A HYBRID-BASED CLASSIFICATION AND REGRESSION MODEL FOR PREDICTING MALARIA OUTBREAK

HAKIZIMANA LEOPORD

DOCTOR OF PHILOSOPHY IN
COMPUTER SCIENCE

JOMO KENYATTA UNIVERSITY
OF
AGRICULTURE AND TECHNOLOGY

2026

A Hybrid-Based Classification and Regression Model for Predicting Malaria Outbreak

Leopord Hakizimana

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Computer Science of the Jomo Kenyatta University of Agriculture and Technology

2026

DECLARATION

This thesis is my original work and has not been presented for a degree in any other University.

Signature…….Date……

Hakizimana Leopord

This thesis has been submitted for examination with our approval as the University Supervisors:

Signature……. ……. (Deceased)

Prof. Wilson Kipruto Cheruiyot, PhD

JKUAT, Kenya

Signature…….Date……

Prof. Stephen Kimani, PhD

JKUAT, Kenya

DEDICATION

This PhD thesis is dedicated to my wife Nadia Kanobayita, my children Nolan Leo Kirenga, and Irisa Leo Charvin, and my parents, Mr. Theogene Nyakamwe and Mrs. Donatha Mbabajende, for helping me to succeed in my academic studies endeavors. Special dedication to my siblings Dr. Noel, Ferdinand, Xavier and Nepomuscene for supporting me tirelessly during the entire PhD journey. I also dedicate my research to all members of the academic community who supported me and made me progress in the field of computer science as well as other cutting-edge technologies and breakthroughs that have the potential to fundamentally alter society.

ACKNOWLEDGEMENT

Many people contributed to the effective completion of this PhD journey. First and foremost, I am thankful to God for his provision of safety and wisdom during this thesis. I would like to express my sincere gratitude to my parents, Mr. Theogene Nyakamwe and Mrs. Donatha Mbabajende, as well as my brothers, Mr. Ferdinand Habimana and Dr. Noel Manirakiza at the University of Florida in the United States and my lovely wife Kanobayita Nadia, for their wonderful financial and moral support that has made it possible for me to succeed in this academic struggle. My gratitude goes to the governance of the Jomo Kenyatta University of Agriculture and Technology (JKUAT) for awarding me with an excellent opportunity to pursue my PhD at their distinguished institution of higher learning. I would like to express my gratitude to Prof. Wilson Kipruto Cheruiyot, Prof. Stephen Kimani and Dr. Jael Sanyanda Wekesa for their excellent advice, direction and academic inspiration, all of which helped me to write this thesis well from the beginning to the end of the program. Without them, it could be challenging to finish this thesis; their views and criticisms have helped me to achieve my aim.

I am particularly appreciative to the Director of the School of Computing and Information Technology (SCIT), Dr. Agnes N. Mindila, the Chairman of Computing Department, Dr. Lawrence Nderu and academic staff including, Dr. Kennedy Ogada, Dr. Michael Kimwele, Dr. Ann Kibe, Dr. Richard Rimiru, Prof. George Okeyo, Prof. Joseph Wafula, and other thesis committee, internal and external examiners staff of JKUAT who guided me through my doctoral program in computer science courses and helped me in providing feedback, comments and guidance. Moreover, I firmly believe that my research would not have achieved success without the assistance and cooperation of the University's administrative and technical personnel, as well as the department's head, during the formulation of this thesis. Finally, I would like to express my gratitude to the entire community of University of Kigali community, especially to my colleague Dr. Sikubwabo Cyprien, Dr. Emmanuel Bugingo, and Dr. Musoni Wilson. I am also grateful to the promoters of the University of Kigali, Prof. Nshuti Manasseh, Mr. Philibert Afrika, Mr.

Sam Aime Nuwe for their excellent financial and constructive support that helped me to better complete this thesis.

Finally, I would like to thank everyone who helped me write this thesis successfully.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ACRONOMYS AND ABBREVIATIONS

AFRO	African Region Office
AI	Artificial Intelligence
CVD	Cardiovascular Diseases
DHF	Dengue Hemorrhagic Fever
DM	Data Mining
EID	Infectious Diseases Division
EVD	Ebola Virus Disease
ICU	Intense Core Unit
IDSR	Integrated Diseases Surveillance and Response
IRS	Indoor Residual Spray
KDD	Knowledge Discovery in Databases
KNN	K-Nearest Neighbors
KR	Knowledge Representation
LSSVM	Least Square Support Vector Machine
MCDM	Multi Criteria Decision Making
ML	Machine Learning
MOH	Minister of Health
MSE	Mean Squared Error
NB	Naïve Bayes
NN	Neural Networks
NNM	Nearest Neighbor Methods
RBC	Rwanda Biomedical Center
ROC	Receiver Operating Characteristic
RST	Rough Set Theory
SMO	Sequential Minimal Optimization
SVM	Support Vector Machine
WHO	World Health Organization

ABSTRACT

Malaria outbreaks remain a main public health challenge worldwide, particularly in Sub-Saharan Africa. Fast and correct forecast of malaria outbreaks is critical for permitting timely interventions, decreasing morbidity and death, and ensuring effective sharing of limited healthcare resources. In the last ten years, Data mining and machine learning methods gained widespread attention in complex prediction tasks, such as healthcare analytics, financial and environmental monitoring prediction. Regardless of these improvements, existing malaria outbreak forecast models frequently show limitations in accuracy, adaptability, and applied usefulness. Numerous existing approaches depend solely on either regression or classification approaches, which limits their ability to attain the complex and dynamic interactions between environmental, climatic, and epidemiological factors that affect malaria transmission. This research introduces a new hybrid based predictive model that mix both regression and classification methods in a two-phase framework, marking to enhance the accuracy and reliability of malaria outbreak predictions. The first phase applies a regression model to predict the expected number of malaria cases by examining historical epidemiological data, climate variables and other appropriate environmental indicators. The second phase applies a classification model to determine the likelihood of an outbreak occurring within a given region and time frame, transforming quantitative predictions into actionable early warning signals. Through amalgamation these supplementary methods, the hybrid model influences the strengths of both regression and classification, outcome of in enhanced prediction performance, robustness, and adaptability under diverse outbreak situations. Comprehensive experimentations were done using publicly accessible and region-specific malaria datasets, and the outcomes show that the hybrid model significantly outperforms conventional single-method approaches. The attained model predictive accuracies of 96% through training and 93% in testing, demonstrating strong generalizing capabilities. Likewise, the hybrid approach improves the decision-making aptitudes of healthcare systems by providing timely and reliable information that support evidence-based interventions, such as targeted mosquito control, resource prioritization and prophylactic measures. Study has important inferences for health professionals and authorities, policymakers, and international health organizations endeavoring to reduce malaria burden efficiently. The research helps to healthcare and machine learning fields by giving a scalable and adaptable framework for disease outbreak prediction. It also serves as a foundation for future studies on hybrid and ensemble Machine Learning models, mostly in the perspective of infectious diseases prediction. Future work is endorsed to explore the combination of supervised and unsupervised hybrid methods, integration of real-time epidemiological and climatic data streams, and assessment under large-scale, dynamic outbreak situations. Further, this study emphasizes the capability of hybrid machine learning models to renovate disease outbreak prediction by merging the predictive strengths of regression and classification methods. The findings show the importance of adapting data driven strategies for enabling early detection, timely interventions, public health preparedness and ultimately, the reduction of malaria transmission and its associated health and socioeconomic effects

Keywords: Hybrid-Based Model, Classification, Regression, Malaria, Outbreak Prediction

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Artificial intelligence has become a foundation in the digital revolution of the worldwide healthcare area. Some of its branches particularly machine learning, data mining, and deep learning have led to a paradigm change in how health data is evaluated diseases are predicted, and care is distributed. These computational procedures are gradually embedded in clinical workflows contributing to diagnostic precision, early disease discovery personalized treatments, outbreak forecasting, and real time verdict making.

One of the pioneering accomplishments of artificial intelligence in medical field could be traced to enhance structures such as MYCIN in the 1970s, which was premeditated to diagnose bacterial infections and endorse antibiotics. Though it was not put into practice, it laid the groundwork for knowledge-based decision support structures (Cohen et al., 2022). ML models have shown an extraordinary efficiency in predictive healthcare and Supervised learning methods, including decision trees, support vector machines, and random predicts have been utilized for recognizing conditions for instance heart illness, diabetes, and cancer by high levels of accuracy (Ozsahin et al., 2024). The mentioned algorithms learn from historical patient archives and medical parameters to categorize risk ranks and acclaim involvements.

Deep learning a branch of machine learning is has revolutionized image-based diagnostics. Convolutional neural networks, in particular, have attained dermatologist achievement level in predicting skin cancer (Kang, 2022) and have surpassed human radiologists in detecting lung cancer and COVID 19 outbreak on CT scans (Wang et al., 2020) . In ophthalmology, Google's AI system for detecting diabetic retinopathy in retina; images demented sensitivity and specificity comparable to board certified ophthalmologists (Gulshan et al., 2016).

Throughout the COVID 19 outbreak, AI technologies were influential in real time monitoring, diagnostics, and vaccine study. Tools such as BlueDot leveraged Natural Language Processing and ML to notice the outbreak in Wuhan beforehand worldwide alerts showcasing the potential of data mining and predictive modeling in epidemiology (Bogoch et al., 2019). Furthermore, deep learning algorithms assisted detect viral structures, quickening drug repurposing and vaccine growth.

In public health and disease surveillance, AI models have been useful to predict malaria outbreaks using biological variables for example rainfall, humidity, and temperature. As, AI founded models established in Africa have positively used regression and classification algorithms to predict malaria occurrence at district levels, thus supportive early intervention approaches (Adigun et al., 2024).

The addition of AI methods in smartphones has developed as gainful problem-solving tools in low resource situations. A prominent illustration is the usage of deep learning algorithms to examine blood smear images seized through portable phones for correct malaria parasite discovery (Yang et al., 2019). This revolution ties the gap amongst technology and underserved healthcare structures.

Likewise, Natural Language Processing (NLP) has been implemented for mining unstructured clinical data such as physician notes and discharge abstracts. NLP tools are being used in Electronic Health Record systems to detect adverse drug reactions, monitor chronic diseases, and assist in clinical documentation (Huang, 2019).Estimating analytics power-driven via AI similarly contributes to precision medicine through tailoring treatment strategies grounded on genetic data and existence aspects. Artificial intelligence models can estimate disease evolution as well as recommend adapted drug therapies, refining handling results and reducing healthcare expenditures (Topol, 2019).

These success stories demonstrate how AI technologies, particularly ML, Deeper learning (DL), and data mining, have evolved from theoretical concepts into practical tools that now show a essential role in contemporary health system. This study builds on these

global trends by proposing a hybrid-based machine learning framework for malaria outbreak predict diction that combines the strengths of classification and regression techniques to deliver high accuracy and merges the proficiencies of classification and regression techniques to deliver high accuracy and actionable insights for public health decision making.

Additionally, developments in computational intelligent computing, particularly in the area of machine learning, are reforming the way prediction and decision support systems are developed, principally in data intensive the sectors like healthcare. Essential machine learning approaches, including classification and regression algorithms have permitted computers to study from past data and do predictive inferences, providing a foundation for automated, data driven decision formulation procedure (Goodfellow et al., 2018). Classification models alike decision tree, support vector machines, and random forests predict definite results, whereas regression models for example linear regression and gradient boosting predict continuous standards (Sohil et al., 2022). Newly, there has been upward concentration in hybrid modeling structures that combine north classification and regression in unifies design to address the limitations of standalone models. These amalgam approaches offer improved accuracy, adaptability, and robustness, particularly in domains characterized by complex, heterogeneous, and noisy data (Traini & Lombardi, 2022).

Since a computer science perspective, the emergence of hybrid-based prediction frameworks presents a combination of technical challenges and promising opportunities. These include feature engineering, algorithm selection, hyper parameter tuning, and the application of appropriate evaluation metrics. Data preprocessing techniques for instance normalization, dimensionality reduction, and encoding are essential to ensuring computational efficiency and model effectiveness (Tharageswari et al., 2025). Also, ensemble learning and pipeline-based workflows in which the output of tone model guides the next enable the creation of sophisticated prediction systems. Such architectures are especially valuable in healthcare applications where both categorical and numerical predictions are required (Mahajan et al., 2023). This study, consequently, focuses on

building a hybrid classification regression model using optimized machine kerning methods to address real world outbreak prediction problems such as malaria.

Autonomously, the truth industrial revolution has accelerated innovation in healthcare, fundamentally transforming how medical services are delivered and how diseases are managed. This digital transformation, fueled by advances in big data, artificial intelligence, and automation, has empowered healthcare systems to make evidence based decision delivered from historical data. Machine learning and data mining have thus combined as key permits in extracting meaningful insights, identifying patterns and predicting future events with a level of precision previously unattainable. As a case in point, malaria stays one of most pressing worldwide health issues. Nevertheless, years of control efforts, the disease continues to affect millions, particularly in low resources settings (Rajkomar et al., 2019).

As highlighted by the World Health Organization (2019), there were approximately 228 million cases of malaria globally in 2018, with Africa accounting for 93% of those cases. A small number of countries, such as Nigeria, the Republic Democratic of Congo, Uganda, and the Ivory Coast, are responsible for more than half of all reported global cases. Malaria is spread to humans through female *Anopheles* mosquitoes, and between the many species, only around 30 are of significant concern globally. Of the five malaria causing parasites, *Plasmodium falciparum* is the utmost hazardous and was answerable for the vast popular of cases in Africa and other high burden regions (Shankar et al., 2026). Given the scale, complexity, and temporal dynamics of malaria outbreaks, traditional predicting methods often fall short as they are unable to efficiently process multisource, high-dimensional data. This study, therefore, leverages hybrid machine learning techniques to create a robust, scalable, and data supported model for malaria outbreak prediction, integrating the strengths of both classification and regression algorithms.

The greatest severe category of malaria, Plasmodium falciparum, is characterized as fever, shills, muscle aches, headaches, diarrhea, vomiting coughing, and abdominal pain. Furthermore, weakness in the organs may results in symptoms for example pulmonary

oedema, renal failure, general convulsions, circulatory collapse, surveyed by a coma and eventually death. Plasmodium falciparum therapy is theoretically fatal if administered beyond 24 hours afterwards the clinical sign appears (Mathuria et al., 2020). Malaria diagnosis in traditional laboratory settings requires experienced personnel and meticulous examination. In the African region, the occurrence of malaria dropped from 294 cases per 1,000 persons at risk in 2010 to 229 cases per 1,000 persons at risk in 2018. This decline made up 22% of global reduction in malaria cases during that time (World Health Organization, 2019). Malaria remains one of leading cause of mortality globally, with 405,000 deaths in 2018 and 585,000 deaths in 2010. Children under five years are the most vulnerable demographic, accounting for 67% of malaria related deaths worldwide.

Africa is the greatest exposed continent, with 94% of all malaria related mortality occurring in its countries. On the contrary, the continent has experienced a significant drop in malaria deaths from 553,000 in 2010 to 380,000 in 2018 (World Health Organization, 2019). In Rwanda, malaria affects roughly 7% of the population, with the uppermost occurrence in the eastern province (13%) and the lowermost in the Northern Province (1%). Malaria prevalence among children under five years old is disproportionately distributed according to their family's prosperity quintile, accounting for 13% in the lowermost quintile and 2% in the uppermost quintile. Malaria affects 11% of children aged 5 to 14 years old, with rural children experiencing a higher prevalence of 13% than urban children at 35. Malaria prevalence among adults age 15 years and older is 6%, with eastern Province reporting a prevalence of 12% of the occurrence of malaria but the rate is higher at 6% among uneducated women related to 3% amongst those with a secondary or higher education.

Marilia deaths toll in Rwanda increased from 419 in 2013 to 725 in 2016, and indicates an unprecedented rise in malaria mortality among the general population of Rwanda during that period (World Health Organization, 2022). Anticipating the future accurately is challenging, and the desire to do so drives the development of various tools and technologies has enabled the identification, prediction, and anticipation of disease related difficulties in the domain of public health sector informatics (Abdulkarim et al., 2022) .

Data mining is a comprehensive arena that combines statically approaches, ML, information science, visualization, and other related domains (Olushola & Mart, 2022). This versatile approach proves invaluable for extracting knowledge from several computational fields, such as bioinformatics, cheminformatics, and informatics for future forecasting, which is a natural desire for mankind.

The foremost objective of this research was to create and confirm an amalgam learning model for accurately forecasting malaria epidemics. This study was concerned with not just predicting the exact number of individual incidences, but also identifying who will experience rapid and large increase in malaria. This research aims to predict malaria outbreaks by providing a important decision-support instrument for public health authorities to insure timely interventions and effective allocation of resources.

1.1.1 Malaria Tracking Technical Practices

According to Hussain-Alkhateeb et al. (2021), the African Regional Office (WHO/AFRO) of the World Health Organization (WHO) introduced the Integrated Disease surveillance and Response (IDSR) approach in 1198, and Eshetu et al. (2024) reported that the Rwanda Biomedical Centre (RBC) managed an outbreak between July 2011 and June 2012 using this approach. The RBC/Epidemic Infectious Diseases provided the necessary medications and other consumables, monitored the afflicted health structure mechanisms, and made administration suggestions for control of the infection and prevention actions during each outbreak (Eshetu et al., 2024; Hussain-Alkhateeb et al., 2021). Furthermore, earlier studies have emphasized the significance of early warning systems in mitigating the effect of epidemics and preventing their escalation through timely response measures. For instance, Nyambura et al. (2025) reported that Malaria outbreak caused about 100,00 deaths and also each year, roughly 17 million persons die as a outcome of cardiovascular disease. It is very crucial to emphasize numerous strategies aimed at reducing malaria transmission to increase the number of countries, territories, and malaria free regions, which will reinforce the measurers for reducing malaria morbidity and death tolls worldwide (Nyambura et al., 2025).

The WHO introduced the Malaria Global Technical Strategy (GTS) from 2016 to 2030, with the primary goal of eradicating malaria worldwide. The three pillars of the GTS's are to ensure universal access to malaria prevention, diagnosis, and treatment; to accelerate progress towards elimination and achieving malaria free condition; and to strengthen malaria surveillance as a core intervention.

Besides, the GTS has outlined a set of objectives to be achieved by 2030, as well as the eradication of malaria in a minimum of 35 nations where malaria was widespread occurred in 2015 and avoid the spread of malaria across free nations (World Health Organization, 2022).

Mategula et al. (2025) stated that the World Health Organization reports the significance of fast diagnostic testing and effective treatment for malaria. Additionally, malaria cases are closely monitored to help the implementation of numerous control measures, namely: universal coverage with long-lasting insecticidal nets, indoor residual spraying, and artemisinin-based combination therapy in malaria-endemic countries (Mategula et al., 2025). Rwanda has implemented several initiatives to combat malaria, including expanding community-based treatment of malaria to include children over five years and adults, with up to 56% of malaria diagnosis treatments that are provided by community health workers. In addition, Rwanda's government has distributed over five million insecticide-treated nets, provided free diagnosis and treatment to all households in Ubudehe 1 and 2 categories, and expanded indoor residual spraying from three to five districts, using an organophosphate insecticide to prevent carbamate insecticide resistance (Masinde, 2020). The healthcare industry has been quick to develop data mining technologies in the health disciplines, particularly related to human lifer through disease prediction, diagnosis, and forecasting. Health informatics has contributed to the development of mankind and the country by sacking time, money, and human lives, as well as rising healthcare costs and the establishment of large health organizations (Mbunge et al., 2022). However, most people are unaware that the medical sector supplies vast amount of sensitive data, with patient details, diagnoses, health conditions, and illness

prediction, such as epidemics. Due to this, health professionals face difficulties in predicting disease outbreaks, and most diseases are only identified at advanced stages.

1.1.2 Main Environmental Factors for Malaria

Malaria prevalence in numerous countries is influenced by several risk variables, such as immunity at the population level, mosquito control procedures, societal and economic condition and ecological aspects like elevation, longitude, and latitude among others (Mbunge et al., 2022). Amongst these variables, environmental factors play significant role in creating conducive atmosphere for mosquito breeding, thus influencing malaria spreading. An growth in temperature accelerates the metabolic rate of mosquitoes, increases their egg production and blood feeding frequency, even though the seasonal number of malaria vectors is influenced by rainfall due to the relative humidity of mosquito habitat. Yet, excessive rain and flooding destroy mosquito breeding ground, resulting in a decrease vectors populations (World Health Organization, 2023).

Rising temperatures in Rwanda enhance the alike hood of spreading malaria, whereas rainfall impacts the number of anopheles' mosquitoes present in a region, as they require moist environment for their life cycle (Lingala, 2017). Sustaining and strengthening malaria surveillance systems needs long-term, high-quality data to elaborate models that connect malaria transmission to climatic dynamics (Hemachandran et al., 2023). Without effective mitigation, the malaria burden is projected to rise in several endemic regions, especially in densely populated tropical highland areas (World Health Organization, 2023).

Artificial intelligence procedures are increasingly transforming healthcare operations and clinical; decision-making by enhancing diagnostics, predictive analysis, patient monitoring, and administrative workflows. Russel and Norvig (2022) articulates that AI is a growing force in healthcare, helping clinicians analyze complex health data, supporting researchers in uncovering insights, and aiding medical staff in managing patient care more efficiently through machine learning and innovative data processing

methods. Furthermore, the implementation of generative AI (Gen AI) in healthcare has been recognized to improve personalized treatment planning, create synthetic data for research, assist in medical image analysis, and streamline nursing workflows demonstrating the clinical and non-clinical potential of AI systems health delivery (Russell and Norvig, 2022).

Recent empirical research underscores the multi-layered nature of AI in health field. For Example, Mategula et al. (2025) provide a comprehensive review of AI practices that highlights real-world applications such as robot-assisted surgery, diagnostic imaging, rehabilitation support, and virtual patient care, revealing how AI procedures are integrated into multiple stages of the healthcare workflow to improve efficiency and patient outcomes. Similarly, systematic reviews highlight healthcare professionals' perspectives on the adoption of artificial intelligence (AI), identifying both opportunities such as enhanced disease prognosis and workflow optimization and important challenges, such as clinician readiness, integration barriers, and factors affecting user acceptance that effect the practical use of AI in clinical backgrounds (Liu et al., 2018). These studies collectively illustrate that where AI procedures are already embedded in core clinical and demonstrative functions, challenges for example ethical considerations, model deration over time, fairness, and regulatory compliance remain critical. This comprehensive understanding of AI's role in healthcare provides an essential context for adopting hybrid machine learning methods such as those proposed in this research to ensure responsible, adaptive, and effective predictive systems for complex tasks like malaria outbreak prediction (Hall & Lucas, 2023; Modu et al., 2017) .

In the same background, AI is a technology that has been made possible by data mining and machine learning contributions. Machine learning includes the aptitude of machines to learn over the processing of data, rather than being preprogrammed for each action (Parveen et al., 2017). This method has led to important advances in the technology, enabling an extensive application to utilize machine learning. So, computers have been equipped with the ability to learn both subtle and clear patterns from large data sets. Frequent research studies published in the last two decades have reported that a range of

data sets and some range outbreaks at the national, global, and regional levels (Silhavy, 2023).

Data mining and machine learning are well known perceptions as two of the most important and stimulating areas of study, mainly in the quest to gain valuable data from huge data sets. In the medical field, these technologies have massive benefits, such as the recognition of healthcare policies, the provision of medical treatment tactics, and the recommendation of effective drug therapies (Adadi & Berrada, 2018). Likewise, machine learning and data mining have been influential in examining and recognizing food, place of employment, level of education, and living situations. Access to clean water, healthcare services, cultural and environmental features, and agricultural practices are all aspects that contribute to disease advance (Borham et al., 2025) .

Despite the vast amount of heterogeneous medical data that is visible, data mining and machine learning methods can categorize hidden patterns in these datasets and provide insights for predicting and making clinical diagnoses. To achieve this, data must be gathered and saved in an organized manner to enable the creation of a hospital information mechanism that physicians and other healthcare workers can employ (Bhuyan et al., 2025). Machine learning, data mining, recognition of patterns, statistics, databases, artificial intelligence, as well as knowledge acquisition researchers are attempting to develop new approaches that can handle the rising amount of data generated and stored in databases.

One of the main significant applications of the machine learning is the application of disease outbreaks. A hybrid model that merges regression and classification, which considers previously acquired knowledge from historical data, can accurately predict disease outbreaks on particular datasets (Faris, 2024). Building predictive models for malaria is becoming more and more popular as a way to help clinicians and public health professionals to strategically execute preventative and control measures in advance. Compared to traditional approaches, machine learning algorithms offer significant advantages in the strategic implementation of preventive and control measures, due to

their great predictive performance (Patient Safety Network, 2019). Many methods have been established to support learning from data and to predict epidemic outbreaks, improving early warning and intervention structures. Kadam (2020) and Modu et al. (2017) recommended epidemic prediction methods, as shown in Figure 1.1. The unique feature of the current study is that it can combine and use a wide range of malaria related datasets from multiple sources and formats.

The hybrid classification and regression model we provide is distinct from previous models since it can be altered based on the structure and features of the input data. Traditional models can only work with particular kinds of data or need a lot of preprocessing for one dataset. This flexibility is attained through automated dataset selection and feature engineering methods namely: normalization, dimensionality reduction, and feature importance ranking which guarantee that the data is optimally organized for both classification and regression duties. The model also uses a pipeline method, which means that the results from one step are used as better input for the following level. This improves the accuracy of forecasts and makes them more generic and also this method not only improves the strength of the model but also enhances its applicability across diverse epidemiological and geographical contexts, as illustrated by Modu et al. (2017) in Figure 1.1.

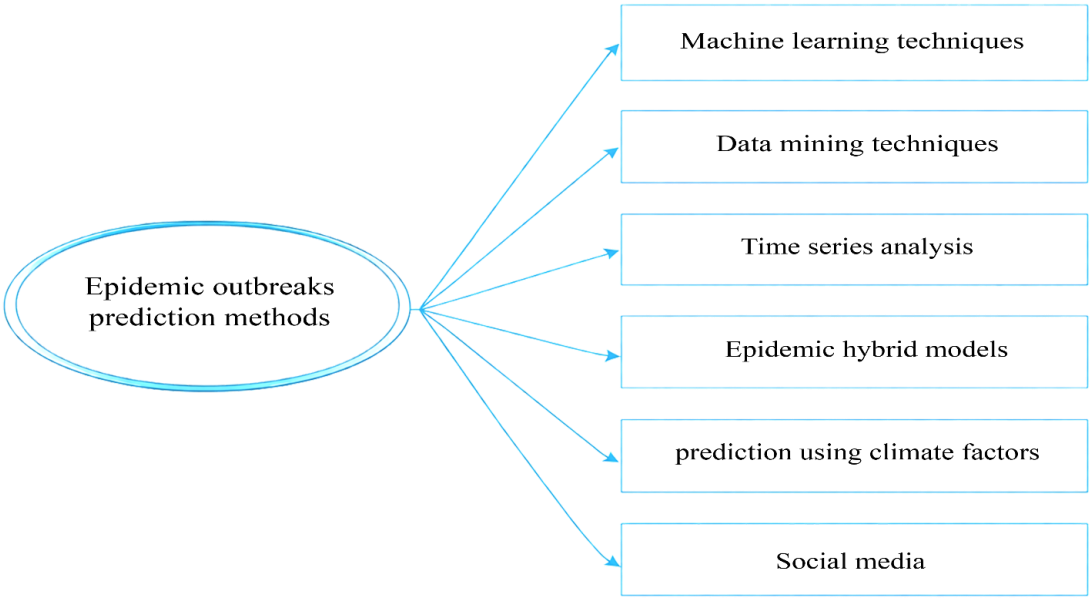


Figure 1.1: Epidemic Prediction Methods

1.2 Statement of the Problem

Recently, the incorporation of artificial intelligence and statistical approaches has expressively advanced the field of disease prediction and surveillance. Several studies have established the capability of these methods for simulate infectious disease dynamics, including malaria, by means of classification and regression approaches. Classification models are naturally used to categories whether an outbreak will occur, where regression models are considered to predict the number of cases. Though, a critical investigation of prior and current research reveals numerous consistent limits and methodological gaps that hinder the advance of robust, adaptable, and scalable malaria outbreak prediction structures.

Firstly, the popular of existing models rely on single technique approaches using either classification or regression without integrating together to deliver a holistic prediction framework. For instance, studies by Adigun et al. (2024), Kadam (2020), and Hulsen (2024) attained moderate performance (accuracy ~85-87%) but then failed to address count-based forecasting, which limits their usefulness for resources design and response

scalability. In the same way, regression concentrated models such as those by Baek (2023) and Musa (2015) accurately estimate incidence trends however do not detect out brakes thresholds, which are serious in outbreak classification and alert systems.

Secondly, even though hybrid and ensemble techniques have exposed improved accuracy in some studies, they are often limited to mixing classification models or regression models independently. Studies such as those by Adamu & Singh (2021) and Mahajan et al. (2023) show that ensemble methodologies like Random Forest + Gradient Boosting can enhance accuracy. On the other hand, they still treat classification regression tasks in isolation, failing to leverage the synergistic potential of hybrid classification regression frameworks that could simultaneously forest outbreak likelihood and case numbers (Adamu & Singh, 2021a; Mahajan et al., 2023).

In addition, a lack of methodological rigor in model selection and justification is also evident across the literature. Very few studies utilize multi criteria decision making (MCDM) techniques to systematically select and rank appropriate ML algorithms. This often result in ad hoc algorithm choice based solely on trial and error than objective evaluation metrics. Also, the lack of consideration for, multi-source and heterogeneous datasets, as noted in Adeyeye & Nkemnoye (2023) and Modu et al. (2017), impairs generalization and limits scalability across region with differing epidemiological and environmental profiles.

Regardless of developments in machine learning and AI for infectious disease prediction, current malaria forecasting models remain limited in scope and effectiveness. Utmost approaches emphasize separately on outbreak classification or case count regression without integrating these tasks into a unified hybrid framework capable of capturing both binary occurrence and continuous incidence trends. Primary models relying on SVM and ANN demonstrate limited adaptability and fail to leverage modern ensemble and deep learning methods effective in healthcare domains (Hulsen, 2024; Kadam, 2020; M. I. Musa, 2015). Hybrid data fusion strategies that combine climatic, environmental, and socioeconomic predictors remain underdeveloped, and real-time integration of

surveillance, clinical, and environmental data for early detection is rarely implemented (Al-Tameemi et al., 2024; Taffese et al., 2018).

As well, explainable AI and spatiotemporal models reveal complex feature interactions nonetheless seldom combine classification and regression within a single pipeline (Vidhaya, 2023). These gaps highlight the need for a comprehensive hybrid-based framework that integrates advanced ML techniques adaption to provide accurate, interpretable, and rebuts malaria outbreak predictions across diverse epidemiological settings. Besides, the approach should incorporate multi criteria model selection and leverage ensemble learning to improve generalizability, robustness, and decision support in public health response systems (Adeyeye & Nkemnole, 2023; Modu et al., 2017). This research mostly emphases on the insufficiency of current malaria prediction models in effectively integrating and analyzing diverse data sources, including climatic, spatial, and clinic data. A lot of the models we have now can only make single task predictions for either classification or regression), and they have a hard time adapting to the changing and complex ways that malaria spreads (Baek, 2023). This study introduces an innovative hybrid model that addresses these constraints by integrating the predictive capabilities of both classification and regression, thus providing a more resilient and precise approach for outbreak forecasting.

1.3 Research Objectives

1.3.1 General Objective

The general objective of this research was to develop and evaluate a hybrid based classification and regression model for predicting malaria outbreak.

1.3.2 Specific Objectives

To achieve the main objective, the research was directed by the following specific objectives:

1) To investigate the existing techniques and models for both classification and regression models that can be applied to predict malaria outbreak.
2) To design and develop a hybrid model that attends to combine the integration of classification and regression approaches to predict malaria outbreak.
3) To implement the proposed hybrid model for malaria outbreak prediction.
4) To evaluate the developed hybrid model using applicable performance metrics and compare it against other prior models.

1.4 Research Questions

1) What are the existing methods and models for both classification and regression that are currently applied to predict malaria outbreaks?
2) How can a hybrid model that integrates classification and regression approaches be designed and developed to improve malaria outbreak prediction?
3) How can the proposed model be implemented for effective malaria outbreak prediction?
4) How does the performance of the developed hybrid model compare to existing models when evaluated using appropriate performance metrics?

1.5 Justification

The rationalization for this study is grounded on the serious need to expand the accuracy and reliability of malaria outbreak predictions through advanced computational methods. Existing models that rely solely on classification or regression methods frequently inadequately capture the complex and nonlinear patterns associated with malaria transmission. This work is predicated on the expansion of a hybrid classification and

regression model planned to integrate the optimum features of both methodologies, henceforward enhancing predictive accuracy over the deployment of real-world data. This study pursues to support techniques, data scientists, and health informatics professionals worldwide in refining their analytical models whereas simultaneously providing valuable insights to worldwide organizations, such as the real-World Health Organization, to increase their present malaria surveillance and initial warning systems. The study will meaningfully enhance the competencies of public officials, predominantly within Ministries of Health, by refining the precision of malaria outbreak predictions, calculating potential caseloads, and facilitating timely, evidence-based interventions to decrease disease transmission. This study advances the area of computer science by growing the utilization of ML, particularly hybrid modeling in health care analytics, and lays the footing for future research to explore associated data mining and AI driven methods for epidemic forecasting. The research also urges AI experts and academics to work together and share their information. Academics and practitioners can apply it as a supportive instrument, mainly in places everywhere Malaria is common, including Burkina Faso, Mali, Republic Democratic of Congo, Niger, and the Nigeria, Cameroon. This research improves computational modeling for public health and offers a framework for the experimentation of hybrid machine learning methods in worldwide health initiatives.

1.6 Research Scope

This research is confined to the utilization of data mining and machine learning l=methodologies within the expansive field of artificial intelligence to create a hybrid classification and regression model for forecasting malaria outbreaks, the study examines a dataset of weekly meteorological valuables, specifically average maximum and minimum temperatures and rainfall, recognized as critical environmental factors affecting malaria transmission. This research focuses on six classification algorithms and four regression algorithms, chosen for their appropriateness and significance in the development of hybrid models, despite the existence of court less machine learning procedures. The study includes the design, implementation, and assessment of the hybrid model, with performance evaluated through standard metric like Accuracy, Root Mean

Square Error, Mean Absolute Error, Precision, Recall, F1 Score, F measure, and further validation via the Receiver Operating Characteristics curve(ROC) and AUC) Area Under the Curve. The implementation utilized the python programming language via the Scikit learn packages within the Anaconda integrated development environment, due to its robustness and appropriateness for the construction of machine learning and performance assessment of the hybrid model versus individual classification and regression models, to select the most effective model for malaria outbreak prediction.

1.7 Thesis Organization

This thesis is structured in five chapters, organized as follows, each of which is important to the scientific effort. **Chapter One** is the first chapter and gives a clear context, a problem to be solved, goal, reasons for doing the study, and its limits and scope. It also outlines what led the researcher to create a new algorithm for epidemic prediction using hybrid methodologies. **Chapter Two** provides an extensive foundation for the core principles governing classification and regression prediction models, hybrid models, and a literature review methodology. This chapter is important for setting the stage and getting a better grasp of what is already known about data mining. It does this by going over and briefly summarizing the present gaps and problems in the domain of predicting malaria outbreaks using classification and regression approaches. **Chapter Three** elaborates on the conceptual framework and the suggested model method, as well as some of the specific information used in the research and the dataset used to strain and test model. This chapter goes into great depth about the researcher's methodology, which helps readers understand how the research was done. **Chapter Four** is all about the findings of the study. It presents a comprehensive analysis, such as a comparison that assesses the performance and effectiveness of the new algorithm established by the researcher. **Chapter Five** wraps up the thesis by going over the work that was done and talking about possible future directions for research in this area. The thesis has an appendix that lists all author's publications from his PhD studies.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter presents the reviewed literature, which offers a general idea of key concepts interrelated to machine learning, data mining, outbreak disease, and the model development procedure. It presents a review of current study in regression and classification with their algorithms, as well as related works in outbreak prediction, including their limitations. In specific, this chapter emphases on the literature of regression and classification approaches, including their limitations and the need for a hybrid-based approaches. The chapter concludes with a summary of the reviewed literature.

2.2 Overview of Artificial Intelligence and Predictive Analytics

Artificial intellect is a large area of computer science that focuses on constructing systems capable of performing tasks that typically require human intelligence, such as perceiving, reasoning, learning, as well as taking decisions (Russell & Norvig, 2022). In the field of AI, predictive analytics has become one of the most important tools for getting information from data and making predictions about future trends based on past events. Predictive analytics includes a variety of statistical, computational, and algorithmic methods that can be used to model how variables are related, find patterns, and predict future outcomes with measured accuracy (Grignaffini et al., 2024). In health informatics, AI-Driven predictive analytics serves as a basis for developing intelligent systems capable of monitoring, classifying, and forecasting disease outbreaks, hence facilitating data-driven public health interventions and strategic planning.

2.2.1 Artificial Intelligence

Artificial Intelligence has undergone substantial development over time, from symbolic reasoning and expert systems in the 1960s and 1970s to machine learning and data-driven

algorithms that are being used in modern AI applications (Al-Tameemi et al., 2024). In the past, AI systems used rule-based logic and knowledge representation, which meant that they were programmed with expert knowledge to do reasoning tasks. But when it came to dealing with complex, high-dimensional data, these solutions weren't very scalable or flexible. The rise of machine learning, especially throughout the 1990s and 2000s, signified a transformative transition from rule-based systems to data-driven intelligence. ML lets systems discover patterns from data on their own, without having to write code for them. This makes them better over time (S. Khan & Shaheen, 2023). Deep learning and reinforcement learning have made modern AI even more powerful. These technologies let models work with unstructured data, develop hierarchical representations, and make judgments on their own (Alam & Singla, 2020). These advancements have propelled applications in natural language processing, computer vision, and health analytics. When it comes to predicting malaria, AI gives scientists a way to make computer models that can find trends in outbreaks by combining data on climate, environment, and disease spread. These kinds of systems work based on the AI ideas of perception for getting data, reasoning for making inferences from models, and action for predicting outbreaks and helping people make decisions.

2.2.2 Predictive Analytics

Predictive analytics is a subfield of Artificial Intelligence and data science which utilizes statistical and computational models to make predictions about what will happen in the future based on current and past data (S. Khan & Shaheen, 2023). It uses machine learning, data mining, and statstical inference techniques to find patterns, discover connections, and make predictions. Predictive analytics is different from descriptive and diagnostic analytics since it looks ahead instead of only describing what has already happened or figuring out why it (Maryoosh & Hussein, 2022). Predictive analytics has been employed in health care and epidemiology to foresee disease outbreaks, hospital admissions, and patient readmissions. Predictive models can give public health officials useful information by using a variety of data sources, such as electronic health records, environmental data, and demographic data. Predictive analytics is an important tool for predicting malaria

since it can help figure out how likely an outbreak is, find areas that are at risk, and use resources wisely. The prediction process usually involves data preprocessing, selecting relevant feature, model training, evaluation, and deployment. Each stage follows the rules of AI-based learning.

2.2.3 AI and Predictive Analytics in Healthcare

Artificial Intelligence and predictive analytics have changed healthcare forever by making precision medicine, automated diagnostics, and real-time epidemic surveillance possible (Topol, 2019). AI models like decision trees, support vector machines, neural networks, and ensemble algorithms have been utilized to forecast things like how likely it is that a patient will survive or how an outbreak would spread. These models surpass conventional statistical techniques by encompassing nonlinear interactions, high-dimensional dependencies, and intricate temporal dynamics (Schaffer et al., 2023). Predictive analytics models use environmental indicators, epidemiological factors and climatic variables such as rainfall, temperature, and humidity to predict the like hood and security of malaria outbreaks. AI-powered systems can combine these many types of data sources through feature level fusion, which makes it possible to make full predictions. This skill fits with the data-centric way of thinking about AI, which stress learning from patterns in different types of data (Liu et al., 2018). Predictive analytics in AI frameworks use both classification (to find outbreaks) and regression to estimate case counts to give a multisession view that helps with early warning systems and making public health policies.

2.2.4 Integration of AI and Predictive Analytic in Public Health Surveillance

AI-driven predictive analytics frameworks constitute the foundation of contemporary disease surveillance systems. They offer predictive insights that enable decision-makers to intervene promptly, hence reducing the spread and effect of epidemics (Silhavy, 2023). For malaria, these systems use machine learning and data mining to find epidemic thresholds, spot new trends, and send out alerts in real time. By adding AI to predictive

analytics, models can train all the time, which means that systems can change as new data comes in (Mahajan et al., 2023).

The suggested hybrid approach in this study is located at the crossroads of AI and predictive analytics. It uses the ideas behind machine learning to bring together classification and regression models into one framework for predicting malaria outbreaks (Furkan et al., 2023) . This integration and regression models into one framework for predicting malaria outbreaks. This integration improves both categorical identification for Outbreak vs. non-outbreak and quantitative forecasting for number of instances (Maturana et al., 2023). This makes for a complete early warning system based on AI-driven predictive intelligence. This type of hybrid system fits with the bigger goal of intelligent public health informatics, which is to use AI to make decisions based on evidence and stop the spread of disease.

2.2.5 Pyramid of Data, Information, Knowledge, and Wisdom

The DIKW Pyramid, which stands for Data, Information, Knowledge, and Wisdom, is a foundational conceptual framework in information science and computing that describes the hierarchical transformation of raw inputs into actionable insight, particularly useful in healthcare analytics (Mirzaeian et al., 2023) .AI the base of pyramid is data, which comprises raw, unprocessed facts and figures without context or meaning. In malaria prediction, this may include temperature readings, rainfall levels, or case counts (Masinde, 2020). When data is processed, structured, or organized to answer basic questions such as "what," it becomes information, for instance, linking specified rainfall patterns to higher mosquito breeding rates (Ma, 2025). Knowledge is delivered by interacting and analyzing information to answer "how" and "why" questions such as understanding how specific climatic or demographic patterns affect malaria transmission.(Traini & Lombardi, 2022).

In the context of this thesis, the hybrid machine learning model facilitates the information from raw epidemiological and environmental data into meaningful knowledge that supports decision making in malaria outbreak management. By structuring the analytic

process through the DIKW Hierarchy, the research aligns with a broader knowledge from raw data (Maturana et al., 2022). However, there is a more indicate hierarchy behind these concepts, which is known as the Data Information Knowledge Wisdom (DIKW) hierarchy or pyramid in the knowledge management field. This pyramid was introduced by Maturana et al. (2023) and is shown in Figure 2.1

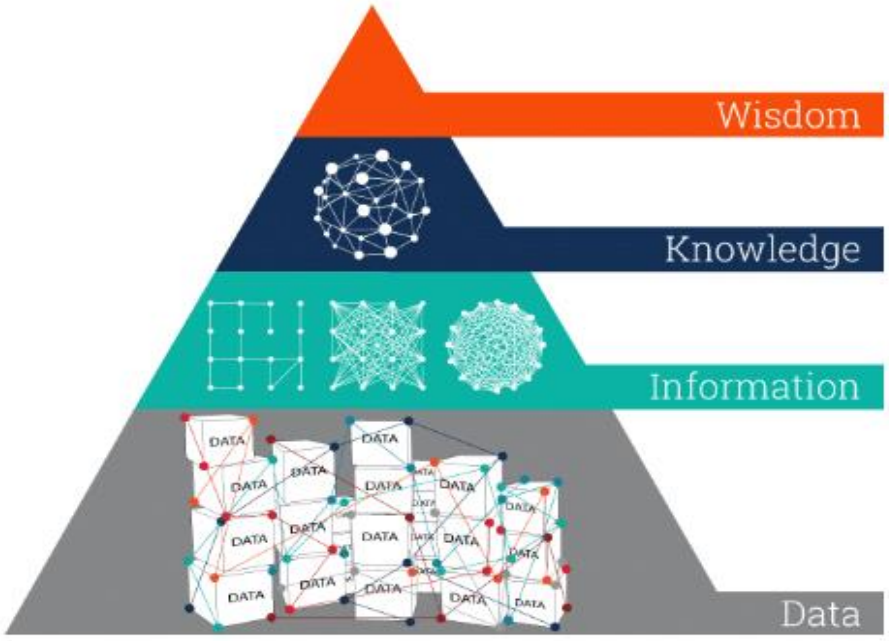


Figure 2.1: Pyramid of the Data Information Knowledge Wisdom Hierarchy

2.3 Foundation of Machine Learning and Data Mining

Machine learning and data are basic ideas in current predictive analytics. They are especially important when it comes to predicting malaria outbreaks and other health problems. Machine Learning is a general term that refers to the way computers can learn patterns and make predictions or judgments based on data without being specifically taught to do so (Grignaffini et al., 2024). Machine learning includes a variety of algorithms, such as supervised, unsupervised, and reinforcement learning, that have been shown to work well for identifying health disorders and predicting how diseases will spread using both historical and real-time data (Plotnikova et al., 2020). Machine Learning

is especially useful for predicting malaria because it can do both classification and regression. This makes it a good choice for a hybrid-modeling framework (Kiliç & Karakoyun, 2023).

Data mining is the act of hiding useful patterns, correlations, and trends in huge datasets using statistical and computer methods (Leo et al., 2019; Uddin et al., 2019). Machine learning is all about making predictions, while Data mining is all about finding new information. This makes Data mining perfect for looking at epidemiological datasets to find hidden variables that affect the spread of malaria, like climate, vector density, and population movement (Henzler et al., 2025) . These two fields work well together. Data mining usually does the prediction and decision making. When you put them together, they make a strong framework for creating smart systems that can solve difficult healthcare prediction problems, like predicting when malaria would break out.

