperatures and the quantity and quality of breeding sites (Githeko & Ndegwa, 2001). While low temperatures can restrict malaria transmission in tropical highlands or at higher latitudes (Parham & Michael, 2010), anomalous rainfall causes an increase in malaria vector breeding and survival rate in semi-arid areas by influencing environmental factors such as vegetation and swamps that provide suitable breeding sites (Abeku, 2007). Africa is considered the most vulnerable continent to the effects of climate change, and African political leaders are beginning to show some positive concern as this was the centre of their discussion at the April 2020 African Union summit held in Addis Ababa. The genetic influence of climatic factors and the complete life cycle of mosquito vectors are key factors considered while developing the malaria early warning system (MEWS). The epidemiology of malaria deals with the reasons for the prevalence of disease and the nature and

causes of its variation. Recently, many studies are beginning to use climatic/environmental, entomological, and morbidity data to build a MEWS that forecasts malaria incidence before its occurrence (Zinszer et al., 2012) in consensus with the "End malaria strategy" by the WHO (S. Kumar et al., 2015) as interannual climate variability is an important determinant of epidemics in parts of Africa. Although international agencies and academic researchers have advocated the development of early warning systems for malaria for many years, practical progress in this area has been relatively slow(Cox & Abeku, 2007). Recent reports show that rates of mortality and morbidity with malaria epidemics is significantly high due to late interventions to prevent its occurrence, and this might be as a result of inadequate epidemic response mechanisms, which reflects on the inability of decision-makers to anticipate or identify epidemic occurrence(Cox & Abeku, 2007).

In 2019, Nigeria (23%), the Democratic Republic of the Congo (11%), the United Republic of Tanzania (5%), Burkina Faso (4%), Mozambique (4%), and Niger (4%) approximately accounted for about half of all malaria deaths worldwide (World malaria report, 2019), as shown in Figure 1.1. Despite the efforts to eradicate malaria disease transmission and infection, the malaria incidence or epidemic rate in these countries remains quite insignificant (Alonso & Noor, 2017). Nevertheless, malaria outbreak is preventable; therefore, implementing an effective climate-based MEWS would positively impact the malaria eradication research and help the decision-makers regulate the climatic factors that support mosquito vectors' breeding sites (Rosenthal et al., 2019).

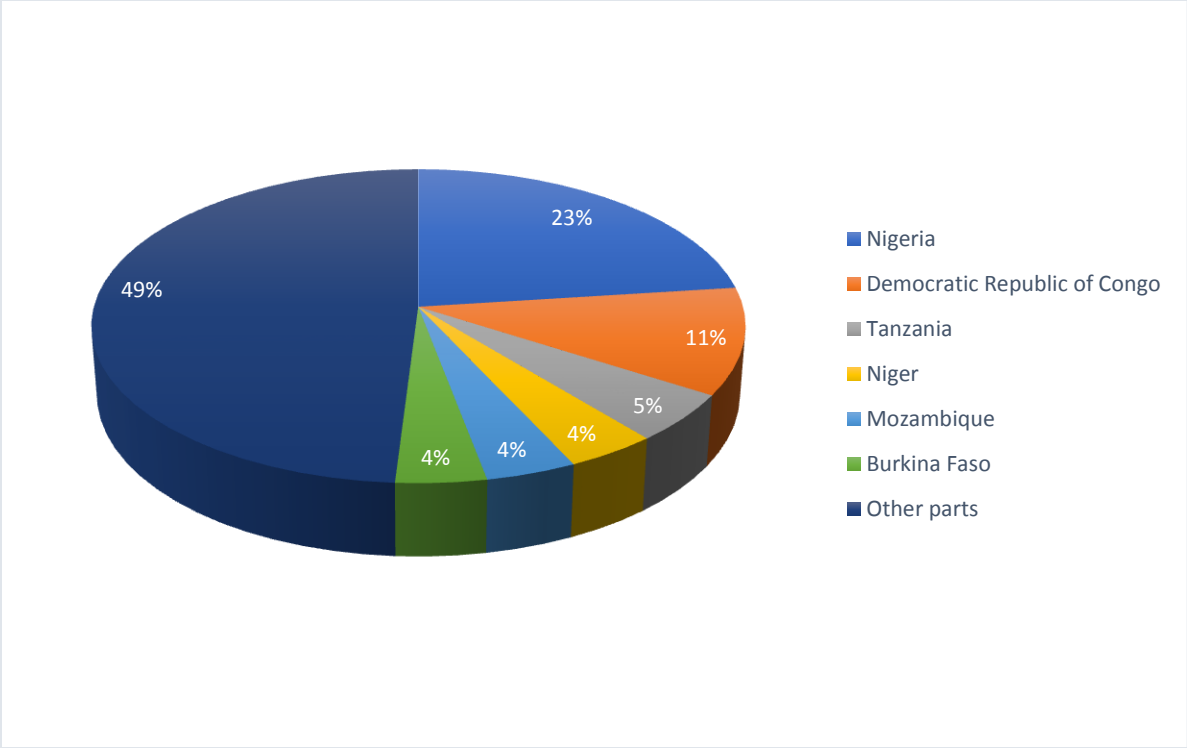


Figure 1.1 Global Malaria death rate (*Data source: WHO global malaria report 2019*).

MEWS is domain-specific due to heterogeneity in the pattern of malaria transmission, hence the need to develop a unique epidemic malaria surveillance mechanism in the regions susceptible to the disease. This research documents an in-depth Machine Learning based study of the six malaria-endemic regions of sub-Saharan Africa; it analyzed the effects of variability in each climate factors such as precipitation, relative humidity, atmospheric temperature, surface radiation, and atmospheric pressure, on malaria incidence in the following countries, including, Nigeria, Burkina Faso, DRC, Niger Republic, Mali, and Cameroon. This research further applied an efficient machine learning technique known as an extreme gradient boosting algorithm to develop a malaria incidence prediction model that predicts annual variations in malaria incidence using climate variability. It is worth noting that malaria instability occurs when the number of

malaria cases drops seasonally below or above the normal average. To the best of our knowledge, no recent studies have considered predicting anomalies in malaria incidence based on climate variability. Therefore, the goals of this research include evaluating the effect of inter-annual climate variability on malaria incidence, develop a malaria incidence classification system using a combination of climatic variables and annual malaria incidence report to help predict and prevent malaria outbreak in the six endemic countries. This research is a data-driven knowledge discovery research that can fill knowledge gaps and assist clinical decision-makers.

Tuberculosis (TB) is an airborne infectious disease whose causative agent was discovered in 1882 by Robert Koch to be organisms of the Mycobacterium tuberculosis complex (Daniel et al., 1994). Infection with *M. tuberculosis* can evolve from containment in the host, in which the bacteria are isolated within granulomas (latent TB infection), to a contiguous state, in which the patient will show symptoms that include cough, fever, night sweats, and weight loss. TB bacillus is classified into drug-sensitive TB (DS-TB) and drug-resistance TB (DR-TB) based on its sensitivity to medication: DS-TB responds to known anti-tuberculosis medicines as the patient recovers from the disease after medication which is not the case with the drug-resistant strain of TB (Pai et al., 2016). Clinical techniques for TB diagnosis include sputum smear microscopy and culture-based methods; they are the current reference standards that requires more advanced laboratory capacity and can also take up to 3 months to provide results. The rapid molecular test is a much faster technique that can simultaneously test for TB and RR-TB and provide quick results. This technique has much better accuracy when compared to the sputum smear microscopy (Guillermo del Rey-Pineda, 2015). However, these conventional clinical methods are costly and require specialized laboratory equipment by a well-trained skilled expert(s)

(Huddart et al., 2016). They are also invasive as they require the collection and proper handling of blood or tissue samples from the patient's internal organs. Collecting these samples might result in harmful effects to patients, such as damage to organs. The predictive diagnosis of TB using non-clinical means is driven by the need to obtain faster results cost-effectively, especially in circumstances where skilled or expert laboratory staff is limited.

TB continues to be a major cause of morbidity and mortality in many low-income and middle-income countries, and drug-resistant TB is often a major concern in many conditions(“Global Tuberculosis Report 2020,” 2020). The incidence rate of TB is reducing, yet not so fast enough to reach the first milestone of the End TB Strategy; that is, a 20% reduction between 2015 and 2020 (Dirlikov et al., 2015). Despite increases in these strategies, about 2.9 million gaps remain between the number of people newly diagnosed and reported and the 10 million people estimated to have developed TB in 2019(“Global Tuberculosis Report 2020,” 2020). Ideally, the WHO in African Region has made good progress, with a 16% reduction. However, the challenges of controlling TB infection and developing more effective strategies to reduce TB and concerted effort is required to achieve the WHO's global TB control strategy. In 2015, WHO released the publication “Digital health” for the End TB Strategy: an agenda for action to end TB (Organization, 2015). The emergence of electronic health records such as diagnostic reports, doctor's prescription, medical images, pharmacy records, and research data from medical journals has made it imperative to digitize the data generated by healthcare industries to improve the quality of treatment, early diagnosis of diseases to avoid risk factors, and better manage information systems in hospitals(Gangopadhyay et al., 2016).

Furthermore, automating some of the components that require human expertise would allow the healthcare industry to build and validate more rapidly. DR-TB is an urgent public health concern in infectious disease as incorrect detection results in the wrong prescription of treatments, leading to increased morbidity and mortality. Early detection of *Mycobacterium tuberculosis* (MTB) is important to decrease mortality. World-wide efforts at minimizing TB include the WHO's "End TB 2016–2020 strategy" to enhance the diagnosis, treatment, and control of TB, especially in developing countries (Cohn et al., 1997). Efficient measures are needed to minimize the prevalence of DR-TB; these include ensuring that TB patients across the world stick to the best available treatment regimen(Laurenzi et al., 2007). Also, developing novel drugs such as Clofazimine (CFZ) for DR-TB is promising, although clinical trials of CFZ show it causes some adverse effects, such as discolouration of the skin and mucous membranes (Widyasrini & Probandari, 2015). The predictive diagnosis of TB using non-clinical approaches is driven by the need to obtain faster results cost-effectively, especially in circumstances where the availability of time and effort of skilled or expert laboratory staff is limited (Dande & Samant, 2018).

This research presents a logistic regression-based system for diagnosing patients into DR-TB or DS-TB by exploring an efficient technique of association rule mining to develop a knowledge-based system that extracts insightful information from the real-world data obtained from the Specialist Hospital, Yola. The system can discover some very interesting disease diagnostics association rules that serve as decision support and provides a good starting point to physicians for TB diagnosis.

1.2 Research Question and Motivation

Advancements in information technology infrastructure in many developing countries have raised hopes that artificial intelligence and its sub fields, including machine learning, might help address challenges unique to global health and quicken achievement of health-related sustainable development goals (He et al., 2019). Consequently, AI-driven health interventions fit into four categories relevant to global health researchers: (1) diagnosis, (2) patient morbidity or mortality risk assessment, (3) disease outbreak prediction and surveillance, and (4) health policy and planning. Most developing countries are beginning to employ AI techniques in solving different health issues with a primary focus on communicable diseases, including tuberculosis and malaria. Different types of machine learning methods are frequently used today for public health surveillance to predict disease outbreak and evaluate disease surveillance data. The motivation for this research is attributed to the continuous accounts of Malaria and Tuberculosis as the most severe disease burden in some African countries such as Nigeria, the Niger Republic, the Democratic Republic of Congo, Burkina Faso, Cameroon, and Mali.

Another reason is the massive growth of health data and the recent emergence of machine learning tools and techniques used to extract important information. For climate-based early warning and forecasting systems, the principal obstacle has been translating promising scientific studies on climate–malaria interactions into robust and reproducible models appropriate for operational use. Due to the severe health impact of malaria epidemics, there is an increasing need for systems that provide forecasting, early warning and timely detection of malaria incidence cases in areas of unstable transmissions, such as the African highlands, so that more effective control measures can be implemented. In addition, a recent publication from the WHO has suggested the adoption of digital technologies for universal health coverage (Organization,

2019). As a matter of supporting the call for action from WHO, this research focuses on applying machine learning to support health care decision making in malaria disease incidence prediction and application of Machine Learning techniques and practices for quick diagnosis of TB even in the absence of a medical expert. This work is evident to explore the practical applications of machine learning techniques in real-life scenarios. In this regard, the following questions are highlighted to guide the objectives and scope of this research:

i. Can we model a system that understands the climatic factors and trends that mostly affect malaria disease transmission?
ii. Can the proposed model detect why and how climatic variables vary across the different countries?
iii. Can the proposed machine learning model correctly predict malaria incidence 2 years ahead?
iv. Can the model detect the most frequent symptoms and relevant association rules for classifying TB in a patient?
v. How much better is the TB diagnosis model when compared to the existing methods?

Each of these questions has been addressed and they formed the basis of efficient implementation of the proposed system.

1.3 Research Contribution

The major contribution of this research to the body of knowledge and the society is to develop an efficient machine learning-based model that can forecast malaria epidemics using climate variability in the six malaria-endemic countries of sub-Saharan Africa. This is achieved by learning about advances in the science of seasonal climate variability in Nigeria, Cameroon, Niger Republic, Burkina Faso, DRC, and Mali and applying the WHO framework for MEWS for

developing plans of action for epidemic preparedness and response for the forthcoming year. The results suggests that a typical environmental condition with drought followed by heavy rainfall and flooding in arid areas in sub-Saharan Africa can lead to explosive epidemics of malaria, which can be prevented through early prediction of the high incidence of malaria and timely vector-control interventions. Malaria transmission is typically seasonal, with significant annual variability, and each country is exposed to a particular and diverse environmental condition. Hence, developing country-specific Early warning and detection systems are needed in these areas to reduce or avert the negative public health and economic impacts of epidemics. Practically, accurate warning signals could help health services to take targeted and specific preventive measures before the onset of epidemics. The second contribution of this research is to develop a machine learning-based TB diagnostic system that can enable physicians to make a quick diagnosis of TB.

1.4 Aim and Objectives

This research aims to apply efficient machine learning techniques for predicting malaria incidence by combining the WHO annual malaria incidence report and climatic variables of the six malaria-endemic countries in sub-Saharan Africa. Secondly, the research aimed to develop a knowledge-based system that can assist physicians in the faster diagnosis of DR-TB. Historical climate variables such as atmospheric temperature, Surface radiation, relative humidity, pressure, and precipitation; obtained from National Centre for Atmospheric Research (NCAR) and the WHO annual malaria incidence report; obtained from the WHO data repository between 1990 to 2017 were used for modeling the system. The climate variables are the predictors, while malaria incidence data are the class variable. Datasets were preprocessed, normalized, and visualized to observe the annual variations and trends in the dataset. We tested the statistical significance of

the climatic factors over malaria incidence at $\alpha = 0.05$. Both datasets were used to train the Extreme Gradient Boosting (XGBoost) model to predict the nature of malaria incidence for each given country using WHO standard for obtaining threshold to determine high and low incidence cases. The second part involves TB data collection from the Specialist hospital Yola, data cleaning, preprocessing to achieve high-quality data, data visualization to detect hidden trends in the dataset, application of the FP-Growth algorithm to generate frequent patterns and association rules that enhanced a fast and reliable classification model based on logistic regression. The results of the system can assist physicians during DR-TB diagnosis. These proposed models can be incorporated into a new or existing Decision Support System (DSS), which can serve as a reliable rapid TB diagnostic tool. While these works did not directly involve working with human subjects, we declare that this research work complies with all sections of the Nigerian National Code for Health Research Ethics of August 2007. Specifically, only anonymous data obtained without incentives were used in this research work.

1.5 Scope and Organization of the Research

This research explored only supervised and unsupervised learning techniques in identifying the factors that influence malaria incidence in the six endemic countries of sub-Saharan Africa and have developed a malaria early warning model capable of predicting the nature of the malaria occurrence in the coming year. FP-Growth and Logistic regression algorithms were used to discover intrinsic patterns from data and generate classification rules for classifying patients into DS-TB or DR-TB classes. This research work is divided into five Chapters, including Chapter 1 that presents the general overview of the research work, research aim and objectives, motivation and research question, and finally, the scope and organization of the research.

Chapter 2 presents the past and current works that have been done in forecasting malaria incidence using climate variability. It further discusses previous existing significant works on malaria incidence and epidemics prediction using machine learning approaches. Also, it discussed different ML techniques that have been implored in the healthcare system to improve TB disease diagnosis.

Chapter 3 provides the system architecture of the proposed malaria incidence prediction model; it presents all the system frameworks, proposed system implementation, experimental datasets, and the result of the proposed model evaluation.

Chapter 4 presents the implementation of the DR-TB classification system, the proposed system evaluation, the evaluation result, and the system implementation summary. It also presents an overview of association rule mining, the FP-Growth algorithm for generating frequent pattern itemset, and association rules for TB diagnosis. Furthermore, it further highlights the implementation tools and the proposed system architecture.

Chapter 5 presents the conclusion and summary of the research; it further provides suggestions and possible ways of extending the proposed works done in this research.

CHAPTER 2

Literature Review

2.1 Introduction

This chapter presents the related works that have been proposed in the literature for improving Malaria disease surveillance and disease outbreak prediction and some practical applications of machine learning on TB diagnosis. Most of the works presented in the malaria disease surveillance and outbreak prediction have proposed implementing a malaria early warning system that can assist policymakers in taking an appropriate informed decision concerning malaria disease outbreak as a means of limiting the spread of the disease. It presents a mathematical model that Sir Ronald Ross proposed on malaria epidemiology, which is the bedrock for other mathematical models on malaria epidemiology. Statistical methods such as Auto-regressive Integrated Moving Average (ARIMA), Seasonal-ARIMA, and ARIMAX have been widely used in predicting the malaria epidemic in different countries. ARIMA is well known for its efficiency in stratifying trends in the time series dataset. Few works have considered using machine learning techniques in predicting malaria incidence. It presented the overview of the study sites, their geographical and environmental details, their current malaria condition, and some traditional methods used in limiting malaria transmission and incidences. We presented the novelty in this work by discussing the state of the art of machine learning in health care and demonstrates how machine learning will transform health care systems through disease prevalence prediction, automation of patient health records, and rapid disease diagnosis. To demonstrate the practical application of machine learning on health systems, the research presented a novel technique in TB diagnosis using association rule mining techniques that is based on a Frequent Pattern (FP) growth algorithm. FP-growth is an efficient and scalable

method for mining the complete set of frequent patterns by pattern fragment growth, using an extended prefix-tree structure for storing compressed and vital information about frequent patterns. Then finally, the Logistic regression-based TB classification model is implemented.

2.2 Overview of the Study Sites and High Incidences of Malaria

This section describes the environmental conditions, the scope, and timing of vector-control interventions in the six malaria-endemic countries.

Nigeria has a tropical climate characterized mainly by the relationship between the dry north-eastern and the moist south-western winds. The main ecological zones are the tropical rainforest along the coast, savannah in the middle belt, and semi-arid zones in the northern fringes (Akinsanola & Ogunjobi, 2014). The total annual rainfall fluctuates from over 3,556mm at the South-east and South-west to 635mm at the Maiduguri North-eastern part of Nigeria with progressive reduction from south to north. In Nigeria, malaria outbreaks has contributed severely to the economic burden of disease in the endemic communities and is also responsible for the annual economic loss ranging up to 132 billion Naira (Onwujekwe et al., 2000). It is estimated that the annual occurrence of 300 000 deaths, 60% of outpatient visits, and 30% of hospitalized patients in Nigeria are all attributed to malaria (Commission, 2008). It has been observed that Nigeria has an annual average temperature of 29^{o}C which supports the complete metamorphosis of mosquito vectors that transmit this disease. Rainfall, relative humidity, maximum and minimum temperatures are the best predictors of malaria incidence in the tropical rain forest and guinea savanna areas of Nigeria. Hence, temperature is the major climatic factor that enhances the prevalence of malaria in Nigeria. Main approaches for limiting malaria disease transmission in Nigeria include efforts such as the distribution of insecticide-treated nets and free medical care

to pregnant women and children below the age of five, yet it is a daunting factor in Nigeria. In Cameroon, the climatic condition is tropical**,** semi-arid in the north, humid, and rainy in the rest of the country. It has two major seasons, dry and rainy season, with an average annual rainfall of about 600mm and an average temperature of 29^{o}C. Malaria remains a public health threat with annual 35% and 67% adult and childhood mortality. In the quest to mitigate the spread of malaria disease, the government has subsidized the price of Artemisinin combination therapy (ACT) and introduced intermittent preventive treatment for pregnant women(IPTp), free distribution of insecticide-treated bed nets (ITNs) as top-down approaches to mitigating malaria disease transmission in Cameroon (Antonio-Nkondjio et al., 2019). Burkina Faso is known for its dry tropical climate, which varies between a short rainy season and a long dry season. Due to its location in the hinterland within the boundary of the Sahara. Its climate is prone to strong seasonal and annual variation with three climatic zones: the Sahelian zone in the north receives less than 600mm average annual precipitation; the North-Sudan zone in the centre has an average annual rainfall between 600 and 900mm; and the South-Sudan zone in the south with an average annual rainfall of 900mm. However, malaria is responsible for 43% of outpatient health issues, 22% of deaths, and 3% of global cases and deaths (Ouedraogo et al., 2018). Rainfall and high temperature are associated with high malaria incidence in Burkina Faso (Rouamba et al., 2019). Since 2008, the country has consistently devoted at least 15% of the annual public budget to healthcare as a way to maintain its commitment to the Abuja Declaration (Malaria, 2000). Similarly, the government has offered various free health services, including malaria treatment and insecticide-treated bed nets for pregnant women and children below five years (Tizifa et al., 2018).

Democratic Republic of Congo (DRC) has a tropical climate known for heavy precipitation, high temperatures, and humidity. It has two main seasons; dry and rainy seasons, although there is always abundant precipitation all year round all over the country. It has an average annual precipitation of more than 2000 mm and an annual average temperature of about 27^{o}C, relatively stable, with little variation between seasons. DRC is the second-leading country globally on malaria cases, accounting for 11% of the 219 million cases and 435,000 deaths from malaria in 2019 (World malaria report, 2019). Despite recent improvements in coverage of malaria interventions, DRC continues to experience challenges in access to preventive and curative malaria interventions and an environment that supports very high transmission rates. Conflict and warfare have also altered the local ecology of many parts of the DRC, leaving agricultural fields uncared-for and susceptible to collecting water in which mosquitoes may breed(Williams et al., 2003). A study also showed that living in a rural location, age, use of non-treated nets, and conflicts increases the chances of malaria transmission in DRC(Messina et al., 2011). Malaria control in the DRC includes screening windows and dirty bins, breeding site management, personal protection with mosquito-treated nets, and pre-treatment with prophylaxis with quinine (Lechthaler et al., 2019). Even with all these precautions and safety measures, DRC remains a malaria-endemic country. Mali has a warm desert climate in the north, the central part in the Sahel has a warm savannah climate, and the southern regions of Mali have a tropical savanna climate. The average annual rainfall in the south of the Sudanese regions is about 762mm, while the northern parts that border the desert only have about 203mm of rainfall annually. The annual number of confirmed malaria cases in Mali was about 1.75million in 2018, and 34.00% of this case occurred in children below the age of five. Several studies in West Africa, particularly in Mali, have highlighted the complex relationship between socioeconomic, hydrological, climatic,

anthropological, and malaria incidence in Mali(Coulibaly et al., 2014). Notwithstanding the concerted efforts of national and international partners to scale up effective malaria control interventions, malaria incidence remains a public health concern in Mali (Druetz, 2018). Niger lies in one of the world's hottest regions, where its annual precipitation hardly reaches 160 mm. The climate is desert in the north, semi-desert in the centre, and semi-arid of the savanna in the south. Niger is one of the poorest countries in the world, a non-coastal sub-Saharan nation (Ministry of Environment and Forests, 2006). Malaria is still a major disease burden in Niger. However, they have adopted some malaria interventions into the public health policy, such as the use of artemether-lumefantrine (AL) combinations as the first-line treatment of uncomplicated malaria, the introduction of free health care for children below the age of five years, the distribution of insecticide-treated nets, the management of malaria cases directly at healthcare points. However, the beneficial effects of these prevention and control measures remain barely perceptible in Niger (Doudou et al., 2012).

Generally, these countries suffer similar problems of high malaria transmission, mortality, and morbidity. Many sub-Saharan African countries have low access to proper treatment, especially on common diseases like malaria (Commission, 2008). This is due to the generally poor state of the health system, high cost of health services, lopsided distribution of health facilities favouring the urban areas, and gross underfunding of the health sector resulting in lack of subsidies and exemptions to the poor. Concerns over the burden of malaria led to the development of several global control strategies and targets such as those under the millennium development goals (MDG) and Roll Back Malaria (RBM), which were set to encourage malaria-endemic communities to control the disease. For instance, Heads of state of African countries made a

commitment in April 2000, at Abuja, Nigeria, with the main agenda to ensure that by 2005, at least 60% of those suffering from malaria have prompt access to affordable and appropriate treatment within 24 hours of the onset of symptoms enabling the home to be the first "hospital" (Malaria, 2000). This is because 60% of epidemics will be detected within two weeks of onset, and 60% of epidemics will be responded to within two weeks of detection. The strategy was termed effective management of malaria nearer the home and was thus adopted as another strategy to combat malaria. Similarly, Another meeting was held in Addis Ababa with the main agenda to enable malaria control services from epidemic-prone countries to gather and review their control programme status and epidemiological trends for the past 3–5 years and identify and map districts they consider to be vulnerable to epidemics (Ouedraogo et al., 2018). The geo-epidemiology of malaria provides the spatial and temporal dynamics of malaria in these countries and hence, helps to deploy, monitor, and evaluate, with a strategic adaptive approach, several sustainable control and elimination interventions.

2.3 The Need for a Malaria Surveillance System

The United States of America's Centre for Disease Control (CDC) published "Guidelines for Assessing Surveillance Systems" in 1988 to encourage the best use of public health resources through the development of efficient and effective public health surveillance systems (Klaucke et al., 1988). Disease surveillance involves disease outbreak detection, epidemic detection, and early warning system and performs the following functions as supporting disease detection and intervention, estimates the impact of a disease, represents the history of a health condition, determines the distribution and spread of illness, generates hypotheses and stimulates research, evaluates prevention and control measures, and facilitating planning (Teutsch & Churchill, 2000). Outbreak detection is defined as a process of detecting an abnormal upsurge in the

frequency of disease above its normal occurrence. Hence, the malaria epidemic is an increase in malaria transmission and incidence beyond the normal experienced. Malaria outbreak is very dynamic and hence a great challenge for the researchers to predict in advance, however, in the absence of the knowledge about the probabilistic attack of high malaria disease, the governments fail to control and provide adequate treatment facility on time. Accurate early prediction of malaria risk or outbreak can successfully reduce the rates of morbidity and mortality resulting from epidemic occurrence.

The first malaria epidemic warning system based on climatic variables was developed as far back as 1911. In their study, Christopher et al. proposed a MEWS based on rainfall for predicting malaria epidemics in Punjab, India (Christophers, 1911). After the inception of such a system, researchers have continued to find the best factors to improve the malaria forecasting system. Recent attempts to predict malaria incidences focus on the role of climate anomalies in epidemic prediction and in response to the effective need to develop MEWS in some malaria-endemic countries. The interaction between climatic factors and their biological influence on mosquitoes and the parasite's life cycle is a key factor to be considered in the development of the malaria early warning system. Another factor to be considered is the location of the study site since the influences of climatic variables on malaria incidence are often inconsistent from one location to another; therefore, building a MEWS that forecasts malaria incidence using domain-specific predictors is appropriate to ensure proper management of resources for prevention activities. Recently, MEWS has been advocated chiefly to maximize the amount of lead time during which decision-makers can plan and implement some interventive measures to reduce its incidence. It detects aberrations in malaria surveillance data and predicts increases in malaria transmission

based on prevailing environmental or climatic conditions. Even WHO, the United Nations Development Programme (UNDP), the United Nations Children's Fund (UNICEF) have unanimously agreed that "roll back malaria strategy" would be achieved with effective implementation of MEWS that facilitate timely responses to prevent and contain the malaria epidemic (Nabarro, 1999).

After the first publication entitled "*Guidelines for evaluating surveillance systems"* and published by Klaucke, Buehler, and Thacker in 1988. A new framework that supplements this previous work has been published by James and Buehler entitled "*Framework for Evaluating Public Health Surveillance Systems for Early Detection of Outbreaks"* by implementing a general framework for an early warning system. This framework is expected to support the evaluation of all public health surveillance systems for the timely detection of outbreaks. It is organized into four categories: system description, outbreak detection, experience, and conclusions and recommendations. This is summarized in Figure 2.1, a conceptual model that facilitates the description of the system. Typically, all MEWS should be built upon the CDC framework. The framework is best applied to systems that have data to validate the system's attributes under consideration, therefore, the description of the surveillance process should address data collection, data processing, statistical analysis for automated screening of data to uncover potential outbreaks. Epidemiologic analysis, interpretation, and investigation involve the rules, procedures, and tools that support decision-making in response to a system signal, including adequate staffing with trained epidemiologists who can promptly review, explore, and interpret the data promptly.

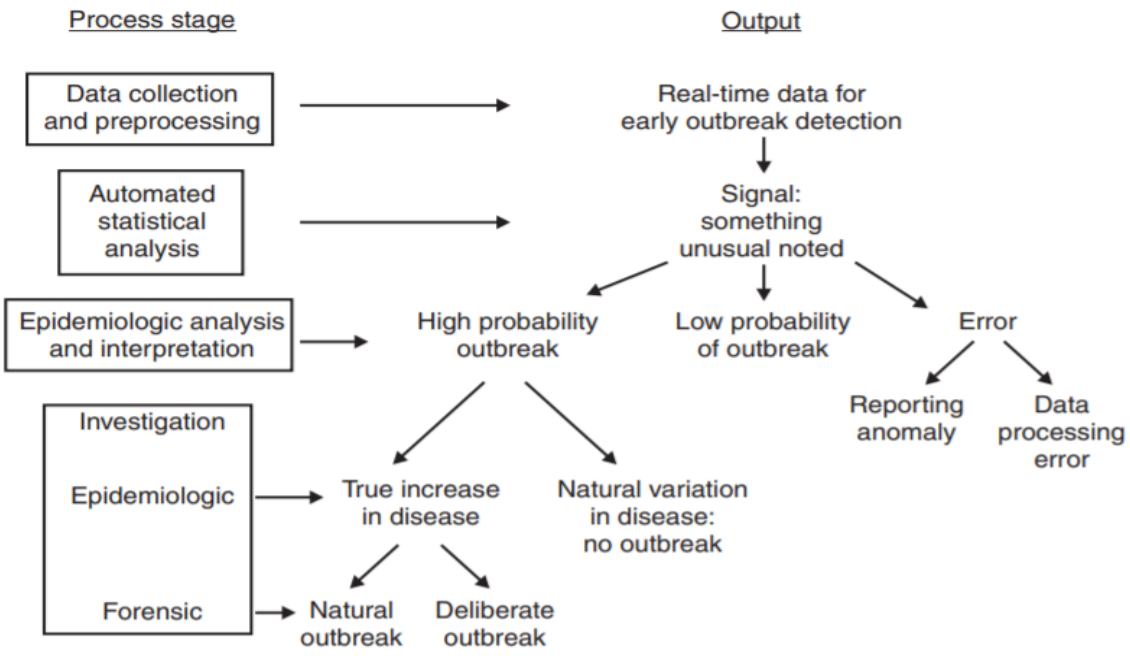


Figure 2.1 Process model for outbreak detection (**Source:** *Framework for Evaluating Public Health Surveillance Systems for Early Detection of Outbreaks* (James W Buehler et al., 2004.))

