

A RANDOM FOREST MODEL TO PREDICT MALARIA OUTBREAKS
A CASE STUDY OF KISII COUNTY

BY
JOYLINE MOGITI NYABUTO

MASTER OF SCIENCE IN DATA SCIENCE
KCA UNIVERSITY

2025

A RANDOM FOREST MODEL TO PREDICT MALARIA OUTBREAKS

A CASE STUDY OF KISII COUNTY

BY

JOYLINE MOGITI NYABUTO

**A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE IN DATA SCIENCE
IN THE SCHOOL OF TECHNOLOGY AT KCA UNIVERSITY**

JANUARY, 2025

DECLARATION

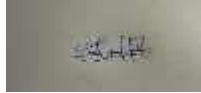
I declare that this is my original work and has not been previously published or submitted elsewhere for an award of a degree. I also declare that it contains no material written or published by other people except where due reference is made and the author is duly acknowledged.

This research paper is wholly original with no submissions for awards to other universities

Student Name: Joyline Mogiti Nyabuto

Registration Number: 23/06874

Signature:



Date: 19/11/2025

This dissertation has been submitted for consideration with my approval as a KCA University Supervisor.

Dr Stephen Njenga



Date: 19/11/2025

ACKNOWLEDGEMENT

First and foremost, I would like to thank Almighty God for giving me the strength, good health, will and for sustaining me throughout my study. I extend my appreciation and gratitude to all those that contributed immensely towards completion of this research project.

I especially would like to thank my family and especially my parents for believing in me and their timeless encouragement, support and enthusiasm. I also want to thank my supervisor Dr Stephen Njenga who guided encouraged and led me through the journey to completing my thesis.

I especially would like to thank my friends who sacrificed their time and gave me invaluable support that saw me through this period.

A RANDOM FOREST MODEL TO PREDICT MALARIA OUTBREAKS

A CASE STUDY OF KISII COUNTY

ABSTRACT

In Sub-Saharan Africa, malaria has remained as a major source of morbidity and mortality, and a constant menace to the population health and socio-economic balance. Seasonal epidemics in highlands areas like Kisii County of Kenya create a huge burden to the healthcare systems, interfere with livelihood, and hamper development in the area. Despite the major advances achieved in reducing malaria rates in the country, the fact that it still is spread in highlands indicates the weakness of the existing control and monitoring measures. Conventional forecasting methods such as logic regression and ARIMA models have been actively used in predicting trends in malaria but their assumptions of linearity, normality, and stationarity makes them less useful to the study of the non-linear and dynamic relationships that exist between environmental, epidemiological, and socio-economic variables that define malaria transmission in highland micro-ecologies.

Machine learning, especially ensemble-based ones, is a promising new direction in the development of malaria prediction. Ensemble models can learn more of complex interactions using heterogeneous data, hence able to adjust to changing epidemiological situations and produce more precise and useful information. This paper thus builds and compares an ensemble learning model to forecast malaria outbreaks in Kisii County that incorporates Random Forest (RF), XGBoost and RF-XGBoost hybrid model. The framework uses ecological, climatic, socio-economic, and epidemiological predictors to show that highland malaria occurs in a multifactorial manner. To resolve the important data issues, methodological advances like the Synthetic Minority Over-Sampling Technique (SMOTE), adaptive class weight, and cost-sensitive learning are used to curb the issue of class imbalance, and temporal feature weighting and sliding-window retraining are used in order to overcome concept drift due to climatic and intervention variability.

Besides the accuracy of predictions, the interpretability and operational relevance are a priority of the study because of the usage of explainable AI tools like analysis of feature importance, SHAP values, and LIME visualization. To assess model performance, we measure such metrics as accuracy, F1-score, ROC-AUC, and precision-recall and complemented by ablation studies to evaluate the strength of and contribution of a specific methodological improvement.