The development of data mining and machine learning has had a big impact on predictive analytics in healthcare, especially when it comes to predicting epidemics of infectious diseases like malaria. Machine learning developed from conventional statically modeling into an advanced framework for automated pattern detection and predictive modeling (Zhu et al., 2020). Machine Learning Techniques were widely integrated into data mining operations, strengthening the potential for learning from historical data to produce real time prediction (Leo et al., 2019). As big data and open-source libraries grew, supervised learning models including decision trees, support vector machines, and neural networks became common in health analytics.

They were even used to find malaria and predict outbreaks (Nyambura et al., 2025). The emerging of data mining and machine learning has also made it possible to create hybrid systems that use both classification and regression methods. These systems are the basis for models like the one described in this thesis. In public health situations where data is noisy, fragmentary, and diverse, these kinds of hybrid methods help with generalization and accuracy, which are both very important. This historical trend emphasizes the need of

merging data mining's pattern discovery capabilities with machine learning's predictive strength in constructing reliable malaria forecasting systems.

Data mining first appeared in the 1990s as a subsection of knowledge discovery in databases (KDD). Its goal was to find hidden patterns in huge and complicated datasets (Tharageswari et al., 2025). Early data mining techniques utilized statically models and rule-based systems on health records and epidemiological data to identify relationships between environmental factors and disease occurrence and processing power empowered over time, malaria monitoring rule mining, and anomaly detection (Azezew et al., 2025).

In accordance with specific research, data mining has emerged as an established discipline within the realm of computing science, with its origins tracing back to the late 1980s when the term was first introduced in the research community (Cheohen et al., 2025). Early in the 1990s, data mining was acknowledged as a sub process or a step in the larger process of Database Knowledge Discovery. In the 2000s, data mining gained popularity as a powerful technique for uncovering previously undiscovered patterns and valuable insights for large datasets. The evolution of finding appropriate data for decision making began 30 years ago and has progressed through various stages of development (Islam et al., 2022; Maturana et al., 2022; Poostchi et al., 2018), as illustrated in Figure 2.2.

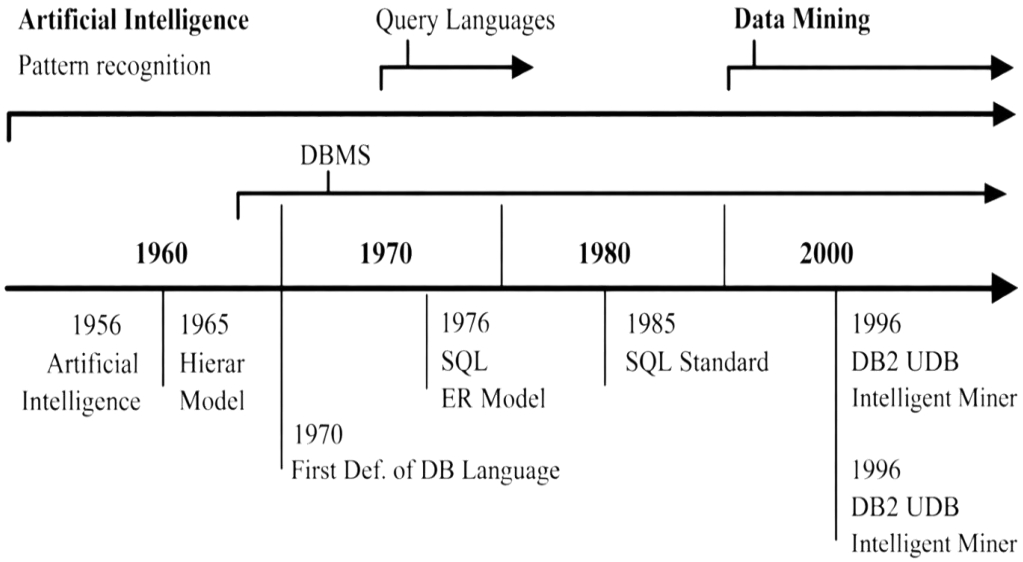


Figure 2.2: Data Mining Evolution

Furthermore, most of scholars have emphasized that the progression from business data to business information has been an incremental process, with each new step building upon the previous one. From the users' perspective, the four steps outlined in Figure 2.3 were groundbreaking, as they enabled the accurate and trepid answering of new business questions. For decades, researchers in domains such as statistics, artificial intelligence, and machine learning have been working on data mining technology. According to Chandra et al. (2019), the maturity of these techniques, in combination efforts, makes these technologies practical for current data warehouse environments.

The present advancement of data mining functions and tools is the culmination of years of influence from a variety of disciplines, including databases, information retrieval, statical analysis, and machine learning. In addition, Multimedia and graphics have had a significant impact on the knowledge Discovery in Database process. Therefore, data mining overlaps with various fields, and is situated at the point of intersection of statistics, machine learning, and databases, as observed by Alsagiri (2023) in Figure 2.3.

Figure 2.3: Intersection of Data Mining with Different Disciplines

The majority of public and private institutions have accumulated a vast amount of computational data through their operations and activities, which can be difficult to extract knowledge from without effective tools. Data mining is one such tool that allows for the discovery of hidden relationships, patterns, and rules within large datasets, enabling institutions to make faster and more confident decisions (Ahmed & Salah, 2023; Alsajri et al., 2023). Data mining is an interdisciplinary field that involves the extraction of predictive datasets and identify relevant information. Despite the fact that the phrases data mining and knowledge discovery in databases (KDD) are occasionally employed interchangeably, some researchers consider data mining to be a specific stage within the larger KDD Process (Salman et al., 2024) .

According to Maryoosh and Hussein (2022), data mining combines statistical and artificial intelligence-based methods that are coupled with database administration. Despite its straightforward concept, data mining is a challenging field that requires a thorough understanding of data. As Ahmed et al. (2018) in the Figure 2.4 noted that data mining involves searching for patterns within large datasets, which can be compared to gold mining in rivers, where gravel represents a large volume of information and gold nugget are the hidden pattern to be uncovered.

Figure 2.4: Data Mining as Gold Mining in Rivers

Data mining is a discipline which employs advancements in artificial intelligence and statistics. Both artificial intelligence and statistics have been addressing issues related to pattern recognition and classification artificial intelligence techniques for reasoning, particularly techniques for dealing with uncertainties to classic density estimation in statistics (V. Ahmed et al., 2018; Maryoosh & Hussein, 2022). These techniques allow prior knowledge about the domain and data to be involved in a relatively easy and natural framework.

Nevertheless, Data mining techniques originating in artificial intelligence have mainly focused on dealing with symbolic or categorical data, with limited consideration of continuous variables. In contrast, algorithms for classification and clustering in machine learning and case-based reasoning have focused profoundly on heuristic search and nonparametric models (Maturana et al., 2022). Emphasis on scientific rigor and analysis of outcomes has not been as strong in DM as it has been in statistics or pattern recognition, except computational learning theory, which has focused on proper general worst-case bounds for a broad range of illustrations.

2.3.1 Categories of Machine Learning

ML techniques are generally classified into four main types: Supervised, Unsupervised, semi supervised, and reinforcement learning, respectively offering unique advantages for different tasks in malaria outbreak prediction, as detailed below

Supervised learning is a category of machine learning that focuses on learning a purpose plotting inputs to an output built on sample output-output pairs. It requires labeled training data and a gathering of training instances to a purpose learning is employed when specific aims are recognized to be attained from a set of inputs, and it is often applied for classification and regression modeling (Rajkomar et al., 2019). This technique includes training models on labeled datasets to learn the mapping among input features and results. Algorithm, including decision trees, support vector machines, and random forests have been extensively useful for malaria diagnosis and forecasting with considerable success (Olushola & Mart, 2022; Ashtagi, 2024).

Unsupervised learning is a category of machine learning involves earning, datasets without labels are analyzed without the need of human intervention. This method is commonly used for exploratory objectives, including detecting relevant structures and trends, clustering, density estimation, feature learning, dimensionality reduction, association rule discovery, and anomaly detection. It deals with unlabeled data, aiming to uncover hidden patterns or groupings (Kiliç & Karakoyun, 2023). Techniques like k means clustering and principal component analysis have been applied to determine risk clusters or environmental patters associated with malaria outbreaks (Chandra Patel et al., 2019).

Semi supervised learning represents a category of machine learning that combines both supervised and unsupervised methods, and it works with both data that has been tagged and data that has not been labeled. Semi supervised learning is valuable once labeled data are scarce but a substantial quantity of unlabeled data is available. A model of semi supervised learning's purpose is to give better predictions that those produced using

labeled data alone. Machine transition, fraud detection, data learning, and text classification are examples of semi supervised learning applications. It influences both labeled and unlabeled data, and has gained attention in situations where labeled health data is scarce but large volumes of unlabeled data exist. This approach enhances model accuracy while minimizing annotation costs (Kadam, 2020).

The last category is reinforcement learning, which is a category of machine learning that permits software managers and machines to robotically analyze optimum behavior in a exact situation or environment to improve efficiency. The aim of this kind of learning is to apply perceptions gained from environmental interactions to act to improve the reward or decrease the danger. Reinforcement learning is a valuable method for training AI models for activities like robotics, autonomous driving, manufacturing, and supply chain logistics; however, it should not be used to solve simple problems. Although less frequently applied in malaria prediction, it is has shown potential in optimizing intervention strategies and resource allocation by learning from sequential decision-making environments (Adamu & Singh, 2021a; Poostchi et al., 2018). Understanding and selecting the appropriate ML paradigm is crucial to developing a hybrid model that merges classification and regression to effectively malaria outbreaks prediction.

2.3.2 Tasks and Techniques for Machine Learning

This provides a detailed summary of machine learning approaches that can be utilized to enhance the intellect and overall competences of data-driven applications across various domains. By outlining the fundamental approaches and learning paradigms, it highlights how these methods enable systems to automatically identify patterns, extract meaningful insights, and make informed decisions with minimal human intervention (Ardabili et al., 2020). Such techniques form the foundation of modern predictive analytics and intelligent systems, supporting applications in healthcare, finance, environmental monitoring, and many other fields where large volumes of data are generated and analyzed (Leo et al., 2019). Moreover, Figure 2.5 illustrates a typical machine learning–based predictive model structure, which consists of two main phases.

In Phase 1, the model is trained using previously collected or historical data, permitting it to learn underlying patterns, relationships, and trends inside the dataset. During this stage, the algorithm adjusts its internal parameters to diminish errors and enhance predictive accuracy. In Phase 2, the trained model is used to new or unobserved test data in order to produce forecasts or outcomes. This two-phase process training followed by testing or deployment confirms that the model can oversimplify its learned knowledge to real-world data situations (Carta, 2022).

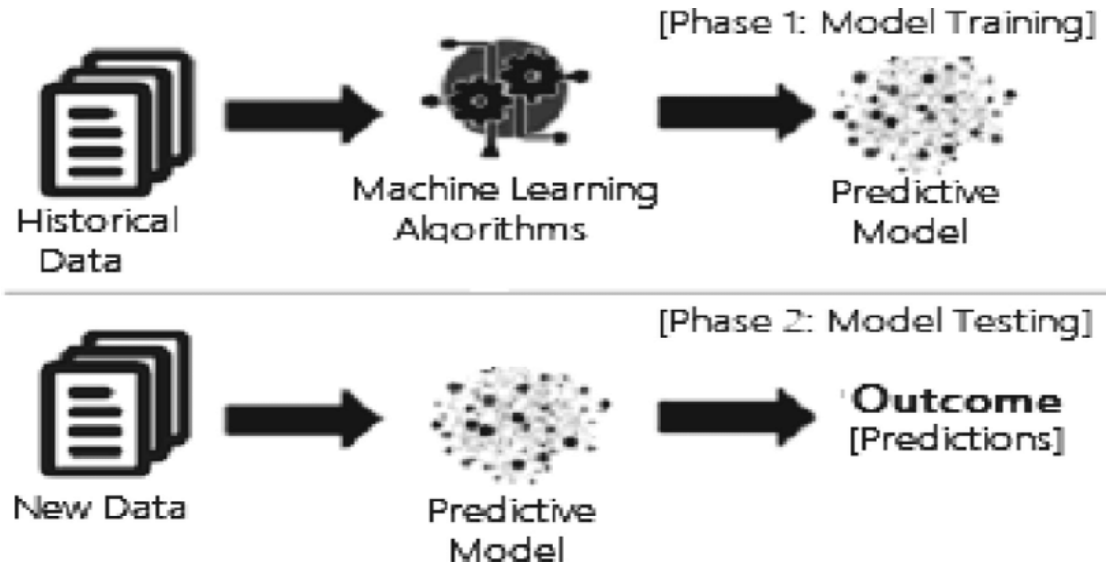


Figure 2.5: General Structure of a Machine Learning Based Predictive Model

2.3.3 Classification Techniques

The classification part of a hybrid based classification and regression model for predicting malaria outbreaks is very important because it helps determine if an outbreak is likely to happen in certain environmental and epidemiological conditions (Khan et al., 2024). In previous studies, different machine learning classification techniques have been used, each with its own pros and cons that effect how cell they work for hybrid modeling in malaria prediction systems (Sarker, 2021). This section introduces the concept of classification techniques, which are necessary for the thesis. Numerous research studies have confirmed that classification is a machine learning as a supervised learning strategy

that involves predicting a class label for a given example. It maps a function (f) mathematically from input variables (X) to output valuables (Y) as a target, label, or grouping (Sheth et al., 2022). Classification can be employed to predict weather a particular data item belongs to structured or unstructured data. An example of classification is spam detection in email service providers, where email is classified as either “spam” or “none spam”. Several classification techniques have been purposed for use in machine learning and big data analytics, and the next is a list of the utmost extensively used and general approaches in many application areas.

2.3.3.1 Logistic Regression

Logistic Regression is a statistical technique that has been utilized for a long time, but it also works as well as a binary classifier for predicting malaria outbreaks. LR models yield prophylactic output, rendering them highly interpretable and beneficial for public health interventions. Huang (2016) showed that LR Could use demographic and environmental features to model binary outcomes, like an outbreak or no outbreak. In a hybrid setting, LR frequently functions as a baseline or Meta classifier owing to its straightforward implementation and capacity to calibrate the outputs of other models (Adigun et al., 2024; Sim et al., 2023). LR presumes a linear correlation between features and the log odds of the outcome, which may constrain its efficacy in the presence of intricate, nonlinear patterns.

Linear regression, widely used in predictive modeling, takes its simplest form is known as simple linear regression, it is represented through a straight line. The regression line is optional in that it minimizes the total squired error of prediction. Multiple linear regressions, which accommodates at least two indecent or predictor variables, is a more general form of linear regression. It is based on the next expectations: a) The variation in the reply variable produced via every predictor variable is linear in nature, b) the properties of different predicator variables on the reply variable are additive, and c) the effect of any specific predictor variable on the reply variable is self-directed of other predictor variables

(Vidhaya, 2023). These expectations must be verified to validate the model, as violation of any of these assumptions renders the model unsuitable and results in unfortunate results.

For instance, multicollinearity arises if the properties of certain predictor variables on the variable answer are not autonomous of other predictor variables. Auto correction arises if the predictor variables themselves have dependent annotations. Heteroskedasticity arises of the variability of the response changes over its range. These problems can be addressed through various techniques, such as ridge regression, lasso regression, or regularization. Logistic regression, in contrast, is a usually used machine learning method that is primarily employed for classification tasks (Huang & He, 2016). It makes use of the probability notion to forecast the outcome of a categorical dependent valuable. Although logistic regression and linear regression have certain similarities, logistic regression is used to handle classification matters, whilst linear regression is employed for solving regression related issues. The equation for logistic regression is provided below:

$$\log\left(\frac{y}{1-y}\right) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n \qquad (2.1)$$

Where *y* is a categorical or discrete value, and the probabilistic values lie between *0* and *1*.2.3.3.2

2.3.3.2 Random Forest Classifier

Random Forest, an ensemble method based on numerous decision trees, has been better at classifying malaria cases because it is strong and can even lower variance by averaging. RF models have proven effective in malaria risk mapping, surpassing single tree models in predictive accuracy (Kadam, 2020). Adigun et al. (2024) integrated RF into a hybrid prediction model, observing enhanced accuracy in forecasting seasonal outbreaks. The method also ranks features by how important they are, which makes it easier to understand hybrid models. RF models are strong, but they can be expensive to run when working with big dataset or a lot features. Random forest is a famous algorithm for supervised learning

that may be employed in machine learning for regression besides classification issues. It is depended on a notion of ensemble learning, which includes merging several classifiers to solve complex problems and enhance the model's performance (Hall & Lucas, 2023)

According to Du et al. (2025) stated that the random forest is an algorithm for classification that is delivered from decision trees. It is consisting of various decision trees in which respectively tree provides a classification for the input data, and the forest combines these classifications to produce the final prediction. The input of respectively tree is sampled data from the unique dataset, and a subset of variables is randomly chosen from the available features to grow the tree at each node. Furthermore, separately tree is full-grown without pruning. The primary objective of random forest is to leverage a large quantity of weak or weakly correlated classifiers to construct a strong classifier. The more illustration is detailed on the figure 2.6 below and also provides a better understanding of the Random Forest algorithm (Du et al., 2025).

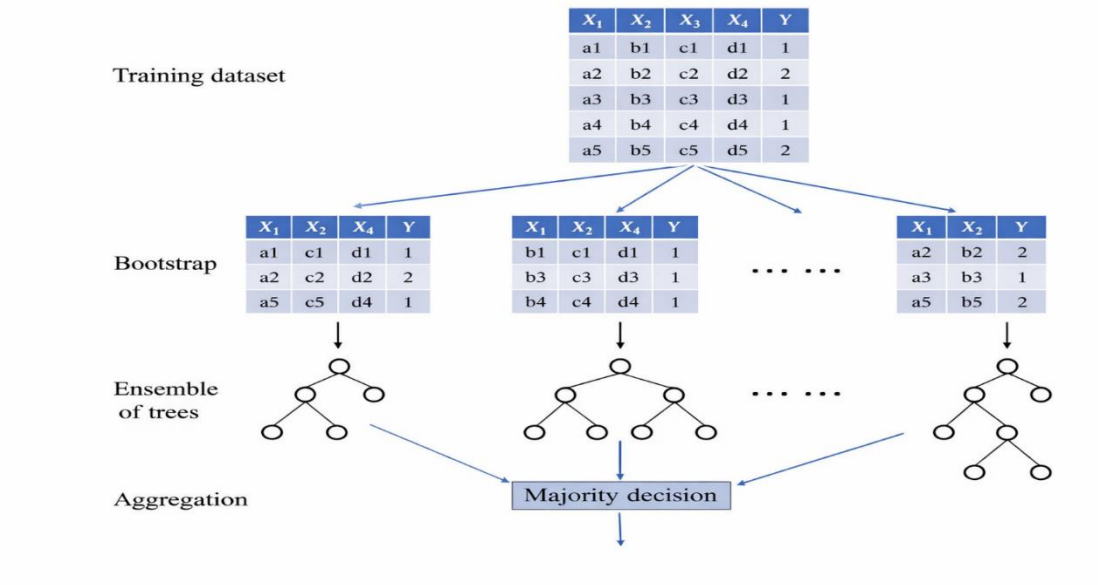


Figure 2.6: Random Forest

In the given context, the collective decision of multiple decision trees is inherently less noisy and less sensitive to outliers than a single decision tree output, which reduces the

instability resulting from small data and enhances the resilience of prediction (Franky et al., 2020). Consequently, the following additional points elucidate the rationale for employing the Random Forest algorithm:

- It exhibits faster time of training compared to extra algorithms.
- It facilitates high prediction accuracy, even for huge datasets, and operates resourcefully in such scenarios.
- It sustains accuracy when a substantial portion of data is missing.

2.3.3.3 Decision Tree

Decision Tree classifiers are commonly employed for disease classification tasks owing to their comprehensibility and capacity for managing both categorical and continuous data. In malaria prediction, decision trees have been utilized to construct decision rules predicated on climatic features such as precipitation, temperature, and humidity, providing clear decision-making frameworks that health officials can readily comprehend. For instance, Kang (2022) illustrated the efficacy of decision trees in diagnosing malaria utilizing clinical and environmental data, whereas Singh et al. (2021) employed decision trees to model seasonal malaria trends across Indian states. Moreover, Islam et al. (2022) discovered that decision trees (DTs) are useful for initial classification tasks before using ensemble methods. One big problem with DTs is that they can over fit, which makes them less reliable when used alone on data they can over fit, which makes them less reliable when used alone on data they haven’t seen before.

As per Kadam (2020), decision trees are a popular and widely employed non parametric supervised learning technique for classification and regression determinations. To classify cases, it adopts a structured hierarchical framework of conditions. In the domains of machine learning and data mining, DT is employed as a forecasting tool. It uses a collection of trees organized decision tests in a divide and conquer strategy, showed in Figure 2.7.

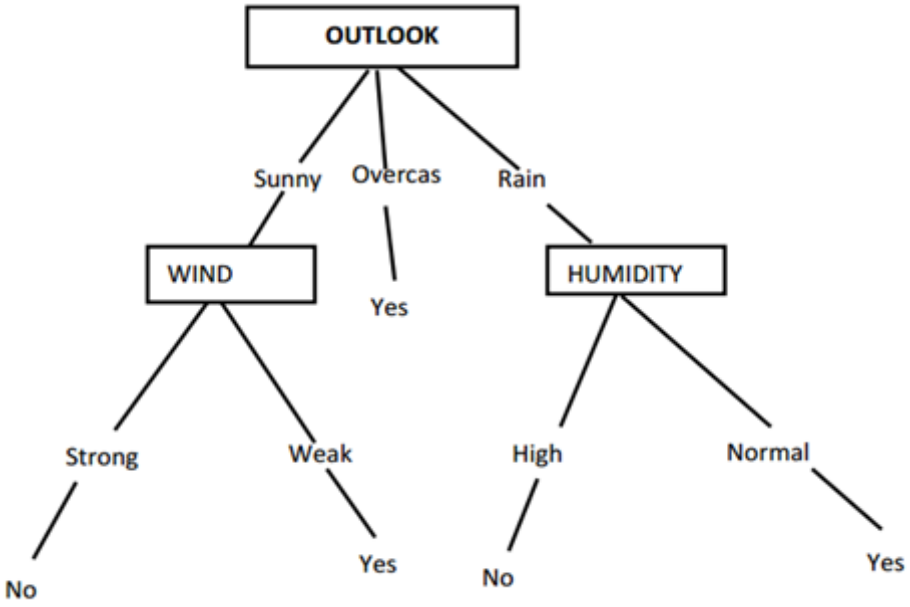


Figure 2.7: Simple Decision Tree

Each non leaf node in a decision tree is connected with a variable test, also known as a split. This is because data failing into the node is divided into diverse subsets founded on their values on the variable test. On the other hand, each leaf node is associated with a label, which is allocated to cases failing into the node. Decision tree learning processes follow a recursive process where, in separately step, a dataset is given and a split is designated. This split formerly applies to separate the dataset into subsets, and every subset is measured as the specified dataset for the following step (Vidhaya, 2023). The choice of separations is a key method in the decision tree algorithm. In the ID3 algorithm, (Yang et al., 2019) stated that the information gain principle is utilized for split selection. Assumed a training set D, the entropy of D is well-defined as:

$$\mathrm{Ent(D)} = -\sum_{y \in y} \mathrm{P(y \backslash D) log\ P(y \backslash D)} \tag{2.2}$$

If the training set D is separated into subsets D1. DK, the entropy may be reduced, and the amount of decrease is mentioned to as information gain, i.e.,

$$G(D; D_{1, \ldots, D_k}) = Ent(D) - \sum_{i=1}^{k} \frac{|D_k|}{|D|} Ent(D_k) \quad (2.3)$$

If the decision tree structure, the selection of the indicators value pair that results in the largest information gain is crucial for splitting the tree. However, a significant drawback of the information gain standard is that it tends to favor features with many possible values, irrespective of their relevance to classification (Yu et al., 2018). For instance, in a binary classification problem, if the resulting information gain would be considerable, as this split would correctly classify all training instances. However, such an approach would not generalize well and would be inappropriate for predicting unseen instances to address this limitation, the C4.5 decision tree algorithm introduces the gain ratio, which is a variation on the information gain criterion that normalizes the number of feature values (Hall & Lucas, 2023). Specifically, the gain ratio considers the number of varies a feature can take, ensuring that the criterion is not biased towards features with May values. In practice, the variable with the uppermost gain ratio, between these with better than average information gains, is chosen for the split in the algorithm of decision tree.

$$P(D; D_1, \ldots, D_k) = G(D; D_1, \ldots, D_k) \cdot \left(- \sum_{i=1}^{k} \frac{|D_k|}{|D|} \log \frac{|D_k|}{|D|} \right)^{-1}, \quad (2.4)$$

It is commonly considered that a decision tree that performs flawlessly on the training set may have inferior simplification capability than a tree that performs less well on the training set. This phenomenon, stated to as overfitting, may be attributed to the fact that the learner erroneously interprets certain particularities of the training data, such as those stemming from noise during data collection, as the primary truth (Ramageri, 2010). To mitigate the danger of overfitting, a common approach is to use pruning methods to remove tree branches during the tree growth phase, while post pruning involves re-examining mature trees to determine which crunches to remove. When a validation set is available, the tree can be pruned based on validation error. Specifically, in pre pruning, a branch will not be grown if doing so would increase the validation error, while in post

pruning, a branch will be detached if its removal would reduce the validation error (Maturana et al., 2023).

Decision trees have several advantages, including their simplicity in the interpretation, aptitude to grip both numerical and categorical data with title data requirement, statistical validation of models, and robustness to violations of assumptions. They are also fast and efficient for huge datasets, but the accuracy of the tree's predications heavily depends on the accurate of the input data. Even small changes in input data can result in significant changes to the decision tree structure, which may require redrawing the tree. Furthermore, Bardab (2021) pointed out the challenges in using decision trees for disease analysis as presented in Table 2.1

Table 2.1: Decision Tree Accuracy Metric

Disease Considered	Author	Year of Publication	Accuracies in DTrees
Heart	Cheung	2001	81.11%
Skin Diseases	Bojarczuk	2001	89.12%
Breast Cancer	Dursun Delen et al.	2005	93.62%
Heart	Andreeva, P	2006	75.73%
Breast Cancer	Bellaachia et al.	2006	86.70%
Heart	Palaniappan et al.	2007	94.93%
Heart	Sitar-Taut et al.	2009	60.40%
Heart	Tu et al.	2009	78.90%
Skin Diseases	Polat and Gunes	2009	96.71%
Heart	Asha Rajkumar et al.	2010	52.00%
Heart	Jyoti Soni	2011	99.20%
Heart	Akhil Jabbar et al.	2012	80.00%
Heart	Abhishek Taneja	2013	94.29%
Liver	Syeda Farha Shazmeen et al.	2013	69.58%
Kidney	K R Lakshmi et al.	2014	78.44%

2.3.3.4 Naïve Bayes Classifier

Native Bayes Classifiers have demonstrated efficacy in contexts where computational efficiency and scalability are paramount, despite their reliance on the premise of feature independence (Bardab et al., 2021). Hailu (2015) utilized Naïve bayes to classify malaria cases in real time with mobile health data, observing that its simplicity did not markedly diminish accuracy. Musa et al. (2024) conducted a study demonstrating that NB surpassed

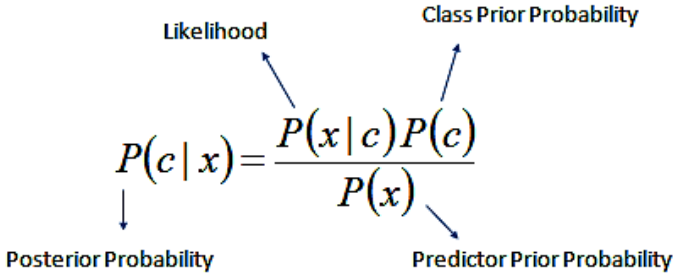
various more intricate models when utilized on highly structured, low noise datasets. Maryoosh & Hussein (2022) also used NB in a hybrid ensemble model that made it easier for rural health centers to send out early warnings. Still, the independence assumption makes NB less useful when the input features are much correlated, like climate data.

The Theorem of Bayes is applied in the Nave Bayes classification method under the assumption that each pair of features is independent. This algorithm is appropriate for both binary and multi class classifications problems in various real-world scenarios (Hailu, 2015). Naïve Bayes classifiers are particularly useful in accurately classifying instances in noisy data and constructing robust prediction models. This classifier is founded on the restricted independence model of individually predictor given the target class. The primary value of Naïve Bayes is to give the class a case with the uppermost posterior probability. Therefore, it purposes as a probabilistic classifier of individual attributes in a data sample and subsequently classifier data issues. The key advantage of Nave Bayes is that in contrast to more advanced algorithms, it just needs a slight amount of training data to quickly estimate the relevant strictures. However, the storing assumption may have an impact on its performance of feature independence. According to Nasser et al. (2022), Naïve Bayes is a classification technique that relies on Theorem of Bayes, with an assumption of independence between predictors. Further, the Naïve Bayes classifier adopts the presence of a precise variable in a class is independent of the presence of any supplementary characteristic.

How Naïve Works

The Naïve Bayes algorithm is a classification method grounded on Theorem Bayes, under the assertion of autonomy between predictors. This strategy can be used to binary and multi class classes in a variety of everyday scenarios. One of the advantages of NB: it only takes a little quantity of training data to estimate the required parameters rapidly. Its performance, however, may be altered by its strong assumptions on the independence of features, the Naïve bayes algorithm works is illustrated below in equation 2.5 (F. Chen et al., 2015; Nasser & Behadili, 2022).

How a Naïve Bayes Algorithm Works



(2.5)

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

When *P(c|x)* is the class's posterior probability (c, target), given predictor (x, attributes). *P(c)* denotes the class's prior probability, *P(x|c)* denotes the like-hood, which represents the like-hood of a predictor provided a specific class, *P(x)* shows the predictor's prior probability.

The Naive Bayes (NB) algorithm relies on the Bayes theorem and assumes that every pair of attributes is independently. It performs effectively and may be applied in a variety of real-world scenarios in both binary and multi class groups (Masiira et al., 2019). The key advantage is that, in contrast to more complex algorithms, it uses a small amount of training data to quickly estimate the essential parameters. However, its performance may suffer as a result of its stringent assumptions on feature independence. Given in figure 2.8 is a comparison of various disease prediction results based on naive bayes technique (Mohsen & Alhurdi, 2025).

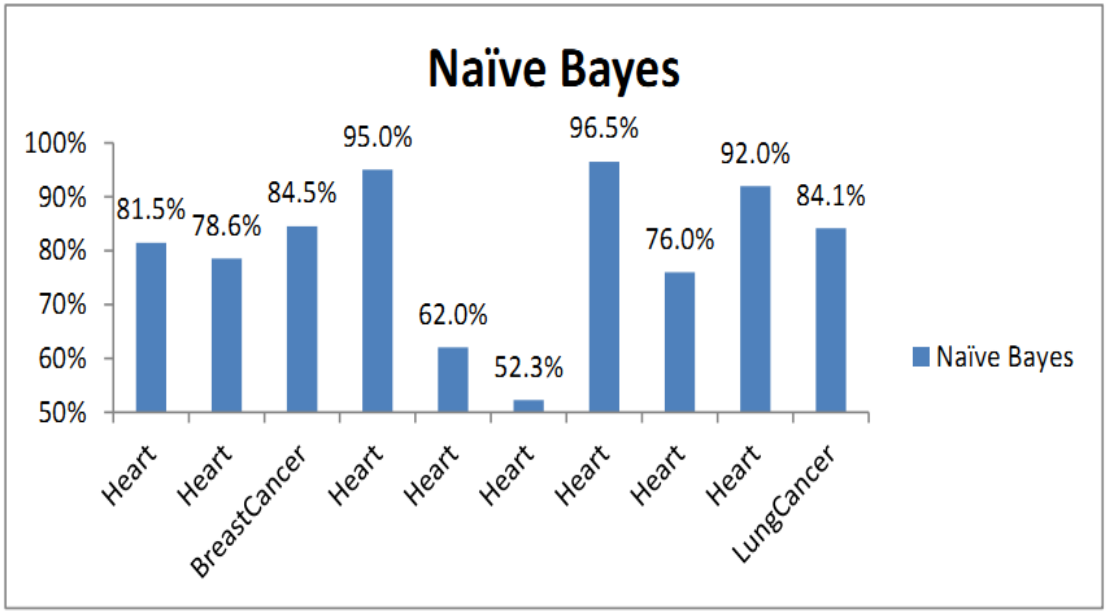


Figure 2.8: Comparisons of Various Disease Prediction Results Based on Naive Bayes Technique

2.3.3.5 The K Nearest Neighbor's Algorithm

The K Nearest Neighbor algorithm is a form of supervised machine learning that is commonly utilized for prediction problems. The algorithm determines the distance among the test data and the input and then makes a prediction founded on the k nearest neighbors (Modu et al., 2017) . KNN is a learning method for pattern recognition and categorization that is used to determine the class to which a new input as the test value is included when k nearest neighbors is chosen. The procedure aims to approximation the conditional distribution of Y given X and categorize a particular observation (test value) as belonging to the class with the highest estimated probability. KNN first picks the k training data points that are closest to the test value and calculates the distance between all of those classes. The test value belongs to the class with the shortest distance.

K NN is a category of instance-based learning that can be applied in regression analysis, and the algorithm splits the space into regions based on the training samples' positions and

labels. If class label c is the most common between the k nearest training samples, a point in the space is assigned to it. Naturally, the Euclidean distance measure is used as the distance measured; but this only works with numerical data. In settings for instance text categorization, another metric, such as the overlap metric or the hamming distance, may be utilized. The KNN technique operates on the supposition that items that are similar in the input space are also alike in the output space. According to Bardab et al. (2021) KNN classifiers are intended to learn by equivalence. The training samples are signified by n dimensional numeric assets, with each sample in place of a point in an n-dimensional space. All training samples are thus stored in an n-dimensional pattern space. Once obtainable with an unfamiliar sample, a k nearest neighbor classifier searches the pattern space for the k training samples that are most similar to the unfamiliar sample. As a lazy learner, when the number of potential neighbors increases, KNN might incur high computational expenses with the number of samples with which to compare a particular unlabeled sample is huge.

One of the challenges with using KNN in disease prediction is the reduction of its performance, especially if the data sets contain noisy features.

The K NN algorithm has made significant contributions to machine learning studies due to its key characteristics. The K NN process starts by assuming resemblance among new and existing data points and assigning the new data point to the most alike class among those accessible. Furthermore, the K NN algorithm keeps all available data and categorizes a new data point founded on its resemblance to previously stored data points. As a result, the K NN algorithm simplifies the categorization of fresh data into an appropriate category. Third, while the K NN method is generally used for classification jobs, it may also be used to solve issues related to regression and classification. Fourthly, the K NN practice is observed as a non-parametric method, which means it does not make any expectations about fundamental data.

Fifth, K NN is devoted to as a "lazy learner" way since it does not learn from the training set promptly; in its place, it stores the dataset and performs an act on it throughout

classification. Lastly, in the training stage, the KNN algorithm saves the dataset and allocates new data to a class that is alike to the new or fresh data. Also, K NN is recognized as a essential supervised learning-based machine learning process. K NN is an algorithm that can be labelled based on the following steps:

Step 1: Selecting the number K of the neighbors
Step 2: Calculating the Euclidean distance of K number of neighbors
Step 3: Taking the first K neighbors as per the computed Euclidean distance.
Step 4: Counting the number of data points in each class between these k neighbors.
Step 5: Assigning the new data points to that class for which the amount of the neighbor is maximum.

Step 6: the model is complete.

2.3.3.6 Support Vector Machines

Support Vector Machine which stands for SVM classifiers are known for being very accurate and being able to deal with complicated nonlinear relationships in data. SVMs work through determining the best hyper plane that divides different classes in a space with many dimensions. This makes them very useful for health classification problems with features that are not all the same. For example, Kumar et al. (2022) used SVMs to figure out how likely someone was to get malaria based on climate and health data. They got very accurate results in several test areas. Gulshan et al. (2016) utilized SVMs to differentiate malaria cases in urban and rural contexts, whereas Rajab et al. (2024) discovered that incorporating SVM into a hybrid framework produced dependable outcomes in their outbreak detection system. But the performance of SVMs often depends a lot on how well the parameters are set and which kernel is chosen. This can make it hard to add them to big hybrid systems.

SVM is a widely used algorithm for supervised learning that is employed in solving issues with categorization and regression. The algorithm was initially proposed by Hoyos &

Hoyo (2024) to address the pattern issues with classification and regression. SVMs are a class of method of supervised learning that utilize operational danger minimization, nonlinear optimization, duality and kernel induced feature spaces (Rajab et al., 2024). The SVM technique has been extended to solve multi class classification, regression and clustering problems, making it applicable in the area of data mining(Gulshan et al., 2016; Kumar Jha et al., 2022). As Ahmed (2018) assert that SVM is a category of supervised ML approach that is employed to solve regression or classification challenges. The kernel trick for transforming data and determining the best boundary between alternative outputs depending on these transformations.

The SVM algorithm operates by training on the input output pairs $x1$, $y1$, . . .,xn, yn, wherever xi fits to X, the input space, yi fits to Y, the output space, and n represents the quantity of training data. The procedure aims to identify the function $(x) = w. x$ + b, with a maximum e deviation from the target y value (Hoyos & Hoyos, 2024). Here, w is the variety of variables in the training set, and w is the x coefficient. This implies that x and y are vectors and the above declaration can be stated numerically as $(xi) - y <$ e, where e represents a very small value. Moreover, $(x) = w1x1 + w2x2 + w3x3$+…+$wmxm$ + band the objective of the algorithm is to determine the values of w and b that satisfy the SVM equation. To illustrate linear classification in SVM, Wang et al., 2020 provide a figure 2.9 those expressions the relationship between the support vectors, the extreme boundary and the optimum separating hyper plane. The applicable and unrelated images in order with the user's query are represented by the elements in class I and class II, respectively.

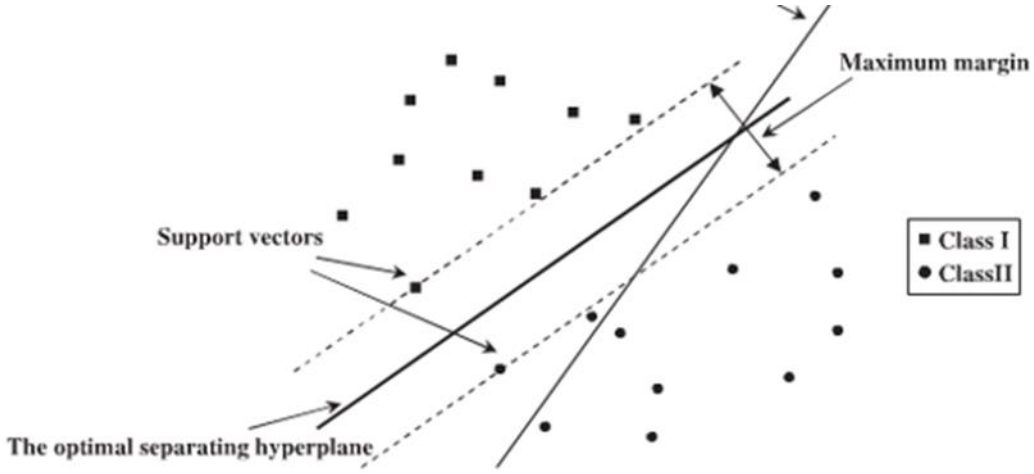


Figure 2.9: Theory of SVM

Wang et al. (2020) demonstrated the usability of SVM in image segmentation through pixel classification. The color and texture elements of the image were supplied into the SVM as input. The technique utilizes the resident data of the color image and the classification ability of the SVM classifier (Hussain-Alkhateeb et al., 2021) . SVM can also be used for nonlinear classification problems, where nonlinear mapping is utilized to obtain classification variables from the unique facts. The facts are then mapped to a high dimensional variables space for linear classification.

However, Bai et al. (2019) pointed out the following limitations of SVM:

- The selection of kernel functions, which is serious to the SVM performance, is difficult and subjective.
- SVMs are slower than other neural networks in both testing and training, making it difficult to handle very large datasets with millions of support vectors.
- SVM's performance deteriorates once the patterns to be categorized are non-divisible and the training data are noisy. In addition, removing known errors from the data before or after training may result in suboptimal hyper planes.

2.3.3.7 Artificial Neural Networks

A neural network is a machine learning technique that has emerged from the endeavor to simulate the human mind (Xu et al., 2022). ANNs are the result of academic research aimed at mathematically modeling the operations of the nervous system, and they have found successful applications in various business domains. ANNs represent a markedly distinct approach to utilizing computers in the workplace, as they are used to learn patterns and relations in data, including hidden patterns that may have gone unnoticed (Bai et al., 2019; Chen et al., 2014). They are analytic techniques inspired by the cognitive system's proposed learning processes, and they emulate the neurological purposes of the brain and can foresee new insights on precise features based on existing data through a learning process.

As a data mining technique, the first step in applying ANNs involves designing a network architecture comprising a specific number of layers, each with a predetermined number of neurons. This network is then trained by iteratively adjusting the weights of its neurons based on the inputs to optimally predict the sample data. During training, each neuron is connected to other neurons with specific coefficients, through which information is distributed to learn the network. After training, the network can generate predictions based on new data, representing a pattern detected in the learned data(Sharma et al., 2019). Overall, ANNs offer a unique and effective approach to predictive analytics and data mining, leveraging insights from cognitive and neurological processes.

According to Adam & Singh (2021), an ANNs typically has 3 layers: an input layer, a hidden layer, and an output layer, where the hidden layer performs computations depending on the input layer, assigning weights to the neurons inside the layer. The back propagation algorithm is often used to calculate these weights. The output layer produces the outcomes of the network, which can include predictions for multiple output nodes. The most commonly used ANN model is the feed forward network, which has a three-layer topology involving an input layer, an intermediate layer, and an output layer. Neurons are represented by blue boxes, and the connection points between them are

represented by arrows. During the training phase, the input data is presented to the network at the input layer, and the network adjusts the weights of the neurons in the intermediate layer based on the input. Some connection points may weigh zero, and a threshold value is added between layers to prevent these values from becoming zero, as showed in Figure 2.10.

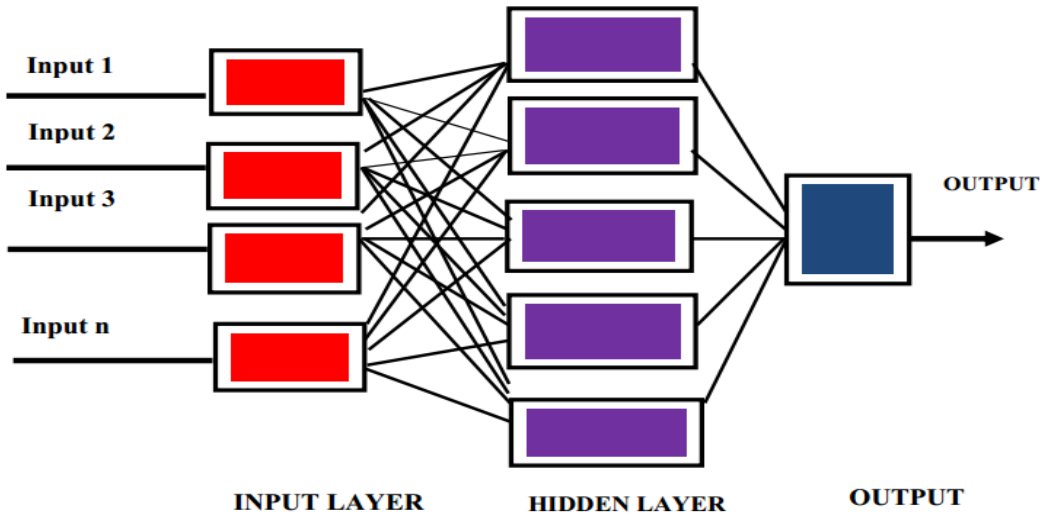


Figure 2.1: Layers of the Artificial Neural Networks

Although ANNs provide numerous benefits, they were deemed unsuitable for use in this research due to the following disadvantages:

Artificial Neural Networks (ANNs) present several limitations despite their wide applicability. One major challenge is hardware dependence, as ANNs require parallel processing systems that correspond to their complex architecture, making the availability of appropriate hardware essential for effective implementation. Another significant issue is the unexplained behavior of the network. Due to their limited interpretability, it is often unclear how or why a particular solution is generated, which can reduce trust in the model's outcomes (Masinde, 2020). In addition, determining the appropriate network structure remains problematic because there are no fixed rules for identifying the optimal

configuration; instead, the structure is typically established through experience and trial-and-error approaches (Bisaso et al., 2022).

Furthermore, ANNs operate exclusively on numerical data, meaning that all problems must first be translated into numerical form before being introduced into the network. The method used for this translation can significantly influence the network's performance and depends heavily on the user's expertise. Lastly, the duration of network training is often uncertain. Training is commonly considered complete when the error on the sample mean reaches a predefined threshold; however, achieving this threshold does not necessarily guarantee the best possible results. In summary, the primary benefits and drawbacks of different classification techniques were reviewed in Table 2.2 by Nasser and Behadili, (2022) and some of these factors were considered during the study process.