A surveillance system is useful for early outbreak detection of public health challenges. An assessment of its usefulness goes beyond outbreak detection to understanding the underlying causes of the outbreak and demonstration of intricate patterns in the observational data. Above all, when an outbreak of a disease is confirmed, prompt interventions can be implemented to regulate disease severity and prevent further spread. In conclusion, well-validated systems to predict unusual increases in malaria cases are needed to enable timely action by public health officials to control such epidemics and mitigate their impact on human health.

2.3.1 Mathematical Models for Malaria Surveillance System

Mathematical models have been used to provide an explicit framework for understanding malaria transmission dynamics in the human population for over a century (McKenzie & Samba, 2004). The first mathematical model of malaria transmission and epidemiology was developed by Sir

Ronald Ross while working at the Indian Medical Service in the 1890s. He studied the life-cycle of the malaria parasite in mosquito vectors and used mathematical models, to investigate and provide an explicit framework for a better understanding of malaria transmission dynamics. The model is used to explain the relationship between the number of mosquitoes and malaria incidences in human. The Ross model outlines the basic features of malaria transmission and puts the main burden of transmission on mosquito-specific features, thereby paving the way for mosquito-based malaria control programmes. The model is currently known as the classical "Ross model" (Ross, 1915). Ross model has been significantly extended by Macdonald, who integrated biological information of latency period of infection in the mosquito due to the development of malaria parasite, and also demonstrated that the survival of adult female mosquito is the weakest phase in the malaria development cycle (Macdonald, 1957). The Ross-Macdonald model is defined in Equation (1a) and (1b).

$$\frac{dx}{dt} = \left({}^{abM}/_{N}\right)y(1-x) - rx \qquad (1a)$$

$$\frac{dy}{dt} = ax(1-y) - \mu \qquad (1b)$$

Where:

x = fraction of infectious humans

y = fractions of female mosquitoes that are infectious

a = number of bites on human by one female mosquito per unit time daily

b = probability of transmitting infection from an infected mosquito population

M= size of the total female mosquito population

N= total population size of a human

r = rate of recovery of infectious human

μ = death rate of female mosquito population

The Macdonald model provided a basis for a huge WHO campaign that encouraged the use of insecticide known as dichlorodiphenyltrichloroethane (DDT) that killed adult mosquitoes, and this led to the successful elimination of malaria transmission among 500 million people in Africa Due to the continuous exposure and the ability of a human to develop immunity to malaria, neglecting immunity while developing malaria models leads to unrealistic predictions and will also affect the vaccination. However, one of the salient advancements in malaria epidemiology using the mathematical model was the inclusion of acquired immunity in the Ross-Macdonald's model proposed by Diets et al. In their work, they proposed a two-class model known as the immune and non-immune classes. The immune class is prone to infection but can neither fall ill nor be infectious, whereas the non-immune class may likely fall sick and recovers with some immunity (Dietz et al., 1974). Dietz et al. model is given in Equations (2a) – (2c)

Given Ross model as $r = \gamma$ (2a)

Macdonald = $r = \begin{cases} \gamma - \lambda \, \gamma > \lambda \\ 0 \, \gamma \leq \lambda \end{cases}$ (2b)

Dietz et al. $r = {}^{r}\Big/\left[exp\left({}^{\lambda}\Big/_{\gamma}\right) - 1\right]$ (2c)

Where:

λ= inoculation rate defined in Equation (1a)

γ = reinfection-free rate of recovery

Furthermore, as part of the Garki project in Nigeria, Dietz and Molineaux applied a complex model, in which they have described in their own words, did a "fairly realistic job" through the simulation of malaria epidemiology at Garki, using entomologic data inputs that provide conditional, comparative forecasts for several specific interventions (Molineaux et al., 1980).

Aron and May introduced the concept of the reproductive number R_0 to the Ross-Macdonald as defined in Equation (3)

$$R_0 = \frac{M}{N}\frac{a^2 b}{\mu r} \quad (3)$$

Where:

a = number of contacts with humans that a mosquito has per unit of time

b = probability of transmission from an infectious mosquito to a susceptible human

i/μ = average duration of the infectious period of the vector mosquito

ab/μ = number of humans that one mosquito infects throughout its lifetime

The product of $\left(M/N\right) and \left(a^2 b/r\mu\right)$ gives the reproductive number. Thus, the reproductive number R_0 is the product of the number of contacts owned by one individual at a given time, the probability of transmission per contact, and the infectious period's duration. Aron and May continued their work by adding additional features to the existing model, and these include the incubation period of the mosquito, seasonal fluctuation of the mosquito density, conditions of malaria immunity in human. They included a continuum model for immunity evaluation where the dynamical variables are the population of asexual blood stages of plasmodium in human, the gametocytes and the level of human immunity. Their partial differential equations and the variables are dependent on time and age, and it improved the Ross-Macdonald model by considering varieties of parasites and level of immunity in human (Aron & May, 1982). Aron et al. analyses the compartmental and continuous models for temporary immunity in humans. In compartmental models, an additional recovered class is added. In the usual Susceptible-Infectious-Recovered-Susceptible (SIRS) or Susceptible-Exposed-Infectious-Recovered-Susceptible (SEIRS) model, is a constant parameter that represents the loss of immunity, the

irregular period of immunity is modelled by making ρ a function of inoculation rate(Aron, 1988) as shown in Equation (4).

$$\rho(\lambda) = \frac{\lambda e^{-\lambda\tau}}{1-e^{-\lambda\tau}} \quad (4)$$

Where:

λ = inoculation rate

τ = average duration of the period of immunity without infection

Koella also extended the Ross-Macdonald model by considering the effect of variability of the parameters, infection rate, period of immunity in human. He further studied the effects of vaccines on malaria transmission(Koella, 1991). Over time, Anderson and May proposed another mathematical model by considering the latent periods in mosquitos and human, the anticipated lifespan of an adult mosquito, the rate of recovery of an infected human, and the malaria prevalence data across all age distributions. They also evaluated the effect of attaching age structure to the Ross-Macdonald model. In the end, they evaluated the various control strategies and discussed the effect of vaccine and the reduction of transmission rate on the malaria age-prevalence profile of the human population (Anderson & May, 1992). Yang described a compartmental model where humans follow an SEIRS pattern and mosquitoes follow a Susceptible-Exposed-Infectious (SEI) pattern. He added temperature as a function of incubation time in the mosquito. Yang defined a reproductive number R_0 for this model and demonstrate through linear stability analysis that the equilibrium at disease-free is stable for $R_0 < 1$. He derived an endemic equilibrium expression that is biologically significant if and only if $R_0 > 1$ (H. M. Yang, 2000). Yang and Ferreira applied the model introduced by (H. M. Yang, 2000) to study the effects of global warming on mosquito abundances. They proposed that using a projected temperature ranging from 1^{o}C to 3.5◦C by the year 2100 and making $R_0 > 1$ would

bring a positive change of making a malaria-endemic area a free endemic zone. Although, they concluded that good economic and social conditions and proper access to an efficient healthcare system are more important than regulating the effects of temperature (H. M. Yang & Ferreira, 2000).

Ngwa and Shu proposed a model that is similar to Yang's model. He stated that human follows an SEIRS pattern and mosquitoes follow an SEI pattern, but only with one immune class for humans. Humans' changes from the susceptible (*S h*) class to the exposed *(E h)* class at some odds when they come into contact with an infected mosquito and then to the infectious class (*I h)*, as in common SEIRS models. Though infectious people can recover with, or without, a gain in immunity; and may return to the susceptible class or move to the recovered *(R h)*. While the mosquito population has three classes as follow: susceptible, exposed, and infectious. Mosquito enters the susceptible class through birth. Susceptible mosquito is probably infected when they bite infectious or recovered humans and then proceed to the exposed and infectious classes. These two species follow a logistic model to grow their population, with humans having higher chances of immigration and mortality (Ngwa & Shu, 2000)

In recent time, some studies have directly applied these mathematical models in malaria epidemiology. A study in Ethiopia has applied a mathematical model using polynomial distributed lag models (PDL) to determine the effects of weather factors on malaria in fairly hot and cold environments using meteorological data and weekly confirmed cases of malaria in the ten districts of Ethiopia from the year 1990 to 2000 (Teklehaimanot et al., 2004). Another scholar proposed a mathematical model that understands the pattern of transmission and spread of malaria better. This was achieved through ordinary differential equations (ODEs) using humans and mosquito's interaction and infection cycle. This model is used to determine the main

factors that are most responsible for the spread of malaria. Their model divides the human population into four distinct classes as follows: susceptible, exposed, infectious, and recovered (immune) and then three classes for mosquito as follows: Susceptible *(S m)*, Exposed *(E m)*, and Infectious *(I m)* (Chitnis, 2005). Another recent study in the Sudanese Savanna has adopted Ross-Macdonald's mathematical approach as the vector-susceptible-exposed-infected-recovered-susceptible (VSEIRS) model. The statistical relationship between Normalized Difference Vegetation Index (NDVI) and incidence of *P. falciparum* infection was evaluated using the ARIMA model. Then, malaria transmission was modelled using a deterministic approach; a model built on the MacDonald equations, specifying states for infected non-contagious and contagious, and then forecasts the evolution of malaria epidemiology (Gaudart et al., 2009). Similarly, Laneri et al. also applied a variant of the VSEIRS model on microscopic datasets of India's Kutch and Balmer districts (Laneri et al., 2010).

Notwithstanding the early contributions by Ross, Macdonald and other scholars on using mathematical models to survey the significance of mosquito vector to mosquito-spread and malaria transmission, mathematical epidemiology has encountered many difficulties in gaining general acceptance by epidemiologists and public health workers. Some of the reasons for this low acceptance lies in the increasing complexity of the models, difficulty in understanding the mathematical models by non-mathematicians, difficulty in interpreting results, and some relevant malaria dynamic features that may be omitted due to computational complexity and an increase in the parameters (Koella, 1991). Therefore, it is important to understand the important parameters in the transmission of the disease and develop effective solution strategies for its prevention and control.

2.3.2 Statistical Model for Predicting Malaria Incidences

The statistical models include the linear models, Auto-regressive Integrated Moving Average (ARIMA), and Seasonal-ARIMA (SARIMA). ARIMA models are designed to account for serial autocorrelation in time series data. It is used to train and then forecast future time points(Zhang, 2003). Regression on the lagged values of the variable of interest is shown by the auto-regressive (AR) part. The moving part shows the linear combination of the error terms that have concurrent values. The integrated (I) part specifies the difference between present values and the initial values. Each feature aimed to make the model fit the data accurately. Therefore, a time-series data with X_t where t stands for integer index whereas X_t are real numbers, and ARIMA(p', q) model is given by Equation (5)

$$X_t - \alpha_1 X_{t-1} - \cdots - \alpha_{p'} X_{t-p'} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \qquad (5)$$

Where:

α_i = Autoregressive parameters,

θ_i = Moving average parameters

ε_t= Error terms

Some variants of the ARIMA model have been used for malaria incidence prediction. Abeku et al. applied the ARIMA model on monthly health clinical data in Ethiopia; the dataset is a mixture of microscopic and clinical confirmed cases (Abeku et al., 2002). Teklehaimanot et al. proposed an ARIMA model to detect the degree of relationship between weather and malaria cases in 10 districts of Ethiopia using weekly weather and malaria cases. The average daily malaria cases were computed using Poisson regression, where rainfall, minimum temperature and maximum temperatures are the explanatory variables in a polynomial distributed lag model in 10 districts of Ethiopia(Teklehaimanot et al., 2004). Gomez et al. proposed an ARIMA model

that predicts malaria incidence in an unstable malaria transmission area by finding the association between disease dynamics and environmental variables in Karuzi Burundi. This study was carried out using time series data that contains monthly reports of malaria cases from local health facilities, together with rainfall and temperature observatory data and the normalized difference vegetation index (NDVI). The data constructed in this work attempts to provide a simple tool to obtain a reliable estimate of the expected incidence of malaria one month in the future based on the observed incidence rate and a combination of climatic factors (temperature, rainfall and vegetation index) for the current month(Gomez-Elipe et al., 2007). Haghdoost et al. (2008) proposed a statistical approach to model the temporal variations in malaria episodes in a hyper-endemic Kahnooj District of Iran. An epidemic early warning system in Plasmodium species-specific models was generated using Poisson regression. Seasonality was modelled using a sinusoidal transformation of time by including both $\sin(2\pi i / 12)$ and $\cos(2\pi i / 12)$ in the regression models, where i represent the number of the month. The results of the trend analysis and periodograms show no trend or periodic oscillation, and the proposed model described a substantial percentage of the observed variability in the malaria rate (R2 adj = 82%, F = 165, df = 2, $p < 0.0001$) with a coefficient of 0.80 for the autoregressive term and 0.99. (Haghdoost et al., 2008). Wangidi et al. (2010) used monthly reported malaria cases from the health centres to Vector-borne Disease Control Programme (VDCP) and the meteorological data from the Meteorological Unit, Department of Energy, Ministry of Economic Affairs. They carried out a time series analysis on monthly malaria cases, from 1994 to 2008, in seven malaria-endemic districts. The method of ARIMAX modelling was employed to determine predictors of malaria of the subsequent month. The ARIMA models of time-series analysis were useful in forecasting the number of cases in the endemic areas of Bhutan. There was no consistency in the predictors

of malaria cases when using the ARIMAX model with selected lag times and climatic predictors. Nevertheless, the ARIMA forecasting models could be employed for planning and managing malaria prevention and control programme in Bhutan (Wangdi et al., 2010)

Akinbobola et al. proposed an ARIMA based model for predicting malaria occurrence in the southwest and north-central parts of Nigeria. This study shows that the relationship between climate and malaria varied from place to place, and one model could not fit all locations. They also observed that rainfall, relative humidity, maximum and minimum temperatures are predictors of malaria incidence in the tropical rain forest and guinea savanna areas of Nigeria. (Akinbobola & Omotosho, 2013). Briet et al. applied ARIMA in health facility data for all Sri Lanka. ARIMA models were used to evaluate the relationship between weather factors and monthly malaria incidence. An ARIMA model was fit first to the predictor variable. The model was then applied to the dependent variable before the two series were cross-correlated to determine whether an association exists.

Modelling with ARIMA involves estimating a series of parameters to account for the inherent dynamics in the time series, including the trends and autoregressive and moving average processes (Briët et al., 2008). (V. Kumar et al., 2014) designed a model to forecast malaria cases using climatic factors as predictors in Delhi, India. Expert modeler of SPSS version 21 software was used to fit the best suitable model for the time series data. The stationarity of the data was checked by autocorrelation function (ACF) and partial autocorrelation function (PACF). Finally, ARIMA (0,1,1) (0,1,0) model was used to forecast the monthly malaria cases for the future from January 2014 to December 2015. This model also included the significant predictors for malaria cases, rainfall, and relative humidity, which were lagged at one month in Delhi. Anwar et al. (2016) applied the ARIMA model to malaria incidence data in Afghanistan, and their model was

able to identify that malaria is always at its peak in July to September and always lesser during January in Sub-Saharan Africa; this suggests that humid temperature and high volume of rainfall are strong factors affecting malaria outbreak and transmission (Anwar et al., 2016).

Another study from Anokye et al. assessed malaria incidence in Kumasi Metropolis and forecasted future incidence using time series malaria incidence data in Kumasi, Ghana, between January 2010 to December 2016. They carried out a retrospective comparative study to observe trends in malaria prevalence using the Quadratic model and then used ARIMA (1,1,2) to forecast the incidence of Malaria. Due to the transmissibility and seasonality of malaria, models with an ARIMA structure have more predictive power compared to other methods (Anokye et al., 2018). Another study in Nkomazi, South Africa, analyzed the malaria incidence datasets and environmental reports in Nkomazi, South Africa, using the SARIMA model and precipitation to predict the possible future incidence of malaria in the next three years(Adeola et al., 2019). A SARIMA model proposed to forecast malaria in Andhra Pradesh, India. The study observed that variations in the malaria trend are linked with the variability of rainfall and temperature. Although there is an occurrence of malaria transmission throughout the year, many malaria cases were recorded during the rainy season (Mopuri et al., 2020). A statistical approach based on dynamic regression models by combining negative binomial models with a time-series model was adopted by Makinde *et al.* They created an association between malaria cases and climate variables in Akure. The findings from their research summarized that an increase in monthly minimum temperature significantly increases the likelihood of malaria transmission, leading to an increase in the number of inpatients and outpatient malaria cases in Akure (Makinde et al., 2020).

ARIMA models are, in theory, the best models for forecasting a time series as they naturally represent temporal patterns, such as seasonality and autocorrelation. However, they lack a good interpretation of weather factor covariance, while SARIMA models require a longer time series. Interestingly, Machine Learning tools and techniques can handle many data more efficiently than most statistical tools, and the advancements in Machine Learning technology have led to a rise in better and quicker prediction, diagnosis and classification of diseases. Quick diagnosis of the disease has helped in designing diagnostic kits, reducing of cost of diagnostics and treatment, and detecting patterns of infections in diseases.

2.3.3 Machine Learning Models for Predicting Malaria Incidence Outbreak

In terms of prediction and the emergence of new algorithms, machine learning has a wider range of applications than traditional ARIMA models in real infectious disease surveillance. As such, this section highlights some previous related works that have been done using machine learning techniques. Gonzalez et al. proposed machine learning and deep learning algorithms to predict mosquito species and age structure of the population using mid-infrared spectroscopy. They used a supervised machine-learning approach to map the pre-selected 17 wavenumbers to mosquito species on any of these classes; An. gambiae or An. Arabiensis. The age classes of the mosquito selected are 1, 3, 5, 7, 9, 11, and 15 days. Their research aims to understand the chronological ages and species of mosquito that transmits malaria to ensure an efficient malaria intervention (González Jiménez et al., 2019). Lee et al. (2020) proposed an ensemble of a machine learning model for diagnosing malaria using patient's information such as nationality, disease, gender, symptoms and present location of the malaria patient in Korea (Lee et al., 2020). Bria et al. have demonstrated the predictive power of machine learning by developing a predictive model to aid in the quick identification of significant symptoms and non-symptoms of malaria for faster

decision making and easy malaria disease diagnosis. Patients' medical records of malaria and other febrile diseases were collected from public hospitals, and seven significant symptoms were identified as the most relevant symptoms of malaria that serve as predictors for quick malaria diagnosis. The symptoms include headache, duration of fever, vomiting, heartburn, severe symptoms, dizziness, joint pain, patients' health history of malaria (Bria et al., 2021). Diagnosing malaria with its symptoms seem to be promising. However, to eliminate or mitigate malaria disease transmissions and outbreaks, a good elimination strategy requires determining the malaria mode of transmission and predicting its possible outbreak rather than diagnosing the disease. Therefore, recent malaria research focuses on using machine learning algorithms to forecast malaria outbreak and incidences using climate factors.

Kalipe et al. proposed machine learning approaches for predicting malaria outbreak in Visakhapatnam. Data collected ranges between 2005 to 2011 involving the information about the two most popular *parasites: P. falciparum* and *P.vivax* and climatic variables such as rainfall, minimum and maximum temperature, humidity. They applied eight different machine learning algorithms, including XGBoost, to compare and select the model with the highest accuracy value. XGBoost algorithm performed better than other algorithms. However, this study focused only on the capability of the machine learning algorithms. The significance of the work done is not well demonstrated in the research (Kalipe et al., 2018). Ranovan et al. proposed applying a Naïve Bayes model for mapping tropical diseases based on data from Twitter to predict tropical disease outbreak in Indonesia. Twitter is a type of Social Network Systems (SNS) that provides a robust platform for microblogging. Researchers have widely used it to identify trending news about people and events. The tropical disease considered in their research includes malaria, dengue, and avian. Their study used two sets of data for training the multinomial naïve Bayes

model, they are feature selection and non-feature selection datasets. They classified the tweets datasets based on the following four conditions: disease name, the location of disease outbreak, health condition of disease victims, the number of cases. The model was able to map the tropical diseases in Indonesia based on data from Twitter with high percentage accuracy (Ranovan et al., 2018). Chekol et al. (2018) proposed an ensemble of machine learning techniques to predict malaria epidemics, they presented a framework that employs several machine learning models for Malaria Epidemic Prediction (MEP) in Ethiopia based on the quantity of relative humidity, rainfall, elevation, mean temperature, and malaria cases. Their proposed model consists of an accurate opaque box model through a support vector regression (SVR) and a transparent box model through an Adaptive Neuro-Fuzzy Inference System (ANFIS). ANFIS is a class of adaptive networks that incorporate both neural networks and fuzzy logic principles. Neural networks were used for adjusting the membership functions of the fuzzy sets. The ANFIS is structured as follows, five inputs with three membership function for each input, 243 rules, and a single output. The model was trained, validated and tested using five years (2013–2017) of historical climatic and malaria data from the study site (Chekol & Hagras, 2018). The model was able to predict malaria epidemics in Ethiopia. A conceptual architecture for the Malaria Surveillance System (MSS) to detect and track disease outbreaks and trends and gives health professionals and organizations sufficient time to initiate preventive intervention strategies to distribute the necessary resources to affected areas effectively and timely manner. However, this work is not yet implemented and evaluated. It is mainly based on the analyses of literature review and provides a basis for the next steps of the text mining and data collection process (Boit & Alyami, 2018). Another research work proposed a Malaria Epidemic Prediction Model using Twitter data and volume of Precipitation in Nigeria. Researchers in public health uses the

crawled data and information from Twitter to detect incidences and manage epidemics and disease outbreaks in some parts of the world. Their research streamed live Twitter messages that have mentioned “Malaria”, preprocessed the Tweets, and then classified them into two different classes as Malaria cases unrelated

Tweets and Malaria cases related Tweets using SVM as the base classifier. Finally, the model was able to predict correctly possible outbreak of malaria in Nigeria (Maurice et al., 2019).

Some basic environmental and social contexts of malaria vary geographically. Hence, the influences of climatic variables like temperature and precipitation often vary from one location to another. In this regard, Davis et al. have considered a genetic algorithm model that forecasts malaria incidence by assembling districts into clusters based on their responses to the environmental variables. Daily environmental data from each study zone, such as precipitation, were used as predictors. Then, a genetic algorithm is used to perform some cluster analysis to map each district with the most influencing climatic factor. In the end, six clusters were formed with each district showing their degree of response to the environmental predictors (Davis et al., 2019). An artificial neural network model was used to predict the abundance of malaria prevalence in four districts of India, using a combination of environmental variables such as vegetation index, rainfall, temperature, relative humidity, and daily clinical report of malaria. The study identified that the malaria prevalence rate is always at its peak during heavy rainy seasons. Their study has shown that ML algorithms are successful for predicting the possible outbreak of malaria in some regions of the world(Thakur & Dharavath, 2019). Santosh et al. (2020) proposed a Long Short-Term Memory (LSTM) based system to predict the prevalence of malaria in India using a big data approach. LSTM is a chain-like structure capable of remembering information and long-term training with four network layers. Their model

identifies and removes the non-required data using the sigmoid function and takes the output (vt-1) at time t-1 and current input (Xt) at time t. The model could predict the possible occurrence of malaria in India(Santosh et al., 2020).

The related works presented in these sections have relatively performed well and correctly predicted the malaria epidemic in different parts of the world. However, they are some underlying issues that need to be addressed. Firstly, most of the works are more concerned with predicting just the incidences and the prevalence of malaria without typically stating a threshold that considers malaria incidence or prevalence as "an epidemic". Secondly, few works have been done using machine learning techniques and algorithms to predict malaria incidence. Thirdly, much attention has not been given to applying machine learning techniques to predict malaria outbreak in the six endemic regions of sub-Saharan Africa. Finally, predicting the exact value of malaria prevalence may not be ideal as it will not always predict the same value for each given year. This research has addressed these issues by considering the six most malaria-endemic countries of Africa using machine learning tools and techniques to predict if there would be an increase or decrease in malaria incidence in a given year, considering the climate variability. It has proposed to apply an eXtreme Gradient Boosting algorithm. The XGBoost model has achieved excellent performance in many fields of medical research and guarantees a better performance model that could predict the nature of malaria incidence using climate variability.

2.4 Application of Machine Learning Techniques in TB Diagnosis

This section presents an overview of some machine learning tools and techniques that have been applied in enhancing TB diagnosis in the past and present time. It presents how association rule mining has been applied in healthcare and different approaches to implementing association rule mining. The most common approaches mentioned are Apriori and ECLAT, only about two

researchers have considered using FP-growth algorithm to discover hidden patterns and information from patients' symptoms and generate relevant classification association rules. The further presents different approaches for classifying TB patients. Finally, it highlights the significance of the proposed work in this research by improving the existing methods.

2.4.1 Association Rule Mining (ARM)

ARM is a concept introduced by Agrawal (Agrawal et al., 1993); It is designed in the context of market-basket analysis for mining frequent item-set data. An itemset in the medical context comprises a set of symptoms in the transactional (patient's database) database. Given T; a transactions database (DB), and $I = I\_1, I\_2, \dots, I\_m$ as a set of binary attributes (items), each transaction t, represented as a binary vector, with $t[k] = 1\ if\ t$ is present, and $t[k] = 0\ if$ t is absent (Agrawal et al., 1993). For instance, association rule (AR) is expressed as Inadequate-Treatment (IT) $|True \Rightarrow DRTB|True$, where IT and DRTB are frequent itemset in the transactional DB and IT $\cap$ DRTB $= \emptyset$. The itemset IT and DRTB are said to be the antecedent and the rule's consequent, respectively. This rule implies that if IT is true, then DRTB will likely occur in the same transaction. Support(minSup) and Confidence(minConf) are two major metrics for evaluating association rules' strength. The support for the rule $X \Rightarrow Y$ is the fraction of the transactions that contain both X and Y, as shown in Equation (6).

$$support(X \cup Y) = \frac{\sup\_count(X \cup Y)}{N} \qquad (6)$$

The Confidence of rule X⇒Y is the proportion of the transaction containing X and contains Y expressed in Equation (7).

$$\text{confidence } (X \Rightarrow Y) = \frac{Sup(X \cup Y)}{sup(X)} \qquad (7)$$

ARM is an effective Machine Learning technique that finds interesting patterns in large databases (Altaf et al., 2017) and is extensively applied in healthcare systems. It has the

capability of conducting intelligent diagnoses and extracts insightful information and knowledge rapidly. The most frequently used methods for generating association rules are Equivalence Class Transformation (ECLAT), Apriori, and FP-Growth algorithms. FG-Growth is the most efficient because it uses a special prefix tree to organize data and its rapidity in finding frequent itemset with less access to the transaction DB (Altaf et al., 2017). In recent times, ARM has gained popularity in the health care domain for discovering hidden patterns in diseases and recurring relationships amongst symptoms of the disease. It has been scarcely applied in TB research, and most of the studies identified the TB co-occurrence and generated ARs on early-stage TB symptoms using the Apriori algorithm but have not considered DR-TB. The Apriori method may also achieve good performance gain; however, it takes a longer time and is too costly to handle due to many candidate sets and frequent scanning of the database(Han Jiawei, 2000).

2.4.2 Applications of ARM in the Health Care System

ARM is widely used in disease investigation. The most relevant existing works focused on association rules discovery: Asha et al. (2011) proposed an associative classification system for classifying TB into Pulmonary TB and Retroviral TB using the Apriori algorithm. The system generated some classification rules that assisted in predicting TB disease (Asha et al., 2011). ARM was applied to the salmonellosis disease dataset of Florida, California, and New York to discover the effects of urbanization on salmonellosis (Raheja, 2012). This dataset comprised numbers of salmonellosis-affected people. The dataset was preprocessed, followed by ARM application using the Apriori algorithm at minimum support and confidence values of 1% and 40%. Researchers found that urbanization affects the high rate of Salmonellosis (Raheja, 2012). Nahar et al. applied the Apriori algorithm to detect factors that influence heart disease rates in

males and females, using a dataset obtained from the UCI Machine Learning repository. The analysis showed that females are less likely to develop coronary heart disease than the male counterpart (Nahar et al., 2013). Ilayaraja et al. applied the Apriori algorithm to detect frequent diseases in a specific location for a given period. About 1246 patients' medical health records containing 29 attributes were used on WEKA ML software to generate the frequently occurring disease and transmission (Ilayaraja & Meyyappan, 2013).

In a study to extract factors influencing the spread of respiratory disease in Kuala Lumpur, Malaysia, the Apriori algorithm was applied to a dataset of 1000 records with seven feature variables to generate sets of rules associated with respiratory illness. This process involves five-step approaches: data selection, data preprocessing, application of Apriori algorithm, result evaluation, and knowledge extraction. The results show that temperature, carbon monoxide (CO), and particulate matter (PM10) strongly affects the outbreaks of respiratory illness in Kuala (Payus et al., 2013). A system that analyzes clinical observations using a simulated dataset. Their design framework is based on online transaction processing (OLTP) and clinical state correlation prediction (CSCP). The CSCP fetches data from OLTP, then processes and analyzes comorbidity using the Apriori algorithm (Rashid et al., n.d.). The practical application of this work is uncertain as it did not involve real-world data. An FP-growth algorithm that generates positive and negative frequent patterns and ARs using a heart disease dataset is proposed by Wang et al. (2017). It generates only the relevant ARs by removing pseudo-patterns and comparing the negative and positive data items, then sets up a threshold for the number of incidences (Wang et al., 2017). ARM is also applied in medicine to detect the association between the disease and the physician's prescription in treating the patient's illness; this is

achieved through grouping disease into different classes using k-means clustering algorithm (C. Yang et al., 2018).

A. et al. (2021) applied ARM using the Apriori algorithm to understand TB attributes and to identify their correlation. This algorithm was applied on a thousand records of datasets with 26 attributes to generate the association rules (AR) for the early-stage diagnosis of TB and the Decision tree for further classification (A. et al., 2021). Most of these works have considered the Apriori algorithm; however, the Apriori algorithm is inefficient in handling many candidates set and involves constant scanning of the DB, especially when mining longer patterns. The FP-Growth algorithm is most efficient considering space and time complexity. Few works have been applied in generating association rules in pulmonary TB and DR-TB. Therefore, we propose an effective FP-Growth algorithm that generates efficient association rules for discovering the hidden patterns in symptoms of first-degree TB and DR-TB to enhance TB diagnosis. The proposed algorithm aims to discover descriptive rules that find frequent symptoms associated with TB and DR-TB. Consequently, these rules will also help determine if the presence of a symptom may result in another.