The hybrid ensemble approach will be based on the ability of the generalization of the Random Forest with the bias-correcting power of the XGBoost to create a transparent/scalable/adaptive ecologically-specific system in the highland environments. Besides technical improvement, the research advances the principles of responsible and ethical AI by focusing on interpretability, fairness, and data governance. The expected deliverable is the locally-calibrated predictive system, which can be used to support timely, evidence-based public health decision-making in the Kisii County, and can be used to offer a transferable methodological framework of malaria surveillance in ecologically similar highland areas across Sub-Saharan Africa.

ACRONYMS AND ABBREVIATIONS

ADASYN Adaptive Synthetic Sampling

AI Artificial Intelligence

ITNs Insecticide-treated nets

ML Machine Learning

RF Random Forest

SDH Social Determinants of health

SMOTE Synthetic minority oversampling technique

SSA Sub-Saharan Africa

SVM Support Vector Machine

WHO World Health Organization

Contents

DECLARATION	3
ACKNOWLEDGEMENT	4
A RANDOM FOREST MODEL TO PREDICT MALARIA OUTBREAKS	5
A CASE STUDY OF KISII COUNTY	5
ABSTRACT.....	5
ACRONYMS AND ABBREVIATIONS	6
CHAPTER ONE	10
1.1. BACKGROUND OF THE STUDY	10
1.2 Problem Statement	14
1.3 Research Objectives	17
1.3.1 Main Objective.....	17
1.3.2 Specific objectives	17
1.4 Research questions.....	17
1.5 Justification of the study	18
1.6 Motivation of the study	22
CHAPTER TWO	24
2.1 Literature Review.....	24
2.2.2 Social Determinants of Health	25
2.2.3 Ecosocial Theory of Disease Distribution	25
2.2.4 Ensemble Learning Theory	26
2.2.5 Computational Learning Theory	27
2.2.6 Cost-Sensitive Learning Framework (CSL)	30
2.2.7 Linking the Theories to Objectives and Empirical	31
2.3 Empirical Review.....	34
2.4 Previous Models that have been used in predicting malaria outbreaks	38
2.5 Data Description	42
2.6 Data Analysis	45
2.6.1 Data Analysis Framework.....	47
FIGURE 1	48
Data analysis framework.....	48
2.7 Data Processing and Feature-Engineering	48
2.8 Model Training and Evaluation	50

2.9 Addressing challenges.....	50
2.9.1 Feature Drift.....	50
2.9.2 Data Imbalance	51
Data Processing Tools.....	53
FIGURE 2	55
Conceptual framework.....	55
Conceptual Framework	55
CHAPTER THREE.....	56
RESEARCH METHODOLOGY	56
3.1 Introduction.....	56
3.2 Research Design.....	56
3.2.1 Research Approach	59
3.3 Ethical Considerations	60
3.3.1 Nature of Data Collected and Level of Risk	60
3.3.2 Dataset Structure and Re-Identification Risk.....	61
3.3.3 Consent and Original Data Collection	61
3.3.4 Biases and Limitations	62
3.3.5 Potential for Harm and Respect for Data Subjects	64
3.3.6 Legal and Regulatory Compliance.....	64
CHAPTER 4	65
FINDINGS AND DISCUSSIONS.....	65
4.1 Descriptive Statistics.....	65
FIGURE 3	66
Total lagged malaria cases by sub-county	66
4.2 Models Evaluation	66
4.2.1 Model performance metrics	68
TABLE 1.....	68
Model performance metrics	68
4.2.2 Confusion Matrix	69
FIGURE 5	71
Confusion matrix.....	71
4.2.3 Receiver Operating characteristic and Area Under Curve.....	71
FIGURE 5	73

Roc curves for the models.....	73
4.2.4 Class distribution before and after SMOTE.....	73
FIGURE 5	75
Class distribution before smote	75
4.2.5 Correlation matrix between the Variables	75
4.2.6 Features' Relative Importance Plot.....	77
CHAPTER 5	80
CONCLUSION AND RECOMMENDATION.....	80
5.1 Comparison with Previous studies.....	80
5.2 Practical Implication	81
5.3 Limitation of the study	83
5.4 Recommendation for Future Studies.....	84
5.5 Conclusion	87
References:.....	89
APPENDIX.....	92
RESEARCH TIMELINE AND BUDGET	92
RESEARCH TIMELINE	92
Budget	93
GANNT CHART	93