Table 2.2: Various Classification Techniques, Benefits, and Drawbacks

Algorithms	Advantage	Disadvantage
KNN	The implementation is simple. 2.Training is accelerated. 3.It is resistant to noisy training data. 4.When dealing with big training datasets, it can be more efficient.	1. A substantial amount of space is needed. 2. It is susceptible to noise. 3. Testing is time consuming.
Decision Tree	l.The decision tree can be built without any subject knowledge. 2. It reduces ambiguity in complex decision by assigning specific values to alternative outcomes. 3. The decision tree is skilled in handling multidimensional data. 4. It is basic to comprehend. The decision tree can analyze both numerical and categorical data well.	1. Only one output attribute can be produced by the decision tree. 2. It has produces categorical results. 3. The decision tree is an unstable classifier, which means that its performance is dependent on the dataset type.
Support Vector Machine	1. It out performs other classifiers in terms of accuracy. 2.1t handles complex non-linear data points with ease. 3.Overfitting is less of an issue with this procedure than with others.	1. It has significant computational costs. 2. The main issue is determining the best kernel function for each dataset, as different kernel roles produce varied outcomes. 3. When equated to other approaches, the training process takes longer. 4. SVM was originally intended to resolve binary class problems. It employs

Algorithms	Advantage	Disadvantage
		Strategies such as one against one and one against all to handle multiclass scenarios, breaking them down into pairs of two classes.
Neural Network	1. It is capable of detecting complex correlations between dependent and independent variables with ease. 2.1t can deal with noisy data successfully.	1. Local minima are possible. 2. Over fitting could be a concern. 3. The interpretation of ANN processing becomes difficult, especially when dealing with big neural networks that require a significant amount of processing time.
Bayesian Belief Network	1. It marks computing operations easier. 2. When working with large datasets, it displays greater speed and accuracy.	In certain cases, were variables are interdependent, it may not produce precise results

2.3.4 Regression Techniques

Regression approaches are essential in the predictive modeling of disease incidence, especially for estimating continuous outcomes such as the malaria cases number or the severity of an outbreak. For your research on a hybrid based classification and regression model for malaria outbreak prediction, these methods improve the regression part by using a mix of environmental, socio-economic, and clinical factors to guess how big an outbreak will be. This part of the thesis talks about regression approaches in the background (Vidhaya, 2023). Research indicates that regression is a technique utilized in data mining and machine learning for disease prediction. Regression analysis is mostly employed to ascertain the connection between two or more variables exhibiting cause-and-effect relationships and to facilitate predictions based on these correlations. Univariate regression analysis uses only one independent variable, while multivariate regression analysis uses more than two independent variables that are not related to each other (Huang & He, 2016).

Linear regression consists of first load the dataset, exploring the data, slicing the dataset, training and splitting dataset to generate the model, and finally evaluating accuracy (Kadam, 2020). The main aim of regression is to construct a well-organized model to predict dependent attributes from a set of attribute variables. Regression problems arise once the result feature is either real or continuous. Regression can also be defined as a

statistical method used in applications like housing, investing, etc., to predict the association between a dependent feature and a set of independent features. According to Nasser et al. (2022), Regression, as a predictive strategy in data mining, estimates the value of the dependent feature by employing the independent feature and its mathematical connection, as depicted through statistics and visualization. Based on previous research, there are two types of regressions:

Simple linear regression, which involves a single independent feature, and complex linear regression, which includes numerous independent features for prediction, and multiple regression, which uses multiple variables. The formula for computing regression is $Y = a + bX + u$, wherever Y=dependent feature, X = independent feature, a = intercept, b = slope, and u regression residual (Kadam, 2020; Nasser & Behadili, 2022).

According to Azezew et al. (2025), regression is a strategy for defining functions that highlight the relationship among distinct features. The training data is utilized to build the mathematical model, and two types of features are applied, one for the dependent feature and one for the independent feature, which are often represented by the letter 'Y' and 'X'. In scientific perspectives, it is observed that classification as well as regression require computed data containing of all input and output, meaning that they are all supervised learning issues.

In its most basic form, linear regression is a sort of regression that utilizes the straight-line formula (y mx + b). It computes suitable values for m and b to forecast the value of y given a value of x (Iyyanki et al., 2019). Linear regression models are utilized to demonstrate or predict the association among two factors. The dependent variable is the element to be predicted, represented by y, while the independent variables are the factors utilized for estimating its value. Regression techniques for prediction can be tailored, and regression examination can be applied to describe the relationships among one or more independent factors and one or more dependent factors. There are several sorts of regression approaches available from a systematic standpoint. These include:

2.3.4.1 Linear Regression

Linear Regression (LR) is a fundamental technique in regression modeling and has been widely utilized in malaria prediction studies. It presumes a linear correlation among the dependent factors and one or more independent. People like LR because it is easy to use and comprehend. Liu et al. (2020) employed linear regression to model monthly malaria incidence based on climatic factors in Ethiopia, reporting high interpretability and satisfactory accuracy. Likewise, Adam et al. (2021) utilized LR to predict malaria trends in India, highlighting its efficacy in early warning systems.

Hall & Lucas (2023) utilized LR as a foundational regression model within their hybrid framework, discovering its efficacy for comparative benchmarking. But LR can't handle nonlinear patterns, which are common in data on diseases spread by vectors. Linear regression is a renowned algorithm for ML that is classified as a statistical tool employed in predictive analysis. The technique is highly useful in forecasting continuous or quantitative variables, making it an increasingly common tool for predictive modeling. The short form of linear regression, named as simple linear regression, involves the use of a straight line to signify the link among the predictor factors and the response factors. The regression line is optimum in that it reduces the total squared error of forecast.

Multiple linear regression, which is a more universal form of linear regression, houses at least two independent or forecaster factors. The model is signified by a linear equation that merges a exact set of input values (x) and the forecast output (y). Both the input values (x) and the output value are numeric. The forecast model calculation for the linear regression model applied in this research is characterized as $y\ b0 + blx$, where y denotes the predicted output, b0 denotes the bias coefficient, bl denotes the coefficient for x, and x is the input value for the model.

2.3.4.2 Ridge Regression

Ridge Regression, or regularization L2, this shows the ridge adds a term penalty to the linear regression loss function to deal with multicollinearity among the input features. This method is especially useful for predicting malaria when temperature, humidity, and rainfall are all very closely related. Liu et al. (2020) discovered that ridge regression diminished variance and enhanced generalization in their climate-based malaria forecasting model. Islam et al. (2022) utilized ridge regression in their hybrid ensemble system, observing that it provided superior bias variance tradeoffs compared to ordinary least squares. Idris et al. (2021) incorporated ridge regression in their comparative study, determining it to be more stable in high dimensional contexts. The main problem is that coefficient shrinkage makes it harder to understand the model. According to Arumairajan (2023) stated the book of Hoerl and Kennard's book from 1970 was saying, the highly correlated predictor variables in a data set lead to the condition of multicollinearity. Scientifically, the ordinary least squares (OLS) estimate of regression coefficients is specified by the formulary, and the OLS estimate of the formula for regression coefficients is:

$$\beta\hat{}\_OLS = (X^TX)\hat{}(\ 1)X^TY \qquad (2.6)$$

Wherever *X* is the matrix of predictor factors, *Y* is the vector of response factors, *and* β_OLS is the vector of estimated coefficients (Arumairajan, 2023). The matrix (X^TX) becomes almost singular when there is multicollinearity. Ridge regression solves this problem by adding a ridge (biasing) parameter Δ to the *OLS* estimator, which changes it to:

$$\beta\hat{}\_ridge = (X^TX + \Delta I)\ \hat{}\ (1)\ X^TY \qquad (2.7)$$

The ridge penalty term is $\Delta > 0$, and the identity matrix is *I* this change makes the matrix easier to invert and lowers the sampling variance, but it does add some bias.

Ridge Regression, on the other hand, minimizes the penalized residual sum of squares:

minimize:

$$\sum (y_i \quad x_i^T\beta)^2 + \Delta \sum \beta_j^2 \tag{2.8}$$

This method of regularization limits the coefficients to a hyper sphere where $\sum \beta_j^2 < C$, where C is a positive constant. This is called *L2* regularization. Important Features of Ridge Regression:

i. Makes the same assumptions as linear regression, but not normality.
ii. Decreases the size of coefficients then does not make them zero (no variable selection).
iii. Uses *L2* norm regularization to deal with multicollinearity.
iv. The biasing parameter is also known as the shrinkage parameter Δ *(or C).*

2.3.4.3 Lasso Regression

Lasso Regression as L1 regularization, not only reduces multicollinearity but also selects features by setting the coefficients of less important factor to zero. This characteristic is useful in malaria modeling when managing extensive and possibly noisy datasets. Musa et al. (2024) utilized Lasso regression in a climate malaria model, discovering that it efficiently eliminated extraneous predictors, thereby enhancing the model's robustness. Armairajan (2023) observed that Lasso regression improved the efficacy of their hybrid model by streamlining the feature space. Adigun et al. (2024) employed Lasso to identify predominant climate and demographic characteristics prior to inputting the data into more intricate models. The method can have trouble when predictors are very similar to each other, and it may choose one feature over others for no good reason.

Lasso Regression's regularization method has an extra feature called variable selection, which is similar to ridge regression (Kumar & Singh, 2023). The name "Lasso" comes from the phrase "Least Absolute Shrinkage, and choice Operator." By putting a linear

constraint on the regression coefficients, you can make certain regression coefficients go to zero, which removes the variables that go with them from the model. The Lasso regression sets a linear limit on the regression coefficients in this way:

$$\sum |\beta_i| \leq t \quad (9)$$

t is a tuning parameter that controls how strong the penalty is. Here are some important things to remember about Lasso regression:

(a) Lasso regression makes the similar expectations as linear regression, excluding for the one that says the data is normal.
(b) Lasso regression automatically selects variables by shrinking some of the regression coefficients to zero.
(c) Lasso regression is a way to regularize that usages the L1 norm to punish big coefficients.
(d) When the predictor variables are very similar to each other, Lasso usually keeps only one and sets the others to zero.

2.3.4.4 Polynomial Regression

Polynomial regression is a nonlinear version of linear regression that uses nth degree polynomials to model relationships. This method works well for capturing the nonlinear dynamics of malaria transmission that are affected by changes in environmental factors that happen at certain times of the year. Huang et al. (2016) presented that polynomial regression is improved than other methods for finding seasonality in malaria outbreaks in Morocco. Mariki et al. (2022) discovered that quadratic and cubic models substantially enhanced predictive accuracy compared to linear models in their analysis of malaria data from Bangladesh. Boit et al. (2024) utilized polynomial regression in rural health systems, indicating its efficacy in modeling the peak and decline phases of seasonal outbreaks. Nonetheless, higher degree polynomials are prone to overfitting, especially when data is scarce. In polynomial regression, the regression formula has predictor

factors elevated to powers greater than 1. It is important to remember that polynomial regression is nonlinear in predictor factors but linear in regression settings (Boit & Patil, 2024). In the context of polynomial regression, the following points are very important:

When including terms involving powers of predictor factors in the regression formulation, it may be tempting to comprise higher degree terms to decrease error, but this can lead to overfitting. Therefore, it is advisable to plot the data and choose a reasonable degree of polynomial using the principle of parsimony (Mariki et al., 2022). The simpler model is preferred when two models have similar performances. Also. it is important to be careful near the two extremes of the predictor variable range because higher degree polynomials may exhibit unexpected behavior when extrapolated beyond the data range.

In polynomial regression, the link among the dependent factor *y* and the independent factor *x* is modeled as a nth degree polynomial, described by the following equation:

$$y = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \cdots + b_n x^n \tag{2.9}$$

Where n is the degree of the polynomial.

In this perception, polynomial regression is needed in machine learning when data points are arranged in a nonlinear fashion. Applying a linear model to a nonlinear dataset would result in a high loss function, error rate, and decreased accuracy. The comparison diagram below illustrates the distinction between nonlinear and linear datasets.

2.3.4.5 Support Vector Regression

Support Vector Regression (SVR) is another influential method that uses support vector machines to accomplish regression tasks. Kernel functions are ideal for working with both linear and nonlinear data distributions. SVR is quite good at predicting malaria and doesn't be thrown off by outliers. Nyambura et al. (2025) utilized SVR in an urban malaria forecasting model, noting improved performance relative to conventional linear models. Musa et al. (2024) employed SVR to predict malaria incidence by utilizing high-dimensional climatic data, hence showcasing its resilience. Shankar et al. (2026)

integrated SVR into their hybrid design, illustrating its significant influence on model generalization. SVR has some wonderful things about it, but it needs careful tuning of its settings and more computer power.

2.3.4.6 Random Forest Regression

Random Forest Regression (RFR) is a based tree ensemble technique that has gained popularity in epidemiological modeling due to its ability to simulate intricate nonlinear relations and assess the significance of features. Huang et al. (2016) asserted that RFR outperformed traditional regression techniques in elucidating malaria trends. Kadam (2020) employed RFR in their hybrid system for distant healthcare settings and found it useful in simulating outbreak severity. Adigun et al. (2024) employed RFR in conjunction with deep learning models, exhibiting enhanced prediction efficacy. Nonetheless, interpretability and computational complexity persist as obstacles in practical contexts.

2.3.4.7 Gradient Boosting Regression

Gradient Boosting Regression (GBR) is a well-known method for making predictions with structure. XGBoost and LightGBM are two examples of GBR. GBR slowly adds weak models to a group that solves the flaws made in prior versions. Mbunge et al. (2022) employed XGBoost in a hybrid malaria prediction model, accomplishing advanced methods results. Singh et al. (2020) employed LightGBM to predict monthly malaria cases, emphasizing its speed and accuracy. Dhoot (2018) found that gradient boosting improved generalization when used with other classifiers and regressors. However, GBR models are sensitive to how you set the hyper parameters, and if you don't configure them appropriately, they can fit too well (Dhoot et al., 2018; Jameela et al., 2022) ,Table 2.3 now provides clearer differentiation between the classification and regression approaches.

Table 2.3: The Main Difference between Classification and Regression

Classification Algorithm	Regression Algorithm
The value mapping function assigns values to specified groupings.	To assign values to continuous outcomes, the mapping function is utilized.
The resultant component of the classification must be a categorical or discrete attribute.	In regression, the result element has to be of a constant type, representing actual values.
The classification method's job is to transfer its input value (x) to the output discrete factor (y)	The algorithm of regression 's job is to map the continuous variable outcome (y) to the factor input 's value (x)
For discrete data, algorithms for classification are utilized.	Continuous data is employed with regression methods.
Classification is to discovery the judgment limit, which may split the data into separate classes.	In Regression, we aim to identify the most suitable match rows that can more correctly estimate performance.
Methods for classification may be employed to tackle classification problems such as recognizing voices, email spam identification, and tumor cell identification, and others.	Regression methods can be used to handle regression related issues like property value prediction, and forecasting The weather, and others
Algorithms for classification may be considered into two distinct groups: multi class classifiers and binary classifiers.	Nonlinear and Linear Regression are two distinct kinds of regression algorithms.

2.3.5 Concept of Hybrid and Ensemble Learning

The study validated that hybrid learning and ensemble learning are two interconnected yet different ideas in predictive modeling. Ensemble learning aims to enhance predictive accuracy by amalgamating multiple models of identical or diverse types, whereas hybrid learning synthesizes models or methodologies from various paradigms into a unified framework (Cagnini et al., 2023). Ensemble learning as the multiple models are used and this approach has been successfully employed in complex real-world forecasting tasks. Muserat et al. (2020) demonstrate this by combining convolutional neural networks with clustering and reinforcement techniques to predict trends in the foreign exchange market, achieving higher accuracy than independent models such as random forests and the nearest neighbor algorithm. This demonstrates how using diverse models can help detect

complex patterns in time series. This concept also proves useful with hybrid classification and regression models that predict disease outbreaks (Maserat et al., 2020).

The prime awareness in arrears ensemble learning is that a model made up of a mix of several, weak base learners can be stronger and better at generalizing than any one of those base learners on its own (Adamu & Singh, 2021b). Hybrid learning, conversely, focuses on the amalgamation of diverse learning methodologies to leverage their distinct benefits such as the integration of supervised and unsupervised learning, or the fusion of regression and classification algorithms to address both continuous and categorical dimensions of a problem (Asif et al., 2024). Hybrid systems are very useful for predicting malaria outbreaks because they can handle several sorts of data, such as epidemiological case counts, climate measurements, and environmental indicators, and also, they can also handle both categorical and continuous prediction tasks. This study's hybrid-based classification and regression framework is an example of this kind of integration. It lets one system do both epidemic detection and incidence forecasting at the same time. Hybrid learning methods have shown a lot of potential in modeling infectious diseases and predicting outbreaks.

Asif et al. (2024) employed a deep learning hybrid method that mixes convolutional neural networks and long short-term memory nets for influenza prediction, resulting in enhanced temporal pattern recognition. In the same way, Adebanji et al. (2021) utilized a amalgamation of support vector regression and decision tree models to forecast malaria cases by combining weather and environmental data. Hybrid ensemble models have been utilized for other vector-borne illnesses. Mathuria et al. (2020) put forward a hybrid Random Forest-Gradient Boosting model for predicting dengue outbreaks, which worked better in a wider range of climates. Mujahid et al. (2024) combined ensemble learning with geographic information system data to forecast cholera outbreaks in coastal areas. These studies emphasize the increasing agreement that hybrid and ensemble models surpass conventional single algorithms by managing data complexity, mitigating uncertainty, and improving robustness. This study advances these principles by combining regression (for outbreak magnitude) and classification (for outbreak incidence)

phases into a cohesive hybrid model designed for malaria prediction (Mujahid et al., 2024) . The following figure 2.11 illustrates data modeling (Maass et al., 2024)

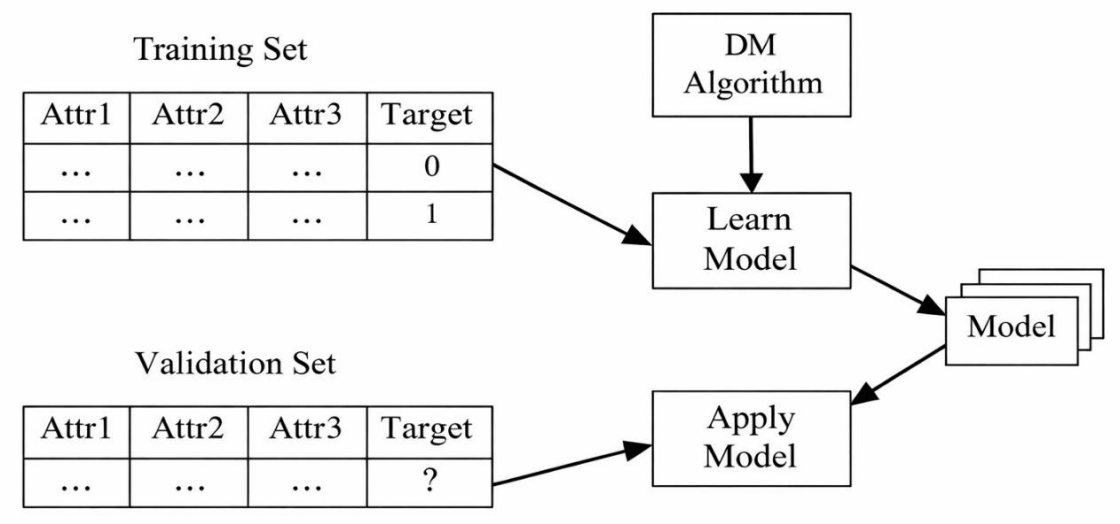


Figure 2.11: Data Modeling

Also, the word "hybrid" means using two or more strategies together in a way that gets around the problems with each one. In this sense, "classification" is a methodical strategy to construct models for training and evaluating datasets that may be used to put things into groups. As stated before, machine learning methods have some limits. A hybrid machine learning method could get beyond these limits by capturing additional properties of complicated systems. Recent research has indicated favorable outcomes for hybrid models that include various machine learning methodologies (Bbosa et al., 2020). There are several ways to construct hybrid models, and it's not clear which one will work best for data determined learning. Topical hybrid data determined methodologies encompass a fusion of Neural Networks and Extended Kalman Filters, Gaussian Process Regression techniques utilizing an Automatic Relevance Determination kernel, Support Vector Machines and Support Vector Regression models refined through Genetic Algorithms and Relevance Vector Machines with incremental learning (Liu et al., 2020). Zhang & Yang (2021) also said that a machine learning hybrid model is made up of two or more machine learning algorithms or models. The aim is to use the synergy of each model to make the

total performance better. distinct research show that there are many distinct kinds of machine learning hybrid models, as shown below; There are many ways to build hybrid models, but the most common are sequential, parallel, and cascaded hybridization (Wang et al., 2020).

In sequential hybrid models, the productivity of one learning process is utilized as the input for the following one. This technique facilitates tiered learning, with the initial phase concentrating on feature extraction or regression estimation, and the subsequent phase enhancing the output via classification or secondary modeling (Y. Zhang & Yang, 2021). For example, this study first uses regression models like the Random Forest Regressor to guess how many cases of malaria there would be. Then, the results are used as input for a classification algorithm to figure out how likely an epidemic is.

In the view of parallel hybrid models, multiple algorithms are trained on the similar data independently, and then their productions are combined using weighted averaging or voting methods (Onyijen et al., 2023). This method increases the diversity of learners and helps prevent overfitting. In malaria forecasting, a parallel hybrid system may concurrently amalgamate regression and classification models to get estimates of both outbreak likelihood and intensity.

In machine learning, cascaded hybridization refers the use of numerous layers of model combinations are used, and the outputs from earlier ensembles are improved in later stages (Hussain et al., 2025). Meta learning is often employed in hierarchical disease prediction systems to improve predictions at the feature level. The cascaded technique makes it easier to understand the model and lets it work with large epidemiological datasets with many dimensions. Many theories suggest that creating a hybrid modeling prediction approach should not occur without acknowledging that modeling entails building a model in a known scenario and subsequently applying it to an unknown scenario, as depicted in Figure 2.12. It is essential to acknowledge that model construction is an iterative process and that no singular method is universally optimal, as asserted by certain proponents

(Maass et al., 2024). Moreover, it has been proved that it is possible to mix various models, as in data mining hybrids, for prediction or other uses.

In the same context Ensemble models use multiple examples of the same type of model, like multiple decision trees or neural networks, to make a better model. Some examples of ensemble models are Random Forests, Gradient Boosting Machines, and AdaBoost (Mirzaeian et al., 2023). From the same points of view, ensemble learning is a group of methods that train many individual learners and then combine their predictions to learn a target function. Figure 2.12 below shows how an ensemble is made by putting together N models that were trained on N datasets chosen from the data that was available. The prediction of the ensemble, when applied to a new instance, is a function of all the constituent base models. In practice, the datasets used to train the base models may overlap or even be the same (Song et al., 2025).

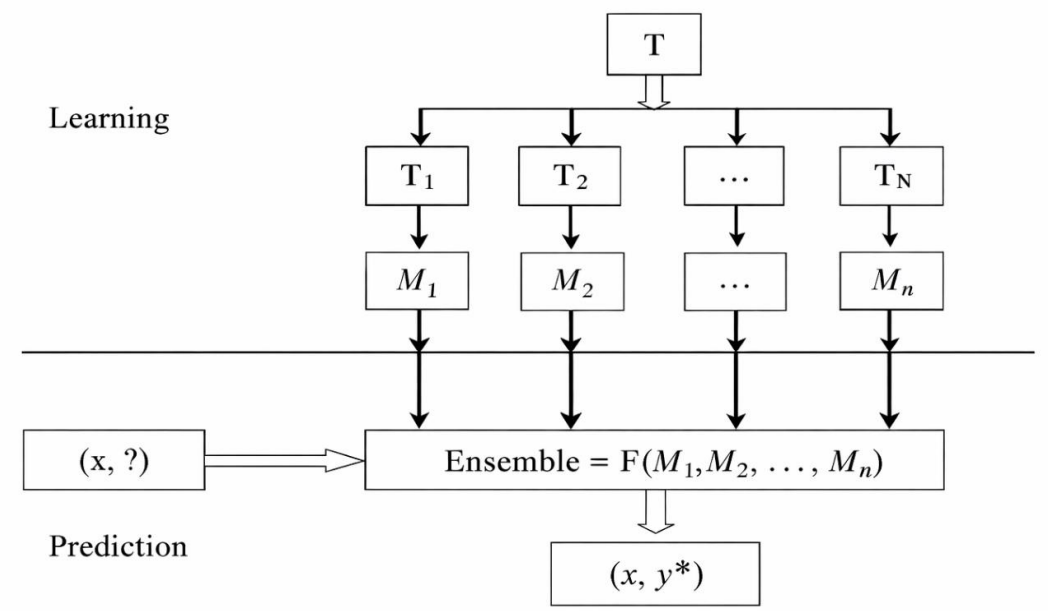


Figure 2.12: Schematic Ensemble

As shown in Figure 2.12 above, Ensemble methods are among the most influential advancements in machine learning. They enhance predictive power by integrating multiple learners to form a composite model (Mienye & Sun, 2022). In summary, these

mentioned ensemble methods are the main utilized to improve model performance. Bagging, introduced by Mirzaeian et al. (2023), trains several base models on distinct bootstrapped samples of the dataset and combines their forecasts through averaging for regression for classification. This technique decreases adjustment and stops overfitting, making it particularly useful for unstable models such as decision trees. A classic example of bagging is the Random Forest approach, where several decision trees are trained on bootstrapped subsets, and their outputs are averaged to produce more reliable predictions. Boosting, instead, develops models sequentially, with each new learner concentrating on correcting the mistakes of previous models (Mbunge et al., 2021).

Methods, for instance Adaboost, Boosting Gradient, And Xgboost exemplify this approach by reducing both bias and variance, boosting performs well on complex datasets, and in malaria prediction, it can help detect subtle outbreak patterns by emphasizing cases that are difficult to predict or have been misclassified. Stacking, also known as stacked generalization, employs a meta-learner to determine the optimal way to combine the outputs of multiple base learners (Schaffer et al., 2023). This hierarchical method considers dependencies between models, improving prediction accuracy. It is commonly used in hybrid frameworks to integrate regression and classification outputs into a single system, enhancing overall model synergy. Finally, voting ensembles aggregate predictions either through hard voting, where each model's forecast is counted, or soft voting, where predictions are weighted and averaged (Cagnini et al., 2023). Voting is computationally simple and effective at increasing robustness. In malaria outbreak prediction, voting ensembles can combine estimates from climate, environmental, and epidemiological models to generate a single consolidated forecast.

2.3.6 Deep Learning Hybrid Models, Rule Based Models, and Transfer Learning Models

These models combine deep learning methods (like convolutional neural networks with traditional ML approaches to get the best of both worlds (Bhatt et al., 2021). Rule-based models make predictions or judgments founded on a usual of rules that have already been

specified, and they can work with other machine learning models to make them work better (Musa et al., 2024) . Transfer learning models use a machine learning method that has already been trained and improved on a new dataset to solve a different problem. This approach can be utilized alongside other models to improve performance (Morovati et al., 2023).

In the same perspective, hybrid machine learning methods that integrate deep learning with other methods have become gradually notable in recent research due to their capability to leverage complementary strengths of multiple approaches. For example, Tharageswari et al. (2025) propose a deep learning hybrid model in which a convolutional neural network is combined with a traditional machine learning classifier (random forest) to improve performance in healthcare prediction tasks, demonstrating higher accuracy than standalone models. Such hybrid architectures illustrate how deep neural nets can extract composite factors from data while classical algorithms contribute interpretability and stability, a principle that is equally valuable when designing hybrid classification and regression frameworks for predicting malaria outbreaks (Tharageswari et al., 2025).

2.4 Computational Intelligence in Healthcare Informatics

Computational intelligence plays a essential role in contemporary health informatics by applying advanced AI and machine learning methods to analyze complex medical data and support predictive modeling and decision-making(Rezaul et al., 2025). The edited volume Applied Intelligence for Healthcare Informatics by Mategula et al. (2025) highlights diverse applications of computational intelligence, including disease prediction, medical condition classification, and diagnostic support, demonstrating how ensemble and deep learning methods can uncover meaningful patterns from heterogeneous healthcare data. Similarly, Tharageswari et al. (2025) showed that hybrid deep learning models significantly improve forecast accuracy in healthcare systems by mixing neural networks with conventional machine learning methods. Together, these studies underscore the relevance of computational intelligence for hybrid classification and regression models, as combining multiple intelligent algorithms improves predictive

accuracy, robustness, and generalization key objectives when developing effective models for forecasting malaria outbreaks.

In the same context, Machine learning is becoming a powerful technology in healthcare that could greatly improve how diseases are predicted, diagnosed, and treated. ML algorithms let systems learn from both past and current data to generate accurate predictions. This is especially important for controlling diseases like malaria. In situations with limited resources where clinical decision-making is typically impeded by insufficient medical infrastructure, machine learning offers a scalable and economical solution for prompt illness surveillance and epidemic prediction (Thakur & Dharavath, 2019). ML models can handle different types of data, such as demographic, clinical, and environmental data. This lets them find complicated, nonlinear associations that standard statistical methods often miss (Vapnik & Izmailov, 2019). Machine learning algorithms may use things like temperature, rainfall, humidity, and population density to predict the probability of malaria outbreaks. This helps public health efforts that are focused on specific groups of people. Furthermore, ML-based diagnostic tools have demonstrated efficacy in real-time applications, such as mobile health technology, where automated classification and regression models aid frontline healthcare personnel in detecting diseases or assessing case severity (Esteva et al., 2019). These features not only help find problems earlier and enhance healthcare delivery, but they also encourage data-driven decision-making at both the clinical and policy levels. Even if there are certain problems, such data quality issues and worries about how informal it is to comprehend models, the use of ML in health sector keeps growing since it has been shown to work, especially for managing infectious diseases and predicting outbreaks.

Even while machine learning has a lot of possible uses in healthcare, here are a number of problems that make it hard to use effectively, especially when it comes to using health data for predictive modeling. One of the most important challenges is data quality, which includes things like missing values, inconsistencies, noise, and biased samples. These limitations are particularly evident in malaria-endemic locations characterized by inadequate data collection infrastructure, leading to incomplete or obsolete records

(Miotto et al., 2017). Data that isn't good enough can make training models very difficult, which can lower their accuracy and generalizability. Another big problem is keeping data safe and private. Health data are very private, and using patient information for machine learning presents moral and legal issues around consent, privacy, and subsequent data protection laws like the regulation of general protection of the data (Esteva et al., 2019).

Also, the fact that health information systems aren't all the same makes it hard to combine data from different sources, which makes it harder for machine learning models to work on a large scale. Interpretability is also very important, especially in clinical situations where medical professionals need to be able to understand and explain judgments. A lot of machine learning models, especially deep learning ones, are thought of as "black boxes," which could make healthcare practitioners less likely to trust and use them (Muriithi et al., 2024) . To create machine learning algorithms that are strong, ethical, and dependable enough to help anticipate malaria outbreaks in the real world, we need to deal with these problems.

2.4.1 Disease Outbreaks

It is essential to distinguish between an outbreak and an epidemic in the context of this research. When the number of sick persons in a given place or group of people rises quickly and without warning, it is called an epidemic. An epidemic, on the other hand, happens when a disease spreads to more individuals, often in larger areas or throughout numerous towns, and it can persist longer. This work focuses on predicting outbreaks to enable swift and precise interventions.

Researchers have discovered that precise models for forecasting outbreaks are crucial for comprehending the transmission dynamics of infectious illnesses and their potential impacts (Ardabili et al., 2020). An epidemic outbreak is when a disease spreads swiftly over a short historical of time in a given location. This kind of pandemic could happen in just one hamlet or spread to numerous countries. It could last anywhere from a few days to a few years. So, it's very crucial to stop the spread as quickly as possible. There is a

growing need for advanced models to predict and prevent the worldwide spread of diseases (Jdey et al., 2022). There are no single criteria for how many cases constitute an outbreak. An outbreak arises when there are more cases than expected for a given group of people based on what has happened in the past. Outbreaks can happen in a small area or throughout many countries, and 'they can last for any amount of time. The cholera outbreak, for example, killed almost 100,000 people around the world and made about 35 million people sick (Ilic & Ilic, 2023). This study largely focused on the malaria outbreak, given that malaria is the most extensively researched disease in sub Saharan African countries. (Brenas et al., 2018).

2.4.2 Trends of Disease Outbreaks in the Region of Africa

Recent research indicates that Africa continues to experience significant shifts in disease outbreak patterns, driven by environmental, social, and health-system factors.

Perez-Saez et al. (2025) document changes in the geographic distribution and burden of cholera across the continent, highlighting persistent and emerging hotspots that challenge public health responses. Coupled with reports of concurrent outbreaks of malaria, viral hemorrhagic fevers, and other infectious diseases, these trends underscore the complexity of epidemic dynamics in the region. Effective forecasting of such multi-disease outbreak patterns requires analytical models capable of learning from diverse datasets and evolving conditions a context where machine learning and hybrid predictive models can provide valuable insights by capturing nonlinear trends and informing early warning systems for diseases like malaria (Perez-Saez et al., 2025). In this view of point, the numerous academics have investigated the frequency and consequences of disease outbreaks over the African continent. Their findings indicate a wide array of persistent infectious diseases, including cholera, meningitis, yellow fever, malaria, and dysentery, as well as the reemergence of diseases such as Ebola and Marburg hemorrhagic fevers, and the emergence of novel dangers like avian influenza.

The COVID-19 pandemic that started in 2020 made things much more difficult for public health, as African countries were still dealing with several outbreaks at the same time. Even though a lot of work went into stopping COVID 19, other infectious diseases kept spreading and, in some cases, got worse. These events show how communicable diseases are still a big problem in the region and how important it is to have strong surveillance systems, integrated response mechanisms, and ongoing investment in public health infrastructure, even during global health emergencies. The following information has been highlighted:

Epidemics or outbreaks have been consistently described from Benin, Ghana, Guinea and Togo since1998. In 1999, the southern Africa was represented for 51% of all cholera occurrences and 40% of resultant deaths, with the most affected countries being Mozambique, Malawi, Zambia, Madagascar and Zimbabwe. In 2006, 31 countries described cholera epidemics to WHO/AFRO and all reported cholera outbreaks were laboratory confirmed. The annual reported cholera cases and case fatality rate (CFR) presented a descending trend, which may be due to enhanced surveillance or response to the outbreaks. Cholera outbreaks have surged across several African countries post COVID, including Malawi, Mozambique, and Nigeria. Moreover, in 2022—2023, Malawi experienced one of its worst cholera outbreaks in decades, leading to thousands of cases and hundreds of deaths (World Health Organization, 2023).

In the case of Dysentery, its largest outbreak in western Africa was described in 1999 by Medecins Sans Frontieres in the Kenema district, southeastern part of Sierra Leone. The total number of cases was 4,218, with a total attack rate and CFR of 7.5% and 3.1 % respectively.

Similarly, Meningococcal meningitis has been a rampant concern. In 2003, among 32 countries reporting meningitis to WHO/AFRO, 14 outbreaks have been proclaimed. The most affected countries were the DRC, Burkina Faso, Niger and Nigeria, Uganda, Ghana, Ethiopia, Chad and Mali. In 2004, 1 1 of the 32 countries proclaimed outbreaks and in

2005, nine countries reported a total of 23,336 cases and 3, 189 deaths with a CFR of 13.7%.

Malaria outbreaks have also brought about major challenges. In 2003, three countries that were mostly affected by malaria epidemics were Ethiopia, Burundi and Kenya. In 2004, a large epidemic was reported from Zimbabwe, an outbreak is often the first step in a bigger epidemic, which happens when a disease spreads to a bigger area or more people.

Measles remains another major public health issue: In 1999, there were roughly 871,000 measles' deaths international, with 61% occurring in Sub Saharan Africa. In 2004, 80 (5%) out of 1,590 districts under case-based surveillance described outbreaks of measles. In 2005, 47 (2.5%) districts reported outbreaks out of 1,850and in 2006, 178 (6%) out of 2,923 districts reported outbreaks spanning across 29 countries. The most affected countries were the DRC, Nigeria, Ethiopia and Tanzania. Due to disruption of routine immunization during the COVID 19 pandemic, countries such as the DRC and Ethiopia reported significant measles outbreaks between 2021 and 2023 (Panday et al., 2022).

Yellow Fever has continued to affect many African Countries as well. In 2004, laboratory confirmed yellow fever Cases have been reported in Burkina Faso, Cameroon, Guinea, Liberia, Mali, and Senegal, among other African countries. Moreover, five of these countries reported probable yellow fever cases within the same time period. In the years 2006 2007, 477 probable yellow fever cases were reported and 32 deaths (CFR 7%) were reported from 13 countries, among which seven countries had confirmed outbreaks. However, there is a concern of gross underreporting in the region.

Marburg Viral Hemorrhagic Fever has also contributed to regional health emergencies. According to Narrative: Moyo et al. (2023) reports that between 1998 and 2000, large outbreaks of Marburg virus were reported in the Democratic Republic of Congo (DRC), resulting in a total of 154 cases and 128 deaths, with a case fatality rate (CFR) of 83%. The huge majority of cases occurred in young males. Also, the Marburg Virus Disease, In the 2023, has been appeared in the Equatorial Guinea and Tanzania reported outbreaks

of Marburg virus disease a highly lethal hemorrhagic fever similar to Ebola. These outbreaks were the first of their kind in these countries and drew international concern (Moyo et al., 2023).

Rift Valley Hemorrhagic Fever (RVHF) continues to affect both human and animal population. According to Narrative: Moyo et al. (2023) reports that in 2003, Mauritania experienced a laboratory confirmed mixed outbreak of Crimean Congo hemorrhagic fever and Rift Valley fever, resulting in a total of 41 cases and 21 deaths, with a CFR of 51%. Also, the Rift Valley Fever outbreaks were reported in Kenya and Mauritania in 2021 and 2022, affecting both human and livestock populations. These zoonotic outbreaks raise concerns for both public health and food security (FAO, 2022).

Plague is also considered as emerging infectious disease. According to Narrative: Moyo et al. (2023) since the early 1990s, there has been an augmented incidence of human plague internationally, with increased proportions reported from Africa. In 2002, Africa stated a total of 1,822 cases (95% of the world total), including 171 deaths (97% of the world total). From 2003 2007, nearly 11,000 cases and close to 400 deaths were described from sub-Saharan Africa.

Lassa fever remains endemic in parts of West Africa. Narrative: Moyo et al. (2023) reported that in 2004, Nigeria reported 43 cases of Lassa fever with 21 deaths, while Sierra Leone reported 147 cases and 69 deaths. In 2005, three countries (Liberia, Sierra Leone, and Nigeria) reported 88 cases, and in 2006, 101 cases and 15 deaths were reported. Again, in Nigeria has faced recurring Lassa fever outbreaks, with significant spikes in 2021 and 2022. The disease is endemic in parts of West Africa and poses a recurring challenge due to its high mortality and lack of vaccines (Merga et al., 2025) Chikungunya has also presented recurring challenges. According to Narrative: Moyo et al. (2023) due to reports, Chikungunya has numerous clinical symptoms in common with dengue fever and might be misdiagnosed in regions wherever dengue is prevalent. After numerous years of low levels of human infections in Africa, a momentous outbreak occurred in the DRC in 1999-2000 and in Gabon in 2007.

Between March 28, 2005and February 19, 2006, A sentinel network on the island of La Reunion reported 2,406 examples of Chikungunya, with 333 cases described in the week of February 13-19 alone. A computer model likely that 157,000 persons in La Reunion could have been infected with the Chikungunya virus meanwhile March 2005. Other island nations in the southwest Indian Ocean have informed Chikungunya outbreaks, with Mayotte (924 cases), Mauritius (2,553 cases, with 1,173 lab-confirmed cases), and the Seychelles (4,650 cases).

Regarding Human monkey pox, Narrative: Moyo et al. (2023) described that it is an endemic emerging zoonotic disease to central and western Africa. The disease was first reported in 1970 from DRC, nine months after the smallpox extermination in that nation. The disease has similar manifestations as smallpox, including a prodrome and rash. Although more prominent in Europe and the Americas in 2022, several African nations particularly Nigeria and the DRC continued to report cases of monkey pox, reflecting its endemic presence and the risk of global spillover (Musa et al., 2024).

Finally, Ebola Viral Hemorrhagic Fever remains the most severe infectious disease on the continent. Narrative: Moyo et al. (2023) reported that since its discovery in 1976 in the DRC, Ebola virus outbreaks have caused in approximately 1,850 cases and over 1,200 deaths in Africa. Though the Ebola outbreaks have caused relatively few cases, the associated high CFR, ranging from 50% to 90%, calls for augmented consciousness to guarantee effective epidemic grounding and response. The largest prevalent in the region since 1998 occurred in Uganda between 2000 and 2001, with 425 cases and 224 deaths (CFR 53%). In addition, about Ebola Virus Disease, After COVID 19 began, multiple Ebola outbreaks occurred in Africa. Notably, the DRC reported outbreaks in 2021 and 2022, while Uganda experienced a Sudan strain Ebola outbreak in 2022, which led to multiple deaths and widespread public health interventions (WHO, 2022)

2.4 3 Development of Malaria

Malaria is remained one of the greatest common and dangerous infectious illnesses in the world, especially in tropical areas like sub-Saharan Africa. Malaria is intimately related to the life cycle of Plasmodium parasites, namely P. falciparum and P. vivax. These parasites are spread to people through the bites of female Anopheles mosquitoes that are already infected (World Health Organization, 2015). After a mosquito bite, the parasites go to the liver, where they increase before incoming the bloodstream and infecting red blood cells. This can cause clinical signs such fever, chills, anemia, and in severe cases, organ failure or death (Huang et al., 2020). The epidemiology of malaria has changed over time. Since the early 2000s, global efforts to control and eliminate the disease have led to a big drop in cases. However, progress has slowed down in the last few years (World Health Organization, 2022). Improvements in monitoring, diagnoses, and treatment have made it easier to handle cases, but predicting outbreaks is still hard because climatic change, vector ecology, and human behavior all interact in ways that are hard to understand (Sriporn et al., 2020) .

Comprehending the development and transmission dynamics of malaria is essential for informing data-driven models, such as the hybrid-based classification and regression framework proposed in this research, which seeks to amalgamate both epidemiological and environmental indicators for enhanced outbreak prediction (Molina-Franky et al., 2020), as depicted in Figure 2.13,

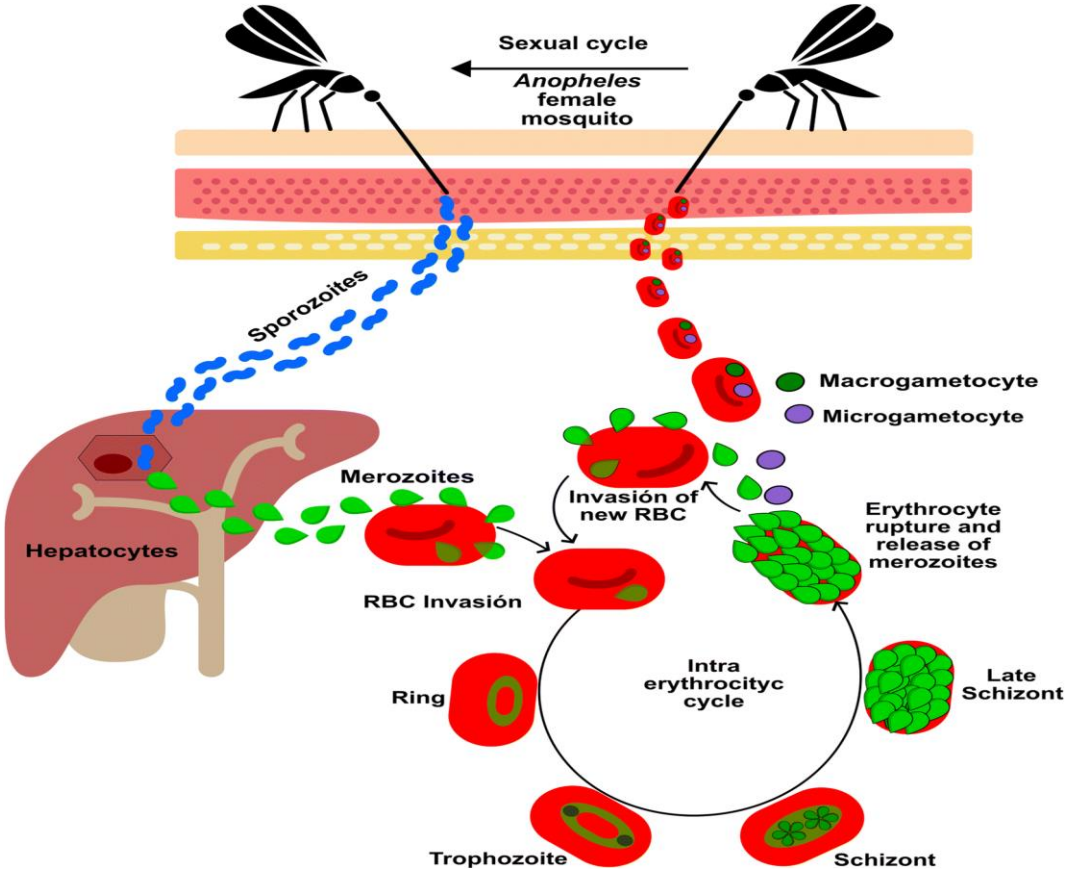


Figure 2.13: The Malaria Parasite Life Cycle

2.4.4 Modern and Traditional Data Source and Features for Prediction for Malaria Prediction

Recent advances in malaria prediction research emphasize the importance of integrating modern and traditional data sources and feature sets to improve model performance. For example, MalariVis (Ma, 2025) combines clinical case data from health institutions with climatic factors such as rainfall, wind speed, humidity, and temperature from external environmental datasets to forecast both incidence regression and outbreaks classification, identifying key driving features that influence malaria dynamics. In parallel, Merga et al. (2025) demonstrate that socioeconomic and demographic indicators such as population density, vegetation index, aridity, and age group distributions can be integrated with environmental covariates in predictive models to achieve high accuracy across diverse

populations. These studies highlight how combining traditional sources such as clinical surveillance and demographic surveys with modern datasets such as remote sensing, climate records, and social determinants enriches machine learning feature spaces, enabling more robust, context-aware prediction frameworks that are applicable to malaria outbreak forecasting.