2.4.3 Classification and Prediction Techniques for TB Diagnosis

Machine learning methods have been widely applied for the early prediction/diagnosis of DR-TB. A Classification and Regression Tree (CART) model for TB diagnosis using smear-negative pulmonary TB for patients in remote areas. This model considered the physical signs and chest X- rays. This predictive model's results had a sensitivity range of 64% to 71% and a specificity range of 58% to 76% (F.S. et al., 2012). However, these studies did not consider patients with DR-TB. Évora *et al*. proposed an ANN-based model consisting of 50 hidden nodes and three-layer fully connected feed-forward Multi-layer Perceptron (MLP) used as a diagnostic Support

tool that detects DR-TB and MDR-TB (Évora et al., 2016). An Adaptive Neuro-fuzzy inference system (ANFIS) is a rule-based system that uses some predefined set of rules while considering the following symptoms sputum, culture, x-ray, duration of symptom, drug susceptibility test (DST), and weight loss to detect pulmonary TB (Shrivastava et al., 2016). A rule-based system may fail if there is a new situation or an exception in the case of new TB symptoms (L.-H. Yang et al., 2018). Existing baseline methods classify drug resistance as present or absent based on several predetermined libraries of variants from the literature. Yang *et al*. applied machine learning techniques for classifying DR-TB DNA sequencing data; the models performed well when tested with real-world datasets(Y. Yang et al., 2018). Support vector machine (SVM) approach was used to detect the effectiveness of the various method of TB treatments by considering the following features; patients' unique numbers, patient number of visits to the hospitals, results of patient's test and type of drug administered to the patient that tested positive TB. This study achieved a prediction accuracy of 95%, aiding physicians in determining the most effective treatment to administer to TB patients(Rakhmetulayeva et al., 2018). Lokeshkumar *et al.* proposed an ensemble of ML techniques to classify MDR-TB in four classes such as Defaulted, Died, Treatment Completed, and Cured. Adaboost algorithm worked best on the datasets when compared to other ML techniques. The main aim of applying ML techniques to disease classification and prediction should not be based on such technique's capability; rather, it should be based on its efficiency in healthcare (Lokeshkumar et al., 2019). This research presents a data-driven approach that verifies the statistical significance of the symptoms of TB to prove the degree of their significance to DR-TB. Find the relationship between TB and DR-TB symptoms. It then presents a logistic regression (LR) model that classifies patients into positive or negative TB and DR-TB or DS-TB classes based on the probability of the attributes. LR is

used to identify the best fit that significantly describes the relationship between the output and input variables.

2.5 State of the Art of Machine Learning Techniques in Healthcare

Computer science today has a large focus on machine learning algorithms, and these are demonstrated massively in the field of knowledge Discovery Data (KDD), and generally, it has been the new focus of the WHO to employ technology to improve the healthcare system, thus the field of health informatics. The advances in digital technologies are becoming essential tools for healthcare specialists to provide the best care for patients(Toh & Brody, 2021). Drug discovery is the next key research area for the healthcare industry (Mathur, 2018). Research in pharmaceutical companies for certain diseases is continuously growing, and machine learning helps speed up drug discovery by analyzing medicinal data and providing prediction models on drug reactions even before they are injected into subjects in a controlled environment. This saves both time and money, as the simulation of drug reactions gives an estimate on likely cure patterns and reactions to the drug (Mathur, 2018).

Machine learning models are rapidly advancing and are useful for predicting and assessing structural performance, identifying the structural condition and informing preemptive and recovery decisions by extracting patterns from data collected via various sources and media. Machine learning tools have become available in diagnosing and predicting diseases, thereby saving costs and improving the likelihood of surviving, especially in some deadly diseases (Elshawi, 2020). In the case of infectious diseases, early diagnosis is highly needed in isolating the subjects to reduce the spread of the disease. Predictive analytics is the next era of application of machine learning in the healthcare industry. The focus would be on predicting the likely number of people who could develop a certain disease at a given time, the age of who may likely

develop a certain disease, and find patterns that may indicate a disease's status. This is done by assessing the large volume of datasets and images faster (Waring et al., 2020). Hence, the research on the classification and prediction model of auxiliary diagnosis based on clinical data has become one of the hot spots in intelligent medicine.

However, most hospitals are not currently deploying machine learning solutions. One of the reasons for this is that health care professionals often lack the machine learning expertise necessary to build a successful model, deploy it in production, and integrate it with the clinical workflow (Sun et al., 2021). However, employing machine learning in the healthcare sector will help ease the stress on physicians since high volumes make them more error-prone, machines can handle large sets of imaging data with a lower error rate, work in place of doctors in their absence.

2.6 Summary of the Related Works and our Major Contributions

From the review, it is observed that few works have been done applying machine learning models in malaria-endemic zones, and regardless of this, none have considered the six selected malaria-endemic countries of sub-Saharan Africa. Therefore, this research follows a systematic data mining approach, applying its tools and techniques to ensure high-quality data and better classification accuracy. Following this, The XGBoost algorithm employed in this research has been used by only a few researchers in the health care sector. The eXtreme Gradient Boosting (XGBoost) model is a machine learning method that can yield high precision prediction results through its strong self-learning ability for these non-linear data. A common term is included in the objective function of the XGBoost model, which helps prevent overfitting, and can control the complexity of the model. The XGBoost model has many advantages in model prediction, such as a fast operational speed, complete feature extraction, a good fitting effect and high

prediction accuracy (Alim et al., 2020). In the other part, this work has also presented an efficient TB diagnostic model that provides the most significant symptoms associated with TB and DR-TB to improve the time taken by physicians to diagnose patients. This is achieved with association rule mining by detecting hidden relevant information from a group of a dataset and providing some classification rules. FP-growth algorithm is used to mine and develop the association rules used for the TB classification/diagnosis. Finally, a Logistic regression-based model is efficient with fast training and prediction time.

CHAPTER 3

Prediction of Malaria Incidence using Climate Variability and Machine Learning

3.1 Introduction

This chapter highlights the different tools and techniques adopted to implement the malaria incidence classification model (MIC), including the design frameworks for implementation, the evaluation metrics for evaluating the performance of the proposed models, experimental result presentation, and concluding remarks.

3.2 Study Site

The study sites include Burkina Faso, Mali, Niger Republic, Nigeria, Cameroon, and DRC with details of their geographical locations as follows: Nigeria is between latitudes 4° and 14°N, and longitudes 2° and 15°E with a population of about 206,139,589. Mali is situated between latitudes 10° and 25°N, and longitudes 13°W and 5°E, with a population of 20,250,833 people. The Niger Republic lies between latitudes 11° and 24°N, and longitudes 0° and 16°E, with a population of 24,206,644. Cameroon lies between latitudes 1° and 13°N, and longitudes 8° and 17°E, having a population of 21,917,602. Burkina Faso is between latitudes 9° and 15°N and longitudes 6°W and 3°E, with a population of 20,321,378. DRC lies between latitudes 6°N and 14°S, and longitudes 12° and 32°E, with a population of 84,068,091. Although there is an overall decrease in malaria incidences, no significant change is observed in these six countries despite the investments made to reduce the transmission (Feachem et al., 2019). Therefore, a good understanding of the effects of climate variability on these regions distinctively will be useful to each of these countries for effective control and decision-making. Figure 3.1 shows the geographical map of the selected six countries, highlighting the major areas that are malaria-endemic.

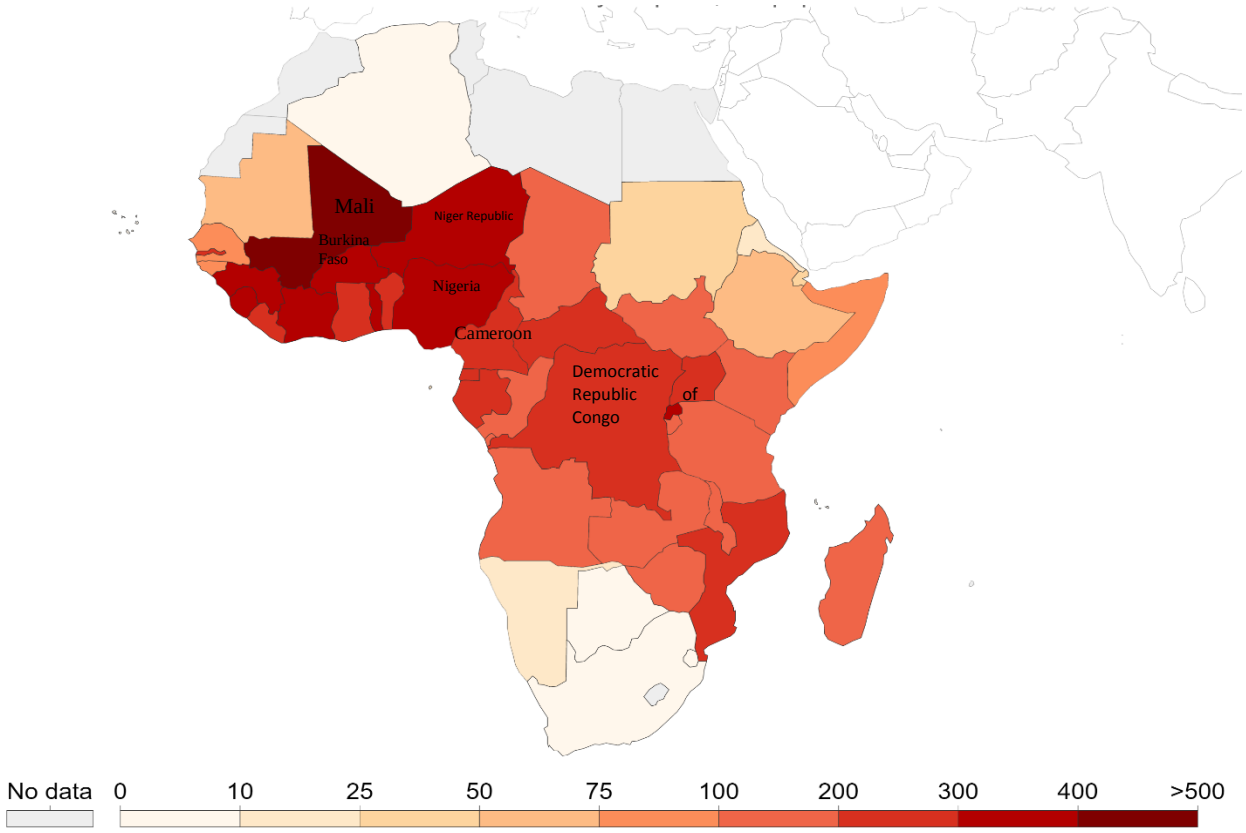


Figure 3.1. Geographical map of the six selected regions and their endemicity to malaria (source: World Health Statistics, 1990 - 2017)

3.2.1 Clinical Data

The confirmed malaria incidence for 28 years ranging from 1990 to 2017 for all the six selected countries, was obtained from the WHO data repository (Roser & Ritchie, 2020). The dataset contains a normalized value of the annual confirmed malaria incidence per 1000 population, which is the annual rate computed by dividing confirmed malaria incidence by its population size. The malaria incidence datasets are confirmed malaria cases recorded in different hospitals

and healthcare centres and then transferred to the WHO to eliminate diseases. Figure 3.2 shows the annual malaria incidence for the six selected regions for the past 28 years.

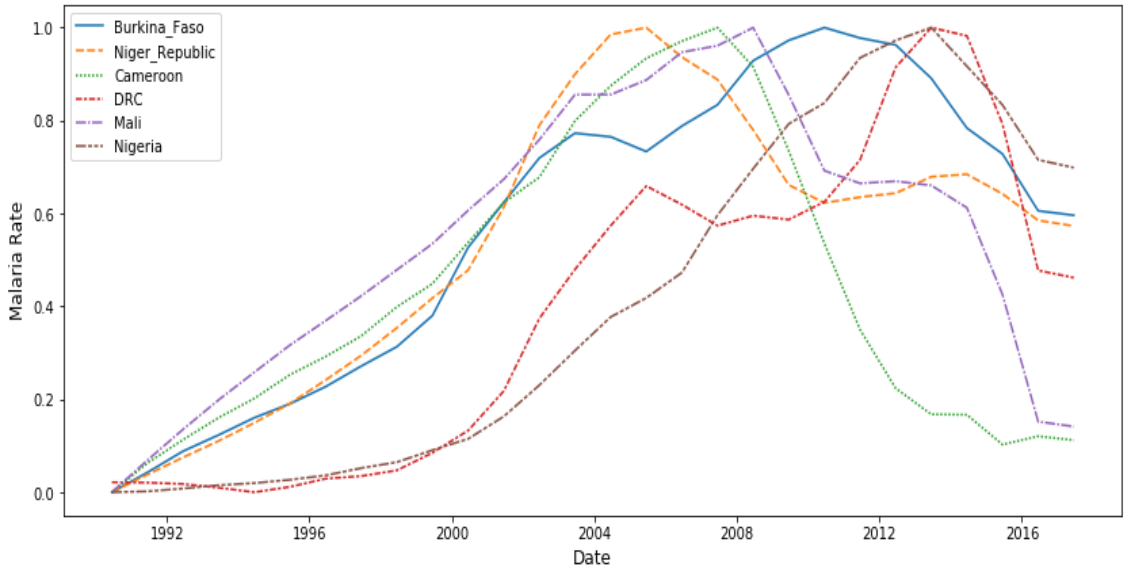


Figure 3.2. Annual malaria incidence case per 1000 population for the six selected regions (Source: Authors)

3.2.2 Climate Data

The climatic dataset was collected from the National Centre for Atmospheric Research (NCAR) (https://ncar.ucar.edu/what-we-offer/data-services, 2019). These are observational data taken for 28 years (1990 to 2017) containing daily records of the earth system observations such as atmospheric pressure, surface temperature, precipitation, surface radiation, and relative humidity. Only the annual records of both climate variables and malaria incidence reports have been considered due to the availability of annual incidence reports of malaria in the six selected countries. The annual precipitation is within 1192 and 1694 mm, the annual temperature is between 25.0 and 29.5°C, the annual relative humidity is between 40.2°C and 45.5°C, the annual

surface radiation ranged between 220 and 240°C, and finally, the pressure is between 99814 and 99820pa across the six selected countries.

3.3 Experimental Dataset

Both climatic variables and malaria incidence reports contain continuous real value numbers per region. The climatic variables have five attributes: precipitation, surface radiation, temperature, atmospheric pressure, relative humidity used as the independent variables (predictors), and the malaria incidence report known as the dependent variable (target class). The increase and decrease in malaria incidence were represented using 1 and -1, respectively. Table 3.1 presents a sample of the raw dataset before preprocessing. These predictors are defined below:

i. Precipitation: It is a product of the condensation of atmospheric water vapour that falls under the force of gravitational pull from the cloud. When some portion of the atmosphere becomes saturated with enough water, it reached up to 100% of relative humidity. Precipitation may exist in the form of rain, drizzle, snow, ice pellets, hail, graupel and sleet.
ii. Relative humidity: Humidity is defined as the total concentration of water vapour in the air. The relative humidity is the ratio of the partial pressure of water vapour to the saturation vapour pressure of water at the same temperature
iii. Atmospheric temperature: Atmospheric temperature is a measure of temperature at different levels of the Earth's atmosphere. It is controlled by many factors, including incoming solar radiation, humidity and altitude.
iv. Atmospheric pressure: Atmospheric pressure is defined as the force per unit area exerted against a surface by the weight of the air above that surface.
v. Surface radiation: The total radiative energy absorbed by the land surface is the major forcing that drives the land surface processes of water, energy, and biology

Table 3.1: Sample of the dataset before preprocessing

Date (mm/dd/yyyy)	Precipitation (mm)	R_humidity ($g.kg^{-1}$)	Atm_Temp (^oC)	S_radiation (W/m^2)	Pressure (pa)	Malaria incidence
6/16/1990	0.812998	44.6705	25.5804	292.177	97208	17720.77
6/16/1991	0.980199	45.8741	25.4737	287.061	97218.8	17708.78
6/16/1992	0.851804	46.5215	25.0838	290.031	97235.7	17658.16
6/16/1993	0.83694	46.6726	25.3986	292.72	97210	17525.48
6/16/1994	1.15679	46.8185	25.4253	297.068	97232.3	17363.39
6/16/1995	0.958874	47.2742	25.5901	292.391	97201.2	17556.83
6/16/1996	0.827338	48.3952	25.5854	294.092	97144.8	17850.52

3.3.1 Data Pre-processing

Error-prone systems due to noise in data can negatively affect the detection of unusual trends. Machine learning principles uphold high accuracy through data preprocessing to obtain high-quality data (Alexandropoulos et al., 2019). An in-depth analysis of both climate variables and malaria incidence report was carried out in collaboration with an ecologist and a health professional to examine the health implication of these variables to malaria transmission and occurrence; it was discovered that pressure has less significance in malaria incidence, which is in harmony with the statistical result. It was normalized using minmax_scaler to unify them into the same scale. Following this, the target variable was transformed from continuous variables into a discrete variable, using the malaria incidence threshold method as proposed by the WHO (Roser & Ritchie, 2020). The annual mean for the past 28 years (n=28) plus two multiplied by the standard deviation (SD) as shown in Equations (8), (9), and (10).

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (8)$$

$$SD = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} \qquad (9)$$

Malaria incidence threshold = $\bar{x} + 2(\text{SD})$ (10)

Where:

$\bar{x} = population\ mean$

$x_i = annual\ incidence\ report$

$n = total\ number\ of\ years$

$SD = Standard\ Deviation$

The following thresholds for malaria incidence were obtained for each country; Burkina Faso: 13185.2, Nigeria: 215096, DRC 2833.52, Mali: 26321.5, Cameroon: 25755.8, and Niger: 2361.83. The class variable is grouped into 1 and -1, which implies that whenever malaria incidence rises above the threshold, it indicates high malaria incidence and is classified as 1, and if it is below these values, it indicates low incidence. Table 3.2 shows a sample of the preprocessed data of Table 3.1.

Table 3.2: Sample of the dataset after preprocessing

Precipitation	R_humidity	Atm_temp	S_radiation	Pressure	Malaria incidence
-0.50512	-0.74055	-0.9594	-0.49644	0.343463	-1
-1.70873	0.486246	-1.11436	0.380656	1.149704	1
-1.22037	0.244946	-2.55947	1.720385	1.352471	1
-0.72998	0.406975	-1.26241	1.640703	2.115261	-1
1.637181	1.333713	0.10851	0.04944	-0.14897	-1
0.561079	-0.20725	-0.02379	-0.20566	-0.48692	1
0.744139	1.126933	-0.7237	0.109499	-0.55933	1

3.3.2 Feature Engineering using Statistical Significance

Although data quality might be a less critical problem for screening common, nonspecific indicators for statistical aberrations, quality should be evaluated and improved to the extent possible. Pearson's correlation analysis detects the degree of relationship between the feature variables and target variable, this aids in selecting only the relevant features that have a strong influence on the occurrence of malaria incidence (Zhao et al., 2020). Pearson's correlation coefficient lies between -1 and +1, where -1 indicates a strong negative correlation, 0 indicates no correlation, while 1 implies a strong positive correlation. Equation (11) expresses the correlation coefficient mathematically.

$$p(x, y) = \frac{\sigma xy}{\sigma_x \sigma_y} \qquad (11)$$

where:

$\sigma_x = standard\ deviation\ of\ x$

$\sigma_y = standard\ deviation\ of\ y$

$\sigma xy = population\ covariance$

Test for collinearity

When predictors exhibit a linear relationship with each other, collinearity occurs (Zaki et al., 2014). Variance inflation factor (VIF) finds the strength of the association between variables. Collinearity occurs if VIF is greater than and no collinearity if VIF is less than or equal to 1, it indicates no collinearity, but if the VIF is greater than 1, it indicates collinearity. VIF is expressed mathematically using Equations (12) and (13).

$$VIF = \frac{1}{1-R_i^2} \qquad (12)$$

$Where$:

$i = The\ predictors\ (x_1, x_2, \ldots, x_n)$

and

$$R^2_{adj} = \left[\frac{(1-R^2)(n-1)}{n-k-1}\right] \qquad (13)$$

R^2_{adj} = Adjusted R squared

n = total number of data sample

k = number of feature variables

3.4 Data Visualization

Data visualization is a technique used to provide quick and effective communication of information in a common way using visual information. This technique helps to identify the hidden relationship amongst variables quicker and highlight areas that need improvement or need more attention. It is used to visualize trends, variabilities and derive meaningful insights from the data, associations, and degree to which each variable affects the other. Figure 3.3 to Figure 3.8 show a line chart for displaying trends and change over time between malaria incidence and climate variability. Figure 3.3 shows the relationship between annual climate variability and malaria incidence in Nigeria. Malaria incidence seems to be stagnant between the year 1990 and 1996. It was at its extremes between 2004 to 2008. On getting close to the peak of the graph, an increase in precipitation, temperature and relative humidity affected the increase in malaria incidence. Between 2010 to 2017, Malaria incidence declined drastically. However, between 2014 to 2017, the pressure and temperature are inversely related to malaria incidence as the increase in pressure and temperature results in an increase in malaria incidence.

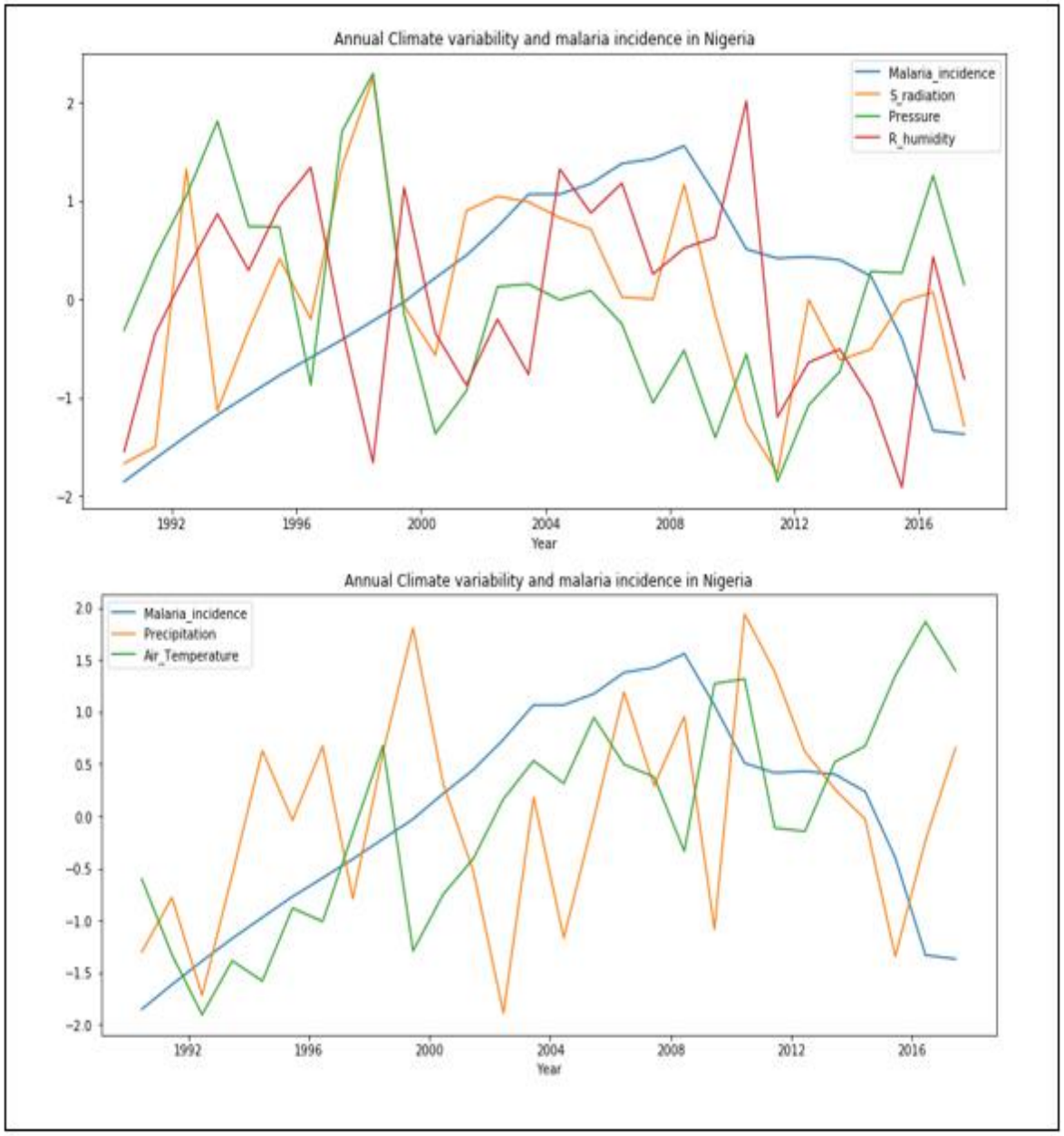


Figure 3.3: Trend in Annual climatic variability and malaria incidence in Nigeria

Figure 3.4 shows the line plot showing the relationship between climate variability and malaria incidence in Mali. The plot shows that malaria incidence was moderate between 1990 through 2002. However, from 2012 to 2015, malaria incidence increased and surpassed the malaria

incidence threshold in Mali. This could be referred to as an epidemic in Mali. Also, the rate of precipitation, atmospheric temperature, pressure and surface radiation, and pressure influenced the rate of malaria incidence, while relative humidity maintained the same non-influential relationship with malaria throughout the line plot. Figure 3.4 also showed that temperature and precipitation moved in the opposite direction from the origin between 1990 to 1996. This shows that as precipitation is increasing, the temperature rate is decreasing. As the precipitation is decreasing, the temperature rate is increasing.

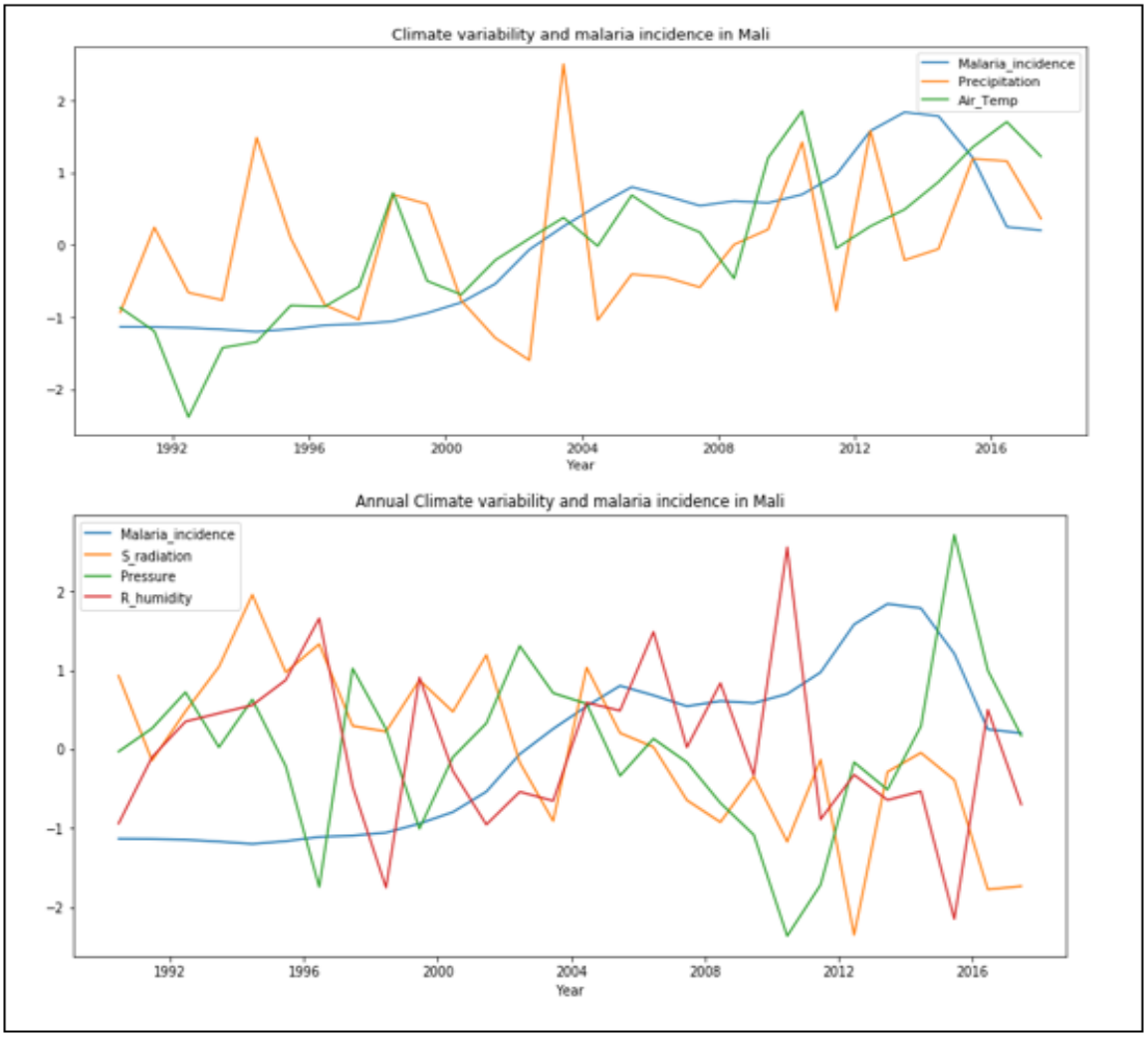


Figure 3.4: Trend in Annual climatic variability and malaria incidence in Mali

Figure 3.5 shows the trend in the relationship between annual climate variability and malaria incidence in DRC. Malaria incidence started increasing from 1995 and reached its peak in 2004 where the country experienced an abnormal increase in the incidence rate of malaria in DRC. During this period of abnormal increase in malaria incidence, there was a slight reduction in the volume of precipitation. There is a slight decrease in malaria incidence at the beginning of 2006. Similarly, it is observed that while the country also experienced a large amount of relative humidity, there was also a similar movement when the malaria incidence started decreasing. This shows that climate variability affects malaria incidence in both ways.