CHAPTER ONE

1.1. BACKGROUND OF THE STUDY

Malaria remains one of the most persistent and complex public health challenges globally, despite decades of concerted international control and elimination efforts. According to the World Health Organization (WHO), an estimated 249 million malaria cases and over 600,000 deaths were reported worldwide in 2022, with sub-Saharan Africa accounting for more than 90 percent of the global burden (WHO, 2023). Beyond its immediate health consequences, malaria exerts profound socio-economic impacts, including lost productivity, increased household health expenditures, school absenteeism, and sustained pressure on already fragile health systems. These multidimensional effects make malaria not only a biomedical concern but also a developmental challenge that disproportionately affects low- and middle-income countries.

In Kenya, malaria has historically been one of the leading causes of morbidity and mortality, particularly among children under five years of age and pregnant women. Over the past two decades, the country has achieved notable progress in reducing malaria prevalence through the scale-up of vector control interventions such as insecticide-treated nets (ITNs), indoor residual spraying (IRS), improved diagnostic capacity, and expanded access to effective antimalarial therapies (Snow and Noor, 2015). These interventions have contributed to substantial declines in transmission in several regions, especially in traditionally high-transmission lowland and coastal zones. However, these gains have been uneven, fragile, and increasingly threatened by emerging challenges, including insecticide resistance, drug resistance, climate variability, population mobility, and declining donor funding for malaria control programs (Noor et al., 2022; WHO, 2023).

Kenya's national malaria elimination target, aligned with global and regional malaria strategies, aims to eliminate malaria as a public health problem by 2030. While this goal reflects strong political commitment, current epidemiological trends suggest that it is unlikely to be achieved without substantial innovation in surveillance, prediction, and response strategies. Transmission persists in ecologically complex settings such as the lake-endemic regions and the western highlands, where malaria dynamics differ fundamentally from those observed in stable lowland areas. These regions experience unstable and epidemic-prone transmission, characterized by pronounced inter-annual variability and episodic outbreaks that are difficult to anticipate using conventional surveillance tools.

The western Kenyan highlands represent one of the most challenging malaria ecologies in the country. Unlike lowland endemic zones, where transmission is relatively stable and predictable, highland areas are marked by sharp gradients in altitude, temperature, rainfall, and land use over short spatial scales. Small changes in climatic conditions can therefore trigger substantial changes in vector abundance and parasite transmission. In such settings, malaria transmission is often seasonal and epidemic in nature, with outbreaks occurring when favorable environmental conditions align with population susceptibility and gaps in immunity. These features make highland malaria particularly sensitive to climate variability and long-term climate change.

Kisii County, located in the western highlands of Kenya, exemplifies these challenges. The county is characterized by undulating topography, elevations ranging from approximately 1,300 to over 2,000 meters above sea level, and highly variable rainfall patterns influenced by orographic effects. These environmental features create localized micro-climates that can support mosquito

breeding and malaria transmission under specific conditions, even though the region is generally classified as a low-transmission zone. As a result, Kisii County experiences recurrent seasonal malaria epidemics that place substantial strain on health facilities, contribute to increased pediatric and maternal morbidity, and disrupt household and community livelihoods (Gikonyo et al., 2019).

The socio-economic context of Kisii County further compounds malaria vulnerability. Many households rely on subsistence agriculture and informal economic activities that are highly sensitive to illness-related productivity losses. Housing quality varies widely, with a significant proportion of homes lacking structural features that reduce mosquito entry, such as screened windows or closed eaves. Access to health services is uneven, particularly in rural and peri-urban areas, and delays in seeking care can exacerbate disease severity and transmission. These socio-economic determinants interact with ecological drivers to shape malaria risk in ways that are not adequately captured by aggregated national-level surveillance systems.