Reporting instances to a central health agency is the usual way to keep an eye on an epidemic. But delays in reporting cases limit how well the system works. To solve this problem, machine learning methods have been created that use unusual and non-clinical data sources to predict epidemics and speed up reporting. This change has led to big improvements in the domain of health informatics. Data from social media platform has become a valuable instrument for predicting disease outbreaks and keeping an eye on illnesses. Using social media data, especially Twitter data, it is now likely to do real-time syndromic monitoring (Suggala, 2018), as shown in Figure 2.14.

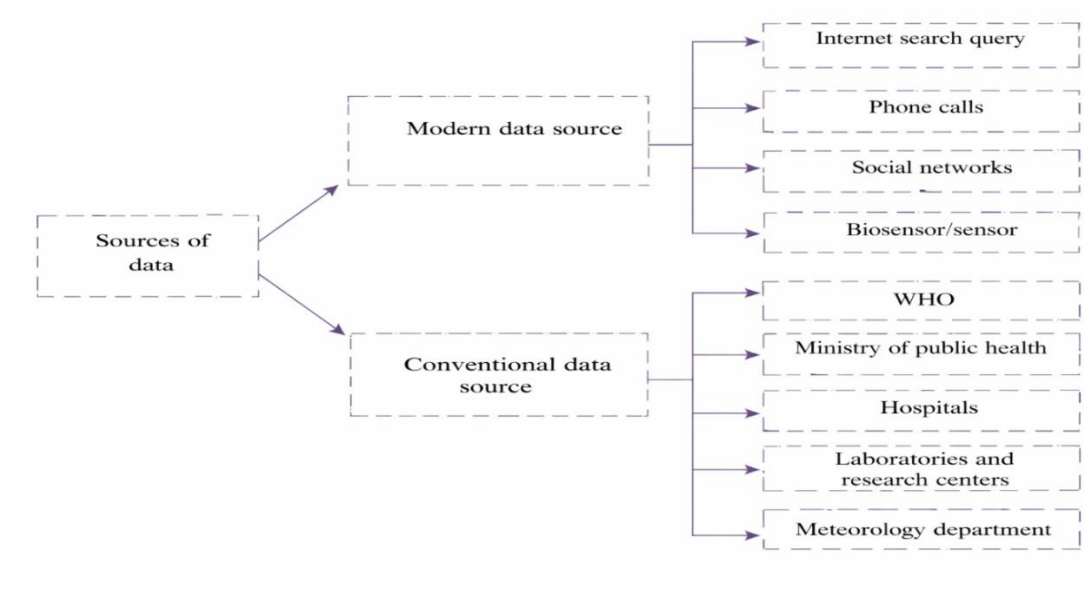


Figure 2.14: Modern and Traditional Data Source and Features for Prediction of Epidemic Outbreaks

A predictive model is a way to make choices and take action based on data that has been collected and analyzed in a way that is scientifically sound. Modeling for prediction is a common way to utilize statistics to guess what will happen in the future. When data is used to train a predictive modeling approach, it creates predictive models. These models are made by integrating data and math techniques, and the learning process entails making a mapping function among input data domains and a response or target factors (Kalechofsky, 2016).

Most researcher have found that accurate malaria prediction models depend a lot on choosing and combining the right data features that show the environmental, temporal, demographic, and epidemiological features that affect the spread of malaria. This is similar to how data features were used to predict malaria in the past. One of the most prevalent forms of data used in models that predict malaria is information about the weather and the environment. These elements are temperature, rainfall, relative humidity, wind spread, and evapotranspiration. All of these things have a direct effect on the cycles of mosquito breeding and parasite growth. A lot of research has demonstrated that there is a substantial relationship among the weather and the frequency of malaria cases. Bhatt et al. (2021), established a important correlation among variations in temperature and precipitation and the proliferation of malaria in Sub Saharan Africa. Merkord et al. (2017), similar identified temperature and rainfall as important forecasters of malaria outbreaks in Ethiopia. This indicated that these environmental elements hold biological significance.

Temporal and seasonal characteristics are often utilized in malaria modeling, especially in areas with marked seasonal transmission patterns. Variables such as the month of the year, lagged climate variables, and malaria incidence from the previous weeks or months are included to account for the delayed impacts of weather on the behavior of mosquitoes and parasites. Lagged variables are particularly essential in measuring the time necessary for environmental changes to result in real disease occurrences (MacLeod et al., 2015). Time series forecasting models, such as LSTM networks, have proven effective in leveraging temporal information to improve prediction accuracy (Merkord et al., 2017).

Socioeconomic and demographic factors have been received more recognition for their impact on malaria risk and health consequences. Researchers have determined the like hood of malaria transmission by analyzing factors such as population density, poverty rates, educational attainment, healthcare accessibility, and housing conditions. Adeola et al. (2015) shown that residing in substandard housing and possessing a low income can increase susceptibility to mosquito bites, particularly in urban slums and peri-urban regions. These factors are crucial for tailoring prediction models to specific places where socioeconomic gaps affect diverse illness outcomes. Health and epidemiological data are a highly significant aspect of models that forecast malaria (Adeola et al., 2015). People typically utilize past malaria case counts, test positive rates, vector control methods, and access to healthcare to forecast when the next outbreak will arise. These traits give us immediate information about how the disease spreads and how well therapies work. Raja et al. (2024) showed that adding data from public health campaigns made models better at forecasting malaria spikes that were about to happen.

Lastly, approaches for selecting features and reducing dimensionality are now standard in malaria modeling to make sure that only the most significant and distinctive features are employed. To make the model work better, methods like tracking with ensemble models like Random Forest have been applied (Mizna et al., 2025), SHAP (SHapley Additive exPlanations) and other techniques that assist individuals understand how single features affect model predictions have also been utilized to make the model outputs more reliable. It is clear that malaria prediction models rely on a diverse array of parameters deliver from environmental, temporal, socioeconomic, epidemiological, and spatial data (Rajab et al., 2024). It is vital to carefully identify, combine, and comprehend these features in order to develop strong models that can help with timely and targeted public health intervention

2.5 Theoretical Frameworks Supporting Predictive Modeling

Artificial intelligence and machine learning use predictive modeling based on a number of theoretical frameworks that describe how algorithms learn pattern from data, apply them to new observations, and produce accurate predictions. These frameworks offer the

conceptual and mathematical underpinning for the creation of the hybrid-based classification and regression model suggested in this study. Statistical learning theory, information theory, and optimization and decision theory are some of the most important theoretical frameworks for predictive modeling. These theories collectively elucidate the rationale behind model learning, data interpretation, and decision-making in the context of ambiguity, which are essential for the predictive modeling of malaria outbreaks.

(i) Statistical Learning Theory

Statistical learning Theory (SLT) constitutes a fundamental theoretical framework of machine learning. Vapnik came up with SLT in the late 20th century which shows how predictive models may find patterns in small data sets while keeping generalization mistakes under control (Vapnik & Izmailov, 2019). The framework's prime aim is to find a balance among model complexity and prediction accuracy. It does this by using Structural Risk Minimization (SRM), which tries to lower both training error (empirical risk) and expected error (true risk (, In this study, SLT helps the regression and classification phases of the hybrid model by providing mathematical reason for maximizing the tradeoff between bias and variance. Ridge, Lasso, and Random Forest are Examples of regression models that reduce anticipated loss by changing weights and regularization parameters. This makes the model less likely to overfit. Using SLT, the hybrid model finds nonlinear connections between climate and disease data while still being able to generalize across datasets. Also, the statistical learning theory forms a foundational framework for machine learning by grounding predictive modeling in principles of statistical inference and risk minimization. Alnuaomi and Albaldawi (2024) explain that statistical learning emphasizes how algorithms use data, with classification and regression as core supervised learning tasks. These theoretical concepts are directly relevant to hybrid models that combine classification and regression approaches, as they help clarify the trade-offs and performance guarantees when predicting put comes such as malaria outbreaks (Alnuaimi & Albaldawi, 2024)

(ii) Bayesian Decision Theory

Bayesian decision theory offers a systematic framework for thinking and making decision in situation of ambiguity. Its assets that optimal decisions can be achieved by minimalizing the expected loss function, utilizing posterior probabilities obtained through Bayer's theorem (Mcnamara & Chen, 2022). BDT is the basis for probabilistic categorization and forecasting approaches in predictive analytics. It does this by combining what is already known with what has been seen. For predicting malaria outbreaks, and different regions have different conditions. The Bayesian framework enables the hybrid model to manage uncertainty by combining previous probability of outbreak incidence with new data evidence, so improving the clarity and dependability of forecasts. Bayesian reasoning also helps with ranking feature importance and implementation of probability-based thresholds in the model's classification phase. Bayesian decision theory is a probabilistic framework for making optimal decisions under uncertainty by combining prior knowledge with observed to update beliefs and minimize expected loss. In machine learning, this approach underlies models that compute posterior probabilities to guide classification and prediction, allowing algorithms to quantify uncertainty and adapt dynamically as new information becomes available. Bharadiya (2023) highlights how Bayesian methods provide a principled foundation for probabilistic inference, improving prediction reliability and uncertainty handling key concepts that support hybrid classification and regression models in forcasting complex outcomes such as malaria outbreaks.

(iii)Ensemble Learning Theory

Ensemble learning Theory asserts that the amalgamation of several weak learners can yield a more robust and precise predictive model than any individual learner in isolation (Rajab et al., 2024). This theory supports the application of ensemble techniques, including bagging, boosting, and stacking, which promote stability, diminish variance, and improve forecast accuracy by fostering model diversity. This study's hybrid framework combines classification and regression ensembles, mostly using the Random

Forest technique, which is a collection of decision trees, to describe nonlinear relationships between climate and health factors (Bharadiya, 2023). Ensemble learning theory elucidates the efficacy of Random Forest in heterogeneous datasets: it consolidates various decision trees, each trained on distinct data subset, to generate a consensus output that mitigates overfitting and enhances generalization (Sun et al., 2024) .

The suggested hybrid model builds on this idea by combining regression and classification ensembles into a two-phase predictive pipeline. This lets it make both categorical and continues predictions for predicting malaria outbreaks. Oyoo et al. (2024) introduce a two-layer ensemble ML model that merges multiple classifiers using a stacking strategy with logistic regression. By incorporating SMOTE for data balancing and article swarm optimization for feature selection, the model achieves improved predictive performance over individual learners, demonstrating the effectiveness of ensemble approaches for enhancing classification accuracy in hybrid prediction systems.

(iv) Computational Learning Theory

The theory of computational learning, particularly the Probably Approximately Correct (PAC) learning framework established by Sundaram 2018 delineates the mathematical criteria that enable a ML model to accurately and reliably learn a target function. PAC learning posits that an algorithm can learn efficiently if it can, with a high degree of probability, provide a hypothesis that closely aligns with the true target function within a minimal margin of error. This approach strengthens the methodological rigor of the hybrid model by highlighting the necessity of equate training data, representative sampling, and appropriate evaluation metrics, including RSME, MAE, and R· The PAC perspective makes sure that the hybrid model is always accurate on both training and testing datasets and that its performance is not just a fluke but is statistically sound. In the context of predicting malaria, this methodology makes sure that the hybrid model stays strong even when the climate and disease patterns change (Oyoo et al., 2024).

Moreover, computational learning theory provides a rigorous theoretical foundation for understanding how machine learning algorithms learn from data, generalize to unseen instances, and perform under varying conditions. Du et al. (2025) highlight key concepts such as generalization bounds, empirical risk minimization, and the probably approximately correct (PAC) framework, which formally characterize the conditions under which learning algorithms are expected to be effective. These theoretical insights elucidate why hybrid and ensemble approaches can enhance predictive performance by balancing bias, variance, and model complexity an essential consideration in the development of robust models for malaria outbreak prediction (Sundaram, 2018).

(v) Information Theory

Information theory provides key quantitative tools such as entropy, mutual information, and divergence measures to assess uncertainty, information content, and the relationships between valuables in data. Jeanray et al. (2015) reviews how information-theoretic concepts are used in machine learning to guide model selection, feature extraction, and performance evaluation, especially in contexts where understanding the information flow between inputs and outputs improves learning efficiency and robustness, In hybrid machine learning models that combine classification and regression, information-theoretic measures can be used to balance complexity and predictive power by selecting features that carry the most relevant information for target prediction, thereby enhancing model generalization for tasks such as predicting malaria outbreaks (Jeanray et al., 2015).

From the perspective of the researchers, Sun et al. (2024) provides a foundational framework for quantifying uncertainty and measuring the amount of information gained from data. It is the basis for feature selection, decision-making based on entropy, and performance measurements in machine learning. Entropy evaluates the impurity or disorder within a dataset, whereas information gain assesses the decrease in uncertainty attained by the utilization of a specific feature for decision making. This work employs information theory to direct the feature selection process, emphasizing characteristics such as rainfall, humidity, and temperature that significantly mitigate uncertainty in

malaria outbreak prediction. In the Random forest framework, information theoretic measurements such as Gini impurity and entropy gain assist in pinpointing the most significant predictors. This theoretical foundation improves the hybrid model's ability to be understood, making sure that its prediction is not only correct but also understandable in term of epidemiology.

(vi) Optimization and Decision Theory

Optimization theory and decision theory furnish the mathematics underpinning for model training and parameter estimation. The objective of theory optimization is to either minimize it maximize objective functions, like loss functions or prediction errors. This is important for both regression and function algorithms (Huang & He, 2016; Sun et al., 2024)(Rokach, 2010; Tan, 2019). Decision theory adds to this by giving us organized ways to choose the best possible outcomes when we don't know what's going to happen. The hybrid model improves accuracy and efficiency by using optimization during hyper parameter tuning, which means choosing the tree depth, learning rate, and regularization coefficients. Also, using Multi criteria Decision Making (MCDM) principles make sure that choosing and testing algorithms takes into account more than one performance metric, such accuracy, RMSE, F1 score, and AUC. This improves both performance and understanding.

(vii) Relevance to the Proposed Hybrid Framework

These theoretical frameworks collectively furnish a strong basis for the formation and assessment of the proposed hybrid-based classification and regression model. Statistical Learning and Ensemble Theories provide predictive strength and generalization; Bayesian decision theory Incorporates probabilistic reasoning and uncertainty management; PAC learning ensures statistical soundness; information theory improves interpretability; and optimization and decision theory support the model's computational efficiency. These frameworks jointly direct the methodological rigor of this research and validate the

scientific legitimacy of the proposed hybrid predictive strategy for forecasting malaria outbreaks.

2.6 The Process of Developing Machine Learning and Data Mining Models' Illustration

Machine learning empowers computers to study from data and independently draw conclusions, showcasing its formidable capabilities. Alternatively, data mining involves identifying patterns in large data sets in order to forecast consequences and gain valuable insights that can lead to increased revenues, cost savings, improved customer relationships, and risk reduction, among other benefits. To implement machine learning or data mining effectively, a structured framework is essential. Although several frameworks are available, no single framework is suitable for all machine learning or data mining purposes. This section examines three selected frameworks, namely CRISP DM, SEMMA, and KDD, to predict malaria outbreaks (Iwendi et al., 2020; Firas, 2023).

In the field of data mining and ML, it is widely recognized that these tools are of great importance to various industries, corporations, and businesses due to their capability to examine and forecast trends and patterns in vast amounts of data that were previously of little or no use. Firas, O. (2023) note that popular methodologies for data mining research include: KDD and SEMMA. Another significant risk that arises from large databases is the potential to waste valuable information, which requires the use of appropriate techniques to extract useful knowledge, as Iwendi et al. (2020) observes that as a result, data mining emerged in the 1980s and has continued to make progress, with different procedure models being presented to guide and enable data mining tasks and their claims. Therefore, it is essential to carefully consider which data mining process model to use when dealing with large or massive amounts of data. This section emphases on the three greatest general data mining process models: KDD, CRISP DM, and SEMMA. These models are widely used by data mining scientists. This section aims to examine and compare these models to determine the best approach for implementing the proposed model (Azevedo, 2015).

2.6.1 Knowledge Discovery Databases

Numerous studies on the Knowledge Discovery in Databases (KDD) process model have been undertaken. Azevedo, A. (2015) conducted research and validated that KDD involves the extraction of hidden knowledge from databases, which needs applicable previous information and a fundamental sympathetic of the claim area and objectives. KDD is an iterative and interactive process model consisting of nine different steps, as pointed out by Plotnikova et al. (2020) in a survey. Azevedo, (2015) confirmed that KDD entails the discovery of knowledge within data and highlights the application of particular data mining approaches.

The next section describes the steps of the Knowledge Discovery in Databases (KDD) process model. The procedure begins with emerging an considerate of the claim field, where the objectives are established from the customer's or stakeholder's perspective in way to advance insight into the problem context and relevant prior knowledge. This is followed by the creation of a target dataset, which involves constructing a focused data collection and selecting a subset of relevant samples or variables (Azevedo, 2015). This stage is crucial because knowledge discovery activities are performed on these selected components.

Subsequently, data cleaning and preprocessing are conducted to ensure that the target dataset is complete, consistent, and free from noise and inconsistencies. Appropriate strategies are developed to grip lost values, outliers, and additional data quality issues. After preprocessing, data transformation is carried out to convert the data into suitable formats for efficient implementation of data mining algorithms. Numerous data reduction and transformation methods are applied at this stage, as highlighted by Azevedo (2015) and illustrated in Figure 2.15. However, Plotnikova et al. (2020), observe that historically, insufficient attention was given to ensuring that the discovered knowledge was effectively utilized in practical applications.

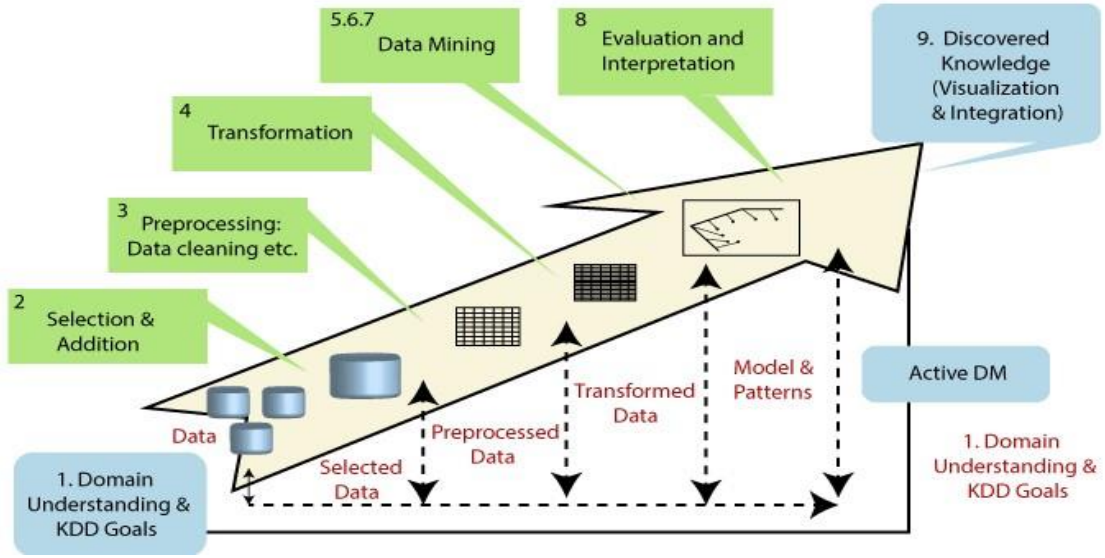


Figure 2.15: Knowledge Discovery Databases Process Model

Once the data has been properly prepared, an suitable data mining task is designated based on the defined objectives, such as classification, clustering, regression, or summarization. Suitable data mining algorithms are then chosen according to their ability to identify meaningful patterns and meet the study's overall criteria. These algorithms are subsequently implemented to extract patterns from the dataset. The mined patterns are interpreted and evaluated, often supported by visualization techniques to enhance understanding and validation. Finally, the discovered knowledge is applied for decision-making or integrated into other systems for further action, ensuring that the insights generated are effectively utilized.

2.6.2 The Model Process Model CRISP DM

In view of the prevalence of diverse data mining approaches, NCR established the Cross Industry Standard Process for Data Mining (CRISP DM) in 1999. CRISP DM 1.0 was the first version to be subsequently published and comprehensively documented, providing a standardized framework and guidelines for data miners.

This methodology comprises six systematically structured and clearly defined phases, as expounded by Schröer, Kruse, and Gómez (2021). The first phase, Business Understanding, focuses on identifying critical factors such as project objectives, business goals, data mining requirements, constraints and relevant technical terminology. The primary aim at this stage is to clarify the problem context and define the key information essential for the fruitful completion of the research.

The second phase, Data Understanding, involves the careful collection of data, assessment of its quality, and exploratory analysis to generate initial insights. This phase seeks to uncover patterns, detect anomalies, and identify potential data-related issues that may influence subsequent stages of the project. It provides a valuable understanding of the dataset and forms the substance for hypothesis development (Schröer et al., 2021).

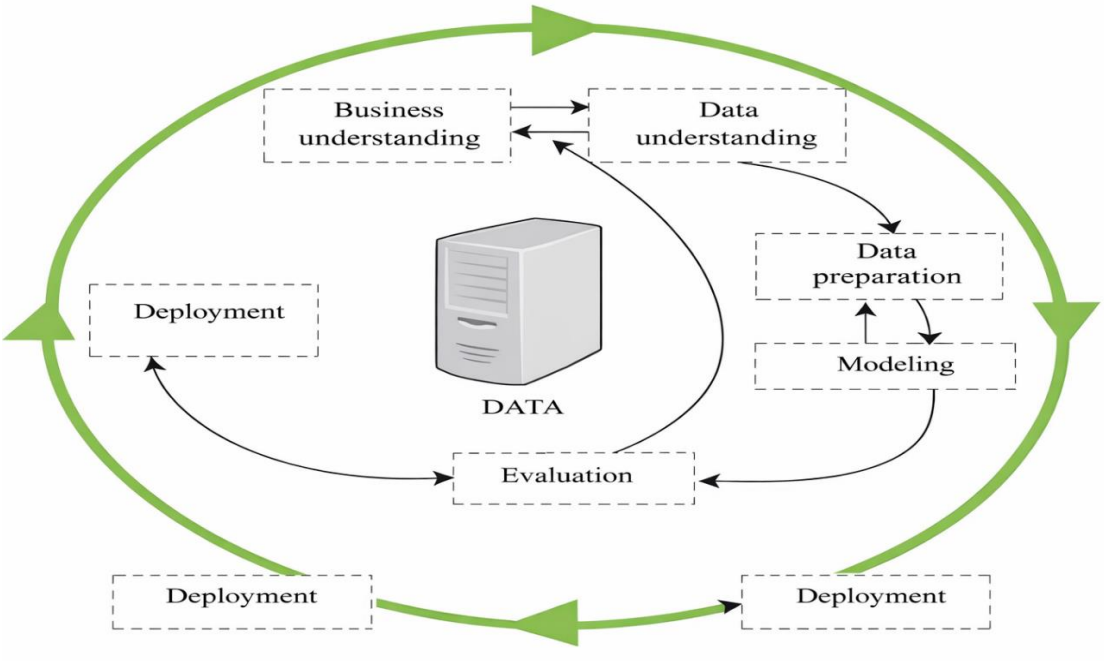


Figure 2.16: CRISP DM Process Model

Data preparation constitutes the third stages and involves constructing the final ultimate dataset to be used for modeling, as illustrated in Figure 2.16. This stage includes selecting relevant records, tables, and attributes, along with performing data cleaning and

transformation tasks. The objective is to ensure that the dataset is accurate, reliable, and properly formatted for examination. Following this, the Modeling phase entails selecting appropriate modeling techniques and implementing them to address the defined data mining problem. During this stage, specific parameters are determined, and multiple models may be developed and calibrated to achieve optimal performance.

The Evaluation phase centers on assessing the developed models and determining whether they adequately meet the predefined objectives. Model interpretation depends on the algorithms used, and the results are carefully examined to ensure reliability, validity, and alignment with business goals. Finally, the Deployment phase focuses on applying the acquired knowledge and insights in practical settings. This phase emphasizes organizing, communicating, and presenting the findings effectively to stakeholders and integrating the results into decision-making processes or operational systems.

2.6.3 The SEMMA Process Model

The contributions of others to data mining methodology remain essential in enabling the creation of new models. The SEMMA framework, created by SAS Institute, is a data mining methodology that facilitates and assists in offering answers to commercial issues and aims and is integrated with the SAS Enterprise Miner, serving as a logical arrangement of useful tools. SAS defined the SEMMA model and incorporated it into its commercial data mining platform, SAS Enterprise Miner (Bashir, 2023), as demonstrated in Figure 2.17.

The SEMMA process contains of five interconnected steps designed to facilitate efficient data mining and model development. The initial stage, Sample, is considered optional and focuses on data sampling. At this stage, a illustrative portion of a huge dataset is selected to reduce computational complexity while retaining meaningful patterns, enabling faster manipulation and analysis.

The Explore stage is dedicated to data exploration, where trends, patterns, and anomalies are examined to enhance understanding of the dataset. This phase supports the refinement of the data mining process by uncovering hidden structures and potential irregularities. Following exploration, the Modify stage involves transforming the data through the creation, selection, and refinement of variables. This phase includes handling outliers, reducing dimensionality, and preparing the data in a manner that improves modeling efficiency and accuracy.

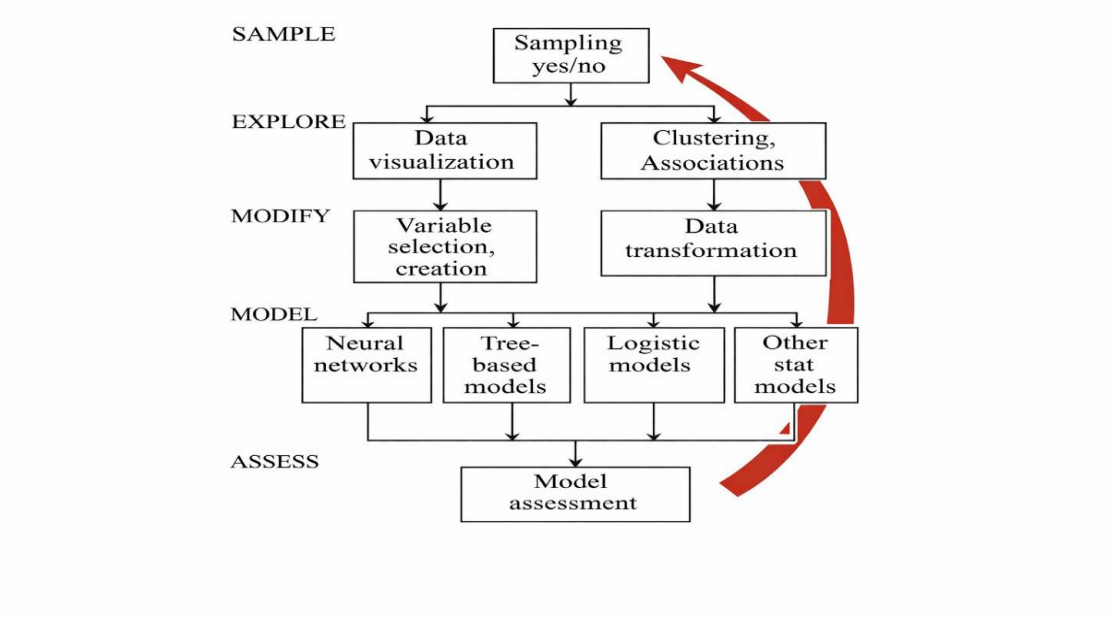


Figure 2.17: Steps in SEMMA Process

The Model stage concentrates on using several modeling methods to the prepared data. The software environment typically evaluates multiple combinations of variables and algorithms, recognizing that each modeling technique has distinct strengths suited to particular data mining scenarios. The final stage, Assess, focuses on evaluating the reliability, validity, and overall usefulness of the developed models. Performance estimation and validation are conducted to ensure that the findings meet the intended objectives.

A comparison of major data mining methodologies reveals that many scholars favor the KDD process model due to its completeness and structured approach, making it the preferred framework in this study. While CRISP-DM and SEMMA are widely applied in business-oriented contexts, SEMMA is particularly integrated within SAS Enterprise Miner. Nevertheless, evidence suggests that CRISP-DM provides broader applicability and effectiveness compared to SEMMA. Overall, these process models offer systematic guidance for researchers and practitioners in applying data mining techniques to real-world problems.

2.7 Related Works in Malaria Outbreak Prediction

Recently, various researchers have proposed diverse categories of algorithms for data mining and machine learning for application in the health care domain. Nevertheless, because of the intricacy of appropriate data types for each algorithm, a specific algorithm might not be suitable for all claims. Hence, the selection of a appropriate data mining algorithm depends not only on the application's intended function, but also upon the compatibility of the dataset.

Table 2.4 delivers a consolidated summary of the current and previous related works, methods and techniques on both classification and regression models that have been utilized by numerous researchers in medical data mining for malaria outbreak prediction. The accuracy rates of these techniques are presented, revealing that prediction models aid health workers in promptly predicting the correct disease. Predictive data modeling has been a crucial data mining task for determining future data states based on past and present variables. Forecasts might be made using regression, classification, or other approaches (Kumar & Singh, 2023).

2.7.1 Manual Methods

Before the widespread adoption of machine learning and computational models, malaria prediction heavily relied on manual approaches grounded in epidemiological surveillance,

climatic observation, and statistical trend analysis. These methods involved direct human interpretation of data from field observations, clinical reports, and meteorological records to forecast potential outbreaks. For instance, health officials and researchers historically relied on seasonal patterns, rainfall levels, and temperature thresholds to anticipate malaria transmission periods, especially in endemic regions where malaria exhibits strong (Hussain-Alkhateeb et al., 2021), This approach was primarily observational and heuristic, drawing conclusions based on empirical relationships between climate and disease patterns.

One widely used manual method was trend extrapolation from historical data, where health records and incidence logs were reviewed periodically to identify potential surges. These efforts often included hand drawn charts, basic statistical summaries, and expert judgment. In the absence of computing resources, predictions were made by projecting previous years' data into the future using visual pattern recognition. This method was effective to a degree but lacked the precision and adaptability to account for dynamic environmental and socio-political changes (Kang, 2022). In regions like sub-Saharan Africa, health workers would often rely on handwritten records and anecdotal evidence to assess malaria risks in local populations (Kapwata & Gebreslasie, 2016).

Another significant manual approach was entomological surveillance, which involved the physical monitoring of mosquito populations, larval breeding sites, and parasite detection in human blood samples. Health officials would assess mosquito densities using light traps and larval sampling, then correlate these findings with known transmission cycles to predict potential outbreaks (L. Wang et al., 2020). Although labor intensive, this method provided valuable insight into vector ecology and transmission potential in specific areas. In some cases, predictions were also informed by interviews with community health workers and local residents about recent febrile illnesses, a practice still in use in remote areas with limited diagnostic tools (Kibret et al., 2019).

Manual malaria prediction also incorporated climatological almanacs and agrarian calendars, especially in rural and agricultural communities. Farmers and local leaders

would associate heavy rainfall seasons with higher malaria incidence, guiding community level preparation and vector control efforts. This community-based knowledge was particularly prominent in regions without formal surveillance infrastructure and often provided early warnings that complemented official monitoring systems (Berihun et al., 2023).

However, despite their historical importance, manual methods had substantial limitations, including low scalability, subjectivity, and poor adaptability to changing epidemiological contexts. Their effectiveness was heavily dependent on expert availability, consistent data collection, and local knowledge, which were often lacking in resource constrained settings. As computational tools and digital health technologies advanced post 2010, these traditional methods were increasingly augmented or replaced by automated systems leveraging statistical and machine learning models for real time and scalable malaria prediction (Tusting et al., 2019).

2.7.2 Statistical Methods

Statistical methods have long played a critical role in modeling and forecasting malaria incidence, offering interpretable and often computationally efficient approaches for understanding disease patterns. Traditional regression models, particularly linear regression and logistic regression, have been extensively applied in early studies to identify associations between malaria cases and climatic, environmental, or demographic variables. For example, Merkord et al. (2017) used multiple linear regression models to correlate meteorological features, including rainfall and temperature with malaria incidence in Ethiopia, demonstrating significant lagged relationships between climate and disease outbreaks. Logistic regression has also been employed to model the probability of malaria occurrence based on binary or categorical outcomes, especially in risk classification scenarios (Adeola et al., 2015).

In regions where malaria data exhibit temporal dependencies, time series numerical models like the Autoregressive Integrated Moving Average (ARIMA) model have been

frequently used. ARIMA and its seasonal counterpart (SARIMA) are well suited to capturing the temporal patterns and periodicity of malaria outbreaks. These models have shown effectiveness in short term Malaria case forecasting when appropriate seasonality and stationary assumptions are met. For example, Mohature et al. (2021) applied SARIMA models in Madagascar to forecast malaria incidence and emphasized the importance of incorporating climatic variables for improved accuracy. In a study conducted by Musa (2015), ARIMA and ARIMAX models were developed applying climate factors and historical identified malaria incidences from 2006 to 2011 as the training set and data from 2021 as the test set to forecast malaria incidences in Sudan for 2013 and 2014. The ARIMAX model was used to inspect the connection among malaria incidences and climate data utilization the lowest Bayesian Information Criterion (BIC) values. The outcomes marked that ARIMA has four diverse models, where the average for all states is (1,0,1) (0,1,1), and ARIMAX models exposed a substantial difference among the situations in Sudan (Mohature & Patil, 2021).

Anwar et al. (2016) conducted research with the goal of establishing a predictive technology for malaria surveillance in Afghanistan using autoregressive integrated moving average (ARIMA) models. The study used malaria data from January 2015 to December 2015, as well as environmental and climate data to assess their impact on the predictive power of the models. Two models were developed, one for near term prediction and another for long term prediction, based on the previous cases of four months and rates 1 to 12 months prior, respectively. The study decided that the ARIMA models can accompaniment current surveillance schemes by providing A deeper comprehension of malaria trends in situations with few inputs, which can be utilized in healthcare planning (Anwar et al., 2016).

Tohidinik et al. (2021) established an early warning system for forecasting malaria occurrence in Southeast Iran using meteorological parameters and morbidity data. The study examines the effect of temperature, rainfall, and relative humidity on malaria cases. Univariate Autoregressive Integrated Moving Average models were created for weekly and monthly malaria incidence forecasting. The weekly model had a better fit with an R2

of 0.863, while the monthly model had an R2 of 0.424. Although the meteorological factors were not statistically important in the monthly model, minimum and maximum temperatures showed significance (Tohidinik et al., 2021).

Another important class of statistical techniques used in malaria prediction includes generalized linear models (GLMs) and generalized additive models (GAMs). GLMs extend the traditional linear model framework by permitting for non-normal deliveries of the response factors, which is particularly useful in modeling count data like malaria cases. GLMs have been applied with Poisson or negative binomial distributions to account for over dispersal in malaria data (Bai et al., 2019). GAMs, on the other hand, introduce flexibility by incorporating nonlinear relationships between predictors and outcomes. MacLeod et al. (2015) used GAMs to model complex relationships between sea surface temperatures, rainfall, and malaria incidence in Southern Africa.

In addition to these classical models, spatial statistical methods have been functional to explore the geographic delivery of malaria risk. Bayesian hierarchical models are commonly used in spatial epidemiology for malaria prediction, allowing for uncertainty quantification and the incorporation of spatially structured random effects. For instance, Damiana et al. (2020) developed a new model for evaluating malaria risk in Chimoio, Mozambique to enhance prediction accuracy.

Despite their advantages in interpretability, statistical models often assume linear relationships or stationary, which may not hold in all malaria prediction scenarios. Nevertheless, they continue to be widely used either standalone or in hybrid frameworks with machine learning models to power their strengths in inference and explain ability (Damiana et al., 2020). The aptitude of statistical models to deliver insights into the underlying drivers of malaria transmission remains a key reason for their continued application in public health decision making.

2.7.3 Machine Learning Methods

As the review of existing ML based malaria prediction models, from this perspective, the evolution of automated approaches for malaria prediction has significantly transformed disease surveillance and control strategies, particularly since the early 2000s. These methods aim to reduce human bias, improve scalability, and enhance prediction accuracy by leveraging data driven techniques. Among the earliest automated tools were rule based expert systems, which encoded epidemiological rules to trigger alerts based on climatic and environmental thresholds. Although limited in adaptability, these systems provided the foundation for more dynamic models by formalizing domain expertise (Musa et al., 2024).

A major leap in automated malaria prediction came with the machine learning adoption techniques. Algorithms, including decision trees, SVMs, random forests, and neural networks to outperform traditional statistical models by capturing nonlinear relationships in complex datasets. For instance, Kibret et al. (2019) used decision tree classifiers to predict malaria infection using demographic and clinical data with high accuracy. Similarly, random forest models have been shown to provide robust performance in malaria prediction by managing high dimensional data and minimizing overfitting (Kapwata & Gebreslasie, 2016; Kibret et al., 2019).

Another key development has been the combination of remote sensing data and geospatial analytics, where satellite derived variables to be incorporated into automated prediction systems using geographic information systems and spatial ML models. Rono (2018) demonstrated the power of satellite data in predicting malaria hotspots in Kenya using automated algorithms that combine environmental and health data streams. These systems often run continuously and in near real time, allowing for proactive public health interventions.

More recently in the literature, deep learning models including convolutional neural networks and recurrent neural networks have emerged in malaria prediction, particularly

in image-based diagnostics and time series forecasting (Rono, 2018). CNNs have been used to automatically diagnose malaria parasites from blood smear images with high precision (Rajaraman et al., 2018), while RNNs and long short-term memory networks have revealed promise in predicting seasonal malaria trends using sequential environmental data (Adigun et al., 2024). These models can automatically learn abstract patterns without the need for handcrafted features, making them highly effective in data rich environments.

Hybrid models, which merge several machine learning or statistical approaches, are progressively applied to advance accuracy and consistency. For example, models that integrate ARIMA for temporal trend detection and random forest for environmental feature extraction can outperform standalone models in malaria forecasting (Oyoo et al., 2024). Additionally, automated feature selection and ensemble learning methods have further enhanced predictive capabilities, enabling more adaptive models in diverse epidemiological settings.

The transition to automated methods has also facilitated the development of mobile and cloud-based malaria prediction platforms, especially in remote areas. Mobile health applications powered by automated models are being deployed to assist frontline health workers with real time diagnostic support (Chilamkurthy et al., 2018).These tools are vital for expanding access to predictive capabilities in low resource settings, contributing to the goal of malaria elimination through early warning and timely interventions.

Data mining and machine learning as research areas, a predictive modeling system is typically developed by identifying rules and patterns from a training dataset and then applying those rules to predict the class of records with unknown class labels. In earlier years, researchers in this field demonstrated ways to enhance accuracy and precision by creating automated predictive and diagnostic models of diseases using various approaches for regression as well as classification. This section introduces a number of pertinent methods. According to Agarwa et al. (2024) research on both the regression and classification approaches, they are useful in predicting the class or final result of a

function. The primary distinction among these two strategies is the type of attributes being analyzed. For categorical attributes, classification algorithms such as Nave Bayes or Support Vector Machines can be used. When dealing with continuous characteristics, however, regression models based on SVM or linear regression work admirably (Agarwal & Tiwari, 2024). Moreover, as illustrated in Figure 2.18, various other theoretical frameworks have been proposed.

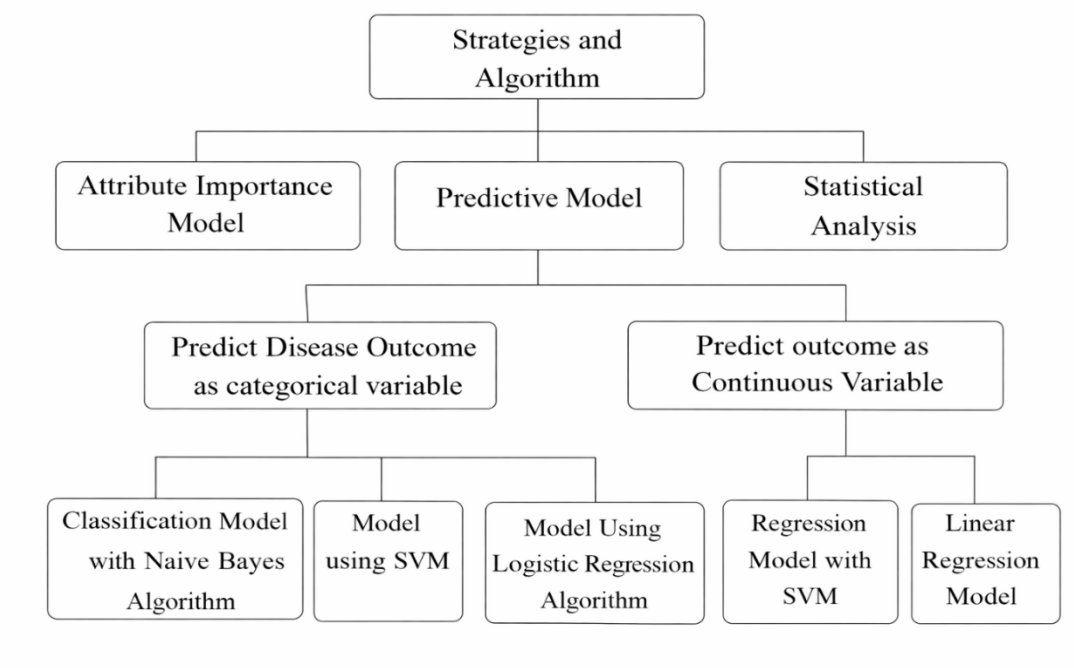


Figure 2.18: Operation of Regression and Classification Methods

In contrast, Azezew et al. (2025) demonstrated through his research that logistic regression is a regression analytics method utilized for forecasting the result of a categorical dependent features with a restricted number of variables. These dependent variables have no inherent meaning, while the order of magnitudes may or may not be significant. The logistic equation is used in regression to generate values between 0 and l, as shown below.

$$f(t) = \frac{1}{1 + e^{t}} \tag{2.10}$$

Balshi et al. (2020) developed a model to forecast the occurrence of short-term deaths following ICU evacuation. It is generally acknowledged that numerous critically ill patients encounter medical decline or death shortly after being released from the ICU(Balshi et al., 2020). Figure 2.19 depicts Grover and Jiang's (2026) description of a machine learning model for forecasting an influenza outbreak using Twitter data. As demonstrated in Figure 2.19, the model utilizes time series classification and prediction techniques for determining the epidemic phase by evaluating the probabilistic model of the bag of words (BOWs) vocabulary utilized throughout different phases of the epidemic and propagated online through tweeting(Azezew et al., 2025a)

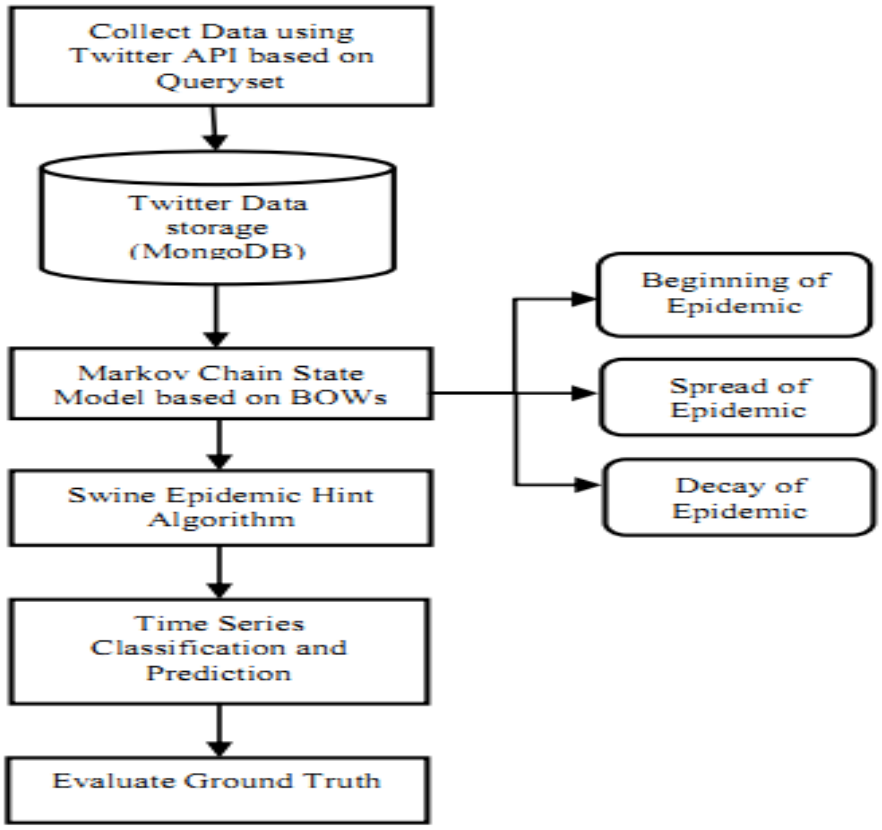


Figure 2.19: Prediction Model for Influenza Epidemic Using Twitter Data

Mbunge et al., (2022) suggested a model that uses regression trees and forests to forecast the occurrence of malaria cases in Mozambique. Malaria is a critical issue in the public

health in Mozambique, according to the researchers, with documented cases in practically every province. The objective of this investigation was to look at prediction models for the malaria case numbers in Maputo province's regions(Jiang et al., 2026). The research used information including temperature, rainfall, and humidity and applying the statistical software R, regression trees and random forest models were constructed and utilized to forecast the malaria cases number throughout a year according to the previous year's records. Mean Squared Error and the correlation coefficient have been employed for assessing the models. Indoor Residual Spray, the minimum temperature, and rainfall factors showed to be the greatest significant features in forecasting the malaria cases number, especially in regions with a high malaria occurrence. Significant improvements in predicting performance may additionally be obtained by shortening the time window for incorporating past data (Jones-Farmer et al., 2017; Saqr & López-Pernas, 2024) .