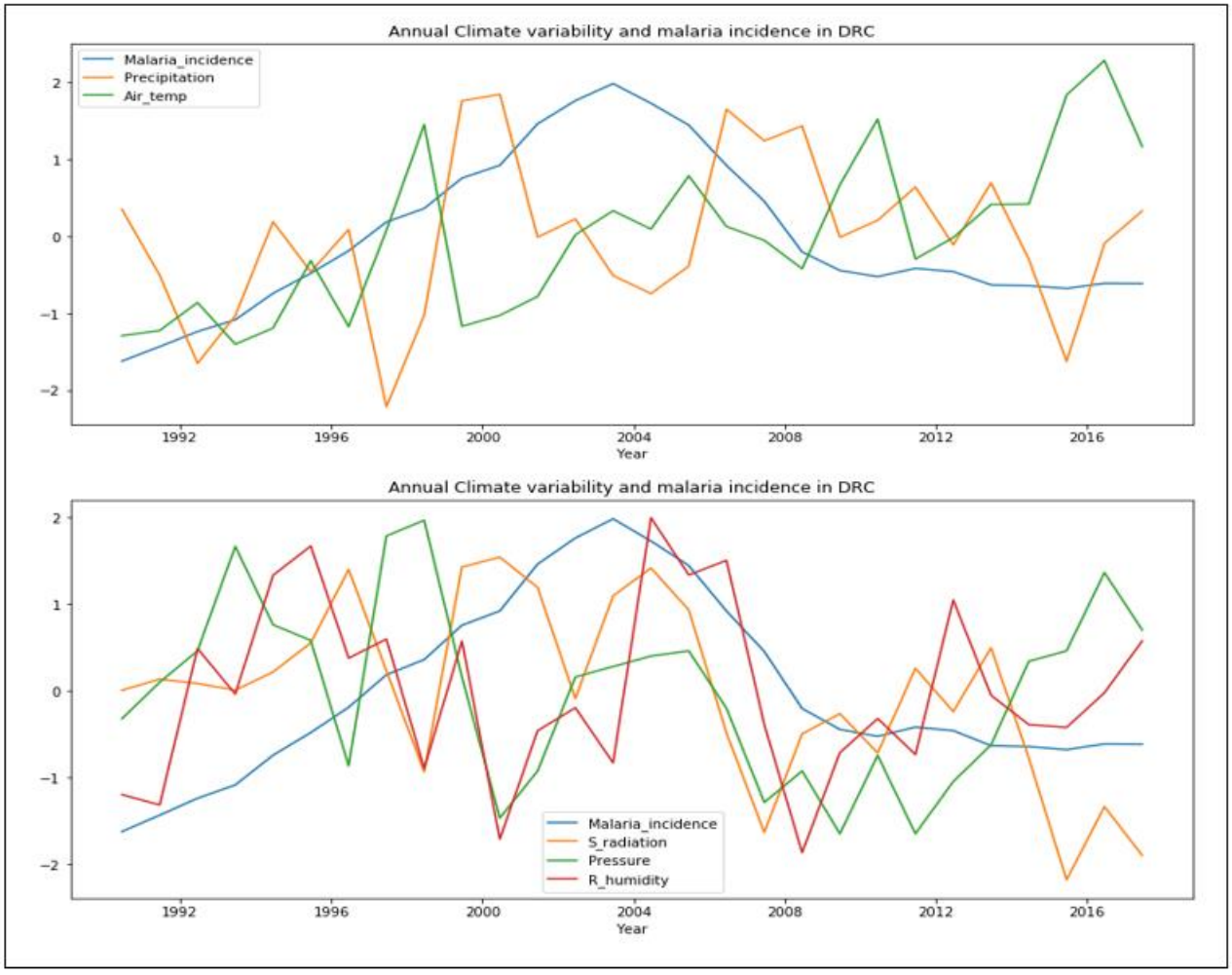


Figure 3.5: Trend in Annual climate variability and malaria incidence in DRC

In Figure 3.6, malaria incidence started rising from 1996 and has never returned to its original position of low incidence as in 1990. Burkina Faso experienced the highest rate of malaria incidence between 2008 to 2013 with the most influencing factors as relative humidity, precipitation and atmospheric temperature. The graph shows that while the malaria incidence case increased, precipitation and air temperature were increasing in a similar direction, except the atmospheric pressure and surface radiation, which moved in a slightly opposite direction. Figure 3.6 Summarizes that an increase in temperature, precipitation, and relative humidity leads to a higher incidence of malaria incidence in a given year.

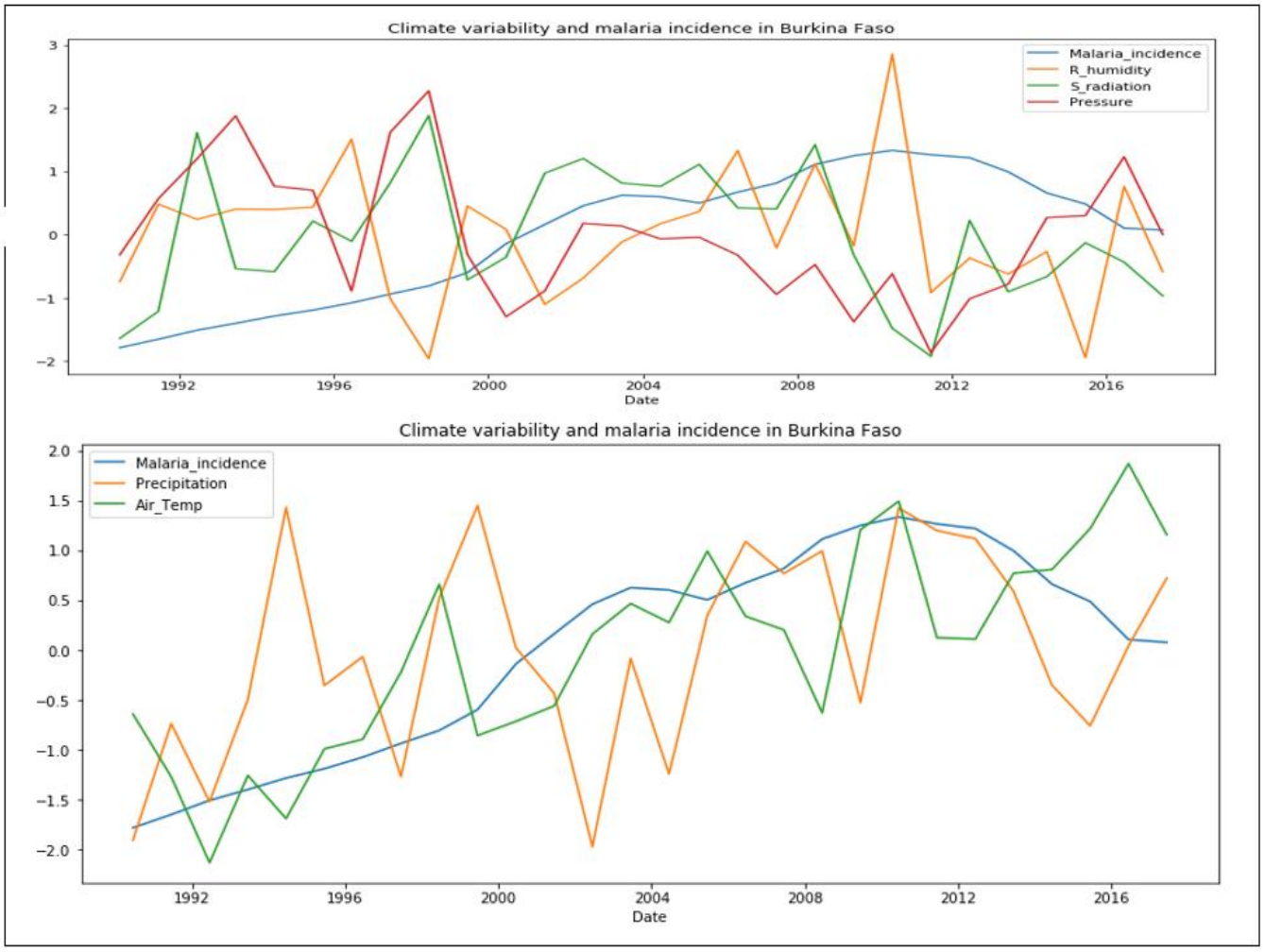


Figure 3.6: Trend in Annual climatic variability and malaria incidence in Burkina Faso

Figure 3.7 shows the variability of climate factors and malaria incidence in Cameroon. Malaria incidence starts increasing from the year 2000 and reached its peak in 2004, when relative humidity, surface radiation, precipitation and air temperature was high. Cameroon experienced what could be referred to as an "outbreak" between the years 2000 to 2008, and it was influenced by a substantial amount of precipitation, relative humidity, temperature and few amount of pressure. However, malaria incidence started decreasing from the year 2008 but has remained unstable. This may be as a result of instability in the occurrence of precipitation, temperature, and radiation.

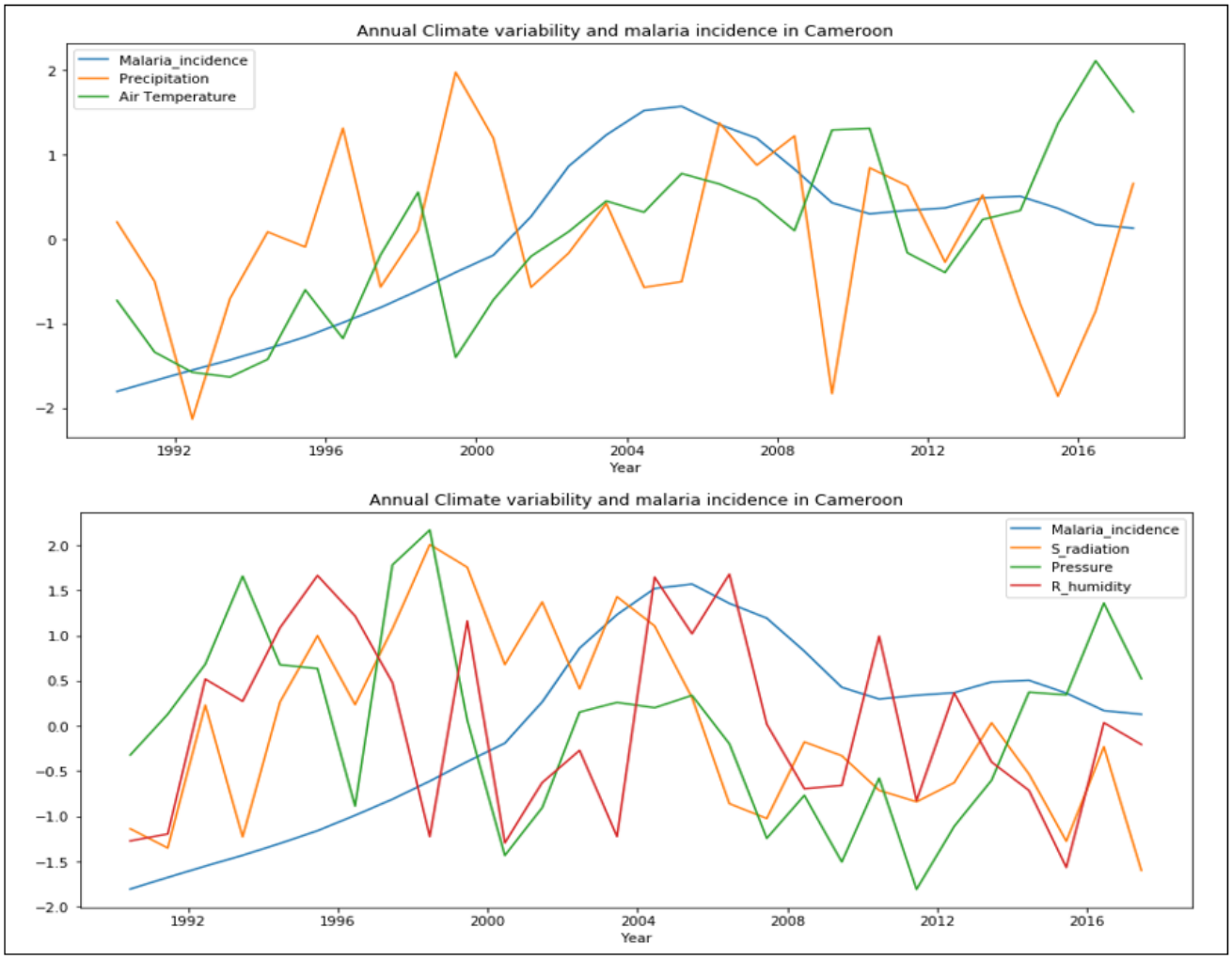


Figure 3.7: Trend in Annual climatic variability and malaria incidence in Cameroon

In Figure 3.8, It is observed that Relative humidity, Pressure, and Air temperature followed the same trends with malaria incidence at some point between 2006 and 2017. This shows that whenever malaria incidence is increasing, relative humidity and pressure increases as well. In the same way, whenever malaria incidence decreases, relative humidity, pressure and air temperature decrease simultaneously. There is a direct opposite variation between malaria incidence and precipitation, and surface radiation. Malaria incidence seems to be stable between 1990 to 2007, and the country experienced an abnormal incidence of malaria between the year 2008 to 2013. At the beginning of 2014, malaria incidences started fluctuating and are mostly affected by temperature. In summary, precipitation, relative humidity and are the principal factors affecting high malaria incidence

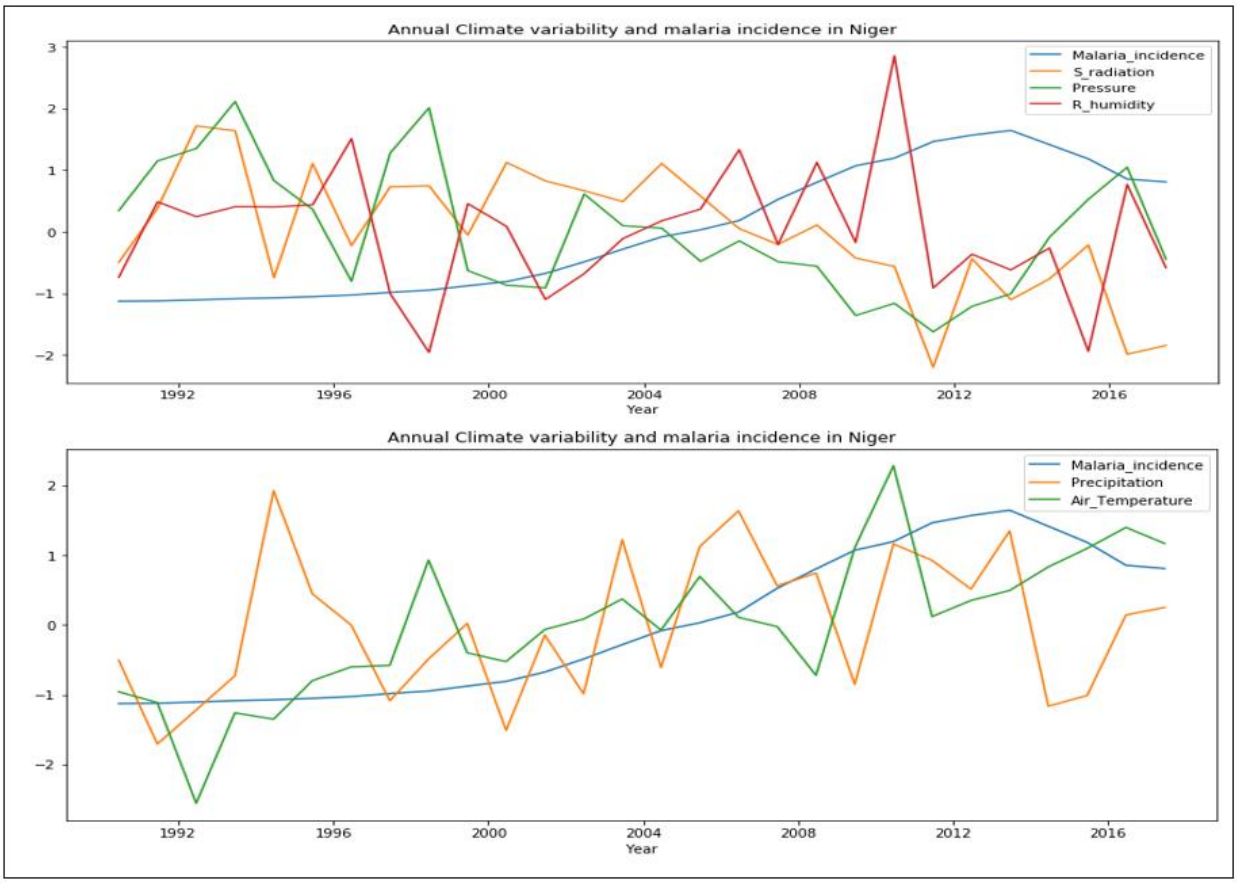


Figure 3.8. Trends in climatic variability and malaria incidence in Niger Republic

3.5 Implementation Frameworks

This Subsection presents the major frameworks for the proposed system implementation. It explains in details the k-means clustering techniques that were used for outlier detection and removal. Then it explains the XGBoost model and how it works. It also explains the performance metrics for evaluating the performance of the XGBoost model

3.5.1 K-means Clustering

Clustering groups large amounts of data points into smaller clusters by adding similar objects to the same cluster based on their distances (Domingues et al., 2018). The purpose of applying the K-means clustering approach here is for outlier detection and removal. Therefore these steps lead through the execution of the k-means clustering approach:

Step 1: k initialization: choose k, and set k=2. Create *k* number of clusters and apply Euclidean distance metric for computing the similarity distance within the input values and then assign the input item to the closest cluster using Equation (14).

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_i^{(t)} \right\|^2 \ \forall j, 1 \leq j \leq k \right\} \qquad (14)$$

Step 2: Update: For each input data that is assigned to a cluster by recomputing the mean value, update the centroid, as shown in Equation (15). Repeat Steps (1) and (2) until the mean value of the cluster converges and then update the centroid again.

$$m_i^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_j \in S_i^{(t)}} x_j \qquad (15)$$

Step 3: Remove the outliers: Remove all the data input that is wrongly clustered to generate a new data size and input. If the new data size is greater than 70%, then move to data classification; else, repeat the k-means clustering process until an appropriate data size is obtained. About 0.012% of outliers were detected and removed from the dataset.

3.5.2 Extreme Gradient Boosting

Extreme Gradient boosting (XGBoost) is a machine learning method used to solve classification and regression problems mostly in the form of trees(Ji et al., 2019). It is scalable and efficient in using memory and drives fast learning through parallel and distributed computing. This model is best suited for fewer samples of datasets. Over-fitting is controlled by the XGBoost model for better performance and makes the model better than the other boosting models. The proposed MIC model is implemented using the XGBoost model, and the algorithm works as follows:

Here, the objective function and the prediction function are created. The training parameters minimizes the objective function to identify the relevant parameters, and so, the parameters obtained and prediction function are used to classify the unknown sample. Given a set of data $D = \{(\ x_i,\ y_i)\}\ (x_i \in R^m, y_i \in R, i = 1,2\ , \dots n)$, having n, m dimensions, the XGBoost model is expressed as in Equation (16)

$$\hat{y}_i = \sum_{k=1}^{k} f_k(x_i),\ f_k \in F(i = 1,2, \dots, n) \qquad (16)$$

Where:

$$F = \ \{f(x) = w_{q(x)}\}(q: R^m \longrightarrow \{1, 2, \dots, T\}, \qquad W\ \in R^T)$$

$$W\ \in R^T = CART\ decision\ tree\ structure\ set,$$

q = tree structure of the sample map to the leaf nodes

T = number of leaf nodes and

W = total score of the leaf nodes.

When constructing the XGBoost model, it is essential to find the optimal parameters following the principle of creating a minimal objective function to building an optimal model. The objective function of the XGBoost model can be divided into an error function term L and a model complexity function term X. The objective function can be written as in Equation (17):

$$Objective\ fxn = L + \Omega \quad (17)$$

$$L = \sum_{i=1}^{n}(y_i - \hat{y}_i) \quad (17a)$$

$$\Omega = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (17b)$$

Where:

γT = regular term of Li

$\frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ = regular terms of L2

$\gamma\ and\ \lambda$ = adjustment parameters helping to prevent the model from overfitting

When using the training data to optimize the model training, it is necessary not to change the original model and add a new function, f, to the model to reduce the objective function as much as possible. The objective function is expressed as in Equation (18):

$$OBJ^{(t)} = \sum_{i=1}^{n}\left(y_i - \left(\hat{y}_i^{(t-1)} + f_t(x_i)\right)\right)^2 + \Omega \quad (18)$$

Where:

$\hat{y}_i^{(t-1)}$ = predicted value

$f_t(x_i)$ = new function added at $tth\ time$

OBJ = scoring function used as an evaluation model

OBJ is used to search for the best tree structure, and the smaller the OBJ value, the better the effect of the model, as this ensures an optimal XGBoost model. By calling the tree recursively, a larger proportion of regression tree structures are obtained (Li & Zhang, 2020),

XGBoost model allows the adoption of cross-validation in every iterative stage of the boosting process, enhancing the precise optimal number of boosting iterations in each run. Hyperparameter optimization tries to find an optimal quickly, or at least an effective, combination of hyperparameter values that maximizes some performance metrics for the given

machine learning task (Bergstra et al., 2011). Hyperparameter optimization is applied to the proposed model to help in the selection of the hyperparameter value that works best on the classification; the following hyperparameters was selected in the training of the MIC model:

i. Learning_rate = 0.2, it is a step size that aids to prevent overfitting, and its values range between [0,1].
ii. Max_depth = 6; this determines the depth of each tree that can grow during each boosting phase.
iii. n_estimators = 100, number of trees to be built.
iv. Gamma = 0.1 regulates a node's splitting based on the predictable decrease in loss after the split.
v. scale_pos_weight = 1, it helps in faster convergence.
vi. min_child_weight = 1, Used to control over-fitting.
vii. Seed = 10, The seed is a random number used for parameter tuning and creating reproducible results.

3.5.3 Performance Metrics and Model Selection

Performance metrics are used to ensure that the learning algorithm has learned enough to predict or classify data accurately without minimal error. The performance of the MIC model is evaluated using the following metrics:

Classification Accuracy

Classification accuracy is the ratio of the sum of accurate predictions to the sum of input samples used as expressed in Equation (19). Accuracy scores range between 0 to 100% or simply 0 to 1. An accuracy score of 1 or 100% means that the classifier has correctly classified or predicted

$$Accuracy = \frac{ratio\ of\ number\ of\ accurate\ predictions}{total\ number\ of\ predictions} \qquad (19)$$

The Area Under Curve (AUC)

AUC is used to evaluate binary classification models. The AUC scores present a good summary of the performance of the receiver operator curves (ROC). A recent study shows that AUC is comparatively better than the classification accuracy (Huang & Ling, 2005) because it avoids the supposed subjectivity in the threshold selection process. It measures if predictions are properly ranked instead of their absolute values, and measures the quality of a model's prediction, regardless of chosen classification threshold. There are some types of problems where accuracy has proven to be insufficient and misleading in helping assess the performance of an algorithm. One of such problems is severe problems where one target variable value has higher occurrences than the other. When the cost of classification rises, relying on only accuracy is not sufficient. The Receiver Operating Characteristics (ROC) is a graphical representation of the false-positive rate (FPR), also known as sensitivity and true positive rate (TPR), also known as specificity, which represents the model's performance. The higher the AUC score, the better the model's performance. The sensitivity and specificity of the model are computed using Equations (20) and (21). On a ROC graph, TP is plotted on the Y-axis and FP is plotted on the X-axis.

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \qquad (20)$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive} \qquad (21)$$

Note that an AUC score of 1 indicates high performance, while an AUC score of 0 indicates the worst performance of a classifier. (Hand & Till, 2001) present a straightforward approach for calculating the AUC of a classifier for binary classification is in Equation (22).

$$AUC = \frac{S_0 - n_0(n_0+1)/2}{n_0 n_1} \qquad (22)$$

Where:

n_0 = number of points in the test set belonging to class 0,

n_1 = number which belongs to class 1.

$$S_0 = \sum r_i$$

r_i = rank of the ith positive example in the ranked list

Cross Validation

Cross-Validation (CV) is a technique used in accessing the ML model's effectiveness, especially while reducing bias and overfitting, which may arise due to the quantity of the dataset (Shao, 1993). Assume that n data points are available for a model selection from a class of models. The n dataset is split into k number of portions. The first portion contains n_c data points to fit the data to the model, while the second portion $n_v = n - n_c$. Validating a model is not only done using $n_v$but can also be done using all the $data, n = n_v + n_c$ and can be divided $\binom{n}{n_v}$ various ways (Shao, 1993). Cross-validation is typically used to select the model with the best average predictive or classification ability and is done using different approaches such as k-fold CV, leave-one-out, calculated based on the different ways of data splitting (Wong & Yeh, 2020). The k-fold CV technique is used in this research, and it involves splitting a specified set of n data points into a k-number of partitions or folds, where each fold is used as a training or testing set. K-fold CV follows an iterative process where the first partition or fold is used to test the model, and the remaining folds are used to train the model for the first iteration. This process is repeated until each fold of the k-fold is used as a testing set. K=5 fold was used for the validation of the proposed MIC model.

Akaike Information Criterion (AIC)

The Akaike information criterion (AIC) is a mathematical method for evaluating how well a model fits data. It is used to select the best model during model comparison (E. J. Wagenmakers

& Farrell, 2004). This is achieved by calculating and comparing the AIC scores of several possible models and then selecting the one that best fits the data. It determines the relative information value of the model using the maximum likelihood estimate and free parameters in the model and is computed using Equation (23).

$$AIC_i = -2logL_i + 2V_i \quad (23)$$

Where

L_i = the maximum likelihood for the candidate model i

V_i = free parameters of a given model

Comparing models with AIC, the AIC of each model is calculated first, and then, the model with the smallest AIC value is preferred. Akaike weight is a measure of the relative likelihood of a model, and it is used to facilitate the interpretation of the results of AIC model comparison procedures. Akaike weights are calculated by computing the difference in AIC models with respect to the best AIC candidate model using Equation (25). Finally, the Akaike weight is computed using Equation (26). The model with the highest Akaike value is selected as the best model

$$\Delta_i(AIC) = AIC_i - minAIC \quad (25)$$

$$w_i(AIC) = \frac{exp\left\{-\frac{1}{2}\Delta_i(AIC)\right\}}{\sum_{k=1}^{k} exp\left\{-\frac{1}{2}\Delta_k(AIC)\right\}} \quad (26)$$

3.6 Malaria Incidence Classification Model

The malaria incidence classification (MIC) model was implemented on Anaconda 3. It is open-source application software that supports machine learning tools and techniques, data science applications and Python 3.6 programming language. The MIC model involves three basic systematic approaches: data preparation and cleaning, k-means clustering, classification, and k-fold cross-validation. The First step is the data preparation task, which includes data collection,

preprocessing to remove noise, test for collinearity to remove variables dependent on each other, statistical analysis, and data visualization to observe trends and relationships in the data. The next step involves the k-means clustering task to detect outlier in the dataset, which may affect the accuracy of the proposed model. When an outlier is detected, such a record is removed from the dataset. In the end, if the remaining dataset is less than 70% of the original dataset, the k-means clustering task is repeated to obtain a $new\ data\ size \geq 70\%$. Finally, the new data inputs obtained from the k-means clustering process serves as an input for training and testing the MIC model. The dataset was slipt into training and testing set using k-fold CV approach. K-fold CV is an iterative process that is repeatedly done k-number of times. The first k-fold is used as a testing set, while the remaining folds are used to fit and train the model. During this process, the performance metric is applied to the dataset to evaluate and record its performance. Lastly, the result of the model evaluation is printed to show their performance. Algorithm 1 shows the systematic approach to the MIC model, while Figure 3.9 presents the system flowchart.

Algorithm 1: MIC model

Step 1: Data preparation and cleaning

Input: complete datasets

Output: A set of clean data with only the relevant features

i. *Get datasets*
ii. *Integrate datasets*
iii. *FOR each* $record_{ij}$ *of the datasets*

IF $record_{ij}$ *is NULL*

Discard record

FOR each variable

IF VIF > 1;

Discard variable;

Step 2: K-means Clustering

Input: k, preprocessed data

Output: dataset with no outlier

i. *Set number of clusters K*

ii. *Initialize the centroids = average of all data points that belong to each cluster*

iii. *Compute the sum of squared distance between data-points and centroids*

iv. *FOR each data point,*

IF data-point is closest to cluster,

Assign data-point to the nearest cluster

v. *Iterate until no change in centroids*

vi. *IF data size < 70%*

Repeat clustering process until datasize >=70%

Step3: Classification and k-folds cross-validation

Input: k-means clustered data, training sets, test set, hyperparameters,

Output: predicted values, classification accuracy scores

i. *Set number of hyperparameters, p*

ii. *Divide datasets into K-folds*

iii. *Perform parameter combination p in P*

iv. *FOR each k_i in k-folds*

Set fold k_i as Test-set

FOR fold $k_j$in, the remaining k-1 folds

set k_j as the validation set

Train MIC model on the remaining k-2 folds

Evaluate the performance of MIC model on k_j

Calculate the average performance over k-2 to select the best parameters p

Train MIC model on the k-1 folds with the best hyper-parameters and

Get Average performance

Evaluate MIC model performance on k fold

v. *Compute the average performance over K-folds*

Stop

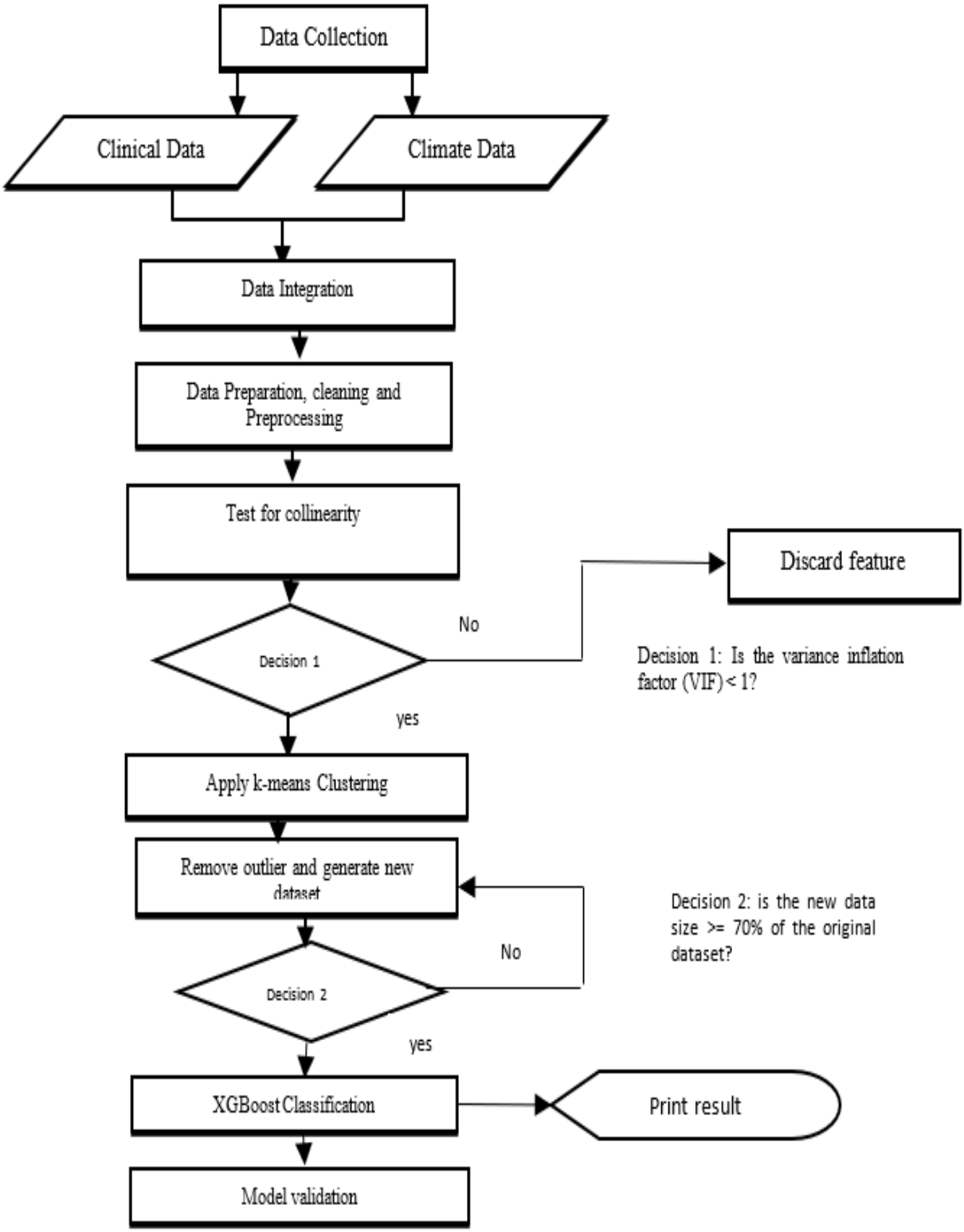


Figure 3.9. Flow diagram of the Malaria incidence classification (MIC) model

3.7 Results and Discussion

This section presents the experimental results of the proposed system. Firstly, it presents and discussed the results of the statistical analysis and experimental result of the MIC model in detail.