Effective malaria control and elimination in epidemic-prone highland settings depend critically on timely and accurate early warning systems that can anticipate outbreaks and support proactive responses. Prediction models are intended to serve this purpose by integrating epidemiological, environmental, and socio-economic data to forecast malaria risk. Historically, malaria prediction in Kenya and similar settings has relied on statistical approaches such as logistic regression, time-series models, and autoregressive integrated moving average (ARIMA) techniques. While these methods have contributed valuable insights, they are limited by strong assumptions of linearity and stationarity that are often violated in complex and changing transmission environments.

One of the key limitations of conventional malaria prediction models is their inability to adequately incorporate local ecological drivers that are particularly important in highland settings. Factors such as topographic rainfall gradients, slope-induced water accumulation, altitude-related temperature variability, and heterogeneity in housing quality strongly influence mosquito breeding and human exposure. However, these variables are frequently omitted or poorly represented in existing models, leading to oversimplified representations of transmission dynamics (Abeku et al., 2003; Githeko and Ndegwa, 2001). As a result, model predictions may fail to reflect localized outbreak risk, limiting their utility for county-level decision-making.

Another critical challenge is concept drift, a phenomenon in which the statistical relationships between predictors and outcomes change over time due to external influences such as climate change, land-use transformation, and the scaling up or withdrawal of control interventions. In malaria prediction, concept drift can erode model performance as historical patterns become less representative of current or future conditions (Zliobaite, 2017; Rabanser et al., 2019). Many existing models are trained on historical data and applied without mechanisms for adaptation, rendering them increasingly inaccurate as transmission dynamics evolve.

Class imbalance represents an additional and often overlooked problem in malaria outbreak prediction. In epidemic-prone settings, non-outbreak periods vastly outnumber outbreak events, resulting in highly imbalanced datasets. Models trained on such data may achieve high overall accuracy by correctly predicting non-outbreak periods while failing to identify the relatively rare but critically important outbreak events. This bias undermines the primary purpose of early warning systems, which is to detect impending outbreaks in time for preventive action (Haixiang

et al., 2017). Despite its importance, class imbalance is rarely addressed explicitly in malaria prediction studies.

Recent advances in machine learning offer promising opportunities to overcome many of these limitations. Ensemble learning methods such as Random Forest and Extreme Gradient Boosting (XGBoost) have demonstrated strong predictive performance in a range of malaria-endemic settings by capturing nonlinear relationships and complex interactions among predictors (Wang et al., 2019; Touray et al., 2024). These models are particularly well-suited to integrating heterogeneous data sources, including climatic, environmental, and socio-economic variables. However, their application to highland malaria prediction remains limited, and when used, these models are often treated as opaque “black boxes” that provide little insight into the drivers of predicted risk.

The lack of interpretability in many machine learning-based malaria prediction models poses a significant barrier to their adoption in public health practice. Policymakers and frontline health workers require not only accurate predictions but also clear explanations of why risk is increasing or decreasing.

1.2 Problem Statement

Despite Kenya recording major improvements in the malaria prevalence rates over the past two decades, the national goal of eliminating the disease by 2030 is unlikely due to the declining donor support, growing insecticides resistance, and persisting malaria in the highland and the lake-infested regions (Snow and Noor, 2015; WHO, 2023). Kisii County can be viewed as an example of such an issue since seasonal epidemics continue to pose significant load to healthcare

infrastructure, result in a wave of pediatric and maternal morbidity, and influence household and community livelihood (Gikonyo et al., 2019).

In the mathematical perspective, the existing methods of predicting malaria have low performance in terms of the use of the highland micro-ecologies because of a variety of reasons. The local ecological drivers that encompass the topographical rainfall gradients and variability of the quality of homes is not usually incorporated in the existing models (Abeku et al., 2003; Githeko and Ndegwa, 2001). Concept drift is also a problem with them, as the statistical properties of predictors change over time due to the fluctuating climate or growing intervention through control (Zliobaite, 2017; Rabanser et al., 2019). Additionally, the majority of the models fail to address the problem of the imbalance of classes that biases forecast on periods without an outbreak and results in the incapacity to identify early warning (Haixiang et al., 2017). Though it is demonstrated that the application of ensemble models such as the Random Forest and XGBoost have stronger predictive components in other malaria-situated areas (Wang et al., 2019; Touray et al., 2024), the use of the two models in predicting highland malaria is limited, and most instances can be regarded as an opaque black box that restricts the interpretability of the models to the stakeholders of the public health (Doshi-Velez and Kim, 2017).