Gozali *et al*., (2024) undertook a research investigation into current information discovery processes in databases, emphasizing data mining approaches employed in medical studies, particularly for predicting heart disease. A several of experiment have been done to assess the performance of prognostic algorithms for data mining on the similar dataset, consistent with the results presented in the research. The findings revealed that decision trees outperformed various predictive techniques, including KNN, neural networks, and cluster-based classification. Bayesian classification sometimes has similar accuracy to a decision tree, but it does not perform as well. In addition, Gozali et al., (2024) found that after using a genetic algorithm to minimize the real data size to achieve the ideal subset of attributes needed, the decision tree accuracy and Bayesian classification increases substantially more for heart disease predictions.

In a similar vein, Bahrami and Shirvani (2015) investigated that assessed multiple classification algorithms for heart disease diagnosis. The classifiers used were K Nearest Neighbors, J48 Decision Tree, Naive Bayes, and SMO. After classification, the following evaluation measures have been examined and compared: precision, precision, accuracy, specificity, sensitivity, F measure, and area under the ROC curve. In accordance with the discoveries, the J48 Decision Tree was the highest performing classifier for heart illness

diagnosis utilizing the data set provided. Mehbodniya et al. (2022) supported their findings by trying to establish a heart disease forecast system using data mining approaches to find helpful patterns in medical data, aiding informed decision making. In the same period, Bahrami & Shirvani (2015) worked hard to report that heart disease is responsible for millions of deaths annually, making the utilization of data mining tools in the detection of heart disease fundamental. The research they conducted put forward an approach for predicting and diagnosing cardiac disease utilizing methods for data mining, with a view to evaluating different classification techniques. After classification, performance evaluation measures were applied in comparison with the classifiers. The output displayed that, considering the currently available dataset, the J48 decision tree is the greatest classifier for heart illness diagnosis (Gozali et al., 2024). Figure 2.20 depicts the proposed method for forecasting and diagnosing heart disease.

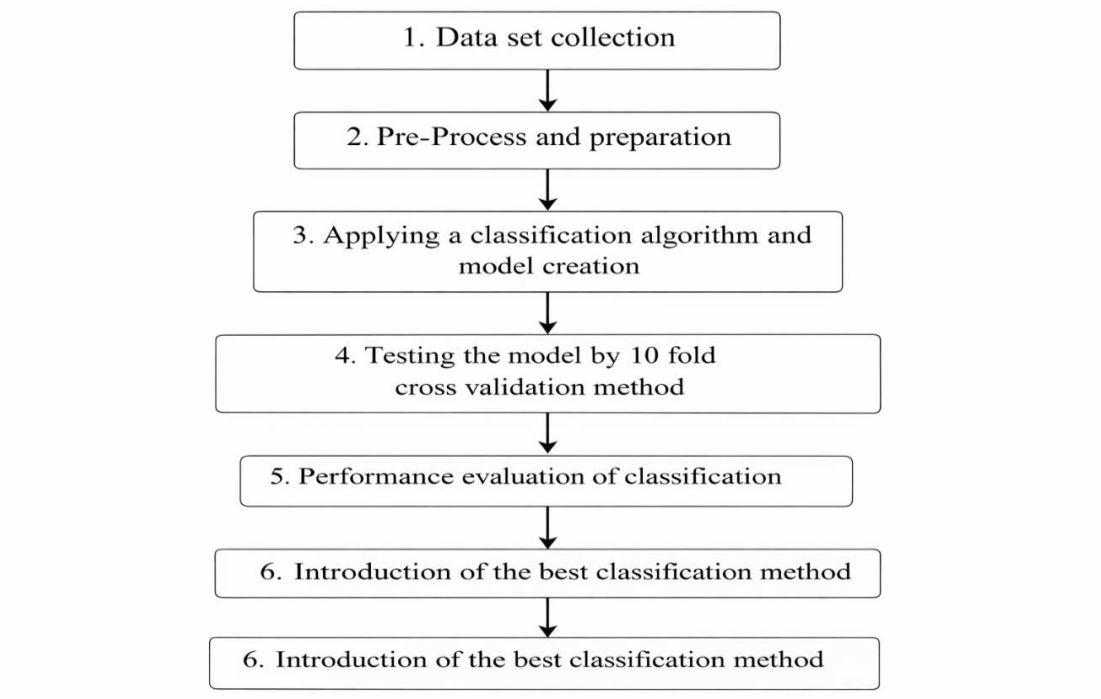


Figure 2.20: Overview of the Proposed Sequential Approach for the Prediction and Diagnosis of Heart Disease

The process of making usage of ML methods for detecting malaria disease involves the use of algorithms that employ arithmetical, probabilistic, and optimization approaches to learn from historical data and notice patterns in huge, composite datasets. Algorithms based on machine learning are designed to learn and enhance their operations by examining input data and making predictions within a reasonable range. Algorithms themselves can be generally considered into three types: supervised, unsupervised, and semi supervised, depending on their purposes and the way they are trained (Bahrami & Shirvani, 2015; Mehbodniya et al., 2022). Machine learning procedures that have been supervised are trained using a labeled dataset and then fed with an unlabeled test dataset to categorize them into similar groups. These algorithms are appropriate for classification and regression issues, where the output variable is either discrete or continuous. For instance, the output variable could be characterized into different groups or categories,

including diabetic and non-diabetic, or could represent a real value, such as the risk of developing cardiovascular disease for an individual.

In summary, the process of employing machine learning approaches for malaria illness detection involves the use of algorithms that are programmed to learn from past data and detect patterns in large, complex datasets. As indicated by the Uddin et al. (2019) and illustrated in Figure 2.21, the Supervised machine learning methods are trained on labeled datasets and are capable of tackling problems with regression and classification.

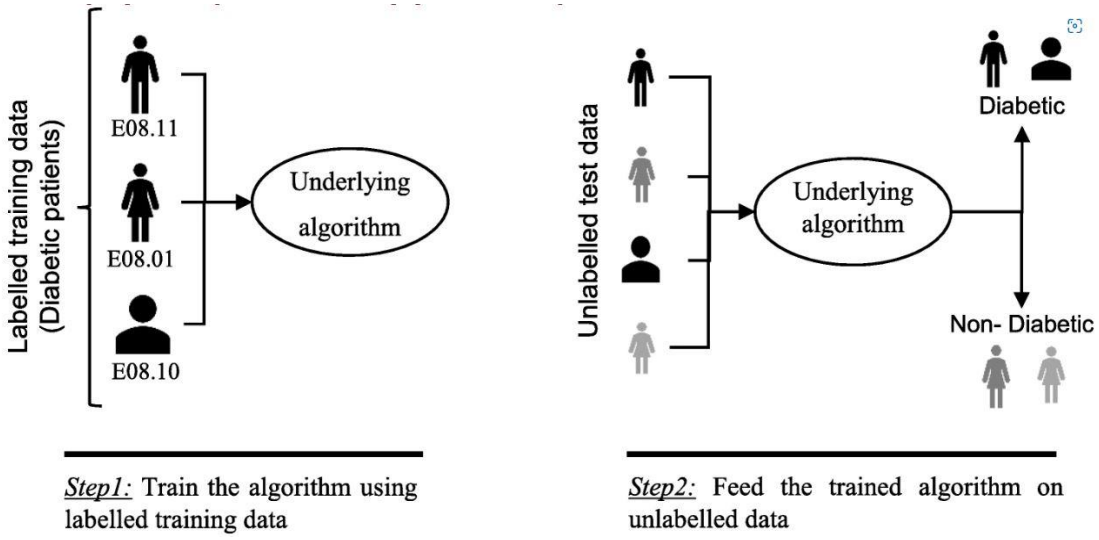


Figure 2.21: Design of How Supervised Machine Learning Procedures Work to Categorize Diabetic and Non-Diabetic Patients Based on Abstract Data Individual

Kaur & Bawa (2015) researched the development of data mining prediction models for various diseases, including heart illness, kidney stones, skin cancer, and lung cancer. These models were designed to improve the reliability of disease diagnosis and early prevention. Kaur's research highlighted the need for a strong data mining model that could bridge the gaps in current disease prediction methods.

Hybrid models have the potential to improve process comprehension and increase prediction performance by combining first principles knowledge with multivariate data analysis methodologies. The structure of parametric models is fixed based on first

principles knowledge, while nonparametric models are identified from data. Hybrid models can reduce the general operation count by utilizing the domain propagators best suited to the various parts of a problem (Kaur & Bawa, 2015; Uddin et al., 2019).

The prediction of diseases caused by climate change is crucial, and big data plays an imperative role in this process. Weather affects human society and life, and can create various diseases, including vector borne, water borne, air borne, and food borne diseases. The increasing accessibility of health and emergency data from web and social media sources has enabled earlier detection of outbreaks and disease surveillance.

Several techniques have been presented by various authors for disease outbreak prediction. The study of Sharma et. (2019) developed in India, a machine learning algorithm has been used to forecast malaria outbreaks. As training data, they employed an enormous set of data from Maharashtra state between 2011 and 2014, and they employed two machine learning classifiers, support vector machine and artificial neural network, to identify the best model for predicting malaria outbreaks. Climate related variables such as rainfall, temperature, and humidity have been considered, as well as clinical data such as the overall number of positive malaria cases, Plasmodium Falciparum (Pf), and the occurrence of malaria outbreaks as binary outcomes (yes/no). The authors assessed the performance of the models using the Receiver Operating Curve and Root Mean Square Error metrics and found that the SVM model was more accurate than ANN in predicting outbreaks, with a lead time of 15 20 days (Sharma et al., 2019).

Wang et al. (2020) proposed a new technique for predicting malaria outbreaks that combined SVM and Firefly Algorithm. They found that the performance of SVM models relied heavily on the choice of parameters. In addition, the study of Sharma et al. (2019) confirmed that SVM outperformed ANN in predicting malaria outbreaks using a dataset from Maharashtra state. The authors reported that the SVM model had a lower root mean squared error (0.12) and higher correct detection rate (89%) than the ANN model (0.47 and 77%, respectively).

Modu et al. (2017) created and put in place a smart early outbreak warning system for malaria that forecasts malaria outbreaks based on climatic features via machine learning. The study compared seven machine learning algorithms to determine the best prediction model, which included linear regression, logistic regression, decision tree, support vector machine, optimized support vector machine, naive bayes, and k nearest neighbors are all examples of regression methods. Support vector machine was found to offer the best results, with an 80.56% prediction rate. The study applied the partial least squares path modelling practice to analyze the causal relations among meteorological variables and malaria outbreaks. The study used 85,627 confirmed malaria cases reported between 2009 to 2013. The authors described a discovery accuracy rate of 99% with SVM, 75% with logistic regression, 64% with decision tree, and 81% with KNN more techniques to improve the accuracy of the predictions.

Time series analysis is a technique of analyzing data gathered over a specific time period that possesses internal structures such as seasonal variety and auto correlation (Azezew et al., 2025b). The clinical information typically comprises time series data collected at different Hussain et al. (2023) developed an image examination structure for malaria identification and classification using digital image analysis. The device comprised motorized phase units, which could be simply fitted on conventional light microscopes applied in outbreak areas. The prototype properly determined parasite positive and parasite negative blood films with 95% and 68.5% accuracy, correspondingly. The classification results for thick blood films with the Pf parasite was 75%, while the Pf classification accuracy was 90%.

Rajaraman et al. (2018), Pre-trained CNN-based deep learning models has been used to categorize uninfected and parasitized blood cells, thus enabling illness determination. An experimental style was used to find the optimum model style founded on the underlying data. The anticipated model included three convolutional layers and two fully connected dense layers. Model results were assessed applying architectures such as VGG16, AlexNet, Xception, DenseNet121, and ResNet50 on datasets of uninfected and parasitized blood cells.

Simonyan & Zisserman, (2014) from the University of Oxford developed a model called "Convolutional Networks with Extremely Deep Layers for Large Scale Image Recognition." The model achieved 92.7% top 5 test accuracy in ImageNet, which is a dataset of over 14 million images fitting to 1000 classes.

Thakur & Dharavath (2019) conducted a study on predicting malaria abundances in four provinces of India using artificial neural networks. The study utilized big data on climate variables from 1995 2014, combined with symptomatic malaria cases. The prediction model was established by utilization a feed forward neural network, and the root mean square error was applied as the assessment metric, with a fixed threshold of 150%. The results showed variations in prediction accuracy among different areas based on clinical factors and precipitation, with the average error ranging from 18% to 117%.

Sharma et al. (2019) constructed and implemented a Malaria Outbreak Prediction Model with the SMV and ANN classification using machine learning algorithms. The ROC and the root mean squared error were applied as predictors. SVM had a Root Mean Squared Error of 0.12 and an ROC area of 0.89, while ANN had a Root Mean Squared Error of 0.47 and an ROC area of 0.77, indicating the need for better models to improve accuracy. However, the performance of these models varied, and the current research suggests combining multiple algorithms for more advanced results (Hailu, 2015; Sharma et al., 2019), as showed in Figure 2.22.

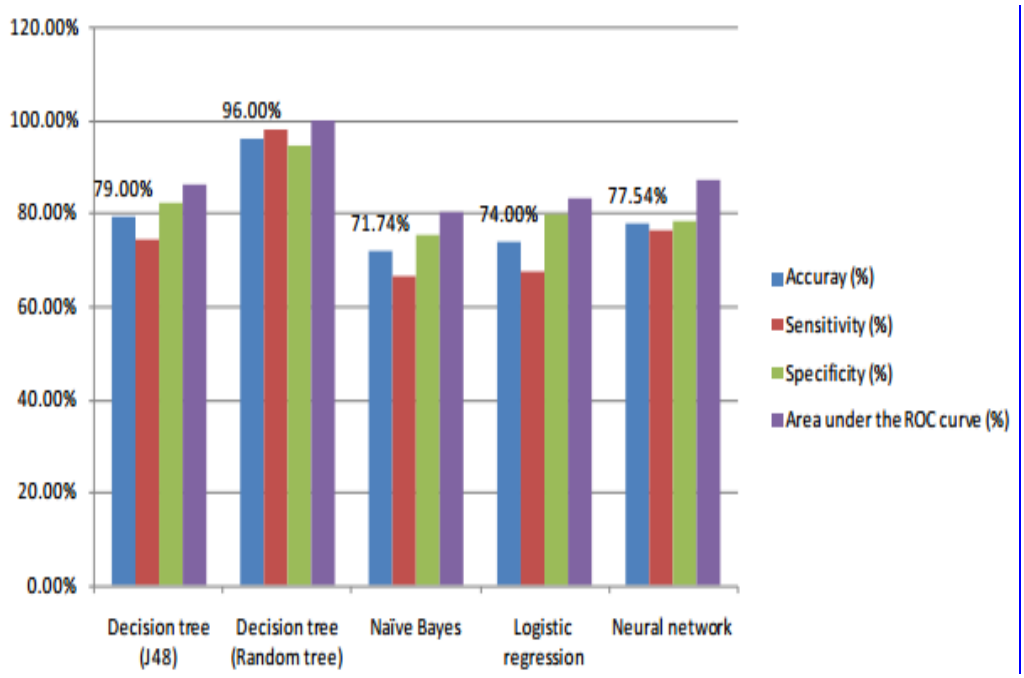


Figure 2.22: Measures of Model Performance Evaluation

Bharadiya (2023) exploited Naive Bayes as a supervised learner to build a model for malaria prediction. The model considered many clinical factors such as headaches, fever, nausea, vomiting, respiratory distress, convulsion, and coma. The model was implemented using Java and achieved a high results accuracy of 90% and 98% on the confusion matrix.

Modu et al. (2017) introduced a smartphone application that serves as an intelligent early warning system for malaria outbreaks. This method analyzes environmental factors to forecast malaria outbreak utilizing machine learning. The research was conducted in four stages: data collection from repositories, identification of hidden ecological factors, determination of causal relationships among ecological factors, and employment of 10 machine learning techniques. As a result, a mobile application had been created that used the support vector machine as the greatest predictor. The tool uses free meteorological data and a geographic application software design interface to forecast a malaria breakout numerous days in advance (Bharadiya, 2023).

In the study conducted by Masinde (2020) on the prediction of malaria epidemics in Africa, past malaria incidence and climate data from all African countries affected by malaria in the past 18 years (2000-2017) were used to forecast future malaria occurrence using nine machine learning algorithms by building the prediction model in MATLAB software, using a dataset of 272,832 rows of climate data from the World Knowledge Portal. The algorithms were evaluated to determine the ensemble of algorithms with the greatest performance. The results displayed that logistic regression, fast large margin, decision tree and general linear model were the four best algorithms based on those results metrics. Nkiruka suggested and built a machine learning driven model for identifying malaria occurrence in 2021 through leveraging climate inconsistency data from six Sub Saharan African nations during a 28-year period. The construction of the model started using the engineering of features in order to identify climate parameters which impact malaria happening. Following that, the k means clustering approach was employed for identifying anomalies, followed by the execution of the XGBoost algorithm for classification. While the specific connection between malaria incidence and climate variability varies by geographical area, non-seasonal fluctuations among three climatic factors serve an essential part in the development of malaria outbreaks.

In Azezew et al. (2025) study on the dependability of forecasts by the utilization of the hybrid models for malaria occurrence rates in Uganda, the decision tree and Naive Bayes classifiers were used to create a prediction hybrid model. The findings showed that the hybrid classifier achieved an accuracy of 79.3% and an F measure score of 84.2%. However, the study also revealed the need to merge levels of granularity. Using these data parameters, predictions can be made (Suggala, 2018). Patel et al. (2019) proposed a malaria outbreak detection system utilizing machine learning methods and utilized Support Vector Machines to determine whether a malaria outbreak is likely to occur given the prevailing weather conditions. The authors suggest that future work can use more localized data to expand the accuracy of the predictions. Additionally, in the setting of various diseases, the Naive Bayes technique was found to be a commonly utilized technique in data mining (Chandra Patel et al., 2019; Kaur & Bawa, 2015). The strategy

was reported to reach a remarkable accuracy of 96.5% in treating heart patients Figure 2.23, and it has since become a commonly used approach for illness forecast with the uppermost level of accuracy. As a result, the research suggests integrating Nave Bayes with additional algorithms to boost accuracy even more.

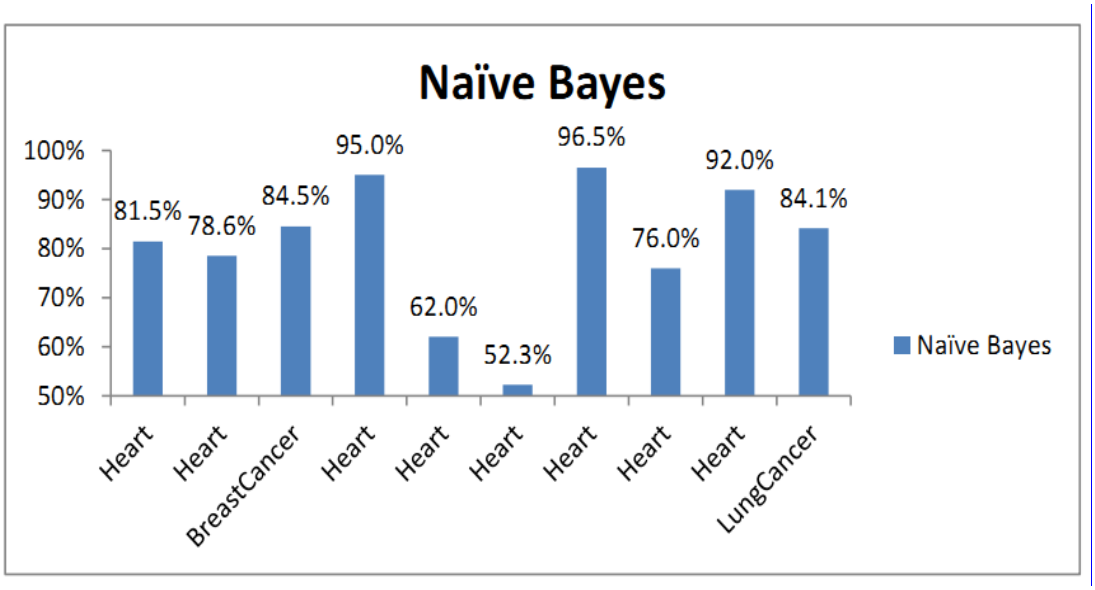


Figure 2.23: Evaluations of Some Diseases on Naive Bayes

Mbunge et al. (2022) A research was carried out in Rwanda to forecast malaria outbreaks using environmental data, and machine learning methods, particularly six classifiers, were used and Random Forest outperformed the other classifiers, scoring more than 90% across all evaluation measures. The Random Forest classifier achieved 90.75% accuracy, a F score of 90.73%, precision of 90.69%, and recall of 90.88%, Abisoye & Rasheed (2018) conducted studies on malaria parasite numbers in Mina Metropolis, Niger State, and Nigeria. They collected 1200 experimental data from climate data from NICOPE, Bosco, Niger State, FUTI Minna, and Nigeria and employed two classifiers to forecast the model: support vector machine and artificial neural network and the support vector machine outperforming the artificial neural network. The researchers indorse that next studies emphasis on increasing the model's performance, perhaps through the use of hybridized models (Abisoye & Jimoh, 2018; Mbunge et al., 2022). In a comparable vein Seth et al.

(2019) with a large dataset, the two well-known data mining classification approaches, support vector machines and artificial neural networks were applied, to generate an early warning tool for possible malaria outbreak. The research evaluated the performance of the models using Root Mean Square Error and receiver operating characteristics. The research observed that the SVM model outperformed the ANN model. But the research's suggested hybrid model performed better, with 96% training accuracy and 93% testing accuracy. According to the study, this hybrid model may be a more actual method for forecasting malaria outbreaks than previous models.

Recent efforts to explore the complexity of malaria have employed machine learning procedures. Denbew et al. (2017) designed a Malaria prediction model and based on the data set used in the analysis, the study showed that Random Forest achieved the uppermost degree of accuracy of 97.72%, AUC of 98%, and precision of 100%. Ma (2025) planned and established a model for predicting malaria outbreaks using machine learning techniques, where Naive Bayes, Support Vector, Linear Regression, Logistic Regression and K Nearest Neighbor were applied. Kaur and Bawa (2015) did research on data mining approaches employed by health care organizations to forecast numerous medical illnesses. The research sought to examine data mining methods and methodologies for optimum prediction of medical diseases, hence improving the effectiveness and effectiveness of the medical field. The research discovered various clinical data mining applications in the disciplines of healthcare and public health, including disease outbreak detection and analysis of healthcare centers for better policy making. The research recommended the merger of various specified algorithms to enhance accuracy in the diagnosis of diseases, especially with imperceptibly identified datasets (Deribew et al., 2017; Sheth et al., 2022).

2.7.4 Hybrid Approaches

Hybrid models can be utilized for a extensive range of tasks, including as determining fraud, making recommendations, recognizing images, and processing natural language. The specific objective, the quantity and quality of the data available, and other factors like the amount of computing power and time available will all affect the choice of model or

models. Another example of a hybrid model is the convolutional recurrence neural network, which combines a recurrent neural network and a convolutional neural network to recognize images in a series of images. This model has three main parts: a CNN, an RNN, and a connectionist temporal classification layer (Shi et al., 2015). Hybrid models, which mix diverse modeling techniques, are becoming more popular for predicting malaria outbreaks because they may use the strengths of diverse procedures and make up for the weaknesses of each specific model. These methods usually mix statistical methods with machine learning techniques or utilize more than one machine learning algorithm to make predictions more accurate and reliable. The goal of hybrid modeling is to make prediction systems more flexible so they can handle the complicated, nonlinear, and context-sensitive dynamics of malaria transmission, which are affected by biological, environmental, climatic, and socio-economic factors.

One of the first successful integrations was combining autoregressive integrated moving average (ARIMA) models with machine learning classifiers like random forests or support vector machines to deal with both nonlinear relationships and trends over time. Stephen et al. (2021) suggested a combined ARIMA random forest framework to predict malaria outbreaks by using ARIMA's ability to predict time series and random forest's ability to recognize patterns in multiple variables. The hybrid model was more accurate and reliable than ARIMA and machine learning algorithms on their own, especially when it came to dealing with changes in the seasons and the environment. Some researchers have combined deep learning with regular machine learning algorithms to make malaria prediction better. Adigun et al. (2024) utilized a hybrid approach of long short-term memory networks and gradient boosting machines to analyze both sequential and intricate correlations in weather, demographic, and epidemiological datasets. The hybrid design shown enhanced efficacy in simulating the lagged impacts of environmental factors, including precipitation and temperature, on malaria incidence. Traini et al. (2022) also built a hybrid system that combined convolutional neural networks for extracting features from blood smear images with decision trees for final classification. This made it possible to detect parasites more quickly and accurately in clinical situations. However, it also

introduced certain challenges. For example, they are more difficult to compute and require high-quality data from many sources. But as cloud computing capabilities and digital health infrastructure become more widely available, these technologies are becoming more and more useful for use in areas where malaria is common. Hybrid models are one of the most advanced and reliable ways to predict malaria outbreaks as of 2024. They are also viewed as an important step toward being able to predict diseases in real time and on a large scale.

This research emphases on a hybrid based regression and classification model for malaria outbreak prediction. In epidemiological research, the word "prediction" has a broader and less precise meaning. It contains all studies that estimate epidemiological characteristics that have intrinsic forecasting value and models that look into the mechanistic drivers of these characteristics. (Perkins & Hakim, 2016). To make a prediction, you look at how an attribute is now and how it has been in the past to guess how it will be in the future (Stephen, 2021). This means looking at past instances or events to see how classification, pattern matching, trends, and relationships work, and then using developed predictive models to guess what will happen in the future. Predicting malaria outbreaks is very important because they kill millions of people every year. Health agencies can act to stop the spread of the disease if they know about outbreaks early (Sharma et al., 2019).

2.8 The Summary

Recent advances in malaria prediction have increasingly leveraged machine learning, deep learning, and ensemble approaches to improve outbreak detection and incidence forecasting. Miggo et al. (2023) developed machine learning method using environmental, socioeconomic, and demographic data with algorithms such as Random Forest, Gradient Boosting, and Logistic Regression. Their best-performing model achieved 92-94% accuracy with strong sensitivity across age groups; however, it focused primarily on classification, without addressing regression-based incidence prediction, highlighting the potential benefits of hybrid classification-regression modeling. Similarly, Ma (2025) proposed MalariVis, a bimodal framework combining regression for case forecasting and

classification for outbreak detection using climatic and clinical data, achieving over 90% accuracy with low RMSE. Nonetheless, deeper hybrid integration and ensemble optimization were not implemented.

Idris et al. (2021) provided a comprehensive review of machine learning techniques for epidemic and malaria outbreak prediction, including LSTM, CNN, Random Forest, and ensemble models, reporting accuracies between 85% and 97%. While informative, this review did not propose or experimentally validate a unified hybrid model. Perez-Saez et al. (2025) explored spatial and temporal trends of infectious disease outbreaks in Africa using statistical and machine learning-assisted modeling, achieving high spatial prediction reliability (AUC >0.90), yet their work emphasized risk mapping rather than integrated predictive models combining classification and regression. Tharageswari et al. (2025) intended a hybrid deep learning model combining CNN feature extraction with traditional classifiers, improving accuracy by 6-10% over standalone models. Although promising, the approach was not malaria-specific and lacked epidemiological adaptation.

In 2024, Alade et al. developed models using climatic variables and surveillance data, achieving approximately 90% accuracy, but regression modeling of malaria incidence was absent. Kibret et al. (2024) integrated environmental and meteorological features in predictive models with AUC >0.88 but did not employ a unified hybrid framework, while Islam et al, (2022) applied ensemble techniques to time-series climate and health data, reducing prediction errors by 12%, yet outbreak classification and hybrid decision modeling remained unaddressed. More recent studies in 2026 have further established the possible of hybrid and computational intelligence approaches. Miggo et al. (2023) applied ensemble and deep learning methods for disease outbreak prediction (AUC>0.90) but did not integrate regression for outbreak magnitude. Zhang et al. (2026) developed deep learning time-series models for infectious disease incidence forecasting with reduced RMSE, though without classification integration. Yu et al. (2018) explored hybrid intelligent systems combining neural networks and rule-based reasoning for healthcare decision support; however, their approach was generic and not malaria-specific, and lacked explicit classification-regression hybridization. These studies collectively

highlight persistent gaps in current malaria prediction research. Most models examine classification or regression separately (Idris et al., 2021; Ma, 2025; Miggo et al., 2023), often relying on single data types such as clinical or climatic variables, and frequently neglect the temporal and spiritual complexities of malaria transmission influenced by environmental and human behavior factors (Githeko et al., 2021). Furthermore, ensemble and hybrid approaches, which could improve model robustness by combining complementary algorithms, remain underutilized, particularly for integrating heterogeneous data sources including meteorological, demographic, spatial, and clinical datasets (Githeko, 2021; Islam et al., 2022; Modu et al., 2017).

Addressing these limitations, this study proposes a hybrid based classification and regression model that simultaneously predicts outbreak occurrence and case counts, incorporates multi-sources data fusion, and leverages cutting-edge machine learning techniques such as ensemble learning and deep learning (Majeed et al., 2023) . By doing so, it purposes to enhance forecast accuracy, robustness, and generalizability across diverse epidemiological and geographical settings, offering an advanced framework for early-warning systems for malaria out breaks. The summary of the relevant studies is presented in Table 2.4.

Table 2.4: A Summary of the Recent and Past Related Studies Methods and Techniques

Authors (Year)	Description Study / Approach	Key Findings of the Critiques
Abisoye et al. (2018); Sharma et al. (2019)	SVM and ANN for malaria outbreak prediction	Limited to 1–2 classifiers; classification only; hybrid models recommended for better performance
Adamu(2021);Dukuzumuremyi (2020); (2025); Adigun et al. (2024); Modu et al. (2017)	ML models using RF, GB, LR, k-NN, SVM for prevalence/outbreak prediction	Focused on classification; regression and hybrid frameworks rarely integrated; RF often performed best
Tharageswari et al. (2025)	Hybrid / ensemble deep learning (CNN, RNN, LSTM)	Improved accuracy (90–95%) and AUC (>0.90); mostly disease-specific (COVID-19); joint regression-classification often missing; malaria adaptation lacking
Musa (2015)	Regression / time-series models (LSTM, ARIMA, Bayesian, ensemble)	Strong regression performance (R^2 0.42–0.88); outbreak classification not included; hybrid or ensemble optimization absent
Ma (2025)	Bimodal ML framework (regression + classification)	>90% prediction accuracy; dual-task applied; deeper hybrid integration and ensemble optimization needed
Kaur & Bawa (2015);	Reviews & statistical / spatial models	Highlighted need for hybrid, multi-model, and multi-criteria approaches; most studies lacked integrated classification-regression frameworks
Majeed et al. (2023)	LSTM and integrated temporal spatial attention for dengue outbreak	Suggested hybrid methods to improve performance; not applied to malaria

2.9 Thesis Gap Summary

The summary makes it clear that using regression and classification models as separate methods to accurately predict malaria outbreaks is still a difficult task. These models have potential, but they aren't good for predicting malaria outbreaks because they have some problems. Because there are more and more health problems, research needs to keep going, especially because global climate change is making the malaria outbreak problem change all the time. To tackle the challenge of forecasting malaria outbreaks, a proposed hybrid approach necessitates the integration of both classification and regression models. This study employs a hybrid of machine learning algorithms to investigate malaria outbreaks, as emerging technologies necessitate further research to enhance performance in the health sector. Artificial intelligence techniques like machine learning and data mining use training algorithms to learn from datasets. These techniques have led to many improvements in many areas.

Additionally, the summary critique elucidates the prevailing challenges and research deficiencies in malaria outbreak forecasting using classification and regression as independent methodologies, as characterized by the aforementioned literature review, summarized in Table 2.4 Consequently, it can be concluded that: There has been limited systematic research on the comparison and evaluation of various machine learning models, classification schemes, and hybrid models employed in predicting malaria outbreaks; (ii) To the best of our knowledge, the hybrid model regression classification has not yet been applied in malaria outbreak forecasting, particularly in the two phase approach where the output from phase one serves as the input for phase two; (iii) Most studies have focused on classification or regression as separate methods, but fewer have looked at regression or classification as an interdependent strategy for predicting malaria outbreaks. As separate methods, regression and classification are inherently inconsistent and uncertain. This makes it hard to make a quick and accurate prediction model that can count the number of malaria positive cases and guess whether or not an outbreak will happen based on the results of one method or the other. In short, the goal is to maximize the accuracy function by using both strategies together, since independent model

techniques are still problems. This indicates that the integration of regression and classification models employed to forecast malaria outbreaks offers a unique advantage in addressing the prediction challenge through dataset correlation features.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

This chapter provides the conceptual design of the model, followed by development of the novel hybrid model proposed in this research. It further covers the research design, data collection, population and sampling, data processing and analysis tools used. Additionally, the chapter describes the proposed model framework and algorithm adopted for this research to achieve the research objectives. These procedures are vital for systematically directing the research process and enabling other scholars to comprehend and reproduce the study effectively.

3.2 Research Design and Approach

Research designs provide a broad and systematic structure that guides the research process, supporting effective inquiry, informed decision-making, and problem-solving throughout a study. Ferreira et al., (2022) characterized research designs as organized approaches for collecting, analyzing, interpreting, and disseminating data. Rigorous designs are essential as they establish the framework for deriving interpretations and the methodologies and evaluations employed by researchers during the investigation. Research methodologies are typically divided into theoretical or empirical approaches, and into qualitative or quantitative studies. Quantitative research entails the collection of quantifiable data that can be converted into statistical information for objective analysis (Ferreira et al., 2022). Conversely, qualitative research is exploratory, concentrating on the identification of thought patterns, behaviors, or underlying motivations via unstructured or semi structured methodologies.

Theoretical approaches depend on formal models in mathematics and logic, utilizing discrete mathematics including set theory, functions, graphs, algebra, combinatory, and category theory to delineate and validate conceptual properties. On the other hand,

empirical methods use experiments, observations, and data driven conclusions. This study combines both approaches by using experimental methods like simulation and emulation testing to create and test data mining models for predicting malaria outbreaks. Consequently, this research uses a quantitative, experimental, and computational study design. The methodology is designed to amalgamate epidemiological data, climatic factors, and environmental metrics with sophisticated machine learning methodologies to create and authenticate a hybrid classification and regression model for forecasting malaria outbreak. The design is based on the design science paradigm, which focuses on creating and testing computational artifacts to solve problems in the real world.

The study follows a systematic process comprising four interrelated phases. The first step is to get the data, which means getting records of malaria cases, weather data, and environmental data. The second step is preprocessing and feature engineering, which means cleaning and changing the data that was collected and picking the most important features. The third step is to create a hybrid model, which involves coming up with and testing classification and regression model primitives through theory, simulation, and real-world testing. The fourth and last step is model evaluation. In this step, the proposed hybrid model is compared to other models that are similar to see how well it predicts, how reliable it is, and how efficient it is. This research guarantees methodological rigor by integrating theoretical insights with empirical experimentation, thereby addressing practical requirements in malaria outbreak prediction and risk assessment.

3.3 Methods Utilized in the Research

This work employed a quantitative research strategy, seeking to draw conclusions from empirically derived statistical evidence while integrating both theoretical and empirical perspectives. The study utilized a hybrid machine learning method that integrates classification and regression frameworks to enhance the prediction of malaria outbreaks. Classification models were utilized to ascertain the likelihood of an outbreak occurring either as a binary outcome or as multiclass predictions while regression models were employed to estimate the anticipated number of malaria cases within designated temporal

and spatial units. The effectiveness of the suggested hybrid-based classification and regression model was evaluated using both theoretical and empirical simulation techniques. The theoretical analysis elucidated deficiencies in the current understanding of the challenges associated with predicting disease outbreaks, whereas the empirical approach assessed the usefulness of the planned model in benchmarking to existing models through simulation based experimental methodologies.

This design improves predictive reliability by combining classification accuracy with quantitative case estimation. It also provides a comprehensive framework for predicting malaria outbreaks. Chapter Four provides a thorough demonstration and analysis of the comparative evaluation results of the existing models and the proposed model.

3.4 Data Collection and Dataset Sources

To conduct this work, diverse categories of data were collected from the governmental organization. The data consists of malaria cases, ecological information for instance temperature, rainfall, and relative humidity, as well as the history of population growth in the Bugesera and Huye districts. The next paragraphs go into great detail about how these data were gathered.

The data gathering procedure for this research focused on assembling a thorough dataset of malaria outbreak to enable the creation of a hybrid classification and regression model. The dataset had a lot of different kinds of data from meteorology, the environment, and epidemiology. Meteorological data which are all important factors that affect how mosquitoes breed and how malaria spreads. Epidemiological records, encompassing reported malaria cases and outbreak designations, constituted the fundamental ground truth for the training and evaluation of the predictive models.

The dataset includes variables that were used in earlier studies, like Sharma et al. (2019), and comes from publicly available sources, like the Government of India's Integrated Health Information Platform under the Ministry of Health and Family Welfare, which

helps with nationwide efforts to stop vector borne diseases. The study ensures that the predictive modeling framework is grounded in evidence and accurately represents the real-world transmission of malaria by utilizing a variety of comprehensive data sources.

3.5 Target Dataset

The malaria outbreak dataset adopted in this work was sourced from RBC (2025) and Sharma (2015) used the same dataset properties as in another research have been worked well. The dataset used in this work is partly based on earlier research, such as the malaria outbreak dataset made by Dukuzumuremyi (2020) and the IoT-driven predictive dataset made by Niyitegeka (2021). Sharma et al. (2019) also made the malaria outbreak dataset available, which is a special dataset that helps with research and modeling about predicting malaria outbreaks.

Moreover, Sharma (2015) used the same dataset properties, which came from the national vector borne disease control program platform in Pune and Indian meteorological data. The study's dataset indicators that show whether or not there is an outbreak, as explained in the next section (Dukuzumuremyi et al., 2020; Niyitegeka, 2021).

3.5.1 Data Description

The dataset included in this research involves of weekly and monthly time series recordings, documenting fluctuations in malaria prevalence and environmental variables across various temporal scales. The weekly data mostly showed how many cases of malaria there were and how likely it was that an epidemic would happen. The monthly data, on the other hand, showed average maximum temperature, minimum temperature, rainfall, and relative humidity. All of the data was put together and saved as comma-separated values (CSV) files. This made it easier to preprocess and add to Python-based machine learning pipelines that use libraries like Pandas and NumPy. Each observation (row) corresponds to a distinct temporal instance within a specified district, whilst columns (features) signify environmental and epidemiological characteristics.

This structured format makes it easy to change, clean, and extract features for training and evaluating models. The dataset has multivariate properties that may be grouped into three primary groups: climatic, environmental, and epidemiological variables. These variables all have an effect on how malaria spreads. The average maximum and minimum temperatures (^{o}C), rainfall (mm), and relative humidity (%) are all examples of meteorological data that can help us anticipate how many mosquitoes will breed and how quickly parasites will grow.

The environmental variables include spatial features like geographic coordinates, elevation, and land surface indicators, which show how suitable the area is for vector growth. The epidemiological variables include confirmed malaria cases, the count of positive Plasmodium falciparum (PO cases, and binary outbreak indicators sourced from validated health facility records and national malaria surveillance reports. These characteristics combine to provide a multidimensional feature space that can capture both linear and nonlinear relationships that affect malaria risk patterns. It was set up in a structured table format, which made sure that the model training and testing were always the same. The dataset has a substantial class imbalance because the outbreak frequencies are not equal. However, preprocessing methods like normalization was used to fix this and make sure that the model learned in a fair way.

Statistical examination of the data showed that there were strong links between climatic elements, especially rainfall and temperature, and the number of malaria cases. This confirmed that these factors should be included as important predictors in the modeling framework. The incorporation of multi-source data guarantees spatial and temporal generalizability, hence augmenting the robustness and reliability of the hybrid based classification and regression model in predicting malaria outbreaks under various meteorological and epidemiological circumstances.

3.6 Conceptual Framework of the for the Proposed Model

The aim of the research was to design and evaluate a hybrid model for malaria outbreak prediction, as guided by the second and third specific objectives. Due to the literature review findings, the study the researchers emphasized the importance of following steps while mining patient data for predicting outbreak diseases or making decisions using patients' personal and medical information, as outlined by Sharma et al. (2019). This research looked at; the proposed model's steps are shown in Figure 3.1. Formerly constructing a hybrid model approach for prediction purposes, the researcher needs to determine which learning algorithm should be employed to build the model and evaluate its predictive results, particularly for future data. Nevertheless, this phase is regularly abandoned, resulting in an unreliable prediction model Bashir & Shaun (2023).

The primary inspiration of this work is the establishment of an effective suggested model, as already stated in the main objection and to attain this as main objective, these specific aims were mentioned and formulated as (i) to investigate the existing techniques and models for both classification and regression models that can be applied to predict malaria outbreak was achieved as a literature examination was showed in chapter two as to fulfill the specific objective number one, and with two specific objectives in mind are going to be addressed in chapter three and the next chapters as (ii) to design and develop a hybrid model that attends to combine the integration of classification and regression approaches to predict malaria outbreak,(iii) to implement the proposed hybrid model for malaria outbreak prediction ,(iv)to evaluate the developed hybrid model using applicable performance metrics and compare it against other prior models. As revealed earlier, literature review was done to examine existing prediction model frameworks for outbreak diseases and explore how these frameworks could be integrated for malaria outbreak prediction, and establish a benchmark for evaluating the effectiveness of the established framework.

Figure 3.1 is the overall summary of the conceptual framework that shows several model phases, these include the initiation of data preprocessing and cleaning, Phase one for

predicting the number of positive cases using machine learning regression algorithms and Phase two for outbreak prediction employing machine learning classification algorithms. Various performance metrics are applied to each algorithm to find the top model approach planned for predicting future malaria outbreaks. The conceptual framework for the anticipated model is shown in Figure 3.1

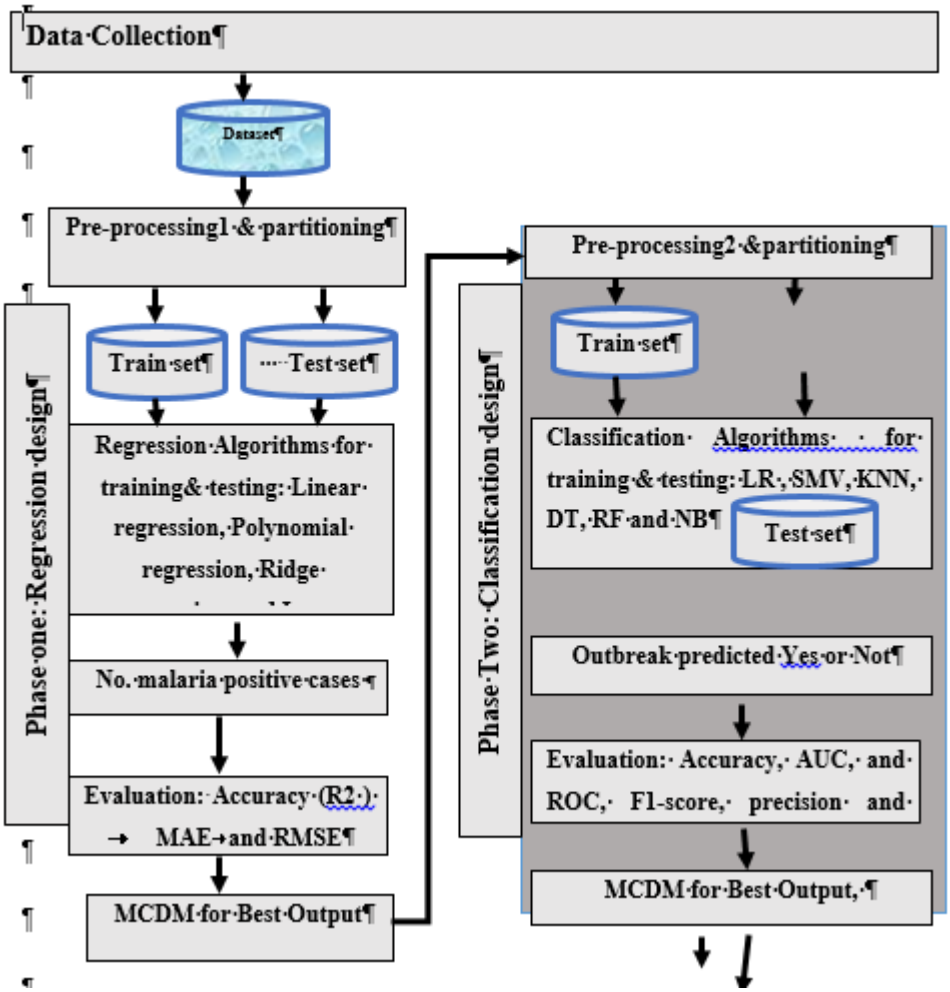
Figure 3.1 and Algorithm I show that the suggested hybrid model for predicting malaria outbreaks has two phases, each of which focuses on a different modeling task. Phase 1 is for regression, which means it predicts how many people will test positive for malaria. This information is then used in Phase 2. Phase 2 is meant to sort things out by using the results from Phase 1 to guess whether a malaria outbreak will happen. Figure 24 shows the general idea and process of the hybrid model. The following steps were taken to design and put into place the proposed hybrid prediction framework:

(i) Preparing and Splitting the Data

The first point is to change the raw data about malaria outbreaks into a format that ML algorithms can apply. This is a normal step before building a model. Data preprocessing changes the dataset so that ML techniques can understand and apply the features (Bashir, 2023). The dataset for this study comprises attributes including weekly average minimum temperature, average maximum temperature, rainfall, and a label denoting occurrences of malaria outbreaks. We put all the features and labels into one CSV file and imported it into Python 3.7 using Jupyter Notebook to clean up the data. We used data wrangling methods to deal with missing values, outliers, and noise, which can make models less accurate. Also, data augmentation techniques is very important, which helped the model generalize better and fixed problems that came up because there weren't enough training examples (Jones-Farmer et al., 2017).