3.7.1 Statistical Significance of Predictor Variables

A hypothesis test shows a linear relationship between feature and target variables strong enough to model the relationships in the sample data at $\alpha = 0.05$. Table 3.3(a)-3.3(f) presents the resulting correlation coefficient matrix and their corresponding p-values across the six countries. The results showed a significant variation across each country using a 95% confidence interval and $p < 0.05$. The general result suggests that malaria incidence has a significant positive linear relationship with air temperature and precipitation and a negative linear relationship with pressure across the six selected countries. Table 3.3(a) shows that a positive significant linear relationship exists between precipitation, surface radiation, relative humidity, and malaria incidence in DRC. In Table 3.3(b), malaria incidence in Niger showed a positive linear relationship with precipitation and surface radiation and a negative relationship with pressure, surface radiation, and relative humidity. Similarly, in Table 3.3(c), Malaria incidence in Mali has a positive linear relationship with air temperature and precipitation and a negative linear relationship with relative humidity, surface radiation, and pressure. Table 3.3(d) shows a significant positive relationship between malaria incidence and precipitation, air temperature, and a negative relationship between pressure, surface radiation, and relative humidity and malaria incidence in Nigeria. Table 3.3(e) shows that malaria incidence has a significant positive linear relationship with precipitation, surface radiation, air temperature, relative humidity, and a negative relationship with pressure in Cameroon. Finally, Table 3.3(f) shows a significant positive linear relationship between air temperature, precipitation, surface radiation, relative

humidity with malaria incidence, and a strong negative linear relationship between pressure and malaria cases in Burkina Faso.

Table 3.3: The Significance Table

Table 3.3a. DRC			
X	Y	r	p-value
Precipitation	Malaria incidence	0.377095	0.047913
Pressure		-0.330831	0.085508
S-radiation		0.197373	0.314065
Air_temp		-0.573343	0.001426
R-humidity		0.053995	0.784939

Table 3.3b. Niger			
X	Y	r	p-value
Precipitation	Malaria incidence	0.310288	0.018058
Pressure		0.019006	0.923524
S-radiation		0.691280	0.000046
Air_temp		0.573343	0.001426
R-humidity		0.660484	0.0000131

Table 3.3c Mali			
X	Y	r	p-value
Precipitation	Malaria incidence	0.211640	0.027963
Pressure		-0.056212	0.776325
S-radiation		-0.613487	0.000517
Air_temp		0.683682	0.000061
R-humidity		-0.056212	0.776325

Table 3.3d. Nigeria			
X	Y	r	p-value
Precipitation	Malaria incidence	0.222421	0.255285
Pressure		-0.495941	0.007276
S-radiation		-0.613487	0.000517
Air_temp		0.338675	0.0477915
R-humidity		-0.056212	0.776325

Table 3.3e. Cameroon			
X	**Y**	r	p-value
Precipitation	Malaria incidence	0.145549	0.459903
Pressure		-0.327888	0.088499
S-radiation		0.048337	0.807032
Air_temp		0.658249	0.000140
R-humidity		0.048247	0.807383

Table 3.3f. Burkina Faso			
X	Y	r	p-value
Precipitation	Malaria incidence	0.399956	0.034961
Pressure		-0.595829	0.000821
S-radiation		0.015477	0.937696
Air_temp		0.694094	0.000042
R-humidity		0.041713	0.833080

The obtained results show that climate variability affects malaria incidence in different countries in diverse ways. Table 3.4 presents a summary of the feature engineering process and shows the statistically significant predictors of malaria incidence. The symbols "+" and "-" indicate the inclusion and exclusion of predictors, respectively.

Table 3.4: Input variables (predictors) for malaria incidence classification.

Input Variable/country	Precipitation	Pressure	Surface radiation	Atm Temperature	Relative Humidity
DRC	+	-	-	+	-
Mali	+	-	+	+	-
Cameroon	-	-	-	+	-
Niger Republic	+	-	+	+	+
Nigeria	-	+	+	+	-
Burkina Faso	+	+	-	+	-

3.7.2 Result of MIC Model

The feature engineering process involves removing irrelevant features by selecting only relevant variables from the original dataset, which enhanced the k-means clustering process and outlier management, as shown in Table 3.5. It is very important to obtain good precision in the machine learning model when dealing with healthcare data. The dataset was divided on a ratio of 70:30, where the training sets contain 70% (18 records) of the data set, and the test set contains 30%(8 records) of the dataset. K-fold CV technique at k=5 was used during the training phase. The CV process is repeated *k* number of times, while the training set is divided into a subset of discrete folds that forms a training set, and each subset is used as a test set to the other four subsets. A single estimation is obtained by taking the average of the k-results. A grid search algorithm was used for optimizing the hyperparameter while fitting the training set to the MIC model to select the best hyperparameters. The grid search algorithm typically tries all the possible combinations of the parameter values and then returns the combination with the maximum accuracy. All the possible combinations of hyperparameters are tested by fitting and scoring each combination of hyperparameters separately; the best hyperparameters are selected at the end. The test set is used

to evaluate the accuracy and AUC scores of the MIC model. Figure 3.4 shows the ROC and AUC scores representing the MIC model's performance across the six countries. Figure 3.10a-10f shows the mean AUC scores across the six countries, and they are as follows 0.97, 0.94, 0.91, 0.97, 0.94, and 0.92 for Mali, Cameroon, DRC, Nigeria, Niger, and Burkina Faso, respectively. The higher the AUC, the better the model can distinguish between the two target variables, low and high incidences. It is seen from the plot that the MIC models for each country have correctly predicted the high incidence cases.

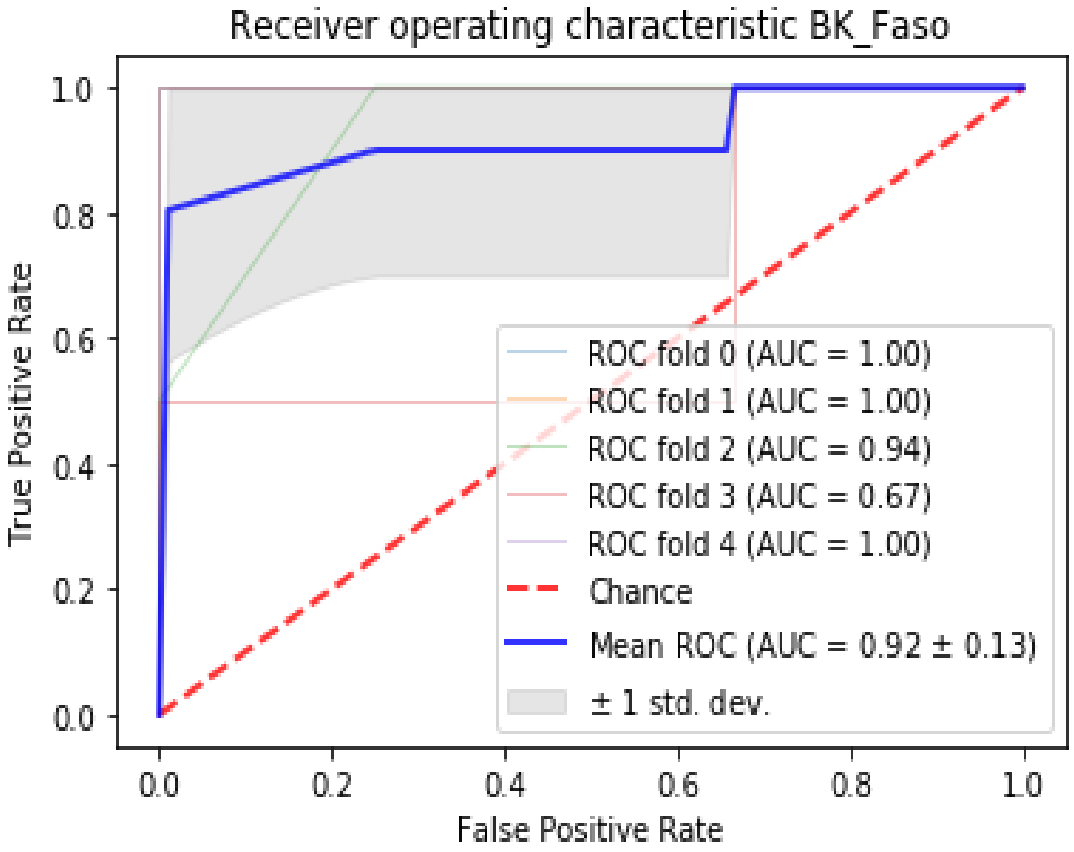


Figure 3.10a: ROC and AUC score of MIC model in Burkina Faso

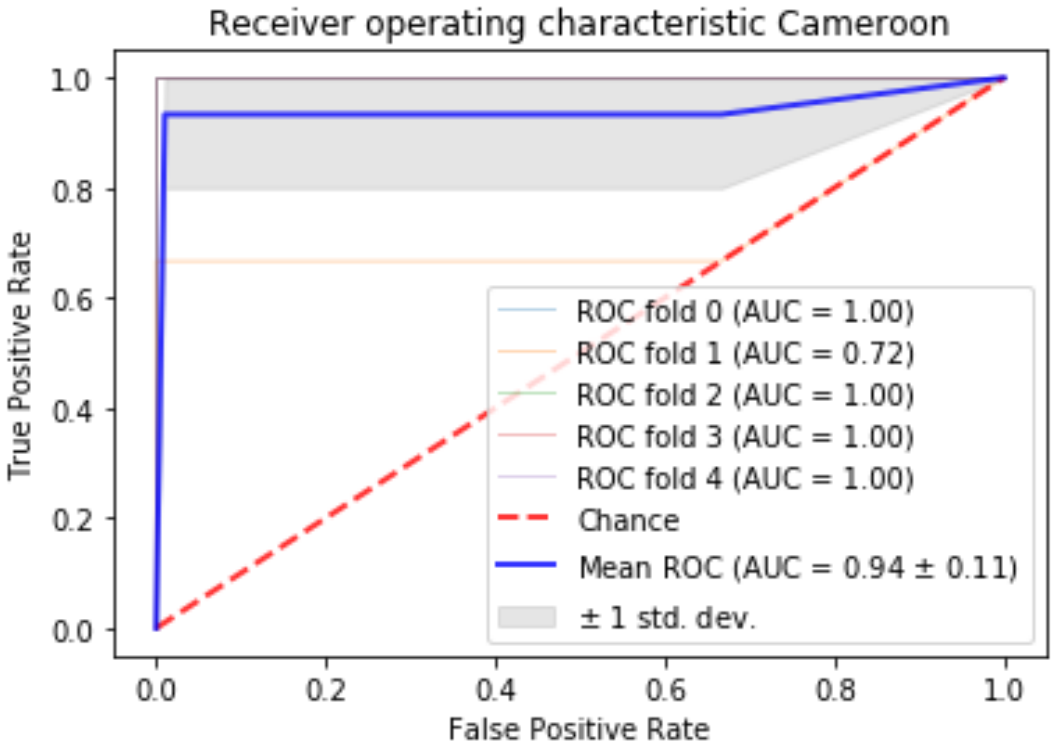


Figure 3.10b: ROC and AUC score of MIC model in Cameroon

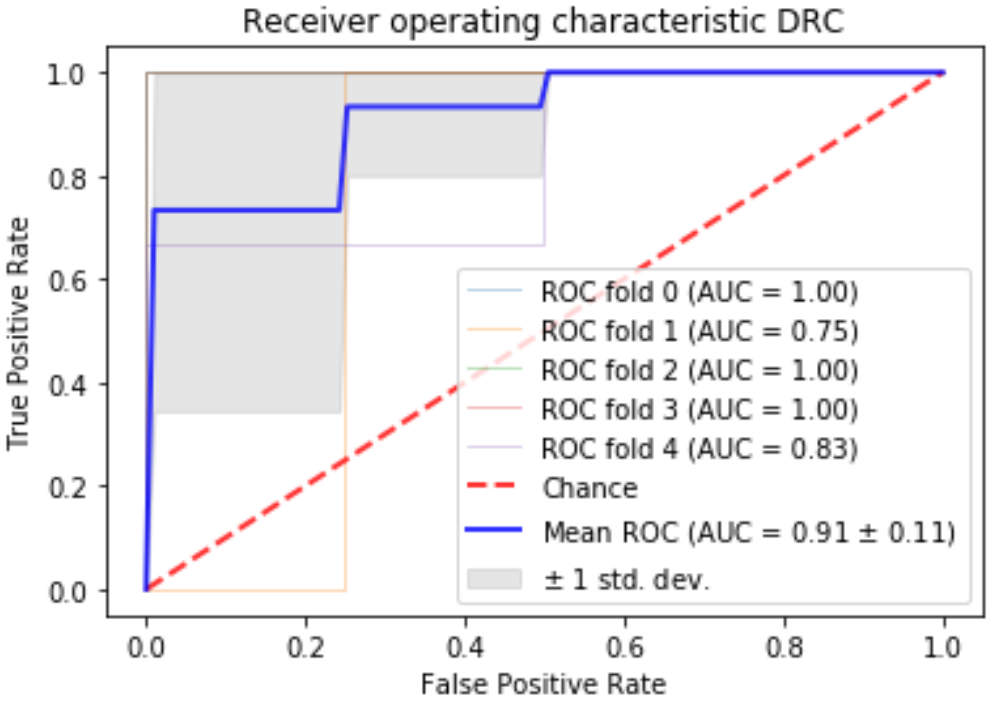


Figure 3.10c: ROC and AUC score of MIC model in DRC

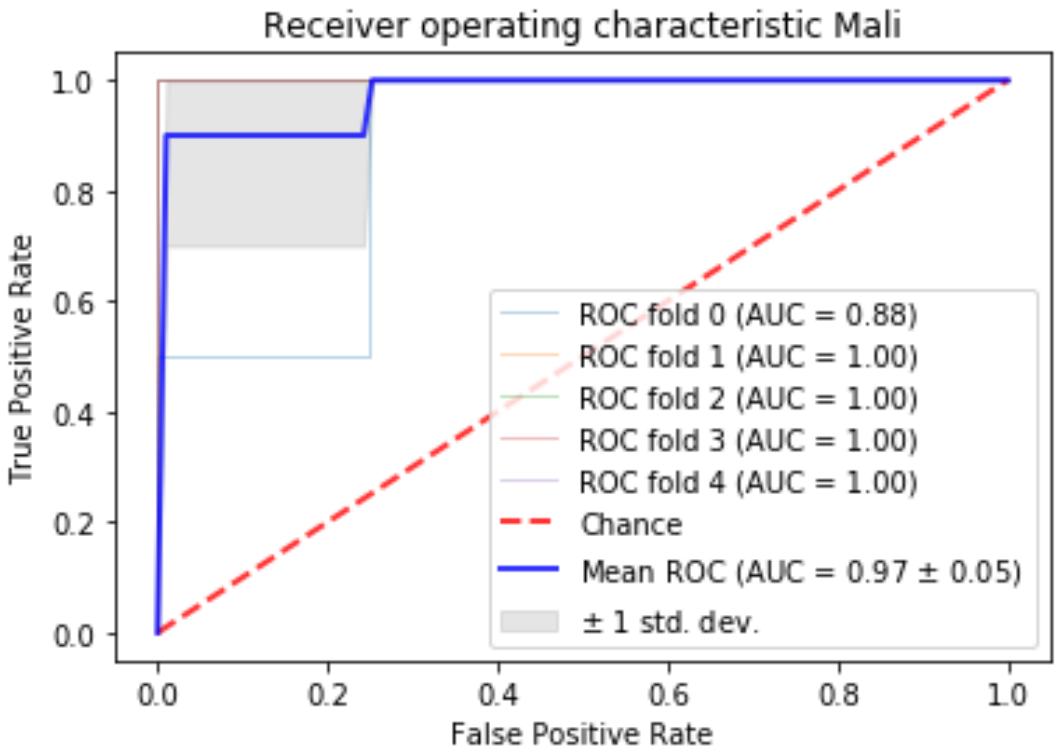


Figure 3.10d: ROC and AUC score of MIC model in Mali

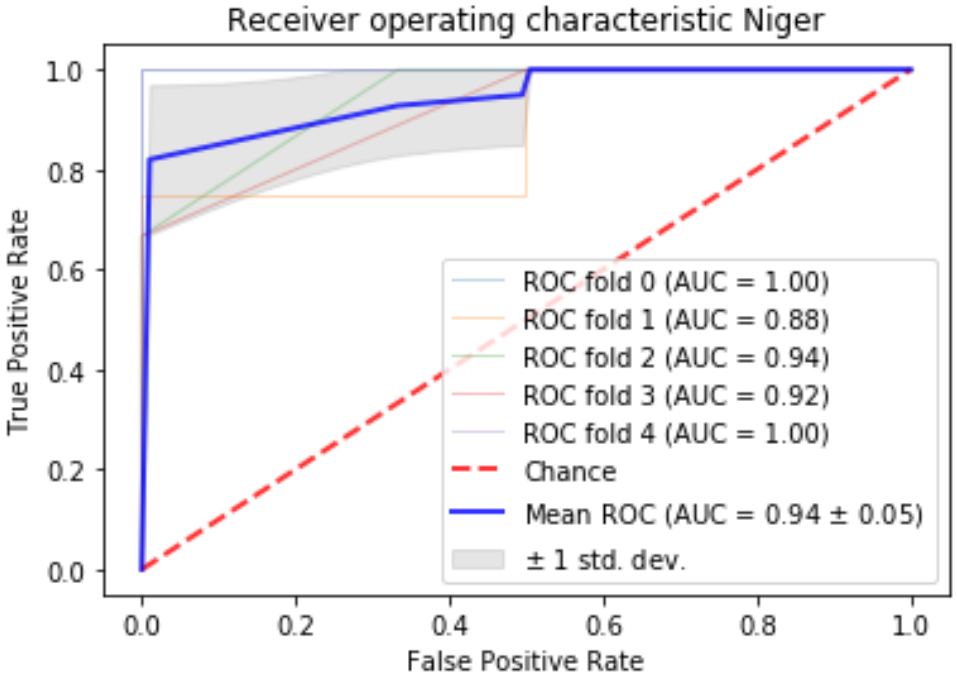


Figure 3.10e: ROC and AUC score of MIC model in Niger Republic

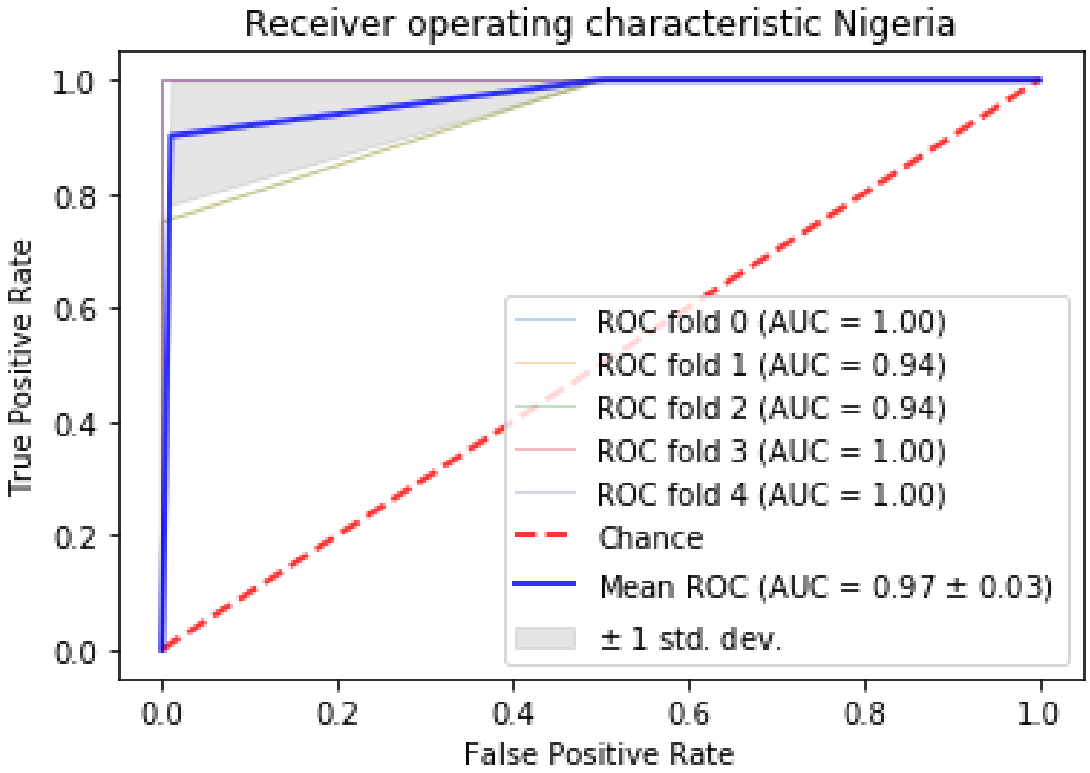


Figure 3.10f ROC and AUC score of MIC model in Nigeria

3.7.3 Results Comparisons

Table 3.5 presents a comparative analysis of the MIC model's average accuracy scores and other ML classification models such as Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression (LR) using the same dataset. Comparisons are made using a combination of the original dataset and feature engineered dataset + k-means clustering. This is to determine the effects of feature engineering and outlier data detection before classification

Table 3.5: Accuracy value of original dataset without feature engineering.

Model/ Country	XGBoost	SVM	Naïve Bayes	LR
Burkina Faso	**0.86**	0.72	0.70	0.77
Cameroon	**0.86**	0.68	0.72	0.70
DRC	**0.81**	0.72	0.71	0.76
Mali	**0.82**	0.70	0.74	0.75
Niger Republic	**0.80**	0.70	0.69	0.76
Nigeria	**0.79**	0.71	0.67	0. 72

Table 3.6: Accuracy values for dataset modeled with feature engineered dataset + K-means clustering.

Model/ Country	XGBoost	SVM	Naïve Bayes	LR
Burkina Faso	**0.97**	0.79	0.74	0.82
Cameroon	**0.94**	0.76	0.76	0.78
DRC	**0.93**	0.78	0.73	0.80
Mali	**0.98**	0.81	0.76	0.82
Niger Republic	**0.95**	0.75	0.73	0.79
Nigeria	**0.98**	0.74	0.71	0.81

Table 3.6 shows the result of the feature engineering and k-means clustering on a dataset; there is an improvement in the accuracy of all the classifiers modelled with the same dataset compared to the previous results of Table 3.5. Although the results vary across each country, the proposed XGBoost model still resulted in the highest classification accuracy score across the six countries

compared to the other classifiers. LR also seems to be promising as it gave a closer accuracy score for some datasets. Furthermore, we observed that the feature-engineered dataset and k-means played a significant role in improving the model's accuracy compared to the original dataset results. It is worthy to note that XGBoost worked best amongst the six different datasets, proving the MIC model to be an efficient model for classifying malaria incidence in the six selected countries.

3.8 Discussion

To validate and select the model that is the best fit for each data in the country, we calculated the Akaike Information Criterion (AIC) and Akaike weight of the models. AIC score is computed by maximum likelihood parameter estimation, while the Akaike weight is obtained by computing the differences in AIC value and an estimate of relative likelihood. The model with the highest Akaike weight is selected as the best classification model. We used the following number of parameters: MIC model: 7 parameters (ref. Sec.4.3), Naïve Bayes: 1 parameter (var_smoothingfloat is a parameter in NB, it has been set to default=1e-9 to signify the fraction of the principal variance of all features added to variances for calculation stability). SVM: 1 parameter (kernel=linear), and LR: 1 parameter (using a parameter: *solver*= 'liblinear'). Table 3.7 shows hypothetical results obtained from fitting four different models with the Akaike weight. It is observed that the MIC model has the highest Akaike weight across the six different datasets; this proved that the MIC model is the best-fitted model for classifying malaria incidence in the six countries of sub-Saharan Africa.

Table 3.7: Akaike weight for the four fitted models.

Country / Model	MIC Model	Naïve Bayes	SVM	LR
B. Faso	0.49689	0.00752	0.03367	0.22177
Cameroon	0.49869	0.00774	0.03369	0.22037
DRC	0.49109	0.00759	0.03339	0.22602
Mali	0.49777	0.00759	0.03395	0.22097
Niger	0.49988	0.00755	0.03327	0.21982
Nigeria	0.49976	0.00739	0.03277	0.22021

3.9 Summary and Concluding Remarks

This chapter has presented a novel Malaria Incidence Classification system, using real-world data to classify malaria incidence based on climate variability. The results suggest that the principal climate variable that influences malaria incidence varies from one country to another in different ways. However, temperature showed a strong statistical linear relationship over malaria variability across the six study sites, resulting in an average of 50% of malaria incidence. Rainfall or precipitation and surface radiation also influenced variability in malaria incidence. Feature engineering helped remove irrelevant features from the dataset, k-means clustering helped in detecting and removing outliers, and finally, optimizing the XGBoost's hyperparameters contributed to achieving high classification accuracy and higher AUC score for the MIC model. The output of this research enhances decision-making towards suitable preparation for future outbreaks of malaria. Through this system, each country's government will understand and regulate those climatic factors that frequently results in high malaria transmission, and consequently, reduce malaria incidence in their countries. It can also enhance

budget making, especially when deploying eradication mechanisms such as sensitization programs and the sharing of insecticide-treated nets or malaria medicines (Nkiruka et al., 2021) .

CHAPTER 4

Drug-Resistant Tuberculosis Classification using Logistic Regression and Frequent Pattern Growth Algorithm

4.1 Introduction

This chapter presents the frameworks used in implementing the TB classification model. Firstly, it presented a detailed description of the dataset used for the experiment. It discussed different techniques to ensuring high-quality data, such as data analysis, preprocessing, and feature selection. This is followed by association rule mining based on an FP-growth algorithm to generate frequent patterns and relevant association rules that can enhance the classification of TB and DR-TB. The feature selection and association pattern mining results are used as input data for the logistic regression model. Finally, the results of the experiments were presented, followed by the summary and conclusion.

4.2 Experimental Dataset

The TB dataset was obtained from the Specialist Hospital's records office, Yola, Adamawa State, Nigeria. Under the guidance of the ethics of the department of Tuberculosis & HIV at the Specialist Hospital Yola, Adamawa State, Nigeria. A relevant patient's record was selected based on symptomatic signs of TB infection and the result of a clinical laboratory test. The first set of data has 769 records and eight features: chills, fever, patient-gender, night-sweat, weight-loss, fatigue, cough-blood, loss-of-appetite, and one target variable that represents the patient's health condition. The second dataset contains 724 records with seven features: patient-gender, HIV-status, contact-DR, chest pain, Illicit-drugs, inadequate treatment, sputum, and a class variable

representing the patient's health condition. The first dataset is used to identify the first frequent TB symptoms, while the next dataset identifies the frequent symptoms of DR-TB and then generates the association rules that would aid in quick TB diagnosis. Generally, the symptoms were used as input variables, whereas the class is used as the output variable.

4.2.1 Data Pre-processing and Feature Selection

Dataset was transformed into electronic format in the form of 0 and 1, which indicates the presence and absence of symptoms, respectively, and preprocessed to remove noise and inconsistencies (Alexandropoulos et al., 2019). The medical relevance of each symptom for classifying DR-TB and DS-TB was analyzed through a consultant in the unit of TB at Specialist hospital, Yola. Phi coefficient is a technique for finding the relationship between the input and output variables. The Phi coefficient r ranges from $-1\ to + 1$, where -1 indicates a perfect negative relationship, 0 indicates no relationship, and $+1$ indicates that a perfect positive relationship exists between variables. Equation (27) expresses the Phi coefficient mathematically. The sample of the dataset are shown in Tables 4.1 and 4.2. The summary of the dataset after preprocessing are presented in Figures 4.1 and 4.2.

$$\phi = \sqrt{\frac{\chi^2}{n}} \qquad (27)$$

Where:

n = Total number of instances

χ^2 = chi-squared statistics

Table 4.1: Sample of phase1 dataset

Patient's Gender	Chills	Fever	Night Sweat	Fatigue	Chough Blood	Weight loss	Loss of appetite	DRTB-status
1	1	1	1	0	0	0	1	0
1	1	0	1	1	0	0	1	0
0	1	0	1	1	1	0	1	1
1	0	1	1	0	1	1	0	1
1	1	0	0	0	1	0	1	1

Table 4.2 Sample of Phase 2 dataset

Gender	HIV status	Chest pain	Sputum	Contact DR	Illicit drugs	Inadequate treatment	TB-status
0	1	1	0	0	0	1	0
1	1	1	1	1	1	0	1
1	0	1	1	0	1	0	0
0	1	1	1	1	0	1	1
1	0	0	1	0	1	1	1

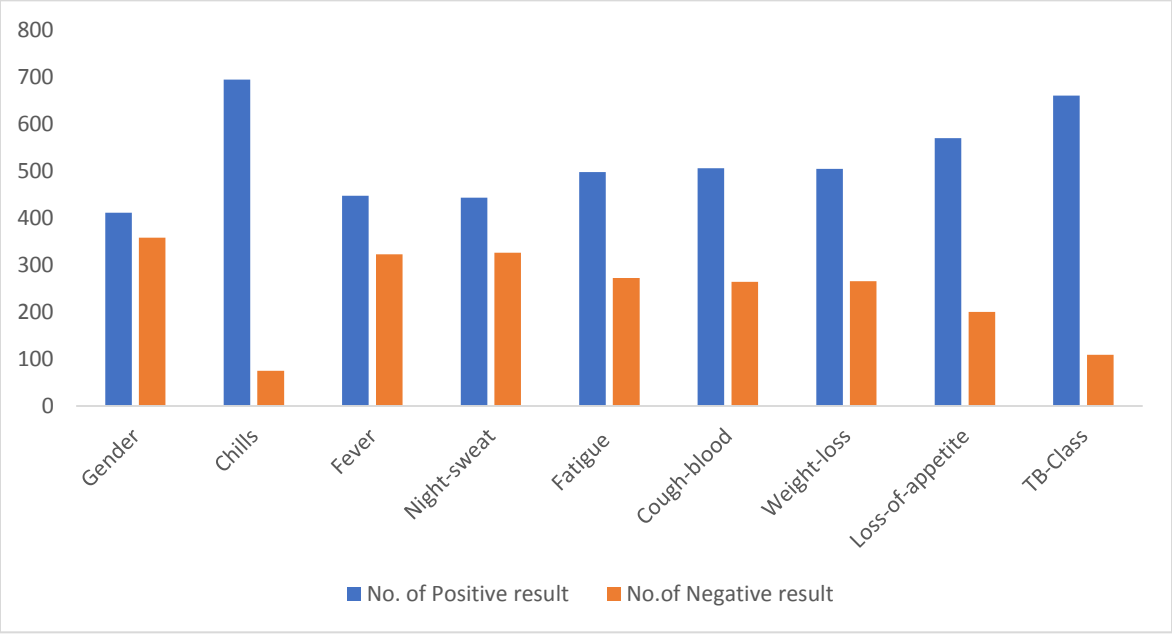


Figure 4.1a: Summary of the proposed dataset after preprocessing

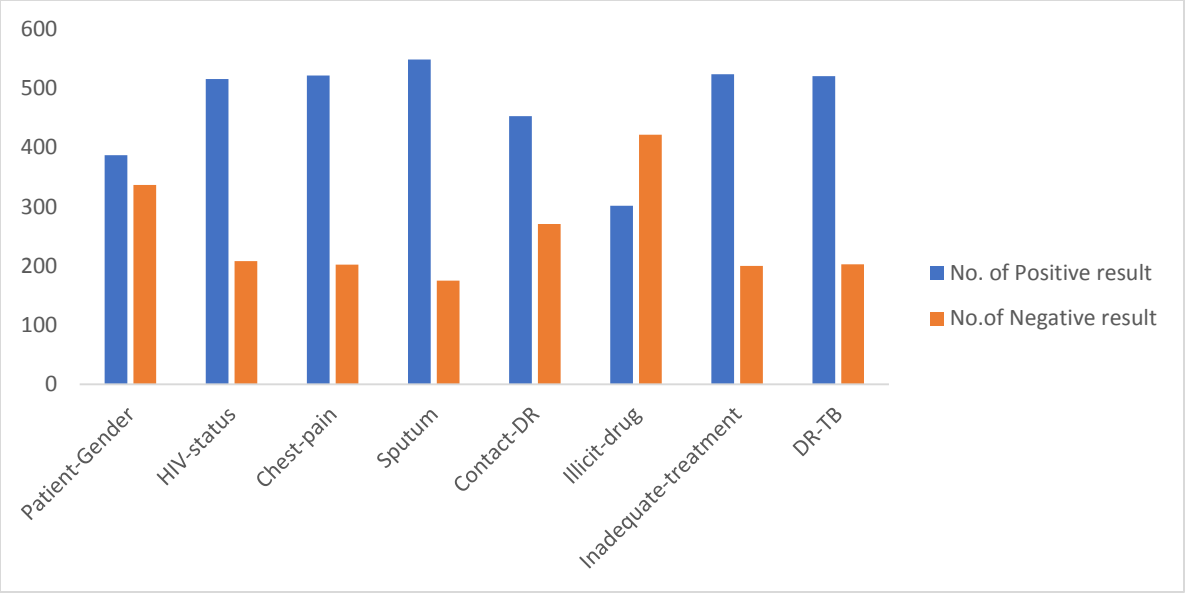


Figure 4.1b: Summary of the DR-TB Dataset after preprocessing

4.3. FP-growth Algorithm for Mining Frequent Patterns and Association Rules

FP-growth algorithm requires three major steps: FP-tree construction, FP-tree mining, and pruning of trees to eliminate non-frequent items (Han Jiawei, 2000).

4.3.1 FP-tree Construction

The FP-tree is in a tree's structure and consists of a root-node known as "null", an itemset containing prefix of subtrees, and a frequent item header-table. The Tree is built by mapping each itemset to a tree path systematically and scanning the database twice. The first scan is done to assemble frequent itemset, while the second scan leads to FP-tree construction. Table 4.2 is an instance of the transactional database with minSup = 30%. Therefore, $\langle Gender = G\rangle$, $\langle HIV\ status = H\rangle$, $\langle Chestpain = C\rangle$, $\langle Illicit\ drug\ = I_d\rangle$, $ContactDR = C_r$, $Inadequate\ treatment = I_t$, $Sputum = S$, $DRTB\_status = T_b$. The DB is scanned initially to detect frequent features with a $Frequency, F \geq minSup$, and then organize them in a compressed structure $as\ \langle H{:}\,3\rangle, \langle C_p{:}\,3\rangle, \langle S{:}\,3\rangle, \langle I{:}\,3\rangle, \langle T{:}\,3\rangle$. The number affixed to those letters

specifies their support count (frequency). Column 3 of Table 4.3 is created through iteration of the Frequent Pattern set and checking if the current item exists in the transaction; if it exists, insert it in the current ordered-Item set of the current transaction.

Table 4.3: Transaction DB

TID	Itemset (attributes)	Ordered Frequent items
1	H, C, I_t	I_t, H, C
2	H, C, Cr, I_d, T_b	I_t, T_b, H, C
3	C, S, Id,	T_b, C, S
4	H, C, S, Cr, I_t, T_b	I_t, T_b, H, C, S
5	S, I_t, T_b	I_t, T_b, S

The root-node is created in the second phase, and the DB is scanned again. Each item is mapped to its occurrence in the Tree through the node-link of the header-table. Nodes with the same name are connected via the same node-link. This process is shown in Figure 4.2, and based on the given explanations, the algorithm that creates FP-tree is developed(Han Jiawei, 2000).

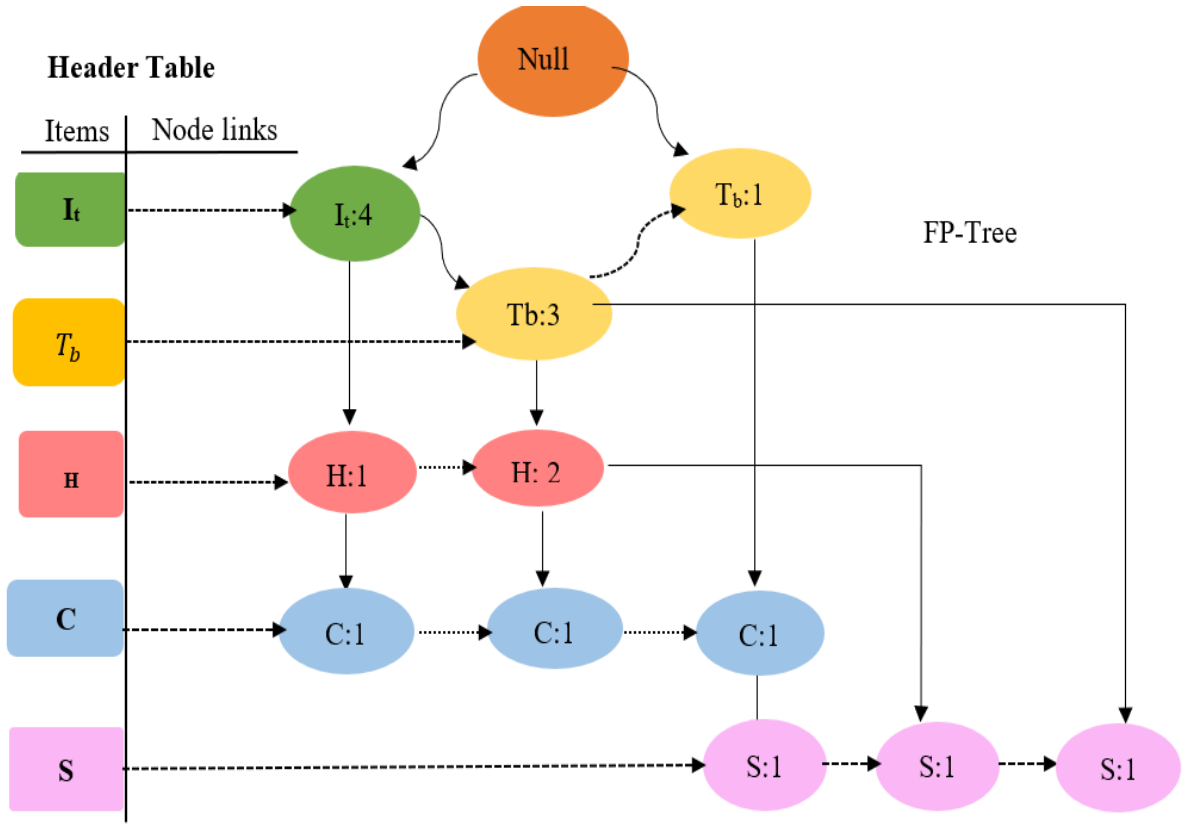


Figure 4.2: The header table and FP-tree

Algorithm (1) to design FP-tree (Han Jiawei, 2000)

Input:
The transaction-database, DB
User-defined minimum-Support, minSup
Output:
Header-table
FP-tree,
Let Transaction =Trans,
Support count = sup_count
Frequent itemset = FiS
First_element = p
Start:

i. *Scan the DB*
ii. *Get the sup_count for each item and add delete items whose sup_count < min_sup and generate FiS*
iii. *Sort FiS using sup_count based on the order of decrease*
iv. *Create header table using FiS*
v. *Create root node of FP-tree, label it as "Null"*
vi. *For all* $Trans \in DB$ *do:*
Create new_Trans, rearrange Trans into the new_Trans using FiS

If p then
Add the corresponding item into new_Trans if the item is in Trans and not in FiS
End if
if new_Trans ¬isempty then,
sort the items in new trans in their order of decrease using sup-count as a factor
insert new_Trans into FP-tree, update header table accordingly.
End if
End for

vii. *Return FP-tree, header table*

4.3.2 Mining Frequent Pattern on FP-tree

After constructing the FP-tree, items are rearranged in the order of their frequencies, seen in Column 1 of Table 4.4. The "Conditional Pattern Base**"** is obtained by collating path labels of all the paths that lead to the node of a given item in the FP-tree. The "Conditional FP-tree" is created by picking the most frequent set of items of all the paths in the Conditional Pattern Base and summing up the support-counts of all the paths in the Conditional Pattern Base.

Table 4.4: Pattern-mining through the creation of conditional-pattern bases

Items	Conditional-pattern base	Conditional FP-tree	Frequent pattern generated
S	{T_b, C:1}, {I_t, T_b, H, C: 1}, {I_t, T_b: 1}	{I_t, T_b: 3}	$\{\langle T_b, S:3\rangle, \langle I_t, S:3\rangle, \langle I_t, T_b, S:3\rangle\}$
C	{H:1}, {I_t, T_b, H:2}, {T_b:1}	{I_t, T_b, H:3}	$\{\langle I_t, C:3\rangle, \langle T_b, C:3\rangle, \langle H, C:3\rangle, \langle I_t, T_b, H, C:3\rangle\}$
H	{I_t:1}, {I_t, T_b:2}	{I_t:3}	$\{\langle I_t, H:3\rangle\}$
T_b	{I_t:3}	{1_t:3}	$\{\langle I_t, T_b\rangle\}$
I_t	∅	∅	∅

Algorithm (2), Mining frequent pattern

Input:

FP-tree in Algorithm 1
header table
predefined minimum-support, minSup
The prefix of conditional Pattern
Output:
Complete set of Frequent patterns
Start

Method FP-Growth (FP-tree, α)
{
$if\ Tree\ has\ a\ single\ path\ P, then,$
$\{For\ each\ combination, \beta\ of\ the\ nodes\ in\ the\ path\ P, do$
$Generate\ patterns, \beta \cup \alpha\ with\ support = minSup\ of\ nodes\ in\ \beta$
$Else$
$for\ each\ a_i\ in\ Tree, header\ do$
{
$Generate\ pattern\ \beta = a_i \cup a\ withsupport = a_i.support;$
Construct β *conditional patter base and* β *conditional FP-tree* $Tree_\beta;$
If $Tree_\beta \neq \emptyset$
Then FP-growth$(tree_\beta, \beta)$ } }

4.3.3 Tree Pruning

The tree is pruned by checking the support count of an item-set, as seen in Column 4 of Table 4.3. The **pattern growth** is generated by concatenating the Conditional FP-tree items to the corresponding item of Column 1 of Table 4.3. Each row of Column 4 is used to form the association rules; a rule is tagged relevant if and only if the $Conf(T_b \Rightarrow S)|minConf \geq minConf$.

4.4. TB Pattern Discovery and Association Rule Model

The minimum support and confidence were set to a minimum threshold of 60% and 80% to generate the relevant frequent patterns and the association rules. Five major steps are involved, as shown in Figure 4.3: Dataset collection, dataset preprocessing, and data transformation into electronic format. The next step involves applying the FP-Growth algorithm for frequent patterns discovery and generating the association rules that meet the threshold $Conf(T_b \Rightarrow S)|minConf \geq minConf$. Finally, the association rules generated are used for fast classification of TB.

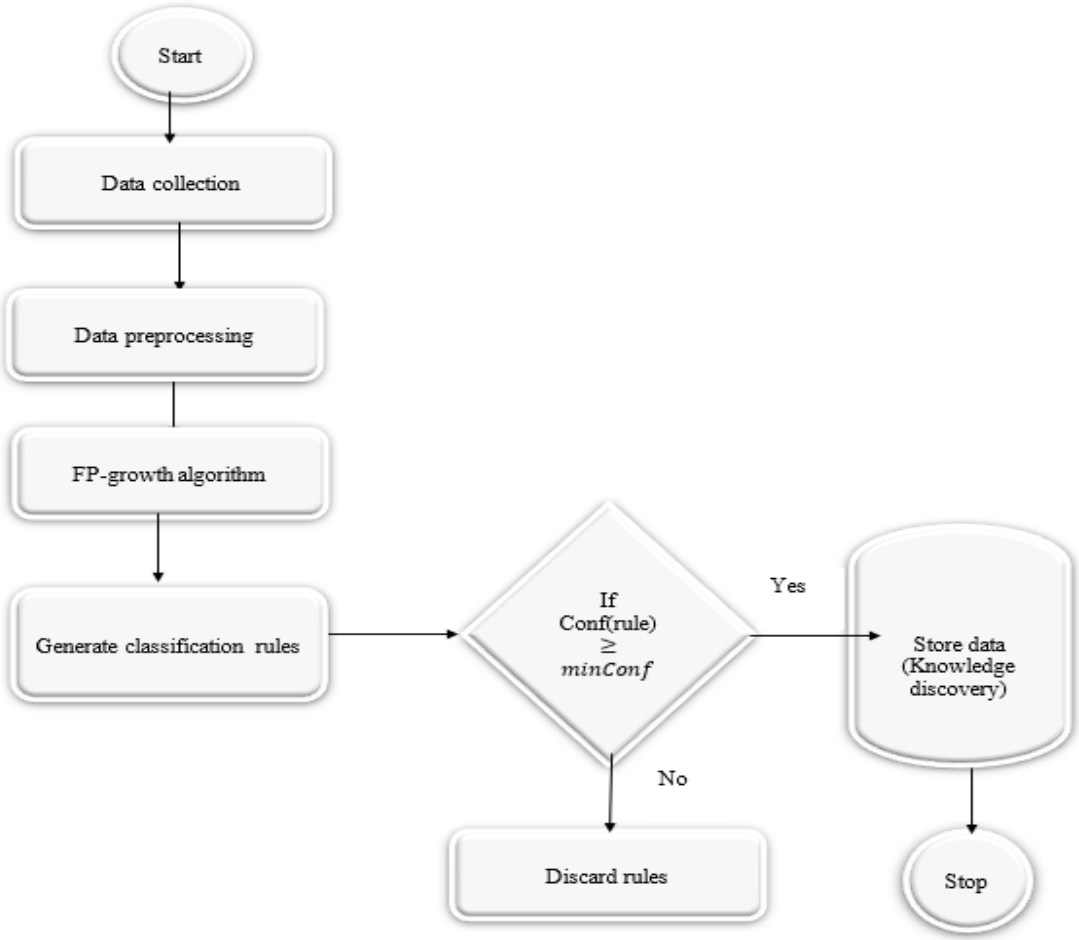


Figure 4.3: DR-TB pattern and association rule discovery model.

4.5 Logistic Regression-based Classification Model

Logistic regression (LR) is based on the concept of likelihood. LR is less prone to over-fitting and very efficient to train(Wassan et al., 2018). Feature Engineering is very vital to the performance of LR, and it performs best in the absence of collinearity in the variables. Output variables only contain binary data in the form of 1 or 0, indicating DR-TB and DS-TB. The LR model, mathematically expressed in Equation (28), is built on the linear regression model and can learn diagnostically by defining relationships between the output and input variables.

$$y = h_0(x) = \theta_x^T \quad (28)$$

To enable classification of a patient's health status into 0 and 1, a function that predicts the binary values is introduced as shown in Equation (29).

$$P(y = 1|x) = h_\theta(x) = \frac{1}{1+\exp\left(-(\theta^T x)\right)} n \equiv \sigma(\theta^T x) \qquad (29)$$

$$P(y = 0|x = 1 - P(y = 1|x) = 1 - h_\theta(x)$$

The sigmoid function is applied to Equation (29) to generate Equation (30) that maps the predicted values to probabilities and keeps the value of $\theta^T x$ within the range of [0, 1]. A threshold (θ) of 0.5 is set, which classifies a patient with a 50% or greater probability to be classified as "DR-TB" and a patient with a probability less than 50% to be classified as "DS-TB." This defines the correct decision boundary between DR-TB and DS-TB.

$$\sigma(t) = \frac{1}{(1+e^{-t})} \qquad (30)$$

4.6 Implementing Drug-resistant TB Classification Model

The DR-TB classification model involves 2 phases. The first phase classifies the supposed TB patients into either positive or negative classes; if phase one classification is positive, the next phase classifies patients into DR-TB or DS-TB. Data-mining tasks such as the data collection, data transformation, and cleaning, feature engineering to select only the relevant features, classification, interpretation, result evaluation, and knowledge discovery as shown in Figure 4.4. Modelling the probability that an input, say any (X), is a member of the default class (Y=1), written as shown in Equation (31).

$$\mathrm{P(X)} = \mathrm{P(Y = 1|X)} \qquad (31)$$

Adding e gives Equation (32).

$$\mathrm{p(X)} = \mathrm{e\text{^}(b0 + b1 * X) / (1 + e\text{^}(b0 + b1 * X))} \qquad (32)$$

add a natural logarithm (ln) to remove e from the other side as shown in Equation (33).

$$\ln(p(X)/1 - p(X)) = b0 + b1 * X \quad (33)$$

The left-hand side of Equation (21) is called the default class's odds, which is computed as a ratio of the probability of the event divided by the probability of no event in Equation (34).

$$odds = e^{\wedge}(b0 + b1 * X) \quad (34)$$

4.6.1 Choice of the Programming Tool Kit

The LRDR-TB model was efficiently implemented using Anaconda 3 that supports Python 3.6 programming language. Anaconda is open-source software that comprising many packages which support machine learning and Data Science applications.

4.6.2 Evaluation Metrics:

The evaluation metrics used in this section has been explained in details in Section 3.6.3 of Chapter three. However, the recall, f-1 score, and precision are explained here as follows:

Precision

Precision identifies the proportion of positive class that is correctly predicted. Precision is the ratio between the True Positives and all the Positive classes. The higher the precision value, the better the model. If a model has a precision value of 1, it means that when it predicts that a certain event occurred, it is correct about 100% of the time. It is computed using Equation (35)

$$Precision = \frac{True\ Positive\ (TP)}{True\ Positive(TP) + False\ Positive\ (FP)} \quad (35)$$

Recall

The recall is the measure of how a model correctly identifies True Positives. It tells how many classes have been correctly identified as positive class. It is calculated using Equation (36)

$$Recall = \frac{true\ positive}{true\ positive + false\ negative} \quad (36)$$

F1-Score

It is the harmonic mean of precision and recall. The highest possible value of an F-score is 1.0, indicating perfect precision and recall, and the lowest possible value is 0 if either the precision or the recall is zero. F1-score is calculated using Equation (37)

$$F_1 = 2 \times \frac{Precision \times recall}{precision + recall} \quad (37)$$

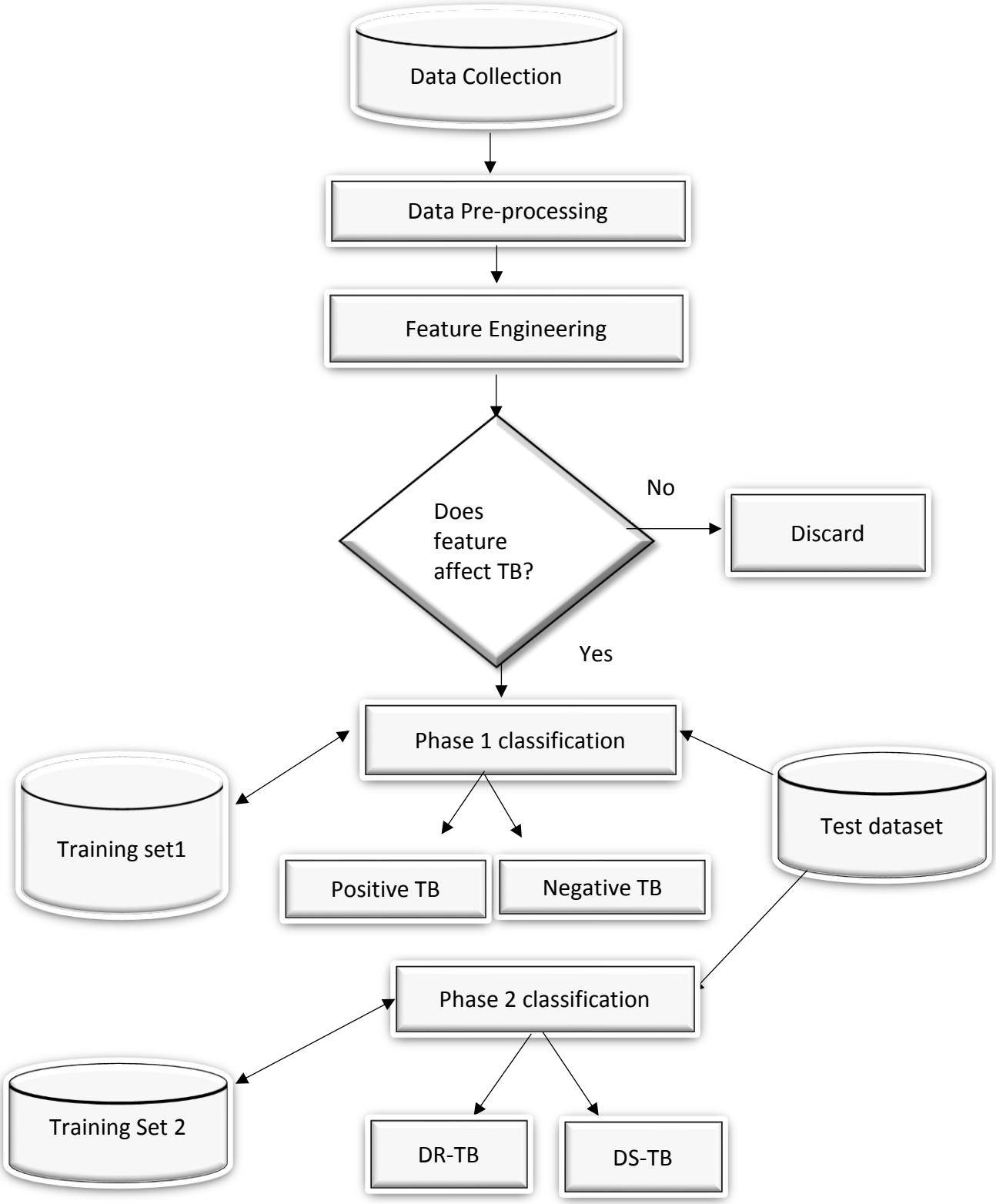


Figure 4.4. LRDR-TB classification system framework

4.7 Results

This section presents and discusses the results obtained from feature engineering, frequent pattern mining, association rules, and classification processes. It also presents the proposed LRDR-TB model classification report and further compares the proposed system's performance with other known classifiers.

4.7.1 Statistical Significance of Climate Variables

The hypothesis test at 95% confidence interval and $\alpha = 0.05$ is shown in Table 4.5 and Table 4.6. Table 4.5 column 2, r signifies the phi-coefficient correlation (ref. Equation 27) between independent variables and the class variable. Chills, cough-blood, and weight loss have a very strong positive association with the presence of TB. Similarly, Table 4.6 column 2 shows that chest pain, contact DR, illicit drugs, and inadequate treatment have a strong positive association with DR-TB as they exhibit high positive correlation coefficients. P-values are a probability that calculates the evidence contrary to the null hypothesis, which states that there is no association between predictor variables and output variables. Probabilities less than $\alpha = 0.05$ provide stronger evidence against the null hypothesis. Column 3 of Table 4.4 shows that chills, cough with blood, weight loss, and appetite loss have $p<0.05$. Column 3 of Table 4.5 shows that chest pain, contact DR, illicit drugs, and inadequate treatment have $p < 0.05$, showing that the relationship between independent and class variables is statistically significant. Column 4 is the LR variable's coefficient, representing the direction and degree of the relationship between the predictor and the output variable. Positive coefficients indicate an event's likely occurrence, whereas negative coefficients indicate that an event is less likely to occur. Using weight loss as a case study, 7.1369 shows that a change in the variable from negative TB to positive TB increases the natural log of the event's odds by 7.3. The results obtained from this process have helped

remove irrelevant features by reducing the number of variables in the original dataset, which helped understand the symptoms that are more significant to DR-TB disease.

Table 4.5: Relationship between predictors and class variable for Phase1 classification

Variable	R	P-value	Coefficient	Std err
Gender	-0.005560	0.261	-0.04904	0.012
Chills	0.083914	0.001	-1.7576	0.015
Fever	0.025384	0.353	-0.4856	0.014
Night sweat	-0.046937	0.167	-0.7087	0.014
Fatigue	0.034670	0.848	0.0821	0.012
Cough with blood	0.691896	0.000	5.0593	0.012
Weight loss	0.758209	0.000	7.1369	0.032
Loss of appetite	0.022535	0.013	0.7382	0.013

Table 4.6: Relationship between predictors and class variable for Phase2 classification

Variable	R	P-value	Coefficient	Std err
Gender	0.033965	0.361464	-1.3097	0.43800
Chest_pain	0.995294	0.001000	-15.607	0.00240
Sputum	-0.014851	0.689948	-2.4769	0.45500
Contact DR	-0.006450	0.010000	-0.1065	0.44700
Illicit_drugs	0.120501	0.001160	-0.7116	0.41100
Inadequate treatment	0.948477	0.000000	23.34	0.00024

4.7.2 Frequent Patterns and Association Rules

This section communicates the association rule mining results, and it first presents the most relevant frequent patterns of pulmonary TB and DR-TB. It then presents the association rules for classifying both types of TB. The association rules are represented using a network graph for easy understanding of how each rule are interrelated with each other.

Generating the Frequent Patterns

The first phase of the experiment involves grouping TB results into positive and negative results. The frequent itemset from the transactional database at a minimum support threshold of 60% is shown in Table 4.7 and Table 4.8. The resulting frequent itemset ranges from itemset of length one to three. This frequent item set indicates the tendencies of how these symptoms would occur. Each item set has a support count associated with it, which shows how frequently the itemset appears in the database. Row 7 of Table 4.7 shows that a patient has a 77% likelihood of suffering from TB and chills simultaneously. Similarly, row 8 shows that loss of appetite and chills has a 67% chance of occurring simultaneously. Also, row 10 shows that TB and cough blood have a 64% likelihood of occurring together. Table 4.8 presents the frequent pattern for the DR-TB at a minimum support threshold of 50%. The most frequent itemset in row 14 shows that chest pain, inadequate treatment, and DR-TB have a 70% degree of association. Also, row 12 shows that DR-TB, Chest pain, inadequate treatment, and HIV have a 70% likelihood of occurring simultaneously.

Table 4.7: Frequent pattern for pulmonary TB

S/N	Itemsets	Support
1	(Chills)	0.902471
2	(TB)	0.858257
3	(loss_of_appetite)	0.739922
4	(Chought_Blood)	0.656697
5	(Fatique)	0.646294
6	(weight_loss)	0.655397
7	(TB, Chills)	0.767230
8	(loss_of_appetite, Chills)	0.665800
9	(loss_of_appetite, TB)	0.638492
10	(TB, Chought_Blood)	0.644993
11	(TB, weight_loss)	0.654096

Table 4.8. Frequent patterns for DR-TB datasets

S/N	Itemsets	Support
1	(DR_TB)	0.719613
2	(HIV_status)	0.712707
3	(Inadequate _treatment)	0.723757
4	(Chest_ pain)	0.720994
5	(Sputum)	0.758287
6	(DR_TB, Inadequate _treatment)	0.711326
7	(DR_TB, Chest_ pain)	0.708564
8	(DR_TB, Chest_ pain, Inadequate _treatment)	0.708564
9	(DR_TB, HIV_status)	0.712707
10	(Inadequate _treatment, HIV_status)	0.704420
11	(Chest_ pain, HIV_status)	0.701657
12	(DR_TB, Inadequate _treatment, HIV_status)	0.704420
13	(DR_TB, Chest_ pain, HIV_status)	0.701657
14	(Chest_ pain, Inadequate _treatment, HIV_status)	0.701657
15	(DR_TB, Chest_ pain, Inadequate _treatment, HIV_status)	0.701657
16	(Chest_ pain, Inadequate _treatment)	0.720994

Generating Association Rules

The frequent Itemsets generated were then used to generate the association rules in Tables 4.9 and 4.10 at the minimum confidence threshold of 80% and 90%. Only the rules containing TB and DR-TB were selected to generate only relevant rules and reduce redundancy. Table 4.9 shows the confidence and lift values associated with each rule for the first datasets. A lift value is a measure of the importance of a rule $X \Rightarrow Y$; a lift value that is greater than 1 indicates a high association between Y and X and less association if otherwise. Using row 7 as an instance shows that $(\{'\text{loss of appetite}', '\text{weight loss}'\} = \text{True}) \Rightarrow (\{'TB'\})$. The loss of appetite and weight loss are the antecedents, and TB is consequent with a confidence value of 0.99774. The rule implies that about 99% of people who suffer from TB also have lost weight and lost appetite. In the same way, row 2 shows the rule $(\{chills', 'cough\ blood'\} = True) \Rightarrow (\{TB\})$. Which implies that 98% of people suffering from TB disease has chills and cough blood concurrently. Another important criterion for selecting these rules is the lift values of more than 1 for each rule in Tables 4.9 and Table 4.10. This implies that these symptoms are frequently associated with TB occurrence, and any patient with such symptoms should be diagnosed with 'TB-positive. Figure 4.5 shows a network graph[1] that summarized the relationship between the associated rules. It is a directed graph that shows how the association rules are interrelated. The graph shows that the item sets $(chills, cough\ blood,)$, $(chills, weight\ loss)$, $(loss\ of\ appetite, weight\ loss)$, $(weight\ loss)$, and $(cough\ blood)$ of TB. $(Fatique)$, $(TB, Fatique)$Are the antecedents of chills.