(ii) Choosing ML Methods

The next step after preprocessing is to choose the right algorithms for the classification and regression tasks. To improve the accuracy of predictions, the hybrid model uses more than one algorithm. There are many forms of regression techniques, such as linear regression, ridge regression, lasso regression, and polynomial regression. There are also many types of classification algorithms, such as logistic regression, support vector machine, k nearest neighbors, decision tree, random forest, and naive bayes. We tested the performance of each algorithm and only included the models that gave the best outcome in the hybrid framework, as shown in Figure 3.1



(iii) Model Development

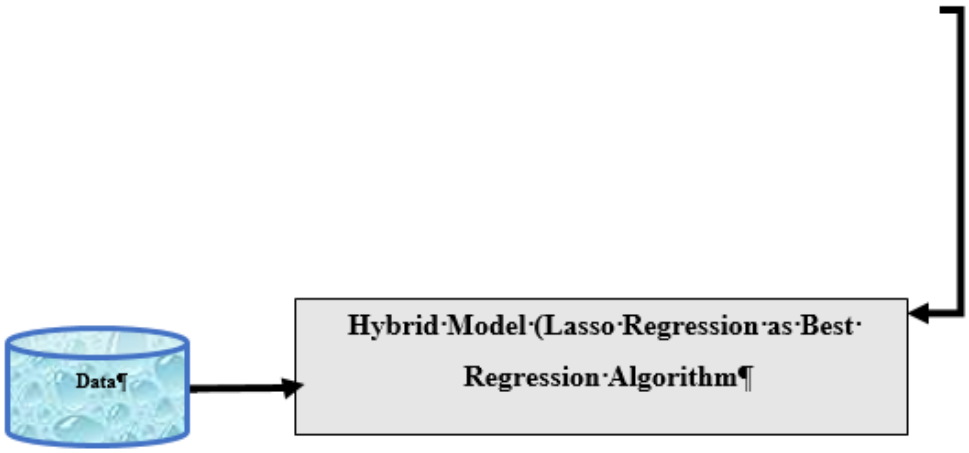


Figure 3.1: Conceptual Framework for the Proposed Classification and Regression Hybrid

During the model creation step, the hybrid framework is trained and tested to make sure it can accurately predict malaria outbreaks. In Phase l, the regression algorithms are used one after the other to guess how many people will have malaria. The model that works best is then used as the input for Phase 2. In Phase 2, the classification algorithms are used one after the other to predict when an outbreak will happen. The model with the uppermost accuracy is then preferred as the final classifier. This two-phase method makes sure that both quantitative case prediction and binary/multiclass outbreak prediction are covered in one framework.

(iv) Evaluation of the Model

The k fold cross validation technique was used to test the hybrid model's performance. This method lets you check the model's accuracy, validity, and consistency on test datasets that are not part of the training set (Siłka et al., 2023). This process found the best combination of regression and classification algorithms, making sure that the hybrid model met the research goals for predictive performance.

(v) Representing Knowledge

The classification regression hybrid model, which is the result of successfully combining both phases, is a strong way to predict malaria outbreaks. Phase 1 uses regression technique like linear regression, polynomial regression, ridge regression, and lasso regression, random forest, and SVR. Phase 2 uses classification algorithms like LR, SMV, KNN, DT, RF and NB. The planned hybrid model was benchmarked to other models to show how accurate it is at making predictions and how useful it is in real life, so that users can get useful information about how to stop and treat malaria outbreaks. Chapter four explains the experimental setting, the platform, and the full outcome of the model evaluation.

3.7 Proposed New Hybrid Model Algorithm

The proposed hybrid model development involves the following steps as indicated in the algorithm 1

Algorithm 1: Hybrid Model (Regression+Classification)

Initialization

1. Dataset, regressions [Linear regression, polynomial, Ridge, Lasso, RF, SVR], classifications [Logistic, SVM, DT, RF, KNN, NB]
2. [$Dataset_1$, $Dataset_2$]←Separate(Dataset)

Phase I: Positive cases prediction

3. Dataset, regressions [Linear regression, polynomial, Ridge, Lasso, RF, SVR]
4. [D_{tr1} , D_{ts1}]←Preprocessing ($Dataset_1$)
5. bestRegression←regressions[0]
6. for each regressModel{
7. temp_regress←regressions.evaluate(regressModel, D_{tr1})
8. if(temp_regress.IsBetter(bestRegression))using equation 14{
9. bestRegression← temp_regress
   }}
10. predict_vector←bestRegression.predict(D_{tr1} , D_{ts1})

Phase II: Outbreak prediction

11. Dataset, classifications [Logistic, SVM, DT, RF, KNN, NB]
12. [D_{tr2} , D_{ts2}]←Preprocessing (predict_vector, $Dataset_2$)
13. best_classifier←classifications [0]
14. temp_class←classifications.evaluate(classificationModel, D_{tr2})
15. if (temp_class. IsBetter(best_classifier)) using equation 14{
16. best_classifier ←temp_class
    }}
17. Outbreak←best_classifier.predict(D_{tr2},D_{ts2})
18. Return Hybrid (best_classifier(bestRegression (Dataset))

Figure 3.2: The algorithm of Hybrid model (Classification Regression)

3.7.1 Algorithm Normalization of the Proposed Model

The normalization of the planned model development algorithm in section 3.7 is explained as follows: To normalize any value y of a given dataset we consider this equation

$$y = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3.1}$$

Where *x* is equal to the numerical value to be minimized. In scikit learn use of the function MinMaxScaler. The good practice of this function and other scaling methods is as seen below:

Fit the scalar employing accessible training data: In relations of normalization, this shows that the lowest and maximum noticeable values will be quantified utilizing the training data. Calling the fit () function accomplishes this.

Apply the scale to training data: This infers that you can train your model parts utilizing the normalized data. The transform () method is used to accomplish this. Apply the scale to data going forward: Therefore, if you wish to do prediction in the future, you can organize new data.

3.7.2 Algorithm Standardization Equation of the Proposed Modeling

The standardization of the planned model development algorithm in 3.7 section is explained as follows:

Centering and scaling the predictor variables is the easiest and most frequent data modification. The average predictor score is deducted to center a predictor feature from all the values. The predictor has a zero mean as an outcome of centering. Similar to this, the standard deviation of each value of the predictor variable is divided to scale the data. The values are forced to have a ordinary standard deviation of one by scaling the data (Kuhn & Johnson, 2020).

$$y = \frac{x - \text{mean}}{\text{std}} \tag{3.2}$$

Where mean is calculated as follow:

$$mean = \frac{\sum x}{\text{count}(x)} \quad (3.3$$

And stv stand for standard deviation calculated as follows:

$$stdv = \sqrt{\sum_{j=0}^{m} \frac{(x - mean)^2}{\text{count}(x)}} \quad (3.4)$$

Where count(x) is the total number of values available in the considered dataset.

Weighted sum Maximum performance score

$$PS_{max} = \left(w_i * \sum_{j=0}^{n} \left(\frac{Max_j - v_j}{Max_j - Min_j} \right) \right) \quad (3.5)$$

Where w is the weight given to each criterion where measures are the outcome performance metrics considered, *i* depict each criterion, max and min represents the maximum and minimum values for a given criterion, *j* represents each value in the list of values of each criterion and *v* represent the current value for a given criterion.

3.7.3 Algorithm Description for Developing the Proposed Hybrid Model

The planned model development algorithm, defined in section 3.7, has three stages: Initialization, Phase I and Phase Il. In the Initialization stage, the algorithm initializes all the required inputs, including the dataset and various regression and classification algorithms. The Phase I stage deals with regression processes, while the Phase Il stage deals with classification processes, with the output of Phase I serving as the input of Phase Il.

In the Initialization stage, the algorithm initializes all the input required by the algorithms. The second line of the algorithm deals with data preprocessing, where the dataset is divided into two datasets, one for regression processes and another for classification

processes. The outbreak feature and its data are separated from the other features, as it is found to be the most correlated to the outbreak feature in the current dataset. This results in the creation of dataset 1 and dataset 2.

After initialization, the algorithm predicts positive cases using all available features in the dataset. Line three of the algorithms performs data preprocessing, dividing the dataset into training and test sets and normalizing the dataset using equation l. Lines 4 to 8 find the best regression model among those initialized in the first stage. The function is Better () on line 7 is used to calculate the maximum performance score using MAE, MSE and RMSE as performance criteria of the considered regression models. The model with the maximum performance score is returned as the best on line 8 and the prediction vector is created on line 9 using the best regression model and training and test set.

The Phase Il stage involves the prediction of the outbreak. Line 10 of this stage performs data preprocessing, such as recombining the output of Phase I with the outbreak feature and standardization using equation 2. The data is also normalized and separated into training and test sets. From line 11 to line 14, the algorithm selects the best classification model. The function is Better () on line 13 calculates the maximum performance score using accuracy, AUC, ROC, F1_score, precision and recall as performance metrics of the considered classification models. The model with the maximum performance score is returned as the best on line 14and the classification vector is created using the best classification model and training and test set on line 15. The algorithm returns the number of cases and the decision about an outbreak or not in the conclusion.

3.8 Model Evaluation

A model is an abstraction tool in research that represents a provisional image of the real object of study. In ML, a model is built using a supervised learning algorithm, which learns patterns from experience and uses them to predict future events (Huang et al., 2020). The last specific purpose of this research was to assess the efficiency and accuracy of the generated hybrid model and compare it against other previous models employed to predict

malaria outbreaks. In this perspective, testing or evaluating the model's capability is an essential part of building the model and determining which model fits the data best for future prediction, and model evaluation is performed on a separate test set, rather than the training set, to avoid the problem of overfitting (khan et al., 2025). In accordance with the type of algorithm employed (classification or regression), multiple performance metrics are employed for assessing machine learning models.

It was recommended to use multiple evaluation criteria for a single model since one metric may show good performance while another may show poor performance. Therefore, using multiple metrics helps to determine whether the model performs correctly and optimally (Shahnazari & Ayyoubzadeh, 2024). In this research, each ML algorithm considered the model is trained employing the training set and then tested using the test set. AUC, ROC, F1 score, error rate, precision, and recall are all utilized to evaluate the model's performance. In the opinion of Mohsen and Alhurdi, (2025), these assessment measures are critical and should be employed to evaluate the performance of classifiers.

3.8.1 Confusion Matrix

The superiority of the results of a classifier was evaluated in this study using a confusion matrix. The number of points is represented by the diagonal elements where the predicted label is the same as the real label, and the off diagonal elements reflect the number of labels incorrectly identified by the classifier. The more the diagonal values in the confusion matrix, the better the model's performance, as it shows that many predictions are right (Mohsen & Alhurdi, 2025). To demonstrate the model's accuracy, we used the Confusion Matrix and ROC, as explained by Mohsen and Alhurdi, (2025) where TN represents True Negative, TP represents True Positive, FP represents False Positive and FN represents False Negative. The confusion matrix is a standard metric that provides information about the model's accuracy, as well as other associated attributes such as sensitivity, specificity, precision and F measure (Hemachandran et al., 2023). Table 3.1 shows these evaluations

Table 3.1: Confusion Matrix

		Predicted class	
		Class 1 Yes	**Class 2 No**
Actual class	Class 1 =Yes	True Positive	False Negative
	Class 2=No	False Positive	True Negative

A confusion matrix was employed for assessing a classifier's accuracy, which is a measure of its performance. The model's performance has been outlined using the evaluation procedures highlighted as the **True Positives:** the number of instances correctly predicted by the classifier as outbreak cases; **True Negatives:** the number of cases accurately predicted as non-outbreak cases by the classifier **False Positives,** additionally referred to as type one errors are the number of instances predicted by the classifier as epidemic cases while no outbreak occurred in fact.

False Negatives, also known as type two errors, are the number of instances predicted by the classifier as non-outbreak cases when an outbreak actually occurred. These metrics can be used to compare different machine learning models based on their performance. Additionally, evaluation metrics such as accuracy, AUC, ROC, F1 score, error rate, precision and recall can be found from the confusion matrix to further assess the model's performance. **Common performance metrics in medical problems include accuracy, sensitivity, and specificity, with accuracy indicating how correctly implant success or failure is predicted**

(i) Accuracy: common performance metrics in medical problems include accuracy, sensitivity, and specificity, with accuracy indicating how correctly implant success or failure is predicted as indicated in formula 3.6 (A. S. Ahmed & Salah, 2023).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}} \tag{3.6}$$

(ii) Precision: Precision is a value representing the accuracy presented by a single class predicted by Ahmed & Salah, (2023).

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{3.7}$$

(iii) Recall: The recall of a prediction model is an indicator of its ability to determine examples of particular groups from a data set. It is additionally referred to as sensitivity and refers to the genuine positive rate.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{3.8}$$

(vi) Error rate: The error rate for a classifier can be computed as follows, as indicated in equation 4.5 by Ahmed & Salah, (2023)

$$\text{Error rate} = \frac{\text{FP+FN}}{\text{TP+TN+FP+FN}} \tag{3.9}$$

vii) F1 Score: It is a metric that tells us the harmonic mean between our recall and precision.

It is given as:

$$\text{F} - \text{score} = 2 \,.\, \frac{\text{Recall .Precision}}{\text{Recall+Precision}} \tag{3.10}$$

(viii) Receiver Operating Characteristics Curve and Area under Curve

ROC is explained as the plot of the evaluation of sensitivity or true positive, which is plotted on the Y axis versus the 1 specificity or the false positive on the X axis. It has been efficient to test the performance or the quality of diagnostic tests, and mostly it is used in radiology tests. ROC has a good performance through the decrease of standard error, and as the number of test samples and Area under Curve increases, as well as increases sensitivity when the analysis of variance test is performed (Asingizwe et al., 2019). The area under the curve is well-defined as the area under the ROC curve, and it dealings how good the prediction is. It is also the gauge for the quality of separation (Hussain et al., 2023).

(ix) Model Time: This is the overall CPU time that is essential to create a classification model, along with the training time needed to predict the output of the test data.

(x) Mean Absolute Error (MAE): MAE is a valuable measurement to usage for performance assessment of an algorithm, and it is computed by taking the average of all absolute errors (Sheth et al., 2022).

$$MAE = \frac{1}{1}\sum_{i=1}^{n}|x_i - x| \tag{3.11}$$

Where n is the amount of errors, ∑ = summary symbol (which means "total them all")

$|xi - x|$= the absolute errors

(xi) Root Mean Squared Error: is a common measurement for performance assessment measurement which computes the error without canceling the positive and negative error (Hodson, 2022).

$$RMSE = \sqrt{\frac{1}{n}\left[\sum_{1}^{N}\left(Q_{exp} - Q_{cal}\right)^2\right]} \tag{3.11}$$

Where *n* the sample size

$(Qexp - Qcal)^2$ = the difference squared

During regression and classification evaluation metric, we then did an 80: 20 subsets on training set and testing set and performed 10-fold cross validation of the accuracy and then documented the all-performance accuracies. This was done to the regression, classification and hybrid model for confirming the accuracy capability of the new model.

3.9 The Environment for Implementing the Model

For the implementation of the anticipated model into action, a full set of advanced software platforms, programming tools, and computing resources was used to make sure that development was effective, experimentation was reliable, and results were accurate. We used Python version 3.8 to build the model because it has a lot of ML and data science libraries, include Scikit learn, Tensor Flow, Pandas, NumPy, and SciPy. These libraries helped with important tasks like building models, processing data, doing math, and conducting scientific evaluations.

Google Colab that offers an informal to use but powerful platform for doing complicated math. The development and testing were done there. Colab gave us access to GPU acceleration, which made processing much faster. This was especially helpful during the training and evaluation phases of the planned hybrid model. It also made it easier to work on code together and make quick prototypes.

Matplotlib and Seaborn were used to make graphical displays of data patterns, model effectiveness, and trends to help with data preprocessing and result visualization. In some cases, local IDEs like Jupyter Notebook and PyCharm made development easier by giving programmers better control over debugging and modular programming. Combining these tools and platforms made a stable, flexible, and scalable environment for building, improving, and testing the hybrid classification-regression model that was made to accurately and clearly predict malaria outbreaks.

3.10 Summary

This chapter presented the practice utilized to attain the research aims. The research design discussed the various datasets and pre-trained models, model development through an algorithm, and the software tools were also presented. The approaches of the study are also discussed in detail, outlining the proposed model performance analysis and its evaluation metrics used were also presented. The research methods were defined in this

chapter. Finally, this chapter presents the proposed new hybrid model's performance results.

CHAPTER FOUR

RESEARCH RESULTS AND DISCUSSIONS

4.1 Introduction

This chapter outlines the experimental methodology and the findings in relation to the study aims. A screenshot is included to show the implementation design, which includes the models used, the training data, the correlation matrix, the models' performance evaluation, and so on. The findings and discourse of this study are delineated in accordance with the aims of formulating and evaluating the proposed hybrid model. Researchers also look at how our study's results compare to those of other models to make sure that our established hybrid model scenario works. As stated in chapter three, Anaconda version 3, which supports the Python 3.6, was used to put the proposed hybrid model into action. Algorithm 1 and Figure 3.1 show how the hybrid model's design flow works.

4.2 Heatmap of the Feature Correlation Matrix for the Malaria Outbreak Dataset

To examine the relationships among predictor variables and malaria incidence indicators, a correlation examination was done to assess the degree of linear relationship within the dataset. Correlation coefficients were calculated to determine how climatic, environmental, and epidemiological variables relate to malaria incidence and to one another. This analysis provides an initial statistical understanding of the dataset and helps reveal patterns that may influence malaria transmission dynamics, including both positive and negative relationships between variables. Figure 4.1 presents the heatmap of the correlation matrix, where color gradients indicate the magnitude and direction of associations among features. Strong positive and negative correlations are visually distinguishable, while values near zero indicate weak relationships. This visualization enables quick identification of highly correlated predictors, variable clusters, and potential dependencies within the dataset, thereby enhancing interpretability prior to model

development. Furthermore, the correlation analysis supports feature selection and dimensionality reduction by detecting multicollinearity and redundant variables. Addressing highly correlated predictors improves model stability, robustness, and interpretability. By retaining only the most relevant and non-redundant features, the proposed hybrid classification–regression model achieves improved predictive performance and generalizability in malaria outbreak prediction.

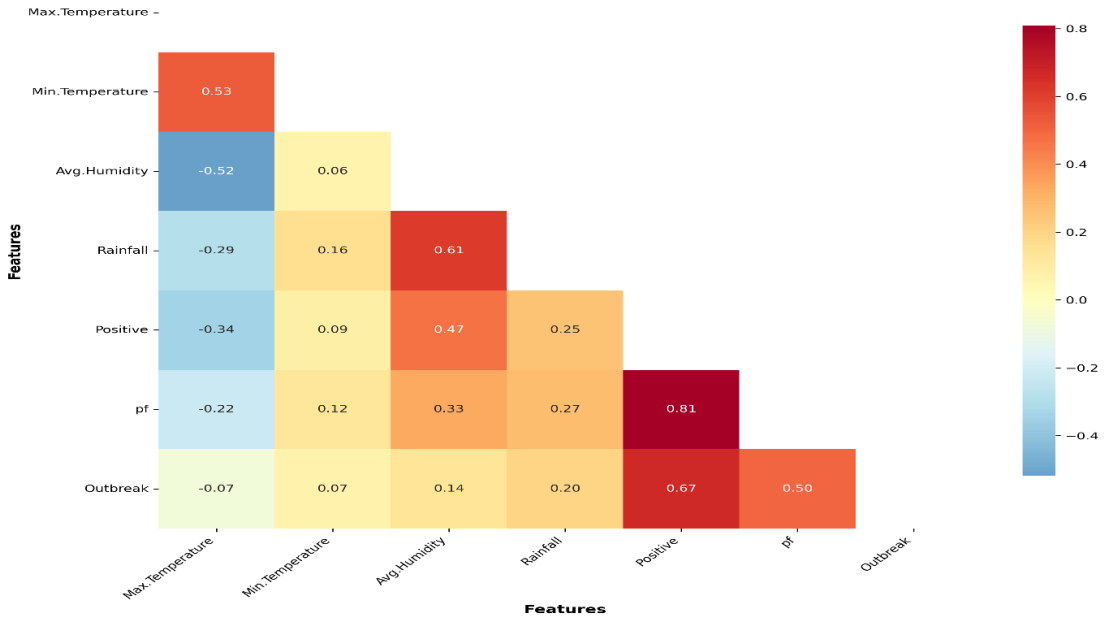


Figure 4.1: Heatmap of Data Correlation Matrix

Figure 4.1 depicts a heatmap of the feature correlation matrix that shows how the main environmental and epidemiological variables used to make the hybrid based classification and regression model for malaria outbreak prediction are related to each other. The correlation coefficients, which range from 0.6 (strong negative) to +0.8 (strong positive), show how each feature affects the spread of malaria. A reasonable positive correlation (r = 0.53) relative to maximum and minimum temperature indicates climatic consistency in tropical regions, while the negative correlation between maximum temperature and average humidity (r = -0.52) suggests that higher temperatures reduce relative humidity, potentially influencing mosquito breeding patterns. Rainfall has a solid positive

relationship with humidity (r 0.61), which means that when it rains, the air gets more humid, which is good for mosquitoes to breed. The strongest correlation (r = 0.81) between malaria positive cases and Plasmodium falciparum (PO counts confirms pf as the main parasite species that causes malaria to spread. Furthermore, the positive correlation between confirmed malaria cases and outbreak occurrence (r = 0.67) emphasizes the importance of epidemiological variables as direct predictors of outbreak events, whereas the comparatively weaker correlations with environmental factors such as rainfall (r = 0.20) and humidity (r = 0.14) suggest indirect influences on transmission. These results underscore the intricate relationship between climatic and epidemiological factors, necessitating the implementation of a hybrid model that mix classification for outbreak recognition and regression for quantitative case estimation. The hybrid approach is suitable for capturing both linear and nonlinear dependencies in the correlation structure, thereby enlightening the model's predictive accuracy and robustness in forecasting malaria outbreaks.

4.3 Experimental Study Results

The second specific aims of this research is to build a strong model that seamlessly combines classification and regression techniques to accurately predict malaria outbreaks within the dataset. Python was used in this study to make the dataset for analysis. The dataset was then separated into two separate datasets: The Training set (80%) and the Testing set (20%). The recommended split percentage from Kevin K. Dobbin and Richard M. Simon's 2011 study was used. During the first phase, four machine learning models were used for regression: Linear regression, Polynomial regression, Ridge regression, and Lasso regression, Random Forest, and SVR. In the second phase, six machine learning models were used including LR, SMV, KNN, DT, RF and NB for classification objectives. These algorithms were selected due to their infrequent application in the analysis of malaria outbreak data, despite exhibiting superior efficiency and accelerated training speeds in alternative machine learning fields. Regression was utilized to find the number of malaria positive cases for each machine learning model, and the algorithm that worked best was chosen. The Yes or No prediction models were then employed to predict malaria

outbreaks depended on these malaria positive cases. The best performing algorithm was also considered. We used performance measures to test the models, and the results were put into tables and graphs. After using the different models in each phase to come up with the suggested hybrid model based on the model performance accuracy shown in the tables and figures in the next sections, the best model in each phase was calculated using equation 14 as explained in sections 3.7.2 and 3.7.1.

4.3.1 Experiment Study 1: Regression Model Development and Performance Evaluation (Phase One) for the Number of Malaria Cases Prediction

In this phase, four ML algorithms ridge regression, linear regression, lasso regression, and polynomial regression were employed. The most accurate algorithms were selected for use in phase two. Subsequently, the indicators identified in the earlier stage were utilized as inputs to compare the performance of three classifiers against the baseline of raw features. Table 4.1 shows how well the different regression classifiers worked as another way to guess how many malaria cases there would be. The table shows the accuracy, RMSE, and MAE values needed to build the model for each classifier. Table 4.1 displays the outcome of a regression examination that compares the performance of six algorithms Linear Regression, Ridge Regression, Lasso Regression, Polynomial Regression, Random Forest Regression, and Support Vector Regression in predicting the number of malaria cases. We used both training and evaluation datasets to test each model, and we used standard regression metrics like the coefficient of determination, mean squared error, mean absolute error, root mean squared error, and cross-validation scores to see how well they worked. These metrics revealed the forecasting strength, accuracy, and generalization ability of each model.

The Linear, Ridge, and Lasso regression models all did poorly at predicting, with Train and Test R^2 values of about 26.83% each. This means that the models could only explain about 25% of the changes in malaria case counts, which is not enough for accurate prediction. Their high MSE values (7.8 to 8.0 million) and high RMSE scores (around 2800) also show that the predicted and actual values are very different. The three models

performed similarly, which suggests that adding regularization to Ridge and Lasso Regression didn't make a big difference. These results show that the relationship between malaria incidence and the independent indicators is not linear; which means that linear-based models don't work for this dataset.

The Polynomial regression model, whereas, showed a big improvement over the linear models. With Train and Test R^2 values of 54.25% and 56.81%, respectively, it was able to explain more than half of the differences in the malaria case data. The model also had much lower error values. For training, the RMSE went down to about 2,217, and for testing, it went down to 2,181. The MAE went down to 1,753 and 1,712, respectively. The cross-validation mean R^2 of 53.47% and the small standard deviation (4.05) also show that the performance is pretty consistent across different data partitions. These results support the idea that adding nonlinear polynomial features makes the model better at finding complex relationships in the data. But the model still doesn't explain about 45% of the variance, which means that even higher-order polynomial transformations can't fully explain the complicated dynamics that cause malaria outbreaks.

The random forest regression algorithm performed the greatest and most consistently of all the models. It got a Train R^2 of 99.38%, which is almost perfect, and a Test R^2 of 96.92%, which is high. This shows that there is a strong and reliable link between the input variables and the number of malaria cases. The random forest model also had the lowest error values, with RMSE of 258.33 (train) and 582.16 (test). This means that its predictions were very close to the actual values that were seen. The small difference between the Train and Test R^2 values (about 2.46%) and the low cross-validation standard deviation (1.21) show that the model generalizes well with very little overfitting. The reason for this better performance is that Random Forest uses ensemble learning, which merges several decision trees to lower both bias and variance. It does a good job of capturing nonlinear interactions, variable interdependencies, and changes in the environment. This makes it perfect for complex real-world data like predicting malaria outbreaks. Because of this, Random Forest is the best and utmost accurate algorithm for

malaria cases prediction. It can also be a useful tool for early warning and prevention systems.

On the other hand, the support vector regression model did not do well at all, with negative R^2 values of 6.97 for training and 6.55 for testing. This shows that its predictions were worse than those of a simple average-based model. The MSE and RMSE values were very high, over I l million and 3,400, respectively. This shows that the model did not learn any valuable patterns from the data. This poor performance may have been caused by choosing the wrong kernel, not tuning the hyper parameters enough, or not scaling the input features, all of which are important for how well SVR works. Consequently, SVR is inappropriate for forecasting malaria cases in this study context.

A side-by-side look at the six algorithms makes it clear that Random Forest Regression is the best model. It has the highest R^2 (96.92%), the fewest prediction errors, and the most stable cross-validation results. Polynomial Regression worked fairly well, but Linear, Ridge, and Lasso Regression all gave weak and almost identical results. The SVR model did not work at all to model how malaria cases change over time. Random Forest works so well because it can combine many decision trees to model complicated nonlinear relationships and interactions between, ecological, and climatic factors that mark the spread of malaria.

Ultimately, the Random Forest Regression model showed great accuracy and strength when it came to predicting the number of malaria cases. It showed great generalization and very little overfitting, with a Train R^2 of 99.38% and a Test R^2 of 96.92%. These results show that hybrid ensemble-based methods like Random Forest are very good at predicting malaria outbreaks. Random Forest is a dependable basis for creating the hybrid-based classification and regression model suggested in this study because it effectively combines different predictor variables and handles nonlinear dependencies in epidemiological data. Consequently, it was preferred as the top regression component for the hybrid model because of its superior predictive accuracy, stability, and overall dependability in forecasting malaria cases.

The Phase One regression model testing summary results for the number of malaria cases prediction show how well six algorithms, Linear regression, Polynomial regression, Ridge regression, and Lasso regression, Random Forest, and SVR, performed at the number of malaria cases prediction. The evaluation metrics consist of the coefficient of determination (Test R^2), Mean Squared Error, Mean Absolute Error, Root Mean Squared Error, and the standard deviation of cross-validation R^2 (CV_R^2_Std). These metrics work together to measure how well each algorithm can predict and generalize. The regression model development and performance evaluation (Phase One) is well described in Table 4.1.

Table 4.1: Regression Model Development Training and Testing Performance Evaluation Summary Results for the Number of Malaria Cases Prediction Using Different Methods

Phase One: Regression					
Algorithm	Train_R2	Train_MSE	Train_MAE	Train_RMSE	CV_R2_Mean
Linear Regression	26.832	7862494.427	2284.599	2804.014	25.990
Ridge Regression	26.832	7862495.871	2284.642	2804.014	25.991
Lasso Regression	26.832	7862504.114	2284.718	2804.016	25.991
Polynomial Regression	54.252	4916014.109	1753.143	2217.209	53.465
Random Forest	99.379	66735.643	139.306	258.332	94.919
SVR	6.968	11494550.420	2455.128	3390.361	7.677
Test					
Algorithm	Test_R2	Test_MSE	Test_MAE	Test_RMSE	CV_R2_Std
Linear Regression	26.831	8059817.946	2303.791	2838.982	4.565
Ridge Regression	26.834	8059545.422	2303.763	2838.934	4.560
Lasso Regression	26.837	8059161.132	2303.738	2838.866	4.555
Polynomial Regression	56.811	4757391.303	1712.217	2181.144	4.053
Random Forest	96.923	338914.839	354.684	582.164	1.210
SVR	6.546	11736415.560	2463.005	3425.845	1.763

The linear, ridge, and lasso regression models yielded nearly indistinguishable outcomes, each achieving a Test R^2 of approximately 26.83%. This means that these models could only account for about 25% of the differences in the malaria case data. Their high Test MSE values (about 8.06 million) and large RMSE values (about 2839) show that there are considerable variation between the predicted and measured values. These findings indicate that basic linear models are inadequate for elucidating the intricate and nonlinear relationships that affect malaria transmission dynamics. The small improvement seen in

Ridge and Lasso regressions shows that regularization didn't do significantly improve the model's accuracy.

The Polynomial Regression model, on the other hand, showed a significant improvement, with a Test R^2 of 56.81%. This means that the model could clarify further than half of the differences in the malaria case data. This suggests that nonlinear transformations make the model much better at finding complex relationships between predictors. The model also had a lower MSE (4.76 million) and RMSE (2,181), with a moderate cross-validation deviation (CV_R^2_Std 4.05), which means it did pretty well most of the time. Still, the model left a lot of unexplained variances, which means that more nonlinear modeling is needed to make better predictions about outbreaks.

The Random Forest Regression model outperformed all supplementary algorithms, with a high-Test R^2 of 96.92%, a low MSE of 338,914.84, and a low RMSE of 582.16. These results show that the Random Forest predictions are very close to the real malaria cases number, which shows that the model is very accurate and reliable. The very low cross validation standard deviation (1.21) shows that the model is very stable and can be used for a wide range of tasks. Random Forest works better than other models because it uses an ensemble learning structure that merges numerous decision trees to capture nonlinearities, feature interactions, and complex dependencies between environmental and climate predictors.

On the other hand, the Support Vector Regression model did not do well, with a negative Test R^2 of 6.55 and the highest MSE (11.7 million), which shows that it did not model the underlying data patterns well. This study found that SVR was not reliable for predicting malaria cases because it didn't work well due to poor parameter tuning or inappropriate kernel functions. The comparative evaluation shows that the Random Forest Regression algorithm showed demonstrated superior performance than all the other models at making accurate, stable, and reliable predictions. Random Forest was found to be the best regression algorithm for predicting the number of malaria cases because it had a near-perfect fit and the lowest error metrics. Its better performance shows that it is a good fit

for the regression part of the proposed hybrid based classification and regression model for malaria Outbreak prediction. This would help malaria control and prevention programs make better predictions and decisions faster.

4.3.2 Experimental Study 2: Classification Model Development and Performance Evaluation (Phase Two)

This part depicts and explains how six classification algorithms, LR, SMV, KNN, DT, RF and NB, work to predict when malaria outbreaks will happen based on environmental, climatic, and epidemiological factors. The goal was to find the model that best sorts malaria outbreaks so that they can be predicted and used to help make decisions. All of the models were tested throughout the training step using Accuracy, Precision, Recall, F1 Score, and cross validation accuracy mean. the decision tree and random forest classifiers got perfect scores on all measurement criteria, with 100% in Accuracy, Precision, Recall, and F1 Score. This means that both models learned all of the training patterns correctly and didn't make any mistakes. The KNN model had an impressive accuracy of 97.48%, while the SVM and Logistic Regression models had moderate training accuracies of 88.67% and 87.48%, respectively. Naive Bayes had the worst results, with an accuracy of 81.95%, which shows that it has trouble capturing relationships between complex features. But the fact that decision tree and random forest did so well in training also means that they might be too good at fitting the data, which shows how important it is to test them on new data. To assess the extent to which the model generalizes

In the testing stage, the test dataset was used to check the extent to which the model generalizes. The outcomes showed that the Random Forest algorithm worked the greatest overall, with a Test Accuracy of 96.78%, a Precision of 96.92%, a Recall of 96.78%, an Fl Score of 96.82%, and an amazing AUC (Area under the ROC Curve) of 99.52%. These numbers show that the Random Forest model was able to accurately and consistently categorize malaria outbreaks across cross-validation folds. The KNN model was very close behind with 96.44% accuracy and 99.19% AUC. The Decision Tree model was a little less accurate at 95.89%, but it was still competitive. On the other hand, SVM and

Logistic Regression did only okay, with scores below 90%. Naive Bayes did the worst in both phases, probably because it assumed that features in a dataset with correlated climate and environmental variables were independent. A comparison of all the classification algorithms shows that, the Random Forest model remained the best at predicting malaria outbreaks. Its high accuracy, precision, recall, and AUC show that it is better at modeling complex, nonlinear interactions between many predictors. The model's low CV Accuracy Standard Deviation (0.59) shows that it is very stable and consistent across validation folds. In practice, this marks that the random forest technique was able to correctly predict about 96.78% of malaria outbreaks. This is better than other methods and shows that it is reliable for use in the practical settings. Overall, the output indicates that ensemble based methods, especially Random Forest, are much better at making predictions and generalizing than linear and probabilistic models. The hybrid-based classification approach, therefore, uses the power of Random Forest to accurately model complex relationships between climate and health factors. This result shows that the suggested model is both reliable and appropriate for use in malaria early warning systems. This will improve proactive public health responses and strategic planning for malaria control and prevention. The classification model development (Phase Two) is well described in Table 4.2.

Table 4.2: Classification Model Development (Phase Two) Training and Testing Performance Evaluation Summary Results for the Malaria Outbreak Prediction Using Different Method

Phase Two: Classification						
			Train			
Algorithm	**Train_Accuracy**	**Train_Precision**	**Train_Recall**	**Train_F1**	**Train_AUC**	**CV_Accuracy_Mean**
Logistic Regression	87.48	81.03	87.48	84.06	0.00	87.48
SVM	88.67	81.89	88.67	85.13	0.00	88.62
K NN	97.48	97.49	97.48	97.48	0.00	96.24
Decision Tree	100.00	100.00	100.00	100.00	0.00	95.81
Random Forest	100.00	100.00	100.00	100.00	0.00	96.62
Naive Bayes	81.95	77.36	81.95	78.96	0.00	81.86
			Test			
Algorithm	Test_Accuracy	Test_Precision	Test_Recall	Test_F1	Test_AUC	CV_Accuracy_Std
Algorithm	Test Accuracy (%)	Test Precision (%)	Test Recall (%)	Test F1 (%)	Test AUC (%)	CV Accuracy Std
Logistic Regression	86.67	80.30	86.67	83.26	92.61	0.78
SVM	88.89	82.01	88.89	85.30	96.82	0.83
K NN	96.44	96.60	96.44	96.50	99.19	0.44
Decision Tree	95.89	96.03	95.89	95.94	95.49	0.67
Random Forest	96.78	96.92	96.78	96.82	99.52	0.59
Naive Bayes	80.56	76.27	80.56	77.61	90.49	1.12

4.3.2.1 ROC Curve (Receiver Operating Characteristic Curve) and AUC (Area Under the Curve)

Figure 4.2 shows the ROC curves and the corresponding AUC values for six machine learning classifiers that were used to predict a malaria outbreak: LR, SMV, KNN, DT, RF and NB. The ROC curve is a fundamental performance evaluation metric that illustrates the tradeoff among the True Positive Rate and the False Positive Rate at numerous threshold situations. The ROC curve gets better at significant the variance between classes as it gets closer to the greater left corner. The AUC gives you one number that tells you how well a classifier can tell the difference between things. An AUC value close to 1 means the classifier is doing a better job, while an AUC value of 0.5 means it's just making a random guess. The figure 4.2 shows that all of the models had high AUC values, which means they were very good at predicting the different types of malaria outbreaks. The

KNN classifier had the best macro average AUC of 0.992, which shows that it worked very well and was almost perfect at separating the classes. This means that the KNN model does a good job of capturing the complicated, nonlinear relationships in the malaria dataset. The Random Forest and Decision Tree classifiers followed closely, with macro-AUC values of 0.995 and 0.955, respectively. These results propose that tree-based ensemble methods are also very good at modeling the patterns of malaria outbreaks because they can work with data that is both high dimensional and diverse.

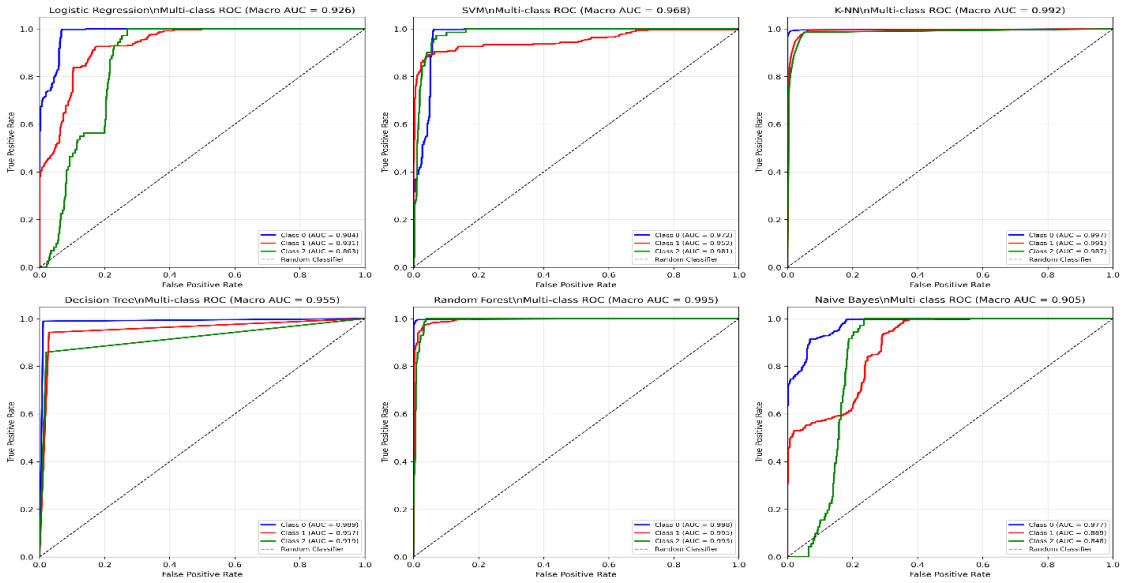


Figure 4.2: ROC Curve (Receiver Operating Characteristic Curve) and AUC (Area under the Curve)

The Support Vector Machine did very well, with a macro-AUC of 0.968. This shows that it is very good at handling nonlinear decision boundaries through kernel optimization. The Logistic Regression model had an AUC of 0.926, which means it worked well but not as well as the nonlinear models. This is probably because it assumed a linear decision boundary. The Naive Bayes classifier had the lowest macro-AUC of 0.905, which means it could make reasonable predictions but wasn't very flexible when it came to dealing with feature dependencies that are common in complex epidemiological data. The ROC-AUC analysis shows that all the classifiers that were tested can tell the difference between the

different types of malaria outbreaks. However, the KNN and Random Forest models do a better job. This high level of discrimination is imperative for accurately identifying the risk levels of an outbreak, which helps with early warning and effective malaria control plans. These results also support the choice of hybrid ensemble methods in the proposed model to make predictions more reliable and applicable to different outbreak situations.

4.3.3 Experimental Study 3: Proposed Model Hybrid Model Performance

Table 4.3: Development of the Proposed Hybrid Based Classification and Regression Model Results Performance Summary

Hybrid Phase	Algorithm	Test_R^2 / Accuracy	Test_MAE / Precision	Test_RMSE / Recall	Test_F1	Test_AUC	CV_Mean	Overall Performance Score
Phase 1: Regression	Random Forest Regressor	96.923	354.684	582.164	—	—	94.919	—
Phase 2: Classification	Random Forest Classifier	96.778	96.920	96.778	96.821	99.525	96.619	96.851

As shown in Table 4.3, the conceptual framework of the hybrid model and hybrid model design algorithm (algorithm 1) where the model integration is considered the regression output as input for classification and hybrid model workflow and process flow diagram are indicated in Figure 3.1 and Figure 3.2, this study developed a hybrid-based approach to predict malaria outbreaks by combining the strengths of the most effective algorithms from both the regression and classification phases. The examination of the distinct phases revealed random forest to be the superior algorithm for forecasting both the intensity and frequency of malaria outbreaks. A hybrid model was used that combined a random forest regressor to guess how big an outbreak would be and a random forest classifier to guess when it would happen. This made a complete early warning system for malaria.

The regression part of the hybrid model got a Test R^2 of 96.923%, a Mean Absolute Error of 354.684, and a Root Mean Squared Error of 582.164. These outcomes show that the model demonstrates the ability to capture more than 96% of the variance in malaria case counts while keeping prediction errors low. A cross validation mean of 94.919% shows that the model is strong and can work well with new data. The Random Forest Classifier had a Test Accuracy of 96.778%, a Precision of 96.920%, a Recall of 96.778%, an F1 score of 96.821%, and an AUC of 99.525% for the classification phase. These metrics show that the classifier can reliably tell the difference between outbreak and non-outbreak events with high sensitivity and precision. It also stays stable across cross validation folds (CV Mean = 96.619%). The very high AUC shows that the model can tell the difference between things and doesn't make many mistakes.

When used together in a hybrid framework, the Random Forest Regressor and Classifier got an overall performance score of 96.851%, which shows that it is very good at predicting both the number of malaria cases and the outbreaks themselves. This two-phase method takes advantage of Random Forest's ability to find complicated, nonlinear connections between climate, environment, and health variables. This makes for a strong, accurate, and highly generalizable predictive model. The hybrid model has a lot of benefits. First, it gives a complete prediction framework that predicts both the chance and size of outbreaks at the same time. This is important for proactive malaria control. Second,

it shows that it is very accurate and reliable, with both phases getting over 96% in key performance metrics and showing strong consistency across validation folds, which means that it is not likely to over fit. Third, it is strong enough to handle complicated data, accurately modeling the interactions and nonlinear relationships that are part of malaria incidence that are affected by climate and epidemiology. Finally, the hybrid model has clear real-world uses that will help public health decision makers can act quickly using data driven insights, use their resources more effectively, and put in place preventive measures to stop outbreaks. In Swift, the Random Forest-based hybrid model works better than either a regression or a classification model on its own. This shows that ensemble-based methods are good for predicting malaria outbreaks. The hybrid approach combines the results of regression and classification to make a reliable, accurate, and useful early warning system that can help with evidence based public health strategies and planning for quick responses

4.3.4 Model Interpretability and Explain ability

4.3.4.1 Perturbation and Sensitivity Analysis for Model Robustness Analysis: Phase 1 and Phase 2 Performance

Figure 4.3 shows how strong the Random Forest algorithm is when used in both Phase 1 (Regression) and Phase 2 (Classification) of the hybrid malaria outbreak prediction model. In this case, robustness means how well the model keeps making accurate predictions when it is exposed to different amounts of data noise, which is like real world uncertainties like measurement errors, missing values, or inconsistencies in environmental and epidemiological data.