[1] Network graphs "show interconnections between a set of entities" where entities are nodes and the connections between them are represented through links or edges

Table 4.9. The seven most relevant association rules for first-case TB

S/N	Antecedents	Consequents	Confidence	Lift
1	({'Cough blood'} = True)	({'TB'} = True)	0.982178	1.144386
2	({'Chills', 'Cough blood} = True)	({'TB'} = True)	0.980044	1.141900
3	({'fatigue'} = True)	({'Chills'} = True)	0.909457	1.007741
4	({'TB', Fatique} = True)	({Chills'} = True)	0.900232	0.997519
5	({'weight loss'} = True)	({'TB'} = True)	0.998016	1.162840
6	({'chills', 'weight loss'} = True)	({'TB'} =True)	1.000000	1.165152
7	({'loss of appetite', 'weight loss'})	({'TB'} = True	0.997442	1.162217

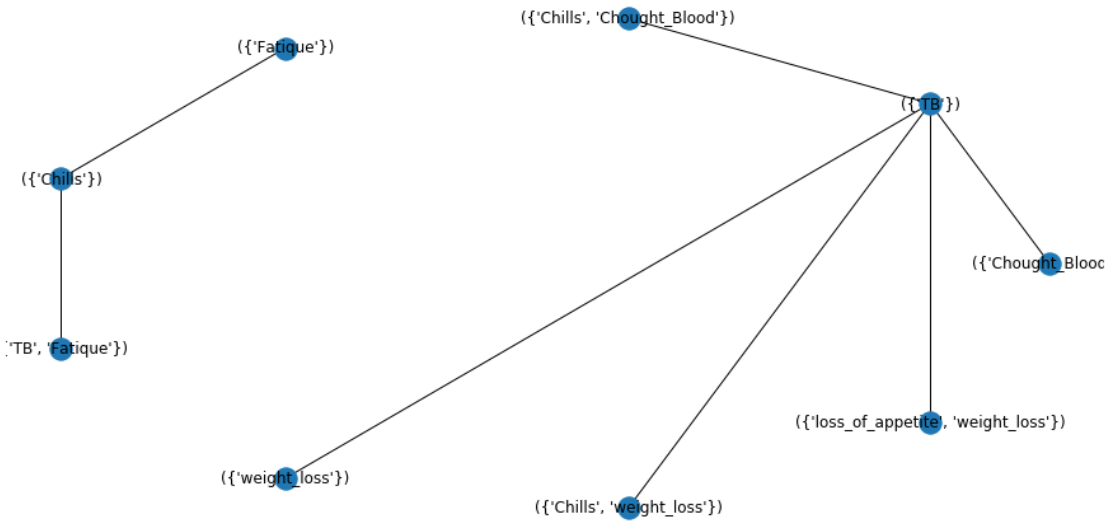


Figure 4.5. Network graph for the first association rule

The frequent pattern in Table 4.8 was used to generate the association rules in Table 4.10 with a minimum confidence threshold of 90% for DR-TB. The table contains the antecedents and consequents of the rules in the form of $X \Rightarrow Y$ together with their lift values. For example, row 4

$(\{DR\_TB\}) \Rightarrow (\{'\text{Inadequate \_treatment}', '\text{Chest\_ pain}', '\text{HIV\_status}' = \text{True})$ shows a confidence value of 0.979 and a lift value of 1.3896. This rule can be clearly stated in a medical term that 97% of patients diagnosed with DR-TB disease also suffer chest pain, HIV disease, which could arise from inadequate treatments. This is supported by the lift value showing high associations between DR-TB (X) and inadequate treatment, chest pain, and HIV (Y). Generally, all the association rules in Table 4.10 contains a lift values greater than one, this shows that both the antecedents and consequents of the rules are highly associated. The relevance of these association rules provides room for quick diagnosis of pulmonary and DR-TB for first hand medication prior to the arrival clinical results. Figure 4.6 provides a network graph for the visualization of the association rules for DR-TB, it is seen that there was a heavy concentration of lines originating from mostly the frequent symptom-sets such as $(chestpain',$ $'inadequate\ treatment',)(chestpain', 'HIV\_status')(inadequate\ treatment, HIV\_status),$ and pointing towards DR-TB, showing their strong connections. In medical terms, the purpose of generating association rule is to provide systematic approaches through which users or physician can figure out how to detect the presence of some sets of symptoms or diseases, given the presence of other symptoms or disease in patient's health database.

Table 4.10 The first ten most relevant association rules for DR-TB

S/N	Antecedents	Consequents	Confidence	Lift
1	({'DR-TB' = True})	({'Inadequate _treatment ', 'Chest_ pain'})	0.984645	1.365676
2	({'Inadequate _treatment '} = True)	({'DR-TB', 'Chest_ pain'} = True)	0.979008	1.381679
3	({'DR-TB', 'Inadequate _treatment '} = True)	({'HIV_status'} = True)	0.98837	1.365613
4	({'DR-TB'}= True)	({'Inadequate _treatment ', 'Chest_ pain', 'HIV_status'= True)	0.975048	1.389635
5	({'DR-TB' = True})	({'Inadequate _treatment ', 'Chest_ pain', 'HIV_status'}= True)	0.97318	1.381534
6	({'HIV_status'}=True)	({'DR-TB', 'Chest_ pain', 'Inadequate _treatment '}= True)	0.984496	1.389425
7	({'Chest-pain', 'Inadequate-treatment'} =TRUE)	({DR-TB}=TRUE)	0.97318	1.381534
8	({'Chest_pain',} = TRUE)	({'Inadequate-treatment', 'DR-TB'}= TRUE)	0.97318	1.381534
9	({'DR_TB'} = True)	({'Chest_ pain', 'Inadequate _treatment '}= True)	0.984645	1.365676
10	({'DR_TB', 'Chest_ pain'} = True)	({'Inadequate _treatment '} = True)	1	1.381679

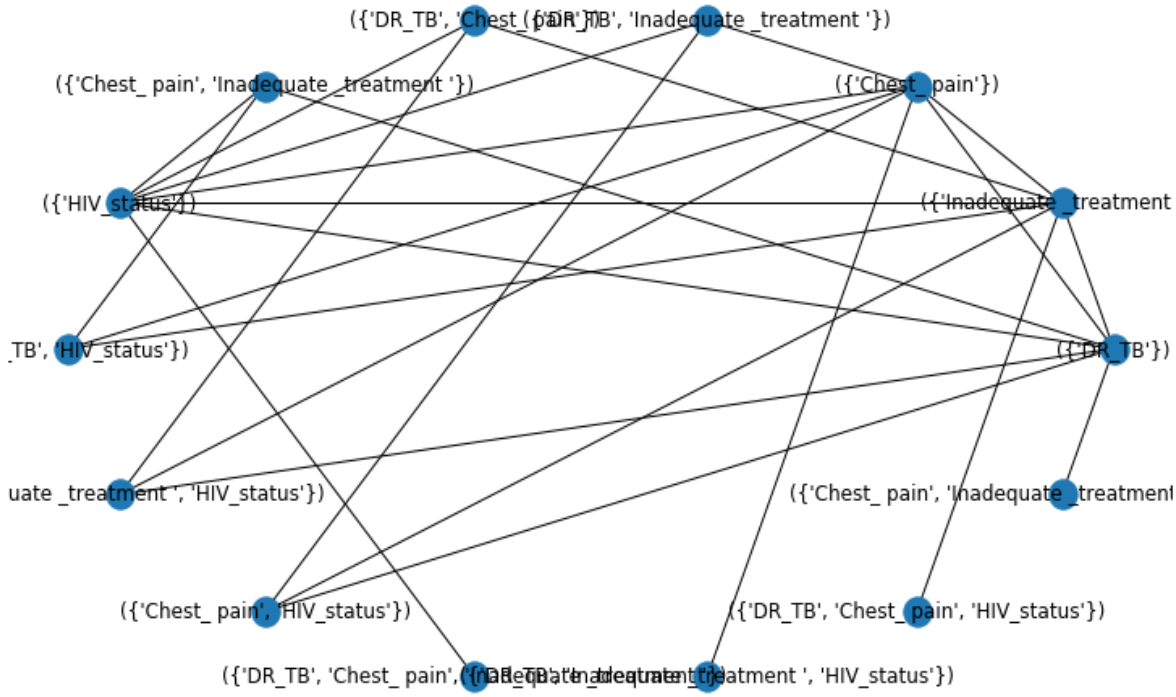


Figure 4.6. Network graph for the Association Rule

4.7.3 Classification Report

The proposed LRDR-TB classification model is evaluated by performing an offline experiment that simulates the entire system's functionality. Dataset was divided on a ratio of 70:30, where 70% records comprise the training set and 30% is the test set. The cross-validation (CV) technique minimizes biases in data as well as overcoming overfitting that may arise because of data volume (Wong & Yeh, 2020). The training set is divided into a subset of 10-folds to form a training set, and each subset is used as a test set to the remaining nine subsets; the test set was used to evaluate the performance of the LRDR-TB classification model while considering its AUC score as shown in Figures 4.7 and Figure 4.8. The CV process is done repeatedly ten times, and a single estimate is computed by getting the average of the ten results.

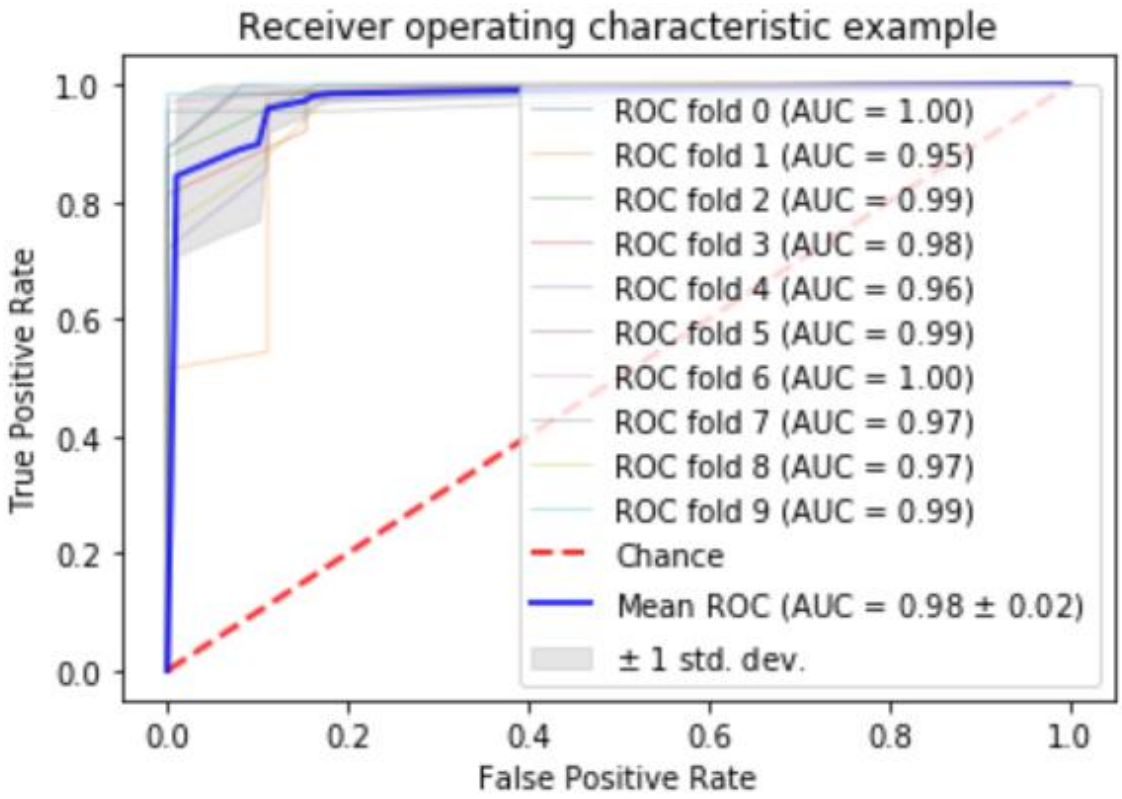


Figure 4.7. AUC score and ROC plot for Phase 1 classification

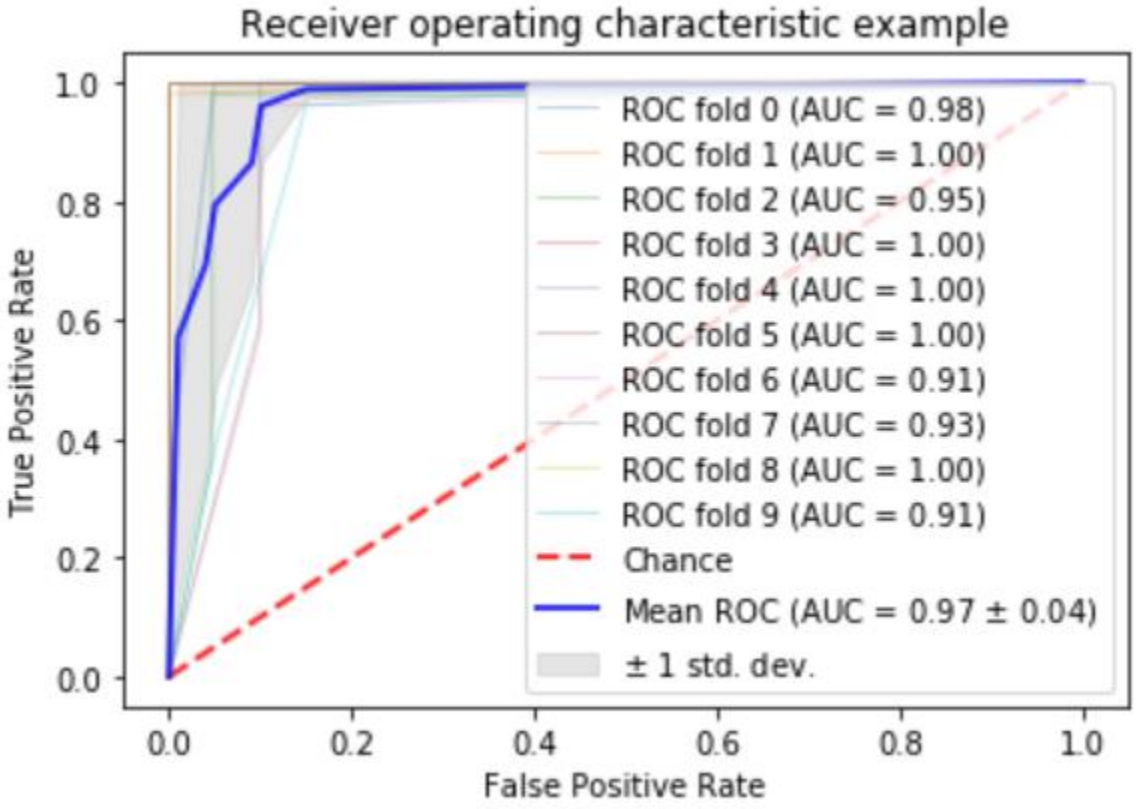


Figure 4.8.AUC score and ROC plot for Phase 2 classification

Figure 4.7 and Figure 4.8 show the ROC plots and the AUC scores of the 10-fold CV. ROC displays a quick representation of model performance across all possible thresholds. A single estimation for AUC is 0.97 and 0.98 for Phase1 and Phase2 classification, respectively. This showed that the proposed model has correctly classified TB into positive, negative, DR-TB, and DS-TB classes. The classification reports are shown in Figure 4.9 and 4.10. In Figure 4.9, precision shows that 98% of the correctly predicted class belongs to the positive class, recall shows that 99% of the positive classes were correctly identified, and f1-score, which combines idea about the two metrics; precision and recall, shows that 98% of the positive prediction is correct. The average precision/recall/f1-score is known as **the Macro average**. Phase2 classification, the classification report is shown in Figure 4.7, the precision result shows that 98% were predicted correctly into the positive class, recall tells that 98% of the positive classes

were correctly identified, and f1-score also shows that 98% positive class is correctly predicted. This shows that the LRDR-TB model has correctly classified data into positive and negative classes and further into DR-TB and DS-TB.

	precision	recall	f1-score	support
0	0.97	0.95	0.96	65
1	0.98	0.99	0.98	153
accuracy			0.98	218
macro avg	0.97	0.97	0.97	218
weighted avg	0.98	0.98	0.98	218

Figure 4.9a. Classification report of Phase1 classification

	precision	recall	f1-score	support
0	0.91	0.89	0.90	35
1	0.98	0.98	0.98	196
accuracy			0.97	231
macro avg	0.95	0.94	0.94	231
weighted avg	0.97	0.97	0.97	231

Figure 4.9b. Classification report of Phase2 classification

4.7.4 Comparing Other Classifiers

The performance results of LRDR-TB were compared with the other three classification models using different variations such as original TB datasets and the feature-selected dataset from the association rules. Table 4.11 presents other classifiers and their accuracy scores. From this table, it is observed that selecting only relevant data for prediction generally improved the accuracy of almost all the classifiers but extensively with LR, which performs well with only the relevant variables. Applying eXtreme Gradient Boosting (XGBoost) on the original dataset seems to be

promising, yet the result shown in Table 4.11 shows a reduction in accuracy from 96% with the original dataset to 90% with feature engineered datasets. LRDR-TB model works best when only the relevant features are critically selected, and overall, the LRDR-TB model, compared to the three other classifiers, outperformed the other three classifiers through the values in Table 4.11, proving LR to be the best model that fits the TB dataset for the DR-TB classification.

Table 4.11: Comparing other classifiers

Model	Phase 1 Classification On original dataset	Phase 1 working with Classification rule dataset with feature selection	Phase 2 Classification On original dataset	Phase 2 Classification With feature selection
LRDR-TB	0.95	0.97	0.94	0.98
SVM	0.89	0.93	0.92	0.93
NB	0.92	0.94	0.95	0.95
XGBoost	0.96	0.90	0.94	0.92

Akaike information criteria (AIC) is a technique used for selecting the best model during comparison (Wagenmakers & Farrell, 2004). Table 4.12 shows the hypothetical results, obtained from fitting four different models with the AIC scores and Akaike weights for comparing the classifiers using their accuracy scores. The first column of Table 4.12 shows the number of free parameters used for each model, and the second column shows the AIC value, whereas the third column shows the Akaike weight. AIC value is obtained by maximum likelihood parameter estimation. From an inspection of the AIC values, it is obvious that the LRDR-TB model is the preferred model since it has the lowest AIC value and lowest Akaike weight amongst the four candidate models.

Table 4.12 AIC score and Akaike weight

	Phase 1 Classification			Phase 2 Classification		
Model	**Parameter**	**AIC value**	**Akaike weight**	**Parameter**	**AIC value**	**Akaike weight**
LRDR-TB model	3	2.673	0.993508	2	18.279	0.997591
SVM	1	12.799	0.006285	1	6.052	0.000201
NB	2	19.629	0.000207	1	23.069	0.002207
XG_boost	4	8.4237	0.038369	4	12.654	0.062094

4.8. Discussion

This chapter focused on the application of machine learning techniques in modelling an efficient system that can assist physicians in the quick diagnosis of TB diseases. The aim is to model a system that correctly classifies TB into positive and negative classes and further classifies them into DR-TB with minimum error. The results shown in Figure 4.4 present the LRDR-TB model's classification accuracy scores, which resulted in 97% and 98% accuracy for the two phases of classifications. However, this research did not rely only on the classification accuracy, other metrics such as Precision, recall, F1-score and the AUC. The values obtained from these metrics show that the LRDR-TB model performed better when compared to other classifiers. It is also observed in phase 1 classification that chills, cough with blood, weight loss, and loss of appetite have a strong relationship with TB, and similarly, chest pain, contact DR, illicit drugs, and inadequate treatment have a strong association with DR-TB in phase 2 classification. The result of the statistical analysis on both datasets has significantly improved the prediction accuracy of the LRDR-TB model.

4.9 Summary and Concluding Remark

This research has presented a prototype of a non-invasive diagnostic support system for pulmonary TB and DR-TB. The gold standard for clinical diagnosis of DR-TB is expensive and often requires an expert's presence, which may not be readily available, especially to the people living in the remote areas of developing countries. This work has presented decision support that is cost-effective which does not require clinical expert know-how. It explored the efficiency of the FP-Growth algorithm and generates association rules for TB disease classification at minimum confidence of 80%. The association rules show that cough-blood, weight-loss, loss-of-appetite, and chills are the most frequent symptoms related to pulmonary TB. Similarly, HIV, Sputum, chest pain, illicit-drugs, and inadequate-treatment are the most frequent symptoms associated with DR-TB. This technique is best used for decision supports by physicians during TB diagnosis. Finally, the performance of the proposed LRDR-TB model was compared with other machine learning algorithms, and the results showed that the LRDR-TB model performed better than others. Both models showed high reliability in predicting TB diagnosis by both models, and the results suggest that LR performance is good for medical diagnosis of Tuberculosis (TB) and DR-TB. Health specialists can use this model for easy and fast diagnosis of DR-TB even in the absence of a medical expert(Bridget et al., 2021).

CHAPTER 5

Conclusion and Suggestion for Future Work

5.1 Conclusion

One of the objectives of this research work includes developing a model to predict malaria incidence in an unstable transmission area by studying the association between environmental variables and disease dynamics. The study was carried out in the six malaria-endemic countries such as Nigeria, Niger Republic, Cameroon, DRC, and Burkina Faso using annual variability data of climatic and malaria incidence data. Malaria incidences, cases, transmission and mortality in Africa is unstable and fluctuates. Although there seems to be a little decrease in malaria incidence and mortality, it is still one of the major causes of mortality and child mortality in Africa. While many factors play a role in malaria transmission, climate variability is one of the driving factors that influence the fluctuations in malaria incidence in Africa. Temperature, precipitation and relative humidity are the major climatic factors whose instability cause some irregular pattern in malaria transmission. Previous work has also evaluated the uncertainty in the deterministic model forecasts. This research has shown that a reliable assessment of a long-term relationship between climate variability and malaria incidence is important to reduce this incessant malaria incidence.

Tuberculosis (TB) is one of the leading ten causes of death worldwide. Drug-resistant TB is an urgent public health concern in infectious disease, and it threatens the treatment and control of tuberculosis. One critical challenge in tackling the global TB epidemic is a timely diagnosis and correct treatment, and most of the TB deaths can be prevented if it is detected at an early stage. Correct and rapid determination of Mycobacterium tuberculosis (MTB) resistance against available tuberculosis (TB) drugs is essential for controlling and managing TB. Advanced

technology like AI tools would make the TB diagnosis process comparatively non-invasive and help obtain rapid diagnostic results. This research demonstrates the application of machine learning as a reliable approach to drug resistance prediction, identifies features associated with TB and DR-TB, and reviews their predictive ability towards TB diagnosis to assist clinical decision making.

This research has successfully applied three machine learning algorithms, such as XGBoost, Logistic regression, and FP-Growth algorithm, which have been considered for predicting malaria incidence and diagnosing TB disease. This research's novelties include a malaria incidence classification model that predicts early annual malaria incidence in sub-Saharan Africa's six endemic countries. This model has successfully shown the statistical significance of climate variables to malaria incidence at "$\alpha = 0.05$" and has shown that these climatic factors vary from one country to another. It is shown that ambient temperature at 28^{o}C is the most common climate variable that affects malaria incidence across the six countries, and if there is an increase in temperature, malaria incidence tends to increase and decrease when the climate factors are decreasing. The statistical results also showed that a high volume of precipitation and surface radiation contribute to increased malaria incidence. The XGBoost classification algorithm successfully classified the real-world data. K-fold CV technique (at k=5) was used to test the model's performance which gave AUC scores such as Mali:0.97, DRC: 0.91, Niger: 0.94, Burkina Faso:0.92, Nigeria: 0.97, Cameroon: 0.94, through hyperparameter optimization. A comparative analysis with other models showed that the MIC model performed well.

Recent application of machine learning algorithms in Tuberculosis diagnosis has scarcely considered DR-TB. This work's novelty lies in applying the logistic regression model for diagnosing DR-TB using a two-way approach. First, the knowledge discovery approach using

the FP-growth algorithm. It is an efficient ARM technique that identifies recurring relationships, co-occurrences of disease symptoms and generates important diagnostic rules for pulmonary TB and DR-TB. These were achieved through systematic approaches such as Dataset collection from the Unit of TB and HIV at the Specialist Hospital, Yola. These are data set extraction from TB patient's database, data transformation, and knowledge extraction, which provides vital information supporting the medical diagnosis of TB. Minimum support and confidence at 60% and 80% guaranteed the relevance of the association rules, and in summary, it is shown that weight loss, chills, cough blood, loss of appetite, and fatigue are strongly associated with pulmonary TB. In contrast, illicit drugs, chills, weight loss, sputum, and inadequate treatment are frequent symptoms associated with DR-TB. The association rules generated from this step was used as input variables for the classification phase, which involves diagnosing patients into positive or negative class and further diagnosing patients into DR-TB or DS-TB classes. The proposed models were evaluated using some performance metrics such as classification accuracy, precision, recall, F1-score, and AUC. The classification accuracy has 0.98 and 0.97 for phase1 and phase2 classifications. The precision and F1-score have the same score of 0.98 for both phases of classification. Recall has a value of 0.99 and 0.98 for phase1 and phase2 classifications, and finally, the AUC values for phase1 and phase2 are 0.98 and 0.97, respectively.

In summary, these proposed systems can assist in diagnosing patients into DR-TB or DS-TB using some set of predefined symptoms. These symptoms are weight loss, chills, cough blood, loss-of-appetite for pulmonary TB diagnosis and Chest-pain, inadequate treatment, HIV, Illicit-drugs and Sputum for DR-TB classification. The LR classifier's main advantages remain in its capability to show the degree of influence of predictors to target variables and its ability to work

effectively with fewer attributes. when applied appropriately to these diseases, will assist the physicians and policymakers to make a pre-informed decision and thus mitigate the disease spread towards reaching the goal of WHO for "End Malaria and TB strategy". However, the best true way to examine the expediency of the proposed systems is the extent to which they become routinely useful to both physicians and decision makers.

5.2 Suggestion for Future Work

The main obstructions to the implementation of MIC model are related to wider health-system issues such as shortage of good quality malaria time series dataset, the availability of only the annual dataset on malaria incidence is insufficient for exploring the efficiency of the proposed MIC model. Therefore, to improve the proposed MIC model's capability, obtaining a time-series dataset of malaria incidence probably with similar resolutions to the climate observations can seasonally stratify important information about malaria transmission and seasonality; this will enhance the real-time prediction of the system. Furthermore, obtaining a larger amount of dataset will strengthen the proposed MIC system's capability and would aid in extending MIC into a real-world application that can run on mobile smart systems.

The larger the amounts of the dataset, the better it becomes in precision and reducing bias and overfitting. The LRDR-TB model would be extended by training and testing the classification models with many datasets. Also, this research will be extended by developing other sophisticated models for the diagnosis of other diseases.

Most hospitals and health centres, especially in African countries, do not have the capacity and the equipment for long-term electronic storage of patients' health records. Therefore, proper utilization of big data and cloud computing that enhances data capturing, sharing, storing,

analysis, and managing the patient's data should be adopted in the health systems to improve ARM applications' scaling for better health data-analysis and also improve research and development in the health sector.

REFERENCES

A., M., S., K., & D, G. (2021). Analysis of Tuberculosis Disease Using Association Rule Mining. In *Advances in Intelligent Systems and Computing*. Springer Singapore. https://doi.org/https://doi.org/10.1007/978-981-15-3514-7_74

Abebe, R., Hill, S., Vaughan, J. W., Small, P. M., & Schwartz, H. A. (2019). Using search queries to understand health information needs in africa. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*, 3–14.

Abeku, T. A. (2007). Response to malaria epidemics in Africa. *Emerging Infectious Diseases*, *13*(5), 681–686. https://doi.org/10.3201/eid1305.061333

Abeku, T. A., De Vlas, S. J., Borsboom, G., Teklehaimanot, A., Kebede, A., Olana, D., Van Oortmarssen, G. J., & Habbema, J. D. F. (2002). Forecasting malaria incidence from historical morbidity patterns in epidemic-prone areas of Ethiopia: A simple seasonal adjustment method performs best. *Tropical Medicine and International Health*, *7*(10), 851–857. https://doi.org/10.1046/j.1365-3156.2002.00924.x

Adeola, A. M., Botai, J. O., Olwoch, J. M., Rautenbach, H. C. J. d. W., Adisa, O. M., de Jager, C., Botai, C. M., & Aaron, M. (2019). Predicting malaria cases using remotely sensed environmental variables in Nkomazi, South Africa. *Geospatial Health*, *14*(1). https://doi.org/10.4081/gh.2019.676

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association in Large Databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data - SIGMOD '93*, 207–216.

Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). *Interpretable Machine Learning in Healthcare*. 559–560. https://doi.org/10.1145/3233547.3233667

Aikins, M. K., Pickering, H., Alonso, P. L., d'Alessandro, U., Lindsay, S. W., Todd, J., & Greenwood, B. M. (1993). A malaria control trial using insecticide-treated bed nets and targeted chemoprophylaxis in a rural area of The Gambia, West Africa: 4. Perceptions of the causes of malaria and of its treatment and prevention in the study area. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *87*, 25–30.

Akinbobola, A., & Omotosho, J. B. (2013). Predicting Malaria occurrence in Southwest and North central Nigeria using Meteorological parameters. *International Journal of*

Biometeorology, *57*(5), 721–728. https://doi.org/10.1007/s00484-012-0599-6

Akinsanola, A. A., & Ogunjobi, K. O. (2014). Analysis of rainfall and temperature variability over Nigeria. *Global Journal of Human Social Sciences: Geography & Environmental GeoSciences*, *14*(3), 1–18.