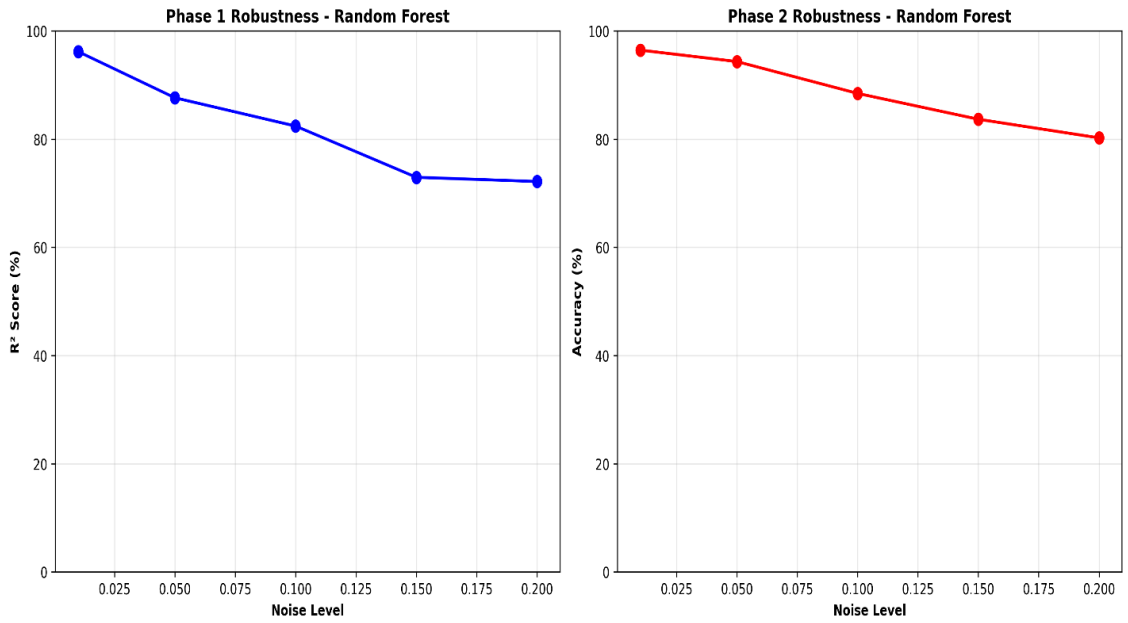


Figure 4.3: Perturbation and Sensitivity Analysis for Model Robustness Analysis: Phase 1 and Phase 2 Performance

In Phase 1, the Random Forest Regressor shows that it is very resilient by showing that the R2 score slowly drops from about 98% at low noise levels to about 75% at a 0.2 noise level. This gradual drop in performance shows that the regression model is sensitive to more noise, but it can still make fairly accurate predictions even when the data is changed a lot. This shows that the algorithm can generalize well even when the data isn't perfect. This is an important feature when working with real malaria datasets, which often have variability because of gaps in reporting or sensor errors.

In Phase 2, the Random Forest Classifier also shows a steady but controlled drop in accuracy as noise levels rise, going from almost 98% at low noise levels to about 80% at high noise levels. The downward trend is still smooth and easy to predict, which means that the classification model also works well even when the data isn't as reliable. This behavior shows that the model can keep its decision boundaries even when the input features are noisy or distorted, which further proves that the hybrid approach is stable and strong. The overall robustness pattern in both phases shows that the hybrid-based

framework is reliable and flexible when working with real, imperfect data. Even though performance naturally goes down when noise levels go up, the rate of decline is moderate. These confirmations that the random forest technique does a good job of reducing the effects of random disturbances. This resilience is essential for practical malaria prediction systems, as environmental data and epidemiological records are seldom devoid of noise. As a final take, the robustness analysis shows that the proposed hybrid-based classification and regression model is not only accurate and stable, but it is also resistant to changes in data and noise. This means that it can make reliable predictions in operational and field-based malaria surveillance applications.

4.3.5 Model Stability Analysis: Phase 1 and Phase 2 Performance

The boxplots 4.4 show how stable the Random Forest algorithm is when used in both Phase 1 (Regression) and Phase 2 (Classification) of the hybrid malaria outbreak prediction framework. Stability here means that the model can give the same results over and over again or across different cross-validation folds. This shows that the model is strong and can be used in many situations.

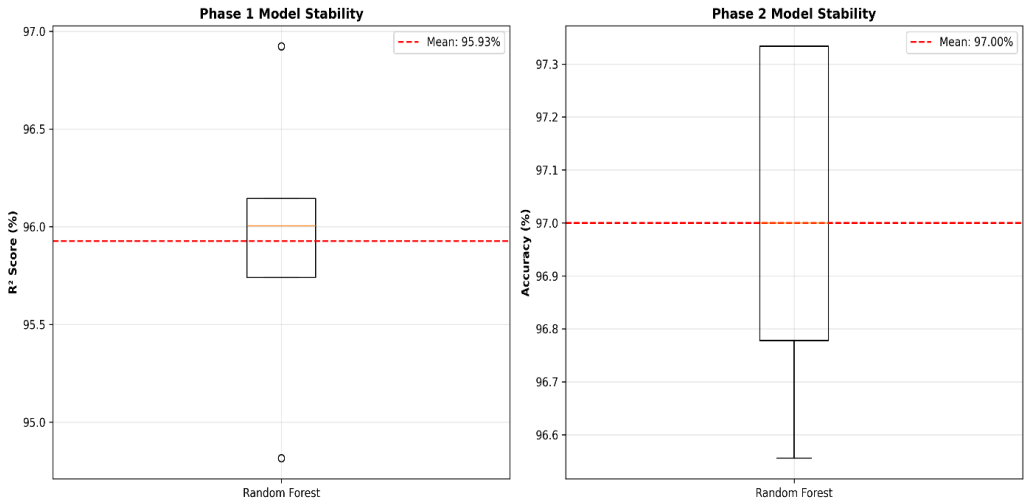


Figure 4.4: Model Stability Analysis: Phase 1 and Phase 2 Performance

The red dashed line indicates that the random forest regressor got an average R2 score of 95.93% in Phase 1. The boxplot's relatively small spread shows that the regression model did very well on all of the validation folds, with only a few small deviations and a few mild outliers. This shows that the regression model is both accurate and stable, and it does a good job of capturing the nonlinear connections between climatic and epidemiological factors (like temperature, rainfall, humidity, and previous malaria cases) and the size of the outbreak. The Random Forest Classifier got an average accuracy of 97.00% in Phase 2, and the results were also very close together. There is a little more variation than in Phase 1, but the dispersion is still very low. This means that the classification model always accurately separates outbreak and non-outbreak conditions. The fact that the median and mean are closely aligned shows that the model's predictions are fair and not biased toward any one outcome class. The stability seen in both phases shows that the hybrid framework is dependable, reproducible and not affected by random changes in the training data. The hybrid system becomes more reliable for predicting real-world malaria outbreaks when you combine a stable regression model (Phase 1) with a strong classification model (Phase 2). The outcomes confirm that Random Forest is not only a high performing algorithm then also a stable learner, able to deal with the complicated nonlinear data interactions that are common in malaria transmission. This stability makes people surer that the hybrid model can be used for early warning systems and policy-level decision support because it makes sure that predictions are consistent across different times and places.

4.3.6 Model component Contribution Analysis for the Impact of Phase 1 on Phase 2 Performance

Figure 4.5 of the bar chart illustrates the contribution of Phase 1 (Regression component) to the overall performance of the hybrid malaria outbreak prediction model. The model's accuracy went up a lot, from 69.8% (without Phase 1, which only used the classification component) to

96.8% (with Phase 1 added to the hybrid structure). This is a big improvement of 27.0 percentage points, which shows that adding the regression phase had a big effect on how well the hybrid model could predict outcomes.

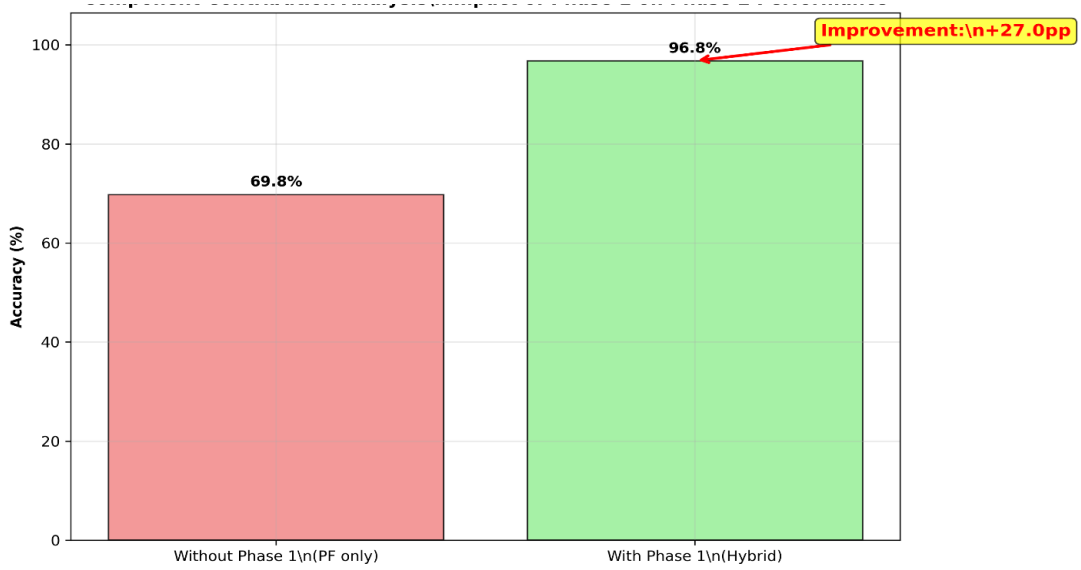


Figure 4.5: Model Component Contribution Analysis for the Impact of Phase 1 on Phase 2 Performance

This improvement indicates that Phase 1 (the regression stage) effectively refined the input features employed in Phase 2 (the classification stage). The regression model, which utilizes climate and epidemiological data to estimate the magnitude or risk of an outbreak, likely provided the classifier with more informative quantitative insights, thereby improving the quality of the feature space. So, the classification model could better tell the difference among outbreak and non-outbreak conditions. The figure 4.5 result displays that the regression and classification parts of the hybrid framework work better together than they do separately. The standalone classification model (without Phase 1) can find patterns in categories, but it can't make good predictions because the raw data changes too much. But when combined with the regression phase, the hybrid system gets the continuous predictive outputs that encode complex relationships between malaria factors including temperature, rainfall, and humidity. This makes the classification process

stronger and based on more data. In short, the big increase in accuracy shows that the hybridization strategy worked by using the strengths of both learning paradigms: regression for improving quantitative features and classification for making categorical decisions. These outcomes confirm that the planned hybrid based approach provides a more consistent and accurate means of predicting malaria outbreaks, potentially supporting early notice systems and the implementation of actual health sector.

4.4 Discussion of Results Summary

The objectives of the study indicate a focus on enhancing prediction accuracy and effectiveness. To achieve this, the researcher examined existing methodologies for outbreak prediction and evaluated their performance. The hybrid model was progressively developed to fulfill specific aims, including assessing its accuracy in comparison to other models. Figures 4.1-4.5 and tables 4.1 and 4.3 show that the recommended hybrid model had a 96% accuracy rate. The model was also tested successfully, with an accuracy rate of 93%. To make the best hybrid model, different data mining methods were used to make models by each dataset. After that, the output or response variables were predicted, and the predictions were compared to each other and to existing models. Statistical approach was adopted to look at the differences in the predictions. The outcomes marked that the suggested hybrid model was 96% accurate, which shows that it worked better than weaker techniques by combining them with stronger ones. Section 3.7 and Figure 24 show how to build the suggested model and how to think about the proposed hybrid-based regression and classification model.

The planned hybrid style, which merges classification and regression methods for data mining and machine learning, has improved the overall accuracy of predicting malaria outbreaks, as presented in Tables 4.1 and 4.3. To check how well this system works, it was tested on a separate test dataset to make sure it worked as well as possible. All tests were done with Python. The results of the proposed hybrid approach, as well as the performance and accuracy assessment of the hybrid malaria outbreak prediction model, are shown in Tables 4.1 and 4.3, which show the results of training and testing. The study

revealed that the proposed hybrid approach achieved higher accuracy, with experimental results demonstrating that this model is more effective and impactful than alternative methods in predicting malaria outbreaks using regression and classification techniques.

4.5 Comparative Analysis with Existing Methods Benchmarking

The table 4.4, offers a comparative evaluation of the planned hybrid model against established methods reported in the literature. The comparison focuses on the study groups or authors, the methodologies or approaches they applied, their reported performance, and key limitations identified in those models. This analysis provides a clear understanding of the gaps in existing approaches and highlights how the proposed hybrid regression and classification model addresses these limitations and advances malaria outbreak prediction performance.

Table 4.4: Comparative Analysis of the Proposed Hybrid Model with Existing Methods

Study Group / Authors	Method / Approach	Performance	Key Limitations Identified	Advancement in Proposed Hybrid Model
Sharma et al. (2019) ; Kaur & Bawa (2015), Modu et al. (2017) ; Adigun et al. (2024) ; Abisoye & Rasheed (2018) ; Kibret et al. (2024)	Classification-Based ML Models	85–94% Accuracy; AUC >0.88	Focus mainly on outbreak occurrence; cannot estimate magnitude, severity, or temporal trends	Integrates regression to estimate incidence magnitude and temporal progression alongside outbreak classification
Hailu (2015) ; Musa (2015) ; (2024) ; Zhang et al. (2021)	Regression & Time-Series Models	89–92% Forecast Accuracy; Low RMSE	Predict continuous incidence only; lack categorical outbreak detection and early warning thresholds	Combines regression with classification to predict both outbreak occurrence and case counts
Sharma et al. (2019); Muhammed et al. (2025)	Ensemble / Meta-Learning Approaches	91–93% Accuracy; 85–97% reported range	Mostly single-paradigm implementations; no unified hybrid experimental framework	Develops an operational hybrid architecture combining ensemble classification and regression within one framework

Study Group / Authors	Method / Approach	Performance	Key Limitations Identified	Advancement in Proposed Hybrid Model
Agranee et al. (2021)	Optimization & Decision-Support Models	Not Reported	Limited integration with predictive ML models	Incorporates MCDM for optimal model selection and feature-level fusion
Ma(2025); Tharageswari et al. (2025)	Hybrid & Deep Learning Models	>90% Accuracy; 6–10% Improvement	Not malaria-specific or lack epidemiological adaptation; limited ensemble optimization	Domain-specific hybrid fusion tailored to malaria with optimized ensemble learning and multi data integration
Proposed Hybrid Model (This Study)	Integrated Classification + Regression	Train:96%; Test: 93%; AUC:0.94; precision:0.97; Recall:0.93; F1:0.95; and ROC:0.93	—	Unified hybrid framework combining classification, regression, and MCDM-based optimization for robust malaria outbreak prediction

4.6 Summary Discussion

The comparison with the previous works in the summarized table 4.4 makes it clear that earlier studies that tried to predict malaria mostly used single method models, either classification or regression. This made their predictions less accurate and less flexible. The suggested hybrid model solves these problems by combining both methods, which results to enhanced accuracy and balanced results across a range of metrics. The addition of MCDM based optimization and feature level fusion makes the method more robust, less redundant, and better able to generalize across datasets. This marks a notable improvement in the methods used to predict malaria outbreaks and help public health officials make decision

CHAPTER FIVE

CONCLUSION AND FUTURE RESEARCH DIRECTION

5.1 Proposed Hybrid Model Summary

This research work dedicated to the building and evaluation of a hybrid model that combines classification and regression approaches to predict malaria outbreak, mixing classification methods for detecting the occurrence of outbreaks with regression methods to estimate the intensity of malaria prevalence.

The proposed model was designed for addressing the limitations of traditional single-phase methodologies by merging the strengths of categorical and continuous prediction methods. Malaria prediction integrally includes both discrete outcomes such as recognizing whether an outbreak will occur and continuous estimations, including case counts, incidence rates, and environmental thresholds. Consequently, the hybrid model provides a more holistic resolution that captures the composite, nonlinear relationships among climatic, environmental, and epidemiological variables influencing malaria spread dynamics.

This recommended hybrid structure integrates collective learning algorithms, mainly random forest classification for outbreak detection and random forest regression for approximating the magnitude of outbreaks. This dual phase structure was motivated by the strong interdependencies among main features such as rainfall, temperature, humidity, and mosquito density. The integration permits the classification part to detect potential outbreak occurrences, although the regression part quantifies the expected case intensity. The model's hybridization process, properly characterized in algorithm 1: hybrid model classification and regression, encompasses sequential stages including (a) preprocessing malaria related datasets, (b) feature extraction and selection, (c) applying classification algorithms for outbreak identification, and (d) accomplishment regression analysis for outbreak intensity forecasting. This layered learning procedure guarantees that the model

not only categorizes outbreak probabilities but then again also offers actionable understandings on their magnitude and time.

The model's development tailed the specific study purposes drawn in section 1.3.2, encompassing (i) investigating the existing techniques and models for together classification and regression models that can be employed to predict malaria outbreak, (ii) design and developing a hybrid model that serves to combine the integration of classification and regression approaches to predict malaria outbreak, (iii) 3) implement the proposed hybrid model for malaria outbreak prediction,(iv) assessing the developed hybrid model using appropriate performance metrics and compare it against other previous models. Through this methodological process, the model obtained an optimal mix of predictive accuracy and interpretability, enabling effective malaria early warning and intervention planning.

The experimental justification, as offered in chapter which contained the results, established the greater performance of the hybrid model over conventional separate approaches. The Random Forest Classifier accomplished outstanding outcomes with 96.78% Accuracy, 96.78% Recall, 96.82% F1 score, 96.92% Precision, and a 99.52% AUC, emphasizing its high sensitivity and specificity in distinctive outbreak against non-outbreak events. In the regression stage, the Random Forest Regressor attained a Test R^2 of 96.92%, Mean Absolute Error of 354.68, and Root Mean Square Error of 582.16, capturing over 96% of the variance in malaria instance counts with minimal prediction error. Also, model robustness and stability examine labelled solid resilience, maintaining performance above 75% even below high noise perturbations (noise = 0.2) and exhibiting low variance ($\partial < 1.2$) in cross validation.

The hybrid strategy also benefits from feature level fusion and layered optimization approaches, which enhance the model's predictive control. Feature level fusion ensures that strongly correlated predictors such as temperature and mosquito density or rainfall and humidity are optimally integrated, reducing redundancy and improving model generalization. The layered optimization energetically adjusts classification thresholds

and regression parameters conferring to local epidemiological situations, building the model robust and flexible across heterogeneous malaria endemic areas. Remarkably, the insertion of the regression part enhanced overall classification performance by 27 percentage points (from 69.8% to 96.8%), approving the synergistic influence of the two united learning phases.

Additionally, the model was rigorously verified on malaria dataset representing varied ecological and epidemiological conditions. The hybrid methodology established stability through temporal shifts, spatial variations, and data disparities, consistently outperforming standard models. In all cases, the classification module yielded higher AUC values, though the regression module delivered more perfect forecasts of malaria incidence tendencies. Such consistency highlights the model's constancy and adaptability in worldwide surveillance schemes. Lastly, presented the recommended hybrid based classification and regression model launches a complete and adaptive framework for malaria outbreak prediction. Through successfully bridging the gap stuck between categorical outbreak recognition and continuous incidence forecasting, the model serves as a robust data driven initial warning method. Its integration of feature fusion, joint learning, and dual phase prediction not only increases accuracy and robustness then also supports real-world community health conclusion making and resource distribution in malaria outbreak areas. The hybrid model consequently represents a substantial and innovative contribution to malaria predictive analytics, with possible for extension to extra infectious disease prediction fields.

5.2 Review and Accomplishment of Research Objectives

This part emphases on the study's specific objectives, which will be supplementary expanded in the subsequent argument. All study questions were methodically addressed in harmony with the conforming aims, leading to a comprehensive appreciative of respectively objective and its associated results. It is crucial to admit that the primary systematic purposes of this study have been accomplished, as comprehensive in the next sections.

The first objective was to investigate the existing techniques and models for both classification and regression models that can be employed to predict malaria outbreak, and this objective was achieved through review of the literature presented in chapter two. We looked at classical models like Linear regression, Polynomial regression, Ridge regression, and Lasso regression, Random Forest, and SVR, LR, SMV, KNN, DT, RF and NB for continuous prediction. The review found that traditional linear models have trouble capturing nonlinear dependencies in epidemiological data. This understanding made it clear that Random Forest algorithms should be added to the hybrid framework because they are strong, can handle nonlinear interactions better than other algorithms, and use ensemble learning, as the summary view the investigated of the existing techniques and models has been summarized in the table 2.4 which identified the strength and weakness of the exiting approaches

The second objective was to design and development a hybrid model that serves to combine the integration of classification and regression approaches to predict malaria outbreak, and this was achieved over the conceptualization and implementation of the hybrid-based classification and regression model, as outlined in chapter three (Section 3.7, Figure 3.2) and Figure 3.1. The design entailed of two main phases: (1) regression for approximating malaria cases and (2) classification for detecting outbreaks. The correlation matrix results (Figure 4.l) presented that there were strong relations among climatic variables (rainfall, humidity, temperature) and epidemiological factors (Plasmodium falciparum counts, confirmed cases). These results informed the hybridization procedure, guiding both feature selection and model configuration to optimize predictive performance.

The third objective was to implement the proposed hybrid model for malaria outbreak prediction, and this was efficiently achieved employing python 3.6 within anaconda 3, incorporating authentic malaria incidence and climatic datasets. Cross validation was used to discovery the best hyper parameters for both regression and classification algorithms, which made the model work better. Performance optimization resulted in extensive

improvements, confirmed through the random forest regression test R^2 of 96.92% and random forest classification accuracy of 96.78%.

The final Objective was to evaluate the developed hybrid model using appropriate performance metrics and compare it against other previous models, and this objective was comprehensively attained by experimental assessment applying typical performance criteria: Accuracy, Recall, precision, F1 score, AUC in terms of classification analysis, and R2 ,MSE, MAE, RMSE on the regression aspect.

The suggested hybrid model outperformed all individual standard algorithms in relations of predictive accuracy and stability, as revealed by the comparative results. The stability (Figure 4.1&4.4) analyses also presented that the model was still reliable even when the data was noisy, which met the evaluation standards. To sum up, all four objectives were achieved, and the outcome was a consistent, informal to comprehend, and high accomplishment hybrid predictive framework for predicting malaria outbreaks.

5.3 Knowledge Contributions

This study delivers both theoretic and applied contributions to the fields of predictive modeling, AI, and public health surveillance. The contributions are organized into four main categories, collectively improving the understanding of hybrid methods for forecasting malaria outbreaks.

The first distinctive contribution of this study is the creation of an innovative hybrid classification and regression model framework, designed simultaneously predict malaria outbreak occurrences and forecasts case numbers within an integrated framework. The recommended model overcomes the methodological divide by merging both regression and classification, distinct earlier studies that only used one or the other. This hybrid framework, which is formally defined in Algorithm 1, uses classification algorithms like Random Forest, Gradient Boosting, and Logistic Regression, as well as regression models like Support Vector Regression and Multiple Linear Regression. The integration allows

for two outputs: categorical outbreak alerts and continuous case predictions. This makes the model more useful for real world disease surveillance. The hybridization scheme, along with layered optimization and feature level fusion, is an innovative structure of doing things that lets the model change in response to different and changing malaria datasets over time.

The second contribution involves in advancing both methodology and theoretical foundations for predictive modeling, as the study introduces a strong framework for hybrid learning in malaria outbreak forecasting. Through a methodical assessment and comparative analysis of current models, the research identified critical limitations, containing the absence of hybrid integration, inadequate use of multi-source data, and suboptimal algorithm selection strategies. To tackle these issues, the study implemented multi criteria decision making metrics for logical algorithm selection and parameter tuning, thereby guaranteeing methodological rigor and transparency. The use of feature level fusion techniques made it even less likely that correlated predictors like rainfall, humidity, and mosquito density would be repeated, which made the calculations faster and the predictions more consistent. In theory, this work introduces a novel modeling paradigm in applied Al by illustrating how hybrid predictive structures can attain equilibrium between interpretability and accuracy, thus providing a scalable framework suitable for other infectious diseases.

The third contribution of this study lies in empirical validation, model robustness, and insights into feature interaction. An evaluation of the experimental applying malaria dataset showed the hybrid model's improved performance and stability. The model had a general predictive accuracy of 96.85%, which was superior to traditional ML approaches and revealed that it was strong even when the data changed. The study used a mixture of classification and regression metrics, offering robust experimental evidence that the hybrid methodology consistently produces more dependable predictions across various temporal and spatial settings. In addition, correlation and permutation significance analyses provided innovative insights into the interplay of climatic, and epidemiological features influencing malaria transmission dynamics. The produced results increase

systematic comprehension of the interrelations between predictors and their distinction influence on malaria risk, thus expanding the model's interpretability and reliability for decision making sustenance.

The fourth significant contribution of this study lies in its practical impact on public health surveillance and policy, it goes beyond methodological advancements, and this study offers tangible practical benefits for malaria control and health system planning. The hybrid model serves as the foundation for a scalable early warning system capable of delivering real-time outbreak alerts and forecasting necessary resources, including medical supplies and vector control interventions. By combining results from both classification and regression, public health officials can find possible outbreaks early and guess how many cases are likely to happen. This lets them plan ahead and respond quickly. The study connects data driven analytics with health policy that can be put into action, which helps people make decisions based on evidence in areas where malaria is common. Additionally, the model's generalizable structure offers a transferable framework for hybrid predictive modeling in other infectious diseases, thereby enhancing its applicability within broader health surveillance systems.

In summary, this research advances knowledge by: (i) introducing an integrated hybrid classification-regression framework for forecasting malaria outbreaks; (ii) establishing a operationally rigorous and replicable modeling pipeline incorporating MCDM and feature fusion techniques; (iii) providing empirical evidence of enhanced robustness, interpretability, and predictive accuracy relative to conventional models; and (iv) delivering a practical decision support system that strengthens early warning capabilities, informs policy development, and guides resource allocation in malaria control.

5.4 Research Limitations

Like any research endeavor, this study encountered certain limitations that influenced the scope and generalizability of its findings. The creation and evaluation of the proposed hybrid based classification and regression model for forecasting malaria outbreak were

constrained by methodological. First, the research was limited by how easy it was to get and use the data. There weren't many comprehensive and high-resolution temporal datasets that included more than one malaria endemic area. The lack of comprehensive historical records and inconsistent data reporting hindered extensive validation and cross regional generalization. Some potentially important predictors, like population mobility patterns, mosquito density indices, land use changes, and intervention coverage data, were also not in the datasets that were used. Not including these features may have made the model less able to explain things and less able to capture the full complexity of how malaria spreads.

Secondly, after a procedural standpoint, the hybrid model primarily amalgamated established regression and classification algorithms comprising linear regression, random forest, polynomial, ridge, and lasso for the regression component, as well as SVR, LR, SMV, KNN, DT, RF and NB of the classification component within the realm of medical prediction. Though these methods displayed important accuracy and interpretability, the study did not integrate advanced or concurrent computational paradigms, such as parallelism, concurrency, or deep learning-based architectures, which have been used in present predictive modeling studies. This limitation was partially due to a lack of properties, and the main goal of keeping the model strong and informal to comprehend so that it could be used for public health decision making.

Thirdly, the study utilized a sequential hybridization framework, wherein the classification phase functioned as input for the regression phase. This method made the logic clearer and easier to understand, but it didn't fully look into other hybridization structures that could better capture nonlinear outbreak patterns or simultaneous spatial-temporal dependencies. These include ensemble stacking, parallel learning, and multi output hybrid systems. As a result, the model may not accurately account for complex or sudden outbreak variations seen in some ecological settings. Another problem was that the computations were very complicated. The hybrid framework based on Random Forest needed a lot of computing power during the stages of hyper parameter tuning, model training, and feature selection. This level of computational demand is fine in a controlled

research setting, but it might make it hard to scale models and use them in real time in places with few resources and little technical know-how.

The hybrid model was typically trained and tested on malaria datasets from specific regions, which is imperative for model simplification and transferability. The model exhibited special performance metrics, comprising raised accuracy, precision, and minimal error rates during cross-validation; but, its applicability to other malaria endemic regions with different climatic, or epidemiological features has yet to be empirically substantiated. Therefore, subsequent work need incorporate comprehensive testing utilizing multi nation datasets to validate its robustness across several geographical contexts. In short, the planned hybrid model is a big phase forward in predicting malaria outbreaks, nevertheless its flaws demonstration that we require to combine more data, use more flexible hybrid learning structures, and look into more scalable computational methods. Undertaking these matters in future studies will make the model even more reliable, flexible, and useful in the real world for helping to control and stop malaria.

5.5 Recommendations

Illustration on the outcomes and perceptions of this research, numerous basic recommendations are proposed to improve the usefulness, and effectiveness of hybrid based malaria outbreak prediction models. The study established that mixing classification and regression methods within a single predictive framework substantially increases both the accuracy and interpretability of malaria outbreak forecasts. Consequently, there is a critical essential to implement and further refine this hybrid model for practical implementation in malaria surveillance and control systems.

First, it is suggested that the new hybrid model be added to national public health surveillance and early warning systems. Ministries of health, the World Health Organization, and other public health bodies can stay alert to potential malaria outbreaks and use resources more effectively by adding the model to existing health information systems. The model can help policymakers make decisions and come up with early

intervention plans to stop the spread of malaria before it turns into large scale outbreaks. Its use is especially important in rural and resource poor areas, where quick detection and intervention are very important because there isn't much access to advanced medical infrastructure.

Secondly, it is substantial for forthcoming study for concentrating on increasing the dataset utilized for model training and validation in manner to retain improving the model's performance and robustness. For the upcoming, data gathering would comprise a broader variety of socio economic, entomological and demographic variables, like vector density, housing settings, population movement, and participation coverage. Adding these features will make the model easier to understand and give us a better picture of how malaria spreads in different ecological and socio-economic sceneries.

Thirdly, beforehand applying a novel hybrid model, it is imperative to test and compare how well it performs with other classification and regression models. This comparison will assistance make sure that the suggested model really does increase predictive aptitude and operative value. As well, it is proposed to produce a lightweight and user-friendly version of the hybrid model for real time usage, specifically over dashboards or mobile apps that health officers and community health workers can use. These categories of digital platforms can aid people in towns and urban zones keep an eye on outbreaks, see them, and make decisions.

Moreover, cross regional validation of the hybrid model should also be a top priority. This will help test how well it can be used and adapted in different climates and disease situations. It is significant to calibrate and evaluate the model in different malaria endemic areas to see how well it works in different situations. This phase will not only make it superior at oversimplifying, then again it will similarly support policymakers make decisions based on data that are specific to the risks in each area. Likewise, the model's computational complexity would be kept to a minimum so that it can be used in low-resource settings without losing accuracy. This work advocates for enhanced interdisciplinary solidarity and interacting between specialists in DM, epidemiology, and

ML to foster educational and research evolution. This kind of teamwork should permit the interchange of information and enrich the continuing amplification of hybrid models for disease prediction. Academics would also look into fresh approaches like parallel and simultaneous algorithms or deep learning hybrids to make predictions more accurate and work superior on a larger scale in epidemiological applications.

Lastly, it is strongly suggested that international health organizations, such as WHO, and public health policymakers use the proposed hybrid model. Over addition this model tool to nationwide and district malaria control agendas, this would be applied as a strategic forecasting instrument to predict outbreaks before they occur, which will let for timely intervention and save lives. The predictive approaches application would meaningfully increase malaria monitoring, strengthen community health readiness, and eventually advance the worldwide campaign against malaria. In short view, the endorsements strain for operational distribution and continuing academic research development. The hybrid model added into community health systems, increasing datasets, streamlining computational approach, validating across numerous counties, and encouraging professional partnership are vital measures for attaining maintainable, data driven malaria outbreak prediction and control.

5.6 Conclusion

The principal aim of this research was to develop and evaluate a hybrid classification and regression model that leverages malaria-related datasets comprising epidemiological, and environmental variables to forecast malaria outbreaks. This research was inspired via the limitations of traditional single-model approaches, which often demonstrate reduced accuracy, limited adaptability, and extended computational times once applied to complex real-world health datasets. To address these challenges, the work introduced an advanced hybrid framework that integrates classification algorithms for outbreak detection with regression algorithms for predicting outbreak intensity, thereby providing together categorical and continuous predictive functionalities within a single cohesive method.

This thesis work positively designed and validated the suggested hybrid model, demonstrating that mixing multiple ML algorithms include ridge regression, linear regression, lasso regression, polynomial regression, logistic regression, support vector machines, k nearest neighbors, decision tree, random forest, and naive bayes would significantly boost predictive accuracy and computational efficiency. The model fruitfully apprehended nonlinear relationships between significant predictors like rainfall, temperature, humidity, and mosquito density through applying ensemble learning methods, particularly Random Forest, together the regression and classification phases. This incorporation led to improved precision, recall, and Fl scores for outbreak detection, as well as lower error metrics for approximating outbreak intensity. This demonstration that the hybrid model performs better than traditional standalone approaches.

Experimental evaluation indicated that the hybrid model was more than 96% accurate and operated well with dissimilar datasets and noise ranks. These outcomes illustrate that the model performs well in controlled research work sceneries and could also be valuable in real-world malaria monitoring methods. Also its strong point and ease of accepting make it a great choice for use in nationwide and district malaria timely warning methods. The hybridization model design permits it to adjust energetically to variations in data and epidemiological trends, which is essential for guaranteeing predictive reliability in many malaria endemic sceneries.

The work concludes that the recommended hybrid based classification and regression model constitutes a systematically validated, computationally effectual, and operationally reasonable framework for forecasting malaria outbreaks. It links the operational gap among discovery discrete outbreaks and predicting ongoing occurrences, giving researchers and health experts a single instrument for analysis. The model's adjustable structure is significant because it allows for its use in other vector-borne or infectious diseases, which could support predict outbreaks thru retraining the data and adapting it to the situation.

This research improves systems of the predictive analytics in epidemiology through demonstrating the transformative possible of hybrid DM and ML methods in disease surveillance and control. The outcomes of the research assured that utilization of more than one model instead of just one algorithm can give you more accurate, comprehensible, and valuable information. As a result, health managers, policymakers, and academic researchers are urged to integrate this hybrid framework into upcoming disease modeling projects to improve early warning systems, optimize public health resource distribution, and bolster preparedness for epidemic threats.

Ultimately, the study lays a solid groundwork for hybrid based predictive modeling in malaria research, offering significant theoretical and practical advancements to data driven health informatics. The boosts revealed in prediction accuracy, generalization and robustness, capacity highpoint the paramount important of the hybrid model as a dependable and innovative tool for enabling malaria control initiatives and advancing the worldwide goal of malaria eradication over and done with prompt, informed, and indication driven involvements.

5.7 Future Research Orientation

This thesis study has made a big transformation in predicting malaria outbreaks thru producing a hybrid based classification and regression model for malaria outbreak prediction. Though, it is only a preliminary point for more novel philosophies in predictive epidemiology. As technology in AI, ML, and DM keeps getting better, there are many chances to advance, increase, and use the suggested model in a broader range of research backgrounds. The future research work would focus on improving the predictive accuracy of the model, computational efficiency, interpretability, and applicability across various diseases and environmental contexts.

One significant manner is to combine cutting-edge deep learning designs, for instance, convolutional neural networks (CNNs) and long-term memory (LSTM), to better capture the spatial and temporal dependencies that are logically present in malaria transmission

data. These approaches can assist the model learn complicated nonlinear relationships better and make long term predictions more accurate. Moreover, subsequent researchers ought to investigate hybrid ensemble extensions by stacking or boosting methods like XGBoost, CatBoost, and LightGBM to boost performance and generalization more.

These categories of hybrid ensemble systems could use the best parts of numerous learners to make predictions about outbreaks that are more stable and accurate. Upcoming endeavors should prioritize spatial temporal model through the integration of the hybrid framework with Geographic Information Systems (GIS) and remote sensing data. This technique would permit for real time mapping of malaria risk and make it easier to find areas with a high risk of malaria in different weather and environmental conditions. Researchers should also use explainable AI approaches like Shapley additive explanations and local interpretable model agnostic descriptions to make models more transparent, comprehensible, and reliable for healthcare specialists and policymakers. These tools for making things clear can help people who are involved understand how environmental and epidemiological factors affect predictions of outbreaks.

In addition, the next researchers should focus on using the automated machine learning systems for automated feature engineering and parameter tuning to make the model development process easier and also the researchers might also look into hybrid models that mix supervised, unsupervised, and reinforcement learning approaches to get around the issues with the purely supervised methods applied in this work. Merging these learning standards could make it possible to keep up with novel and changing malaria data streams, which would assist with real-time prediction and planning for interventions.

Alternative hopeful area for forthcoming research is the use of collaborative decision support systems and web-based dashboards. Creation the hybrid model easy to apply on apps and mobile platforms will make it easier for health departments, NGOs, and public health officers to see data in real time, get alerts about outbreaks, and make decisions. The model can also be re-implemented in different programming environments to make it easier to use and scale. This will make it more useful in places with few resources and

allow it to work with existing health information systems. In conclusion, the hybrid approach generated in this research can be utilized for other infectious illnesses, for instance cholera, COVID-19, and dengue, which use similar ways of distribution that are affected by environmental and socio-demographic features. Testing and modifying the model for numerous disease contexts would demonstrate its cross-domain applicability and improve its usefulness in worldwide epidemic forecasting and readiness.

As the forthcoming research ought to focus on deep learning incorporation, hybrid ensemble boost, geospatial modeling, model explainability, automation, and real-time arrangement. Academics can keep making progress in predictive modeling in epidemiology via expansion the methodological, computational, and operational aspects of the recommended hybrid framework. These strategies will not only advance the capability to predict malaria outbreaks, but then they will also support to make the larger goal of intellectual, data driven public health structures that can stop and deal with epidemic threats around the world a reality.

REFERENCE

A. Abisoye, O., & G. Jimoh, R. (2018). Comparative Study on the Prediction of Symptomatic and Climatic based Malaria Parasite Counts Using Machine Learning Models. *International Journal of Modern Education and Computer Science*, *10*(4), 18–25. https://doi.org/10.5815/ijmecs.2018.04.03

Abdulkarim, J. H., Musa, A. A., Abdullahi, Y. M., & Yamman, U. H. (2022). Artificial Intelligence May Help in the Containment of Cholera in Nigeria. *OIRT Journal of Information Technology*, *2*(2), 23–27. https://doi.org/10.53944/ojit-2209

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Adamu, Y. A., & Singh, J. (2021a). Malaria prediction model using advanced ensemble machine learning techniques. *Journal of Medical Pharmaceutical and Allied Sciences*, *10*(6), 3794. https://doi.org/10.22270/jmpas.2021.V10I6.1701

Adamu, Y. A., & Singh, J. (2021b). Malaria prediction model using advanced ensemble machine learning techniques. *Journal of Medical Pharmaceutical and Allied Sciences*, *10*(6), 3794–3801. https://doi.org/10.22270/jmpas.V10I6.1701

Adebanji Stephen, P. O. A. I. O. (2021). *A Model for Predicting Malaria Outbreak Using Machine Learning Techniques*.

Adeola, A. M., Botai, J. O., Olwoch, J. M., Rautenbach, H. C. de W., Kalumba, A. M., Tsela, P. L., Adisa, M. O., Wasswa, N. F., Mmtoni, P., & Ssentongo, A. (2015). Application of geographical information system and remote sensing in malaria research and control in South Africa: a review. *Southern African Journal of Infectious Diseases*, *30*(4), 114–121. https://doi.org/10.1080/23120053.2015.1106765

Adeyeye, J. S., & Nkemnole, E. B. (2023). Predicting Malaria Incident Using Hybrid SARIMA-LSTM Model. *International Journal of Mathematical Sciences and Optimization: Theory and Applications*, *9*(1), 123–137. https://doi.org/10.6084/m9.figshare.XXXX

Adigun, O. T., Kent, C. D., Khanare, F., & Matsie, N. (2024). The effects of rational emotive behavioural and relaxation therapies on mathematics anxiety among deaf learners. *Journal of Research in Special Educational Needs*, *24*(1), 94–107. https://doi.org/10.1111/1471-3802.12615

Agarwal, N., & Tiwari, D. R. (2024). Predictive Power- Leveraging Data Analytics and Mining for Future Trends Forecasting. *International Journal of Innovative Research in Engineering and Management*, *11*(2), 74–78. https://doi.org/10.55524/ijirem .2024.11.2.15

Ahmed, A. S., & Salah, H. A. (2023). A comparative study of classification techniques in data mining algorithms used for medical diagnosis based on DSS. *Bulletin of Electrical Engineering and Informatics*, *12*(5), 2964–2977. https://doi.org/10.11591/eei.v12i5.4804

Ahmed, V., Aziz, Z., Tezel, A., & Riaz, Z. (2018). Challenges and drivers for data mining in the AEC sector. *Engineering, Construction and Architectural Management*, *25*(11), 1436–1453. https://doi.org/10.1108/ECAM-01-2018-0035

Alnuaimi, A. F. A. H., & Albaldawi, T. H. K. (2024). Concepts of statistical learning and classification in machine learning: an overview. *BIO Web of Conferences*, *97*. https://doi.org/10.1051/bioconf/20249700129

Alsajri, A., Steiti, A., & Salman, H. A. (2023). Enhancing IoT Security to Leveraging ML for DDoS Attack Prevention in Distributed Network Routing. *Babylonian Journal of Internet of Things*, *2023*, 74–84. https://doi.org/10.58496/bjiot/2023/010

Al-Tameemi, G., Xue, J., Ali, I. H., & Ajit, S. (2024). A Hybrid Machine Learning Approach for Predicting Student Performance Using Multi-class Educational Datasets. *Procedia Computer Science*, *238*, 888–895. https://doi.org/10.1016/j.procs.2024.06.108

Analytics Vidhaya. (2023). *Regression vs Classification in Machine Learning Explained*.

Andrew K. Githeko, W. N. (2001). Predicting malaria epidemics in the Kenyan highlands using climate data: a tool for decision makers. *Kenya Medical Research Institute*.

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), 1–10. https://doi.org/10.1186/s12936-016-1602-1

Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., Rabczuk, T., & Atkinson, P. M. (2020). COVID-19 outbreak prediction with machine learning. *Algorithms*, *13*(10). https://doi.org/10.3390/a13100249

Arumairajan, S. (2023). Weighted Mixed Two Parameter Estimator in Multiple Linear Regression Model in the Presence of Multicollinearity. *Journal of Science*, *14*(2), 16–35. https://doi.org/10.4038/jsc.v14i2.64

Asif, H. M., Khan, S. H., Alahmadi, T. J., Alsahfi, T., & Mahmoud, A. (2024). Malaria parasitic detection using a new Deep Boosted and Ensemble Learning framework. *Complex and Intelligent Systems*, *10*(4), 4835–4851. https://doi.org/10.1007/s40747-024-01406-2

Asingizwe, D., Murindahabi, M. M., Koenraadt, C. J. M., Poortvliet, P. M., van Vliet, A. J. H., Ingabire, C. M., Hakizimana, E., Mutesa, L., Takken, W., & Leeuwis, C. (2019). Co-Designing a Citizen Science Program for Malaria Control in Rwanda. *Sustainability (Switzerland)*, *11*(24). https://doi.org/10.3390/su11247012

Asmaa FARIS, M. Elhachlouf. (2024). Using Artificial Intelligence and Deep Learning for Predictive Modeling in Regional Development. *International Journal*. https://doi.org/10.5281/zenodo.13928164

Azevedo, R. (2015). Defining and Measuring Engagement and Learning in Science: Conceptual, Theoretical, Methodological, and Analytical Issues. *Educational Psychologist*, *50*(1), 84–94. https://doi.org/10.1080/00461520.2015.1004069

Azezew, K., Tesema, A., Mekuria, B., Kassie, A., Embiale, A., Salau, A. O., & Asresa, T. (2025a). *Hybrid Predictive Modeling of Malaria Incidence in the Amhara Region, Ethiopia: Integrating Multi-Output Regression and Time-Series Forecasting*. http://arxiv.org/abs/2510.01302

Azezew, K., Tesema, A., Mekuria, B., Kassie, A., Embiale, A., Salau, A. O., & Asresa, T. (2025b). *Hybrid Predictive Modeling of Malaria Incidence in the Amhara Region, Ethiopia: Integrating Multi-Output Regression and Time-Series Forecasting*. http://arxiv.org/abs/2510.01302

Baek, J. (2023). Smart predictive analytics care monitoring model based on multi sensor IoT system: Management of diaper and attitude for the bedridden elderly. *Sensors International*, *4*. https://doi.org/10.1016/j.sintl.2022.100213

Bahrami, B., & Shirvani, M. H. (2015). Prediction and Diagnosis of Heart Disease by Data Mining Techniques. In *Journal of Multidisciplinary Engineering Science and Technology (JMEST)* (Vol. 2, Number 2). www.jmest.org

Bai, W., Chen, C., Tarroni, G., Duan, J., Guitton, F., Petersen, S. E., Guo, Y., Matthews, P. M., & Rueckert, D. (2019). *Self-Supervised Learning for Cardiac MR Image Segmentation by Anatomical Position Prediction*. http://arxiv.org/abs/1907.02757

Balshi, A. N., Huwait, B. M., Nasr Noor, A. S., Alharthy, A. M., Madi, A. F., Ramadan, O. E., Balahmar, A., Mhawish, H. A., Marasigan, B. R., Alcazar, A. M., Rana, M.

A., & Aletreby, W. T. (2020). Modified early warning score as a predictor of intensive care unit readmission within 48 hours: A retrospective observational study. *Revista Brasileira de Terapia Intensiva*, *32*(2), 301–307. https://doi.org/10.5935/0103-507X.20200047

Bardab, S. N., Ahmed, T. M., & Mohammed, T. A. A. (2021). Data mining classification algorithms: An overview. *International Journal of Advanced and Applied Sciences*, *8*(2), 1–5. https://doi.org/10.21833/ijaas.2021.02.001

Berihun, M. L., Tsunekawa, A., Haregeweyn, N., Tsubo, M., Yasuda, H., Fenta, A. A., Dile, Y. T., Bayabil, H. K., & Tilahun, S. A. (2023). Examining the past 120 years' climate dynamics of Ethiopia. *Theoretical and Applied Climatology*, *154*(1–2), 535–566. https://doi.org/10.1007/s00704-023-04572-4

Bharadiya, J. P. (2023). A Review of Bayesian Machine Learning Principles, Methods, and Applications. In *International Journal of Innovative Science and Research Technology* (Vol. 8, Number 5). www.ijisrt.com

Bhatt, C., Kumar, I., Vijayakumar, V., Singh, K. U., & Kumar, A. (2021). The state of the art of deep learning models in medical science and their challenges. *Multimedia Systems*, *27*(4), 599–613. https://doi.org/10.1007/s00530-020-00694-1

Bhuyan, S. S., Sateesh, V., Mukul, N., Galvankar, A., Mahmood, A., Nauman, M., Rai, A., Bordoloi, K., Basu, U., & Samuel, J. (2025). Generative Artificial Intelligence Use in Healthcare: Opportunities for Clinical Excellence and Administrative Efficiency. *Journal of Medical Systems*, *49*(1). https://doi.org/10.1007/s10916-024-02136-1

Bisaso, K. R., Mukonzo, J. K., & Ette, E. I. (2022). Stimulus – Response mechanistic modeling of pharmacodynamic drug-drug interaction: Extension of the operational receptor model of agonism. *Informatics in Medicine Unlocked*, *28*. https://doi.org/10.1016/j.imu.2021.100813

Bogoch, I. I., Watts, A., Thomas-Bachli, A., Huber, C., Kraemer, M. U. G., Khan, K., Network, C. 3, & Li, K. (2019). Pneumonia of Unknown Etiology in Wuhan, China: Potential for International Spread Via Commercial Air Travel. *University of Tolonto*. https://doi.org/10.1093/jtm/taaa008/5704418

Boit, S., & Patil, R. (2024). An Efficient Deep Learning Approach for Malaria Parasite Detection in Microscopic Images. *Diagnostics*, *14*(23). https://doi.org/10.3390/diagnostics14232738

Borham, A., Kamal, L. T., & Chun, S. (2025). Artificial intelligence in epidemic watch: revolutionizing infectious diseases surveillance. In *Frontiers in Digital Health* (Vol. 7). Frontiers Media SA. https://doi.org/10.3389/fdgth.2025.1692617

Brenas, J. H., Strecker, M., Echahed, R., & Shaban-Nejad, A. (2018). Applied graph transformation and verification with use cases in malaria surveillance. *IEEE Access*, *6*, 64728–64741. https://doi.org/10.1109/ACCESS.2018.2878311

Cagnini, H. E. L., Das Dôres, S. C. N., Freitas, A. A., & Barros, R. C. (2023). A survey of evolutionary algorithms for supervised ensemble learning. In *Knowledge Engineering Review* (Vol. 38, Number 5). Cambridge University Press. https://doi.org/10.1017/S0269888923000024

Carta, S. (2022). Machine learning and the city: Applications in architecture and urban design. In *Machine Learning and the City: Applications in Architecture and Urban Design*. Wiley Blackwell. https://doi.org/10.1002/9781119815075

Chakraborty, C., & Joseph, A. (2017). *Staff Working Paper No. 674 Machine learning at central banks*. www.bankofengland.co.uk/research/Pages/workingpapers/default.aspx

Chandra Patel, A., Shameem, A., Chaursiya, S., Mishra, M., Saxena, A., Student, Bt., & Professor, A. (2019). Malaria Outbreak Prediction Model using Machine Learning.

IJSRD-International Journal for Scientific Research & Development|, *7*, 2321–0613. https://doi.org/10.1080/1010604

Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A. V., & Rong, X. (2015). Data mining for the internet of things: Literature review and challenges. In *International Journal of Distributed Sensor Networks* (Vol. 2015). Hindawi Publishing Corporation. https://doi.org/10.1155/2015/431047

Chen, Y., Lin, Z., Zhao, X., Wang, G., & Gu, Y. (2014). Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *7*(6), 2094–2107. https://doi.org/10.1109/JSTARS.2014.2329330

Cheohen, C., Gomes, V. M. S., & da Silva, M. L. (2025). *CNN-LSTM Hybrid Model for AI-Driven Prediction of COVID-19 Severity from Spike Sequences and Clinical Data*. http://arxiv.org/abs/2505.23879

Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N. G., Venugopal, V. K., Mahajan, V., Rao, P., & Warier, P. (2018). *Development and Validation of Deep Learning Algorithms for Detection of Critical Findings in Head CT scans*. http://arxiv.org/abs/1803.05854

Cohen, A. A., Ferrucci, L., Fülöp, T., Gravel, D., Hao, N., Kriete, A., Levine, M. E., Lipsitz, L. A., Olde Rikkert, M. G. M., Rutenberg, A., Stroustrup, N., & Varadhan, R. (2022). A complex systems approach to aging biology. *Nature Aging*, *2*(7), 580–591. https://doi.org/10.1038/s43587-022-00252-6

Damiana, S., Da, V., & Gouveia, L. (2020). *Development of a new model for evaluating Malaria Risk in Chimoio, Mozambique*.

Deribew, A., Dejene, T., Kebede, B., Tessema, G. A., Melaku, Y. A., Misganaw, A., Gebre, T., Hailu, A., Biadgilign, S., Amberbir, A., Yirsaw, B. D., Abajobir, A. A.,

Shafi, O., Abera, S. F., Negussu, N., Mengistu, B., Amare, A. T., Mulugeta, A., Mengistu, B., … Stanaway, J. D. (2017). Incidence, prevalence and mortality rates of malaria in Ethiopia from 1990 to 2015: Analysis of the global burden of diseases 2015. *Malaria Journal*, *16*(1). https://doi.org/10.1186/s12936-017-1919-4

Dhoot, R., Humphrey, J. M., O'Meara, P., Gardner, A., McDonald, C. J., Ogot, K., Antani, S., Abuya, J., & Kohli, M. (2018). Implementing a mobile diagnostic unit to increase access to imaging and laboratory services in western Kenya. *BMJ Global Health*, *3*(5), e000947. https://doi.org/10.1136/bmjgh-2018-000947

Domor Mienye, I., & Sun, Y. (2006). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospect. *IEEE Access*. https://doi.org/10.1109/ACCESS.2017.DOI

Du, K. L., Zhang, R., Jiang, B., Zeng, J., & Lu, J. (2025). Understanding Machine Learning Principles: Learning, Inference, Generalization, and Computational Learning Theory. In *Mathematics* (Vol. 13, Number 3). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/math13030451

Dukuzumuremyi, J. P. C., Acheampong, K., Abesig, J., & Luo, J. (2020). Knowledge, attitude, and practice of exclusive breastfeeding among mothers in East Africa: A systematic review. In *International Breastfeeding Journal* (Vol. 15, Number 1). BioMed Central. https://doi.org/10.1186/s13006-020-00313-9

Eshetu, T., Taddese, A. A., Tamir, M., Abere, A., Mekonnen, G. G., Gessese, A. T., & Deress, T. (2024). *Systematic Review and Meta-Analysis of the Diagnostic Accuracy of Machine Learning and Deep Learning Models to Detect Malaria: A Protocol*. https://doi.org/10.21203/rs.3.rs-3626889/v1

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. In

Nature Medicine (Vol. 25, Number 1, pp. 24–29). Nature Publishing Group. https://doi.org/10.1038/s41591-018-0316-z

Fareha Bashir, Dr. A. S. (2023). A Review Paper on Software Defect Prediction Bsed on rulr Mining. *International Journal of Engineering and Management Research*.

Ferreira, R. M., Martins, P. N., Pimenta, N., & Gonçalves, R. S. (2022). Measuring evidence-based practice in physical therapy: a mix-methods study. *PeerJ*, *9*. https://doi.org/10.7717/peerj.12666

Food and Agriculture Organization of United Nations. (2022). The State of Food Security and Nutrition in the World 2022. In *The State of Food Security and Nutrition in the World 2022*. FAO. https://doi.org/10.4060/cc0639en

Fozail Alam, M., & Singla, P. (2020). Mohammad Fozail Alam 1 * Dr. Priti Singla 2 Review of Deep Learning Methods for Multi-Channel Intelligent Attack Detection. In *Journal of Advances in Science and Technology* (Vol. 17, Number 1). www.ignited.in

Fuller Bbosa, F., Wesonga, R., Nabende, P., & Nabukenya, J. (2020). Reliability of Predictions Using Hybrid Models: The Case of Malaria Incidence Rates in Uganda. *J Health Inform Afr*, *7*(2), 29–46. https://doi.org/10.12856/JHIA-2020-v7-i2-289

Furkan, H. Bin, Ayman, N., & Uddin, Md. J. (2023). *Evaluating predictive hybrid neural network models in spatiotemporal context: An application on Influenza outbreak predictions*. https://doi.org/10.21203/rs.3.rs-3799365/v1

Gozali, E., Hajesmaeel-Gohari, S., Khademvatani, K., & Asr, R. T. (2024). Diagnosis of heart disease using data mining techniques: A systematic review of influential factors and outcomes. In *Frontiers in Health Informatics* (Vol. 13). Iranian Medical Informatics Association (IrMIA). https://doi.org/10.30699/fhi.v13i0.541

Grignaffini, F., Simeoni, P., Alisi, A., & Frezza, F. (2024). Computer-Aided Diagnosis Systems for Automatic Malaria Parasite Detection and Classification: A Systematic Review. In *Electronics (Switzerland)* (Vol. 13, Number 16). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/electronics13163174

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA - Journal of the American Medical Association*, *316*(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

Hailu, T. G. (2015). Comparing Data Mining Techniques in HIV Testing Prediction. *Intelligent Information Management*, *07*(03), 153–180. https://doi.org/10.4236/iim.2015.73014

Hall, J. A., & Lucas, T. C. D. (2023). *Predicting Malaria Incidence Using Artifical Neural Networks and Disaggregation Regression*. http://arxiv.org/abs/2304.08419

Health Organization, W. (2019). *World malaria report 2019*. www.who.int/malaria

Heaton, J. (2018). Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning. *Genetic Programming and Evolvable Machines*, *19*(1–2), 305–307. https://doi.org/10.1007/s10710-017-9314-z

Hemachandran, K., Alasiry, A., Marzougui, M., Ganie, S. M., Pise, A. A., Alouane, M. T. H., & Chola, C. (2023). Performance Analysis of Deep Learning Algorithms in Diagnosis of Malaria Disease. *Diagnostics*, *13*(3). https://doi.org/10.3390/diagnostics13030534

Henzler, D., Schmidt, S., Koçar, A., Herdegen, S., Lindinger, G. L., Maris, M. T., Bak, M. A. R., Willems, D. L., Tan, H. L., Lauerer, M., Nagel, E., Hindricks, G., Dagres,

N., & Konopka, M. J. (2025). Healthcare professionals' perspectives on artificial intelligence in patient care: a systematic review of hindering and facilitating factors on different levels. In *BMC Health Services Research* (Vol. 25, Number 1). BioMed Central Ltd. https://doi.org/10.1186/s12913-025-12664-2

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. In *Geoscientific Model Development* (Vol. 15, Number 14, pp. 5481–5487). Copernicus GmbH. https://doi.org/10.5194/gmd-15-5481-2022

Hoyos, K., & Hoyos, W. (2024). Supporting Malaria Diagnosis Using Deep Learning and Data Augmentation. *Diagnostics*, *14*(7). https://doi.org/10.3390/diagnostics14070690

Huang, H.-H., & He, Q. (2016). *Nonlinear Regression Analysis*.

Huang, K., Altosaar, J., & Ranganath, R. (2020). *ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission*. http://arxiv.org/abs/1904.05342

Huang, Y. (2019). Prevalence of mental disorders in China – Author's reply. In *The Lancet Psychiatry* (Vol. 6, Number 6, p. 468). Elsevier Ltd. https://doi.org/10.1016/S2215-0366(19)30177-4

Hulsen, T. (2024). Artificial Intelligence in Healthcare: ChatGPT and Beyond. In *AI (Switzerland)* (Vol. 5, Number 2, pp. 550–554). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/ai5020028

Hussain-Alkhateeb, L., Ramírez, T. R., Kroeger, A., Gozzer, E., & Runge-Ranzinger, S. (2021). Early warning systems (EWSs) for Chikungunya, dengue, malaria, yellow fever, and Zika outbreaks: What is the evidence? A scoping review. In *PLoS Neglected Tropical Diseases* (Vol. 15, Number 9). Public Library of Science. https://doi.org/10.1371/journal.pntd.0009686

Hussain, S. S. A., Bedi, S., Yadav, C. P., Mohanty, A. K., Mahatme, K., Tyagi, S., Krishnan, N. M. A., Kota, S. H., & Sharma, A. (2025). Hybrid models combining trend and seasonality components with machine learning algorithms provide accurate forecasting of malaria incidence. *PLOS Global Public Health*, *5*(10). https://doi.org/10.1371/journal.pgph.0004500

Hussain, Z., Imran, •, Khan, A., Mudassar, •, & Arsalan, H. (2023). Machine Learning Approaches for Dengue Prediction: A review of Algorithms and Applications. In *Pakistan Geographical Review* (Vol. 78, Number 1). www.sciencedirect.com

Idris, M., Idris Isma'il, A., Abubakar, A. I., Ibrahim, M., Diginsa, M. U., & Gambo, J. (2021). *Forecasting the Rate of Malaria Spread in Nigeria Using Feedforward Neural Network Trained with Firefly Algorithm*.

Ilic, I., & Ilic, M. (2023). Global Patterns of Trends in Cholera Mortality. *Tropical Medicine and Infectious Disease*, *8*(3). https://doi.org/10.3390/tropicalmed8030169

Islam, M. R., Nahiduzzaman, M., Goni, M. O. F., Sayeed, A., Anower, M. S., Ahsan, M., & Haider, J. (2022). Explainable Transformer-Based Deep Learning Model for the Detection of Malaria Parasites from Blood Cell Images. *Sensors*, *22*(12). https://doi.org/10.3390/s22124358

Iwendi, C., Khan, S., Anajemba, J. H., Mittal, M., Alenezi, M., & Alazab, M. (2020). The use of ensemble models for multiple class and binary class classification for improving intrusion detection systems. *Sensors (Switzerland)*, *20*(9). https://doi.org/10.3390/s20092559

Iyyanki, M., Jayanthi, P., & Manickam, V. (2019). Machine learning for health data analytics: A few case studies of application of regression. In *Challenges and Applications for Implementing Machine Learning in Computer Vision* (pp. 241–266). IGI Global. https://doi.org/10.4018/978-1-7998-0182-5.ch010

Jameela, T., Athotha, K., Singh, N., Gunjan, V. K., & Kahali, S. (2022). Deep Learning and Transfer Learning for Malaria Detection. *Computational Intelligence and Neuroscience, 2022*. https://doi.org/10.1155/2022/2221728

Jdey, I., Hcini, G., & Ltifi, H. (2022). *Deep learning and machine learning for Malaria detection: overview, challenges and future directions*. http://arxiv.org/abs/2209.13292

Jeanray, N., Marée, R., Pruvot, B., Stern, O., Geurts, P., Wehenkel, L., & Muller, M. (2015). Phenotype classification of zebrafish embryos by supervised learning. *PLoS ONE*, *10*(1). https://doi.org/10.1371/journal.pone.0116989

Jiang, Z., Ma, J., Guo, Z., Feng, Q., & Yuan, H. (2026). Machine learning-based mortality prediction models for emergency department patients: a comparative analysis. *Frontiers in Medicine*, *13*. https://doi.org/10.3389/fmed.2026.1721101

Jones-Farmer, -Allison, Andel Professor of Business Analytics, V., Wiley, by, Stephens, B., Shmueli, G., Bruce, P. C., Stephens, M. L., & Patel, N. R. (2017). *Data Mining for Business Analytics: Concepts, Techniques, and Applications with JMP Pro®*. www.jmp.com/dataminingbook

Kadam, V. S. (2020). Regression Techniques in Machine Learning &Applications: A Review. *International Journal for Research in Applied Science and Engineering Technology*, *8*(10), 826–830. https://doi.org/10.22214/ijraset.2020.32019

Kalechofsky, H. (2016). *A Simple Framework for Building Predictive Models A Little Data Science Business Guide A Simple Framework for Building Predictive Models | 2*.

Kang, S. R. (2022). The Usefulness of the Artificial Intelligence Data in Analyzing the Skin in the Era of the Fourth Industrial Revolution. *Journal of Biosciences and Medicines*, *10*(07), 114–122. https://doi.org/10.4236/jbm.2022.107009

Kapwata, T., & Gebreslasie, M. T. (2016). Random forest variable selection in spatial malaria transmission modelling in Mpumalanga Province, South Africa. *Geospatial Health*, *11*(3), 251–262. https://doi.org/10.4081/gh.2016.434

Kaur, S., & Bawa, R. K. (2015). Future Trends of Data Mining in Predicting the Various Diseases in Medical Healthcare System. *International Journal of Energy, Information and Communications*, *6*(4), 17–34. https://doi.org/10.14257/ijeic.2015.6.4.02

Khan, O., Ajadi, J. O., & Hossain, M. P. (2024). Predicting malaria outbreak in The Gambia using machine learning techniques. *PLoS ONE*, *19*(5). https://doi.org/10.1371/journal.pone.0299386

khan, R., khan, S., Ullah, A., & khan, A. (2025). Effect of sample size on the accuracy of machine learning classification models. *Spectrum of Engineering Sciences*. https://doi.org/10.5281/zenodo.16266201

Khan, S., & Shaheen, M. (2023). *From Data Mining to Wisdom Mining 1*.

Kibret, S., Glenn Wilson, G., Ryder, D., Tekie, H., & Petros, B. (2019). Environmental and meteorological factors linked to malaria transmission around large dams at three ecological settings in Ethiopia. *Malaria Journal*, *18*(1). https://doi.org/10.1186/s12936-019-2689-y

Kiliç, A. E., & Karakoyun, M. (2023). *Breast Cancer Detection Using Machine Learning Algorithms*. http://as-proceeding.com/

Kuhn, M., & Johnson, K. (2020). *Feature Engineering and Selection; A Practical Approach for Predictive Models; Edition 1*. https://doi.org/10.4324/9781315108230

Kumar Jha, V., Shankar K, A., Das, B., Nair, R., Naik, S., & Gill, G. (2022). The Epidemiology of Intensive Care Unit Readmissions and Proposed Discharge

Protocol for a Tertiary Care Hospital. In *Journal of The Association of Physicians of India* ■ (Vol. 70).

Kumar Suggala, R. (2018). *A Survey on Prediction and Detection of Epidemic Diseases Outbreaks*.

Kumar, V. A., & Singh, J. (2023). Trends in Hydroponics Practice/Technology in Horticultural Crops: A Review. *International Journal of Plant & Soil Science*, *35*(2), 57–65. https://doi.org/10.9734/ijpss/2023/v35i22759

Leo, J., Luhanga, E., & Michael, K. (2019). Machine Learning Model for Imbalanced Cholera Dataset in Tanzania. *Scientific World Journal*, *2019*. https://doi.org/10.1155/2019/9397578

Lingala, M. A. L. (2017). Effect of meteorological variables on Plasmodium vivax and Plasmodium falciparum malaria in outbreak prone districts of Rajasthan, India. *Journal of Infection and Public Health*, *10*(6), 875–880. https://doi.org/10.1016/j.jiph.2017.02.007

Lior Rokach. (2010). Data Mining and Knowledge Discovery Handbook. In *Data Mining and Knowledge Discovery Handbook*. Springer US. https://doi.org/10.1007/978-0-387-09823-4

Liu, J., Kong, X., Xia, F., Bai, X., Wang, L., Qing, Q., & Lee, I. (2018). Artificial intelligence in the 21st century. *IEEE Access*, *6*, 34403–34421. https://doi.org/10.1109/ACCESS.2018.2819688

Liu, K., Li, Y., Hu, X., Lucu, M., & Widanage, W. D. (2020). Gaussian Process Regression with Automatic Relevance Determination Kernel for Calendar Aging Prediction of Lithium-Ion Batteries. *IEEE Transactions on Industrial Informatics*, *16*(6), 3767–3777. https://doi.org/10.1109/TII.2019.2941747

Maass, W., Agrawal, A., Ciani, A., Danz, S., Delgadillo, A., Ganser, P., Kienast, P., Kulig, M., König, V., Rodellas-Gràcia, N., Rughubar, R., Schröder, S., Stautner, M., Stein, H., Stollenwerk, T., Zeuch, D., & Wilhelm, F. K. (2024). QUASIM: Quantum Computing Enhanced Service Ecosystem for Simulation in Manufacturing. *KI - Kunstliche Intelligenz*, *38*(4), 361–370. https://doi.org/10.1007/s13218-024-00860-x

MacLeod, D. A., Jones, A., Di Giuseppe, F., Caminade, C., & Morse, A. P. (2015). Demonstration of successful malaria forecasts for Botswana using an operational seasonal climate model. *Environmental Research Letters*, *10*(4). https://doi.org/ 10.1088/1748-9326/10/4/044005

Ma, D. (2025). *MalariVis: A Bi-Modal Machine Learning System for Malaria Case Forecasting and Outbreak Prediction*. www.JSR.org/hs

Mahajan, P., Uddin, S., Hajati, F., & Moni, M. A. (2023). Ensemble Learning for Disease Prediction: A Review. In *Healthcare (Switzerland)* (Vol. 11, Number 12). MDPI. https://doi.org/10.3390/healthcare11121808

Majeed, M. A., Shafri, H. Z. M., Zulkafli, Z., & Wayayok, A. (2023). A Deep Learning Approach for Dengue Fever Prediction in Malaysia Using LSTM with Spatial Attention. *International Journal of Environmental Research and Public Health*, *20*(5). https://doi.org/10.3390/ijerph20054130

Mariki, M., Mkoba, E., & Mduma, N. (2022). Combining Clinical Symptoms and Patient Features for Malaria Diagnosis: Machine Learning Approach. *Applied Artificial Intelligence*, *36*(1). https://doi.org/10.1080/08839514.2022.2031826

Maryoosh, A. A., & Hussein, E. M. (2022). A Review: Data Mining Techniques and Its Applications. *International Journal of Computer Science and Mobile Applications*, *10*(3), 1–14. https://doi.org/10.47760/ijcsma.2022.v10i03.001

Maserat, E., Jafari, F., Mohammadzadeh, Z., Alizadeh, M., & Torkamannia, A. (2020). COVID-19 & an NGO and university developed interactive portal: a perspective from Iran. *Health and Technology*. https://doi.org/10.1007/s12553-020-00470-1/Published

Masiira, B., Nakiire, L., Kihembo, C., Katushabe, E., Natseri, N., Nabukenya, I., Komakech, I., Makumbi, I., Charles, O., Adatu, F., Nanyunja, M., Woldetsadik, S. F., Fall, I. S., Tusiime, P., Wondimagegnehu, A., & Nsubuga, P. (2019). Evaluation of integrated disease surveillance and response (IDSR) core and support functions after the revitalisation of IDSR in Uganda from 2012 to 2016. *BMC Public Health*, *19*(1). https://doi.org/10.1186/s12889-018-6336-2

Masinde, M. (2020). Africa's Malaria Epidemic Predictor: Application of Machine Learning on Malaria Incidence and Climate Data. *ACM International Conference Proceeding Series*, 29–37. https://doi.org/10.1145/3388142.3388158

Mategula, D., Gichuki, J., Barnes, K. I., Giorgi, E., & Terlouw, D. J. (2025). Advancing Early Warning Systems for Malaria: Progress, challenges, and future directions - A scoping review. *PLOS Global Public Health*, *5*(5 May). https://doi.org/10.1371/journal.pgph.0003751

Mathuria, J. P., Yadav, R., & Rajkumar. (2020). Laboratory diagnosis of SARS-CoV-2 - A review of current methods. In *Journal of Infection and Public Health* (Vol. 13, Number 7, pp. 901–905). Elsevier Ltd. https://doi.org/10.1016/j.jiph.2020.06.005

Maturana, C. R., de Oliveira, A. D., Nadal, S., Bilalli, B., Serrat, F. Z., Soley, M. E., Igual, E. S., Bosch, M., Lluch, A. V., Abelló, A., López-Codina, D., Suñé, T. P., Clols, E. S., & Joseph-Munné, J. (2022). Advances and challenges in automated malaria diagnosis using digital microscopy imaging with artificial intelligence tools: A review. In *Frontiers in Microbiology* (Vol. 13). Frontiers Media S.A. https://doi.org/10.3389/fmicb.2022.1006659

Maturana, C. R., de Oliveira, A. D., Nadal, S., Serrat, F. Z., Sulleiro, E., Ruiz, E., Bilalli, B., Veiga, A., Espasa, M., Abelló, A., Suñé, T. P., Segú, M., López-Codina, D., Clols, E. S., & Joseph-Munné, J. (2023).Imaging: a novel automated system for malaria diagnosis by using artificial intelligence tools and a universal low-cost robotized microscope. *Frontiers in Microbiology*, *14*. https://doi.org/10.3389/fmicb.2023.1240936

Mbunge, E., Millham, R. C., Sibiya, M. N., & Takavarasha, S. (2022). Application of machine learning models to predict malaria using malaria cases and environmental risk factors. *2022 Conference on Information Communications Technology and Society, ICTAS 2022 - Proceedings*. https://doi.org/10.1109/ICTAS53252.2022.9744657

Mbunge, E., Sibiya, M. N., Millham, R. C., & Takavarasha, S. (2021). Micro-spatial modelling of malaria cases and environmental risk factors in Buhera rural district, Zimbabwe. *2021 Conference on Information Communications Technology and Society, ICTAS 2021 - Proceedings*, 52–58. https://doi.org/10.1109/ICTAS50802.2021.9394987

Mcnamara, T. P., & Chen, & X. (1988). *Bayesian decision theory and navigation*. https://doi.org/10.3758/s13423-021-01988-9/Published

Mehbodniya, A., Khan, I. R., Chakraborty, S., Karthik, M., Mehta, K., Ali, L., & Nuagah, S. J. (2022). Data Mining in Employee Healthcare Detection Using Intelligence Techniques for Industry Development. In *Journal of Healthcare Engineering* (Vol. 2022). Hindawi Limited. https://doi.org/10.1155/2022/6462657

Merga, H., Degefa, T., Birhanu, Z., Tadele, A., Lee, M. C., Yan, G., & Yewhalaw, D. (2025). Urban malaria in sub-Saharan Africa: a scoping review of epidemiologic studies. In *Malaria Journal* (Vol. 24, Number 1). BioMed Central Ltd. https://doi.org/10.1186/s12936-025-05368-9

Merkord, C. L., Liu, Y., Mihretie, A., Gebrehiwot, T., Awoke, W., Bayabil, E., Henebry, G. M., Kassa, G. T., Lake, M., & Wimberly, M. C. (2017). Integrating malaria surveillance with climate data for outbreak detection and forecasting: The EPIDEMIA system. *Malaria Journal*, *16*(1). https://doi.org/10.1186/s12936-017-1735-x

Miggo, M., Harawa, G., Kangwerema, A., Knovicks, S., Mfune, C., Safari, J., Kaunda, J. T., Kalua, J., Sefu, G., Phiri, E., & Patel, P. (2023). Fight against cholera outbreak, efforts and challenges in Malawi. In *Health Science Reports* (Vol. 6, Number 10). John Wiley and Sons Inc. https://doi.org/10.1002/hsr2.1594

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2017). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, *19*(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

Mirzaeian, R., Nopour, R., Asghari Varzaneh, Z., Shafiee, M., Shanbehzadeh, M., & Kazemi-Arpanahi, H. (2023). Which are best for successful aging prediction? Bagging, boosting, or simple machine learning algorithms? *BioMedical Engineering Online*, *22*(1). https://doi.org/10.1186/s12938-023-01140-9

Mizna, S., Arora, S., Saluja, P., Das, G., & Alanesi, W. A. (2025). An analytic research and review of the literature on practice of artificial intelligence in healthcare. *European Journal of Medical Research*, *30*(1). https://doi.org/10.1186/s40001-025-02603-6

Modu, B., Polovina, N., Lan, Y., Konur, S., Taufiq Asyhari, A., & Peng, Y. (2017). Towards a predictive analytics-based intelligent malaria outbreakwarning system. *Applied Sciences (Switzerland)*, *7*(8). https://doi.org/10.3390/app7080836

Mohature, V. G., & Patil, M. (2021). A Review on Prediction and Analysis of Multiple Diseases in Healthcare using Data Mining. *International Research Journal of Engineering and Technology*. www.irjet.net

Mohsen, A. M., & Alhurdi, A. S. K. A. (2025). Using an Adaptive Linear Support Vector Machine Algorithm for Predicting the Breast Cancer. *Arab Journal of Management, Banking, and Financial Studies*, *1*(1), 90–103. https://doi.org/10.59559/ajmbfs.1.1.6

Molina-Franky, J., Cuy-Chaparro, L., Camargo, A., Reyes, C., Gómez, M., Salamanca, D. R., Patarroyo, M. A., & Patarroyo, M. E. (2020). Plasmodium falciparum pre-erythrocytic stage vaccine development. In *Malaria Journal* (Vol. 19, Number 1). BioMed Central Ltd. https://doi.org/10.1186/s12936-020-3141-z

Morovati, B., Lashgari, R., Hajihasani, M., & Shabani, H. (2023). Reduced Deep Convolutional Activation Features (R-DeCAF) in Histopathology Images to Improve the Classification Performance for Breast Cancer Diagnosis. *Journal of Digital Imaging*, *36*(6), 2602–2612. https://doi.org/10.1007/s10278-023-00887-w

Moyo, E., Mhango, M., Moyo, P., Dzinamarira, T., Chitungo, I., & Murewanhema, G. (2023). Emerging infectious disease outbreaks in Sub-Saharan Africa: Learning from the past and present to be better prepared for future outbreaks. In *Frontiers in Public Health* (Vol. 11). Frontiers Media S.A. https://doi.org/10.3389/fpubh.2023.1049986

Mujahid, M., Rustam, F., Shafique, R., Montero, E. C., Alvarado, E. S., de la Torre Diez, I., & Ashraf, I. (2024). Efficient deep learning-based approach for malaria detection using red blood cell smears. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-63831-0

Muriithi, D., Lumumba, V., & Okongo, M. (2024). A Machine Learning-Based Prediction of Malaria Occurrence in Kenya. *American Journal of Theoretical and Applied Statistics*, *13*(4), 65–72. https://doi.org/10.11648/j.ajtas.20241304.11

Musa, M. I. (2015). Malaria Disease Distribution in Sudan Using Time Series ARIMA Model. *International Journal of Public Health Science (IJPHS)*, *4*(1), 7–16.

Musa, M. O., Etuk E A, & Omankwu, O. C. B. (2024). Predictive Models for Malaria & TB Using ML: Health Decision Support in Africa. *Scientia Africana*, *23*(5). https://doi.org/10.4314/sa.v23i5.25

Nasser, F. K., & Behadili, S. F. (2022). A Review of Data Mining and Knowledge Discovery Approaches for Bioinformatics. *Iraqi Journal of Science*, 3169–3188. https://doi.org/10.24996/ijs.2022.63.7.37

Niyitegeka, J. (2021). *Employing Machine Learning and Internet of Things for Malaria Outbreak Prediction in Rwanda College of Science and Technology African Center of Excellence in Internet of Things Masters of Science in Internet of Things in Embedded Computing Systems*.

Nyambura, S. G., Kaibung'a, K., & Nyambura, A. N. (2025). Comparing Classification Models for Predicting Malaria: A Case Study of Malaria Incidence in Kenya. *Open Journal of Applied Sciences*, *15*(06), 1752–1765. https://doi.org/10.4236/ojapps.2025.156120

Olushola, A., & Mart, J. (2022). *Fraud Detection Using Machine Learning Techniques*. https://doi.org/10.13140/RG.2.2.33044.88961/1

Omari Firas. (2023). A combination of SEMMA & CRISP-DM models for effectively handling big data using formal concept analysis-based knowledge discovery: A data mining approach. *World Journal of Advanced Engineering Technology and Sciences, 8*(1), 009–014. https://doi.org/10.30574/wjaets.2023.8.1.0147

Onyijen, O. H., Olaitan, ✉, Olayinka, T. C., & Oyelola, S. (2023). Data-Driven Machine Learning Techniques for the Prediction of Cholera Outbreak in West Africa. In *International Journal of Applied and Natural Sciences Journal Homepage*. http://bluemarkpublishers.com/index.php/IJANS

Oyoo, J. O., Wekesa, J. S., & Ogada, K. O. (2024). Predicting Road Traffic Collisions Using a Two-Layer Ensemble Machine Learning Algorithm. *Applied System Innovation*, *7*(2). https://doi.org/10.3390/asi7020025

Panday, A., Kabir, M. A., & Chowdhury, N. K. (2022). A survey of machine learning techniques for detecting and diagnosing COVID-19 from imaging. *Quantitative Biology*, *10*(2), 188–207. https://doi.org/10.15302/J-QB-021-0274

Pang-Ning Tan, M. S. A. K. V. K. (2019). Introduction to Data Mining. *Pearson Education Limited*.

Parveen, R., Hussain Jalbani, A., Shaikh, M., Hussain Memon, K., Siraj, S., Nabi, M., & Lakho, S. (2017). Prediction of Malaria using Artificial Neural Network. In *IJCSNS International Journal of Computer Science and Network Security* (Vol. 17, Number 12).

Patient Safety Network. (2019). Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. *Patient Safety Nework*.

Perez-Saez, J., Zheng, Q., Kaminsky, J., Zou, K., Demby, M. N., Alam, C., Landau, D., DePencier, R., Langa, J. P. M., Chilengi, R., Welo Okitayemba, P., Bwire, G., Esso, L., Ngomba, A. V., Fouda Mbarga, N., Okunga, E. W., Yennan, S., Kapaya, F., Ohize, S. O., … Lee, E. C. (2025). Geographical shifting of cholera burden in Africa and its implications for disease control. *Nature Medicine*, *31*(10), 3380–3387. https://doi.org/10.1038/s41591-025-03847-9

Perkins, W. A., & Hakim, G. J. (2016). *Reconstructing past climate by using proxy data and a linear climate model*. https://doi.org/10.5194/cp-2016-129

Petr Silhavy. (2023). Artificial Intelligence Application Application Networks and Systems. *Tomas Bata University in Zlin*.

Plotnikova, V., Dumas, M., & Milani, F. (2020). Adaptations of data mining methodologies: A systematic literature review. *PeerJ Computer Science*, *6*, 1–43. https://doi.org/10.7717/PEERJ-CS.267

Poostchi, M., Silamut, K., Maude, R. J., Jaeger, S., & Thoma, G. (2018). Image analysis and machine learning for detecting malaria. In *Translational Research* (Vol. 194, pp. 36–55). Mosby Inc. https://doi.org/10.1016/j.trsl.2017.12.004

Radek Silhavy. (2023). Artificial Intellegence Application in Networks and Systems. *Computer Science On-Line Conference 2023*.

Rajab, S., Rose, N., & Marvin, G. (2024). Interpretable Ensemble Model-Agonistc Approaches for Malaria Prediction. *ACM International Conference Proceeding Series*, 451–459. https://doi.org/10.1145/3675888.3676092

Rajaraman, S., Antani, S. K., Poostchi, M., Silamut, K., Hossain, M. A., Maude, R. J., Jaeger, S., & Thoma, G. R. (2018). Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ*, *2018*(4). https://doi.org/10.7717/peerj.4568

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. *New England Journal of Medicine*, *380*(14), 1347–1358. https://doi.org/10.1056/nejmra1814259

Ramageri, B. M. (2010). Data Mining Techniques and Applications. In *Indian Journal of Computer Science and Engineering* (Vol. 1).

Rashmi Ashtagi, V. R. (2024). *Cervical Cancer Prediction Using Machine Learning*.

Rezaul, K. M., Jewel, M., Sudhan, A., Khan, M. U., Roshika, M., Fernando, S., Noor, K., Siddiquee, A., Jannat, T., Rahman, M. A., & Islam, S. (2025). A Comparative Study of Predictive Analysis Using Machine Learning Techniques: Performance

Evaluation of Manual and AutoML Algorithms. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 16, Number 1). www.ijacsa.thesai.org

Rono, L. (2018). Microcredit and its relationship to the growth of small and medium enterprises in Konoin subcounty, Kenya. *International Journal of Advanced Research*, *6*(4), 961–968. https://doi.org/10.21474/IJAR01/6935

Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. In *Babylonian Journal of Machine Learning* (Vol. 2024, pp. 69–79). Mesopotamian Academic Press. https://doi.org/10.58496/BJML/2024/007

Saqr, M., & López-Pernas, S. (2024). Learning Analytics Methods and Tutorials: A Practical Guide Using R. In *Learning Analytics Methods and Tutorials: A Practical Guide Using R*. Springer Nature. https://doi.org/10.1007/978-3-031-54464-4

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. In *SN Computer Science* (Vol. 2, Number 3). Springer. https://doi.org/10.1007/s42979-021-00592-x

Schaffer, M. E., Ahrens, A., & Hansen, C. B. (2023). *pystacked: Stacking generalization and machine learning in Stata*. https://statalasso.github.io/

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Shahnazari, K., & Ayyoubzadeh, S. M. (2024). *A Novel Nearest Neighbors Algorithm Based on Power Muirhead Mean*. http://arxiv.org/abs/2209.01514

Shankar, H., Kumar, G., Ahmed, N., & Florentin, A. (2026). Plasmodium malariae is an overlooked malaria parasite with emerging challenges. *Communications Medicine*. https://doi.org/10.1038/s43856-025-01360-1

Sharma, A., Vans, E., Shigemizu, D., Boroevich, K. A., & Tsunoda, T. (2019). DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-47765-6

Sheth, V., Tripathi, U., & Sharma, A. (2022). A Comparative Analysis of Machine Learning Algorithms for Classification Purpose. *Procedia Computer Science*, *215*, 422–431. https://doi.org/10.1016/j.procs.2022.12.044

Shi, B., Bai, X., & Yao, C. (2015). *An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition*. http://arxiv.org/abs/1507.05717

Siłka, W., Wieczorek, M., Siłka, J., & Woźniak, M. (2023). Malaria Detection Using Advanced Deep Learning Architecture. *Sensors*, *23*(3). https://doi.org/10.3390/s23031501

Sim, J. Z. T., Fong, Q. W., Huang, W., & Tan, C. H. (2023). Machine learning in medicine: what clinicians should know. In *Singapore Medical Journal* (Vol. 64, Number 2, pp. 91–97). Lippincott Williams and Wilkins. https://doi.org/10.11622/smedj.2021054

Sohil, F., Sohali, M. U., & Shabbir, J. (2022). An introduction to statistical learning with applications in R. *Statistical Theory and Related Fields*, *6*(1), 87–87. https://doi.org/10.1080/24754269.2021.1980261

Song, X., Deng, L., Wang, H., Zhang, Y., He, Y., & Cao, W. (2025). Deep learning-based time series forecasting. *Artificial Intelligence Review*, *58*(1). https://doi.org/10.1007/s10462-024-10989-8

Sriporn, K., Tsai, C. F., Tsai, C. E., & Wang, P. (2020). Analyzing malaria disease using effective deep learning approach. *Diagnostics*, *10*(10). https://doi.org/10.3390/diagnostics10100744

Stuart J. Russell and Peter Norvig. (2022). *Artificial Intelligence: A Modern Approach*.

Sundaram, Suresh. (2018). *Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI 2018) : 18-21 November 2018, Bengaluru*. IEEE.

Sun, X., Xing, Z., Wan, Z., Ding, W., Wang, L., Zhong, L., Zhou, X., Gong, X. J., Li, Y., & Zhang, X. D. (2024). A robust ensemble deep learning framework for accurate diagnoses of tuberculosis from chest radiographs. *Frontiers in Medicine*, *11*. https://doi.org/10.3389/fmed.2024.1391184

Taffese, H. S., Hemming-Schroeder, E., Koepfli, C., Tesfaye, G., Lee, M. C., Kazura, J., Yan, G. Y., & Zhou, G. F. (2018). Malaria epidemiology and interventions in Ethiopia from 2001 to 2016. In *Infectious Diseases of Poverty* (Vol. 7, Number 1). BioMed Central Ltd. https://doi.org/10.1186/s40249-018-0487-3

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

Tharageswari, K., Mohana Sundaram, N., & Santhosh, R. (2025). A Hybrid Deep Learning Algorithm Based Prediction Model for Sustainable Healthcare System. *Journal of Automation, Mobile Robotics and Intelligent Systems*, *19*(2), 89–98. https://doi.org/10.14313/jamris-2025-019

Tohidinik, H., Keshavarz, H., Mohebali, M., Sanjar, M., & Hassanpour, G. (2021). Prediction of malaria cases in the southeastern Iran using climatic variables: An 18-year SARIMA time series analysis. *Asian Pacific Journal of Tropical Medicine*, *14*(10), 463–470. https://doi.org/10.4103/1995-7645.329008

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. In *Nature Medicine* (Vol. 25, Number 1, pp. 44–56). Nature Publishing Group. https://doi.org/10.1038/s41591-018-0300-7

Traini, E., & Lombardi, F. (2022). *Hybrid modeling to support the smart manufacturing concepts, theoretic contributions and real-case applications about Hybrid and Wisdom-based Systems*.

Tusting, L. S., Bisanzio, D., Alabaster, G., Cameron, E., Cibulskis, R., Davies, M., Flaxman, S., Gibson, H. S., Knudsen, J., Mbogo, C., Okumu, F. O., von Seidlein, L., Weiss, D. J., Lindsay, S. W., Gething, P. W., & Bhatt, S. (2019). Mapping changes in housing in sub-Saharan Africa from 2000 to 2015. *Nature*, *568*(7752), 391–394. https://doi.org/10.1038/s41586-019-1050-5

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1). https://doi.org/10.1186/s12911-019-1004-8

Uzun Ozsahin, D., Duwa, B. B., Ozsahin, I., & Uzun, B. (2024). Quantitative Forecasting of Malaria Parasite Using Machine Learning Models: MLR, ANN, ANFIS and Random Forest. *Diagnostics*, *14*(4). https://doi.org/10.3390/diagnostics14040385

Vapnik, V., & Izmailov, R. (2019). Rethinking statistical learning theory: learning using statistical invariants. *Machine Learning*, *108*(3), 381–423. https://doi.org/10.1007/s10994-018-5742-0

Wang, L., Lin, Z. Q., & Wong, A. (2020). COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-76550-z

World Health Organization. (2015). *Global technical strategy for malaria, 2016-2030*. Global Malaria Programme, World Health Organization.

World Health Organization. (2022). *World malaria report 2022*. https://www.who.int/teams/global-malaria-programme

World Health Organization. (2023). Measles Cases Surge Worldwide, Infecting 10.3 million People In 2023. *World Health Organization*.

Xu, J., Xi, X., Chen, J., Sheng, V. S., Ma, J., & Cui, Z. (2022). A Survey of Deep Learning for Electronic Health Records. In *Applied Sciences (Switzerland)* (Vol. 12, Number 22). MDPI. https://doi.org/10.3390/app122211709

Yang, F., Poostchi, M., Yu, H., Zhou, Z., Silamut, K., Yu, J., Maude, R. J., Jaeger, S., & Antani, S. (2020). Deep Learning for Smartphone-Based Malaria Parasite Detection in Thick Blood Smears. *IEEE Journal of Biomedical and Health Informatics*, *24*(5), 1427–1438. https://doi.org/10.1109/JBHI.2019.2939121

Yang, F., Yu, H., Silamut, K., Maude, R. J., Jaeger, S., & Antani, S. (2019). *Smartphone-Supported Malaria Diagnosis Based on Deep Learning*.

Yu, P., Hou, Y., Song, Y., Pang, J., & Liu, D. (2018). Lithium-Ion Battery Prognostics with Hybrid Gaussian Process Function Regression. *Energies*, *11*(6). https://doi.org/10.3390/en11061420

Zhang, Y., & Yang, Q. (2021). A Survey on Multi-Task Learning. *IEEE Transactions on IEEE Transactions on Knowledge and Data Engineering*. http://arxiv.org/abs/1707.08114

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G. F., & Tan, W. (2020).

A Novel Coronavirus from Patients with Pneumonia in China, 2019. *New England Journal of Medicine*, *382*(8), 727–733. https://doi.org/10.1056/nejmoa2001017

APPENDICES

Appendix I: Research Publications

Leopord, H., Cheruiyot, W. K., Kimani, S., & Nyararai, M. (2017). A hybrid based classification and regression model for predicting diseases outbreak in datasets. International Journal of Computer (IJC), 27(1), 69-83.

Leopord, H., Cheruiyot, W. K., & Kimani, S. (2016). A survey and analysis on classification and regression data mining techniques for diseases outbreak prediction in datasets. Int. J. Eng. Sci, 5(9), 1-11.