Alexandropoulos, S. A. N., Kotsiantis, S. B., & Vrahatis, M. N. (2019). Data preprocessing in predictive data mining. In *Knowledge Engineering Review* (Vol. 34, Issue January). https://doi.org/10.1017/S026988891800036X

Alim, M., Ye, G.-H., Guan, P., Huang, D.-S., Zhou, B.-S., & Wu, W. (2020). Comparison of ARIMA model and XGBoost model for prediction of human brucellosis in mainland China: a time-series study. *BMJ Open*, *10*(12). https://doi.org/10.1136/bmjopen-2020-039676

Alonso, P., & Noor, A. M. (2017). The global fight against malaria is at crossroads. *The Lancet*, *390*(10112), 2532–2534.

Altaf, W., Shahbaz, M., & Guergachi, A. (2017). *Applications of association rule mining in health informatics : a survey*. 313–340. https://doi.org/10.1007/s10462-016-9483-9

Anderson, R. M., & May, R. M. (1992). *Infectious diseases of humans: dynamics and control*. Oxford university press.

Anokye, R., Acheampong, E., Owusu, I., & Isaac Obeng, E. (2018). Time series analysis of malaria in Kumasi: Using ARIMA models to forecast future incidence. *Cogent Social Sciences*, *4*(1). https://doi.org/10.1080/23311886.2018.1461544

Antonio-Nkondjio, C., Ndo, C., Njiokou, F., Bigoga, J. D., Awono-Ambene, P., Etang, J., Ekobo, A. S., & Wondji, C. S. (2019). Review of malaria situation in Cameroon: technical viewpoint on challenges and prospects for disease elimination. *Parasites & Vectors*, *12*(1), 501. https://doi.org/10.1186/s13071-019-3753-8

Anwar, M. Y., Lewnard, J. A., Parikh, S., & Pitzer, V. E. (2016). Time series analysis of malaria in Afghanistan: using ARIMA models to predict future trends in incidence. *Malaria Journal*, *15*(1). https://doi.org/10.1186/s12936-016-1602-1

Araújo, F. H. D., Santana, A. M., & Santos, P. D. A. (2016). International Journal of Medical Informatics Using machine learning to support healthcare professionals in making preauthorisation decisions. *International Journal of Medical Informatics*, *94*, 1–7. https://doi.org/10.1016/j.ijmedinf.2016.06.007

Aron, J. L. (1988). Mathematical modelling of immunity to malaria. *Mathematical Biosciences,*

90(1–2), 385–396.

Aron, J. L., & May, R. M. (1982). The population dynamics of malaria. In *The population dynamics of infectious diseases: theory and applications* (pp. 139–179). Springer.

Asha, T., Natarajan, S., & Murthy, K. N. B. (2011). Associative classification in the prediction of tuberculosis. *International Conference and Workshop on Emerging Trends in Technology 2011, ICWET 2011 - Conference Proceedings*, *January*, 1327–1330. https://doi.org/10.1145/1980022.1980315

Baranauskas, J. A., & Macedo, A. A. (2012). *Using Machine Learning Classifiers to Assist Healthcare-Related Decisions : Classification of Electronic Patient Records*. 3861–3874. https://doi.org/10.1007/s10916-012-9859-6

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *25th Annual Conference on Neural Information Processing Systems (NIPS 2011)*, *24*.

Boit, J., & Alyami, H. (2018). Malaria Surveillance System using Social Media. *Proceedings of the 13th Midwest Association for Information Systems (MWAIS) Conference*.

Bria, Y. P., Yeh, C.-H., & Bedingfield, S. (2021). Significant symptoms and nonsymptom-related factors for malaria diagnosis in endemic regions of Indonesia. *International Journal of Infectious Diseases*, *103*, 194–200. https://doi.org/https://doi.org/10.1016/j.ijid.2020.11.177

Bridget, O. N., Prasad, R., Onime, C., & Ali, A. A. (2021). Drug resistant tuberculosis classification using logistic regression. *International Journal of Information Technology*. https://doi.org/10.1007/s41870-020-00592-9

Briët, O. J. T., Vounatsou, P., Gunawardena, D. M., Galappaththy, G. N. L., & Amerasinghe, P. H. (2008). Models for short term malaria prediction in Sri Lanka. *Malaria Journal*, *7*(1), 1–11.

Buehler, J. W., Hopkins, R. S., Sosin, D. M., Tong, V., & Overhage 1957-, J. M. (Joseph M. (n.d.). *Framework for evaluating public health surveillance systems for early detection of outbreaks; recommendations from the CDC Working Group* (E. P. O. Centers for Disease Control and Prevention (U.S.) Division of Public Health Surveillance and Informatics. (ed.)).

Chekol, B. E., & Hagras, H. (2018). Employing Machine Learning Techniques for the Malaria

Epidemic Prediction in Ethiopia. *2018 10th Computer Science and Electronic Engineering (CEEC)*, 89–94. https://doi.org/10.1109/CEEC.2018.8674210

Chitnis, N. R. (2005). *Using mathematical models in controlling the spread of malaria*. 1–124. http://web.abo.fi/fak/mnf/mate/kurser/dynsyst/2009/uke44/chitnis.pdf

Chowriappa, P., Dua, S., & Todorov, Y. (2014). *Machine Learning in Healthcare Informatics* (S. Dua, U. R. Acharya, & P. Dua (eds.); Vol. 56). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40017-9

Christophers, S. R. (1911). Years, Epidemic malaria of the Punjab: with a note of a method of predicting epidemicNo Title. *Trans Committee Stud Malaria India*, *2*, 17-- 26.

Cohn, D. L., Bustreo, F., & Raviglione, M. C. (1997). Drug-resistant tuberculosis: review of the worldwide situation and the WHO/IUATLD global surveillance project. *Clinical Infectious Diseases*, *24*(1 SUPPL.), 121–130. https://doi.org/10.1093/clinids/24.supplement_1.s121

Commission, N. P. (2008). Nigeria demographic and health survey. *Federal Republic of Nigeria Abuja, Nigeria*.

Coulibaly, D., Travassos, M. A., Kone, A. K., Tolo, Y., Laurens, M. B., Traore, K., Diarra, I., Niangaly, A., Daou, M., Dembele, A., Sissoko, M., Guindo, B., Douyon, R., Guindo, A., Kouriba, B., Sissoko, M. S., Sagara, I., Plowe, C. V, Doumbo, O. K., & Thera, M. A. (2014). Stable malaria incidence despite scaling up control strategies in a malaria vaccine-testing site in Mali. *Malaria Journal*, *13*(1), 374. https://doi.org/10.1186/1475-2875-13-374

Cox, J., & Abeku, T. A. (2007). Early warning systems for malaria in Africa: from blueprint to practice. *Trends in Parasitology*, *23*(6), 243–246.

Dande, P., & Samant, P. (2018). Acquaintance to Artificial Neural Networks and use of artificial intelligence as a diagnostic tool for tuberculosis: A review. *Tuberculosis*, *108*, 1–9. https://doi.org/10.1016/j.tube.2017.09.006

Daniel, T. M., Bates, J. H., & Downes, K. A. (1994). History of tuberculosis. *Tuberculosis: Pathogenesis, Protection, and Control*, 13–24.

Davis, J. K., Gebrehiwot, T., Worku, M., Awoke, W., Mihretie, A., Nekorchuk, D., & Wimberly, M. C. (2019). A genetic algorithm for identifying spatially-varying environmental drivers in a malaria time series model. *Environmental Modelling & Software : With Environment Data News*, *119*, 275–284. https://doi.org/10.1016/j.envsoft.2019.06.010

Dietz, K., Molineaux, L., & Thomas, A. (1974). A malaria model tested in the African savannah.

Bulletin of the World Health Organization, *50*(3–4), 347.

Dirlikov, E., Raviglione, M., & Scano, F. (2015). Global tuberculosis control: toward the 2015 targets and beyond. *Annals of Internal Medicine*, *163*(1), 52–58.

Domingues, R., Filippone, M., Michiardi, P., & Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognition*, *74*, 406–421. https://doi.org/https://doi.org/10.1016/j.patcog.2017.09.037

Doudou, M. H., Mahamadou, A., Ouba, I., Lazoumar, R., Boubacar, B., Arzika, I., Zamanka, H., Ibrahim, M. L., Labbo, R., Maiguizo, S., Girond, F., Guillebaud, J., Maazou, A., & Fandeur, T. (2012). A refined estimate of the malaria burden in Niger. *Malaria Journal*, *11*(1), 89. https://doi.org/10.1186/1475-2875-11-89

Druetz, T. (2018). Evaluation of direct and indirect effects of seasonal malaria chemoprevention in Mali. *Scientific Reports*, *8*(1), 8104. https://doi.org/10.1038/s41598-018-26474-6

Elshawi, R. (2020). *Interpretability in healthcare : A comparative study of local machine learning interpretability techniques*. *September 2019*, 1–18. https://doi.org/10.1111/coin.12410

Évora, L. H. R. A., Seixas, J. M., & Kritski, A. L. (2016). Artificial neural network models for diagnosis support of drug and multidrug resistant tuberculosis. *2015 Latin-America Congress on Computational Intelligence, LA-CCI 2015*. https://doi.org/10.1109/LA-CCI.2015.7435954

F.S., A., L.L., A., A., R.-N., A.L., K., F.C.Q., M., & G.L., W. (2012). Classification and regression tree (CART) model to predict pulmonary tuberculosis in hospitalized patients. *BMC Pulmonary Medicine*, *12*. https://doi.org/10.1186/1471-2466-12-40

Feachem, R. G. A., Chen, I., Akbari, O., Bertozzi-Villa, A., Bhatt, S., Binka, F., Boni, M. F., Buckee, C., Dieleman, J., & Dondorp, A. (2019). Malaria eradication within a generation: ambitious, achievable, and necessary. *The Lancet*, *394*(10203), 1056–1112.

Gangopadhyay, A., Yesha, R., & Siegel, E. (2016). Knowledge discovery in clinical data. In *Machine Learning for Health Informatics* (pp. 337–356). Springer.

Gaudart, J., Touré, O., Dessay, N., lassane Dicko, A., Ranque, S., Forest, L., Demongeot, J., & Doumbo, O. K. (2009). Modelling malaria incidence with environmental dependency in a locality of Sudanese savannah area, Mali. *Malaria Journal*, *8*(1), 1–12.

Githeko, A. K., & Ndegwa, W. (2001). Predicting malaria epidemics in the Kenyan highlands

using climate data: a tool for decision makers. *Global Change and Human Health*, *2*(1), 54–63.

Global Tuberculosis Report 2020. (2020). *World Health Organization*. https://www.who.int/publications/i/item/9789240013131

Gomez-Elipe, A., Otero, A., van Herp, M., & Aguirre-Jaime, A. (2007). Forecasting malaria incidence based on monthly case reports and environmental factors in Karuzi, Burundi, 1997–2003. *Malaria Journal*, *6*(1), 129. https://doi.org/10.1186/1475-2875-6-129

González Jiménez, M., Babayan, S. A., Khazaeli, P., Doyle, M., Walton, F., Reddy, E., Glew, T., Viana, M., Ranford-Cartwright, L., & Niang, A. (2019). Prediction of mosquito species and population age structure using mid-infrared spectroscopy and supervised machine learning. *Wellcome Open Research*, *4*.

Gouda, H. N., Charlson, F., Sorsdahl, K., Ahmadzada, S., Ferrari, A. J., Erskine, H., Leung, J., Santamauro, D., Lund, C., & Aminde, L. N. (2019). Burden of non-communicable diseases in sub-Saharan Africa, 1990–2017: results from the Global Burden of Disease Study 2017. *The Lancet Global Health*, *7*(10), e1375–e1387.

Guillermo del Rey-Pineda, G. G.-E. (2015). *Practical and Laboratory Diagnosis of Tuberculosis From Sputum Smear to Molecular Biology*. http://www.springer.com/series/8911

Haghdoost, A., Alexander, N., & Cox, J. (2008). Modelling of malaria temporal variations in Iran. *Tropical Medicine & International Health*, *13*(12), 1501–1508.

Han Jiawei, J. P. (2000). Mining FrequentPatterns without Candidate Generation. *ACM SIGMOD Record*, *29*. https://doi.org/10.1145/335191.335372

Hand, D. J., & Till, R. J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, *45*(2), 171–186. https://doi.org/10.1023/A:1010920819831

He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, *25*(1), 30–36.

https://ncar.ucar.edu/what-we-offer/data-services. (2019). National Center for Atmospheric Research. *Data*.

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *17*(3), 299–310. https://doi.org/10.1109/TKDE.2005.50

Huddart, S., Nash, M., & Pai, M. (2016). Tuberculosis diagnosis: Challenges and solutions. *Journal of Health Specialties*, *4*(4), 230.

Ilayaraja, M., & Meyyappan, T. (2013). Mining medical data to identify frequent diseases using Apriori algorithm. *Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, PRIME 2013, December*, 194–199. https://doi.org/10.1109/ICPRIME.2013.6496471

Ji, C., Zou, X., Hu, Y., Liu, S., Lyu, L., & Zheng, X. (2019). XG-SF: An XGBoost Classifier Based on Shapelet Features for Time Series Classification. *Procedia Computer Science*, *147*, 24–28. https://doi.org/https://doi.org/10.1016/j.procs.2019.01.179

Kalipe, G., Gautham, V., & Behera, R. K. (2018). Predicting malarial outbreak using machine learning and deep learning approach: a review and analysis. *2018 International Conference on Information Technology (ICIT)*, 33–38.

Klaucke, D. N., Buehler, J. W., Thacker, S. B., Parrish, R. G., & Trowbridge, F. L. (1988). *Guidelines for evaluating surveillance systems*.

Koella, J. C. (1991). On the use of mathematical models of malaria transmission. *Acta Tropica*, *49*(1), 1–25. https://doi.org/10.1016/0001-706X(91)90026-G

Kulkarni, P. (2017). Understanding machine learning opportunities. *Intelligent Systems Reference Library*, *128*, 23–48. https://doi.org/10.1007/978-3-319-55312-2_2

Kumar, S., Kumari, R., & Pandey, R. (2015). *New insight-guided approaches to detect , cure , prevent and eliminate malaria* (Vol. 2013). https://doi.org/10.1007/s00709-014-0697-x

Kumar, V., Mangal, A., Panesar, S., Yadav, G., Talwar, R., Raut, D., & Singh, S. (2014). Forecasting malaria cases using climatic factors in Delhi, India: a time series analysis. *Malaria Research and Treatment*, *2014*.

Laneri, K., Bhadra, A., Ionides, E. L., Bouma, M., Dhiman, R. C., Yadav, R. S., & Pascual, M. (2010). Forcing versus feedback: epidemic malaria and monsoon rains in northwest India. *PLoS Comput Biol*, *6*(9), e1000898.

Laurenzi, M., Ginsberg, A., & Spigelman, M. (2007). Challenges associated with current and future TB treatment. *Infectious Disorders-Drug Targets (Formerly Current Drug Targets-Infectious Disorders)*, *7*(2), 105–119.

Lechthaler, F., Matthys, B., Lechthaler-Felber, G., Likwela, J. L., Mavoko, H. M., Rika, J. M., Mutombo, M. M., Ruckstuhl, L., Barczyk, J., Shargie, E., Prytherch, H., & Lengeler, C.

(2019). Trends in reported malaria cases and the effects of malaria control in the Democratic Republic of the Congo. *PLOS ONE*, *14*(7), e0219853. https://doi.org/10.1371/journal.pone.0219853

Lee, Y. W., Choi, J. W., & Shin, E. (2020). Machine learning model for predicting malaria using clinical information. *Computers in Biology and Medicine*, 104151. https://doi.org/10.1016/j.compbiomed.2020.104151

Li, S., & Zhang, X. (2020). Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm. *Neural Computing and Applications*, *32*(7), 1971–1979. https://doi.org/10.1007/s00521-019-04378-4

Lindsay, S. W., & Birley, M. H. (1996). Climate change and malaria transmission. *Annals of Tropical Medicine & Parasitology*, *90*(5), 573–588.

Lokeshkumar, R., Jothi, K. R., Anto, S., Kiran, R., & Narayanan, H. (2019). Prediction of Multi Drug Resistant Tuberculosis using Machine Learning Techniques. *International Journal of Engineering and Advanced Technology*, *9*(2), 1764–1771. https://doi.org/10.35940/ijeat.b2531.129219

Macdonald, G. (1957). The epidemiology and control of malaria. *The Epidemiology and Control of Malaria.*

Maity, N. G., & Das, S. (2017). Machine learning for improved diagnosis and prognosis in healthcare. *2017 IEEE Aerospace Conference*, 1–9.

Makinde, O. S., Abiodun, G. J., & Ojo, O. T. (2020). Modelling of malaria incidence in Akure, Nigeria: negative binomial approach. *GeoJournal*. https://doi.org/10.1007/s10708-019-10134-x

Malaria, R. B. (2000). *The African Summit on Roll Back Malaria, Abuja, Nigeria, April 25th, 2000*. WHO/CDS/RBM/2000.17). RBM, Geneva.

Manjiri, M., Mastoli, M., Pol, U. R., & Patil, R. D. (2019). *Machine Learning Classification Algorithms for Predictive Analysis in Healthcare*. 1225–1229.

Mathur, P. (2018). *Overview of Machine Learning in Healthcare*. *January*, 1–11. https://doi.org/10.1007/978-1-4842-3787-8

Maurice, N., Aicha, S., Young, H. S., Eon, K. J., Hoon, K., Junseok, P., & Won-Joo, H. (2019). Malaria Epidemic Prediction Model by Using Twitter Data and Precipitation Volume in Nigeria. *Multimedia Society*, *22*(5), 588–600.

McKenzie, F. E., & Samba, E. M. (2004). The role of mathematical modeling in evidence-based malaria control. *The American Journal of Tropical Medicine and Hygiene*, *71*(2 Suppl), 94–96. https://pubmed.ncbi.nlm.nih.gov/15331824

Messina, J. P., Taylor, S. M., Meshnick, S. R., Linke, A. M., Tshefu, A. K., Atua, B., Mwandagalirwa, K., & Emch, M. (2011). Population, behavioural and environmental drivers of malaria prevalence in the Democratic Republic of Congo. *Malaria Journal*, *10*(1), 161. https://doi.org/10.1186/1475-2875-10-161

Ministry of Environment and Forests. (2006). National Meteorological Department, average annual rainfall covering the period 1975 to 2004 in National Adaptation Programme of Action (NAPA). *Change*, *Aout*, 83.

Molineaux, L. (1988). The epidemiology of human malaria as an explanation of its distribution, including some implications for its control. *Malaria: Principles and Practice of Malariology. Volume 2.*, 913–998.

Molineaux, L., Gramiccia, G., & Organization, W. H. (1980). *The Garki project: research on the epidemiology and control of malaria in the Sudan savanna of West Africa*. World Health Organization.

Mopuri, R., Kakarla, S. G., Mutheneni, S. R., Kadiri, M. R., & Kumaraswamy, S. (2020). Climate based malaria forecasting system for Andhra Pradesh, India. *Journal of Parasitic Diseases*. https://doi.org/10.1007/s12639-020-01216-6

Nabarro, D. (1999). Roll Back Malaria. *Parassitologia*, *41*(1–3), 501–504.

Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. (2013). Expert Systems with Applications Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems With Applications*, *40*(4), 1086–1093. https://doi.org/10.1016/j.eswa.2012.08.028

National Academies of Sciences and Medicine, E. (2017). Addressing Continuous Threats: HIV/AIDS, Tuberculosis, and Malaria. *Global Health and the Future Role of the United States*.

Ngwa, G. A., & Shu, W. S. (2000). A mathematical model for endemic malaria with variable human and mosquito populations. *Mathematical and Computer Modelling*, *32*(7–8), 747–763.

Nkiruka, O., Prasad, R., & Clement, O. (2021). Prediction of malaria incidence using climate

variability and machine learning. *Informatics in Medicine Unlocked*, *22*, 100508.

Onwujekwe, O., Chima, R., & Okonkwo, P. (2000). Economic burden of malaria illness on households versus that of all other illness episodes: a study in five malaria holo-endemic Nigerian communities. *Health Policy*, *54*(2), 143–159.

Organization, W. H. (2015). *Global tuberculosis report 2015* (20th ed). World Health Organization.

Organization, W. H. (2019). *WHO guideline: recommendations on digital interventions for health system strengthening: web supplement 2: summary of findings and GRADE tables*. World Health Organization.

Ouedraogo, B., Inoue, Y., Kambiré, A., Sallah, K., Dieng, S., Tine, R., Rouamba, T., Herbreteau, V., Sawadogo, Y., & Ouedraogo, L. S. L. W. (2018). Spatio-temporal dynamic of malaria in Ouagadougou, Burkina Faso, 2011–2015. *Malaria Journal*, *17*(1), 1–12.

Paaijmans, K. P., Read, A. F., & Thomas, M. B. (2009). *Understanding the link between malaria risk and climate*.

Pai, M., Behr, M. A., Dowdy, D., Dheda, K., Divangahi, M., Boehme, C. C., Ginsberg, A., Swaminathan, S., Spigelman, M., Getahun, H., Menzies, D., & Raviglione, M. (2016). Tuberculosis. *Nature Reviews Disease Primers*, *2*(1), 16076. https://doi.org/10.1038/nrdp.2016.76

Parham, P. E., & Michael, E. (2010). Modeling the effects of weather and climate change on malaria transmission. *Environmental Health Perspectives*, *118*(5), 620–626.

Payus, C., Sulaiman, N., Shahani, M., & Bakar, A. A. (2013). *Association Rules of Data Mining Application for Respiratory Illness by Air Pollution Database*. *January*.

Raheja, V. (2012). *Comparative Study of Association Rule Mining and MiSTIC in Extracting Spatio- Temporal Disease Occurrences Patterns*. https://doi.org/10.1109/ICDMW.2012.131

Rakhmetulayeva, S. B., Duisebekova, K. S., Mamyrbekov, A. M., & Kozhamzharova, D. K. (2018). ScienceDirect ScienceDirect ScienceDirect Application of Classification Algorithm Based on SVM for Application of Classification Algorithm Based on SVM for Determining the Effectiveness of Treatment of Tuberculosis Determining the Effectiveness of Treatme. *Procedia Computer Science*, *130*, 231–238. https://doi.org/10.1016/j.procs.2018.04.034

Ranovan, R., Doewes, A., & Saptono, R. (2018). Twitter Data Classification using Multinomial Naive Bayes for Tropical Diseases Mapping in Indonesia. *Journal of Telecommunication,*

Electronic and Computer Engineering (JTEC), *10*(2–4), 155–159.

Rashid, M. A., Hoque, T., & Sattar, A. (n.d.). *Association Rules Mining Based Clinical Observations*.

Rosenthal, P. J., John, C. C., & Rabinovich, N. R. (2019). Malaria: how are we doing and how can we do better? *The American Journal of Tropical Medicine and Hygiene*, *100*(2), 239.

Roser, M., & Ritchie, H. (2020). Malaria. *Our World in Data*.

Ross, R. (1915). SOME A PRIORI PATHOMETRIC EQUATIONS. *British Medical Journal*, *1*(2830), 546–547. https://doi.org/10.1136/bmj.1.2830.546

Rouamba, T., Nakanabo-Diallo, S., Derra, K., Rouamba, E., Kazienga, A., Inoue, Y., Ouédraogo, E. K., Waongo, M., Dieng, S., Guindo, A., Ouédraogo, B., Sallah, K. L., Barro, S., Yaka, P., Kirakoya-Samadoulougou, F., Tinto, H., & Gaudart, J. (2019). Socioeconomic and environmental factors associated with malaria hotspots in the Nanoro demographic surveillance area, Burkina Faso. *BMC Public Health*, *19*(1), 249. https://doi.org/10.1186/s12889-019-6565-z

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210–229.

Santosh, T., Ramesh, D., & Reddy, D. (2020). LSTM based prediction of malaria abundances using big data. *Computers in Biology and Medicine*, *124*, 103859. https://doi.org/https://doi.org/10.1016/j.compbiomed.2020.103859

Seneviratne, M. G., & Shah, N. H. (2019). *Bridging the implementation gap of machine learning in healthcare*. 1–3. https://doi.org/10.1136/bmjinnov-2019-000359

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, *88*(422), 486–494.

Shrivastava, A. K., Rajak, A., & Singhal, N. (2016). Modeling Pulmonary Tuberculosis using Adaptive Neuro Fuzzy Inference System. *International Journal of Innovative Research in Computer Science & Technology (IJIRCST)*, *4*(1), 24–27. http://www.ijircst.org/DOC/7_IRP4510ca15a7a-932d-4ea3-b042-0e6438ef1e2d.pdf

Skorburg, J. A. (2020). What Counts as " Clinical Data " in Machine Learning Healthcare Applications ? *The American Journal of Bioethics*, *20*(11), 27–30. https://doi.org/10.1080/15265161.2020.1820107

Sun, H., Burton, H. V, & Huang, H. (2021). Machine learning applications for building structural

design and performance assessment: State-of-the-art review. *Journal of Building Engineering*, *33*, 101816. https://doi.org/https://doi.org/10.1016/j.jobe.2020.101816

Sutabri, T., Selvam, R. P., Shankar, K., Nguyen, P. T., & Hashim, W. (2019). *Machine Learning for Healthcare Diagnostics*. *6*, 999–1001. https://doi.org/10.35940/ijeat.F1304.0886S219

Tangpukdee, N., Duangdee, C., Wilairatana, P., & Krudsood, S. (2009). Malaria diagnosis: a brief review. *The Korean Journal of Parasitology*, *47*(2), 93.

Teklehaimanot, H. D., Lipsitch, M., Teklehaimanot, A., & Schwartz, J. (2004). *Weather-based prediction of Plasmodium falciparum malaria in epidemic-prone regions of Ethiopia I . Patterns of lagged weather effects reflect biological mechanisms*. *11*, 1–11. https://doi.org/10.1186/1475-2875-3-41

Teutsch, S. M., & Churchill, R. E. (2000). *Principles and practice of public health surveillance*. Oxford University Press, USA.

Thakur, S., & Dharavath, R. (2019). Artificial neural network based prediction of malaria abundances using big data: A knowledge capturing approach. *Clinical Epidemiology and Global Health*, *7*(1), 121–126. https://doi.org/10.1016/j.cegh.2018.03.001

Tizifa, T. A., Kabaghe, A. N., McCann, R. S., van den Berg, H., Van Vugt, M., & Phiri, K. S. (2018). Prevention Efforts for Malaria. *Current Tropical Medicine Reports*, *5*(1), 41–50. https://doi.org/10.1007/s40475-018-0133-y

Toh, C., & Brody, J. P. (2021). Applications of Machine Learning in Healthcare. *Smart Manufacturing: When Artificial Intelligence Meets the Internet of Things*, 65.

Wagenmakers, E.-J. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196. https://doi.org/10.3758/BF03206482

Wang, B., Chen, D., Shi, B., Zhang, J., Duan, Y., Chen, J., & Hu, R. (2017). Comprehensive Association Rules Mining of Health Examination Data with an Extended FP-Growth Method. *Mobile Networks and Applications*. https://doi.org/10.1007/s11036-016-0793-6

Wangdi, K., Singhasivanon, P., Silawan, T., Lawpoolsri, S., White, N. J., & Kaewkungwal, J. (2010). Development of temporal modelling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: a case study in endemic districts of Bhutan. *Malaria Journal*, *9*(1), 1–9.

Wassan, J. T., Wang, H., Zheng, H., Antrim, C., Ireland, N., & Kingdom, U. (2018). *Machine Learning in Bioinformatics*. 1–9. https://doi.org/10.1016/B978-0-12-809633-8.20331-2

White, N. J., & Ho, M. (1992). *The Pathophysiology of Malaria* (J. R. Baker & R. B. T.-A. in P. Muller (Eds.); Vol. 31, pp. 83–173). Academic Press. https://doi.org/https://doi.org/10.1016/S0065-308X(08)60021-4

Widyasrini, E. R., & Probandari, A. N. (2015). *Factors Affecting the Success of Multi Drug Resistance (MDR-TB) Tuberculosis Treatment in Residential Surakarta*. 45–57.

Williams, H. A., Bloland, P. B., Council, N. R., & Population, C. on. (2003). *Malaria control during mass population movements and natural disasters*.

Wojtusiak, J. (2014). *Chapter 8 Rule Learning in Healthcare and Health Services Research*. *56*.

Wong, T., & Yeh, P. (2020). Reliable Accuracy Estimates from k-Fold Cross Validation. *IEEE Transactions on Knowledge and Data Engineering*, *32*(8), 1586–1594. https://doi.org/10.1109/TKDE.2019.2912815

World malaria report. (2019). World malaria report 2019. In *WHO Regional Office for Africa*. https://www.who.int/news-room/fact-sheets/detail/malaria

Yang, C., Tang, G., Xue, C., Zhou, J., Chang, W., Yuan, X., & Zhou, S. (2018). *Mining association rule based on the diseases population for recommendation of medicine need Mining association rule based on the diseases population for recommendation of medicine need*.

Yang, H. M. (2000). Malaria transmission model for different levels of acquired immunity and temperature-dependent parameters (vector). *Revista de Saude Publica*, *34*, 223–231.

Yang, H. M., & Ferreira, M. U. (2000). Assessing the effects of global warming and local social and economic conditions on the malaria transmission. *Revista de Saude Publica*, *34*, 214–222.

Yang, L.-H., Wang, Y.-M., & Fu, Y.-G. (2018). A consistency analysis-based rule activation method for extended belief-rule-based systems. *Information Sciences*, *445–446*, 50–65. https://doi.org/https://doi.org/10.1016/j.ins.2018.02.059

Yang, Y., Niehaus, K. E., Walker, T. M., Iqbal, Z., Walker, A. S., Wilson, D. J., Peto, T. E. A. A., Crook, D. W., Smith, E. G., Zhu, T., Clifton, D. A., Walker, S., Wilson, D. J., Peto, T. E. A. A., Crook, D. W., Grace, S. E., Zhu, T., Clifton, D. A., Walker, A. S., … Clifton, D. A. (2018). Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinformatics*, *34*(10), 1666–1671. https://doi.org/10.1093/bioinformatics/btx801

Zaki, M. J., Meira Jr, W., & Meira, W. (2014). *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159–175.

Zhao, H., Lai, Z., Leung, H., & Zhang, X. (2020). *Feature Learning and Understanding*. https://doi.org/10.1007/978-3-030-40794-0

Zimmerman, P. A., & Howes, R. E. (2015). Malaria diagnosis for malaria elimination. *Current Opinion in Infectious Diseases*, *28*(5), 446–454.

Zinszer, K., Verma, A. D., Charland, K., Brewer, T. F., Brownstein, J. S., Sun, Z., & Buckeridge, D. L. (2012). A scoping review of malaria forecasting: Past work and future directions. *BMJ Open*, *2*(6), 1–11. https://doi.org/10.1136/bmjopen-2012-001992