The lack of interpretability in many machine learning-based malaria prediction models poses a significant barrier to their adoption in public health practice. Policymakers and frontline health workers require not only accurate predictions but also clear explanations of why risk is increasing or decreasing in specific locations. Without transparency, trust in model outputs may be low, and opportunities for targeted intervention may be missed (Doshi-Velez and Kim, 2017).

This challenge is particularly salient in resource-constrained settings, where decision-support tools must align with existing capacities and decision-making processes.

Random Forest models offer a potentially advantageous balance between predictive performance and interpretability. By constructing ensembles of decision trees using randomly selected subsets of predictors, Random Forest can model nonlinear relationships while providing measures of variable importance that help identify key drivers of risk (Breiman, 2001; Fernandez-Delgado et al., 2014). In addition, Random Forest models can be combined with resampling and rebalancing techniques to address class imbalance and can be updated periodically to mitigate the effects of concept drift. These features make Random Forest particularly suitable for application in highland malaria settings and in decentralized health systems such as Kenya's county-based governance structure.

Beyond methodological considerations, malaria prediction in highland settings raises important ethical and equity concerns that align with broader debates on algorithmic fairness and transparency in health artificial intelligence. Inaccurate or biased models can exacerbate existing disparities by failing to detect outbreaks in vulnerable populations or by triggering false alarms that divert scarce resources away from areas of greatest need (Rajkomar et al., 2018; Obermeyer et al., 2019). In sub-Saharan Africa, where health systems are already overstretched, such errors carry significant moral and practical consequences. There is therefore a compelling need for predictive models that are not only accurate but also explainable, equitable.

The existing models cannot thus describe the ecological complexity of highland malaria as they have a tendency to generalize the workings of transmission across wide regions without regard to localized environmental variability. These models cannot capture the heterogeneity of

breeding habitats and patterns of human exposure since the finer variables of these variables, including slope, altitude-induced variations in rainfall, and the density of housing are omitted. This means that predictions made by such systems can hardly match ground-based realities which limits their role in supporting the local decision making and resource allocation by county health departments.

1.3 Research Objectives

1.3.1 Main Objective

The main objective of the research is to develop a random forest model that can accurately predict malaria outbreaks in Kisii County, using historical malaria data.

1.3.2 Specific objectives

1. To analyze ecological, socio-economic, and epidemiological determinants of malaria outbreaks in Kisii County.
2. To develop a Random Forest-based outbreak prediction model incorporating adaptive techniques for feature drift and data imbalance correction.
3. To evaluate the performance of the Random Forest model in predicting malaria outbreaks by comparing its accuracy, precision, and recall against other commonly predictive models employed in previous studies

1.4 Research questions

1. What ecological, socio-economic, and epidemiological factors significantly influence malaria outbreaks in Kisii County?
2. How can a Random Forest based outbreak prediction model be designed and implemented to incorporate adaptive techniques that address feature drift and data imbalance in Kisii County malaria datasets?

3. How does the predictive performance and interpretability of the Random Forest model compare with a hybrid Random Forest -XGBoost?

1.5 Justification of the study

The dynamics of malaria transmission are highly geographically and temporally heterogeneous especially in highland ecology like Kisii County in western Kenya. Highland regions unlike lowland endemic regions, have relatively stable transmission of malaria, but they have intermittent, unstable, and climatically sensitive malaria transmission, which is largely marked by sudden outbreak after environmental or climatic disturbance. They may lead to proportionately high morbidity and mortality, since the population immunity of such places is usually low. Therefore, effective malaria prevention and control in a highland area depends on timely and accurate outbreak prediction.

Although this is necessary, the traditional malaria surveillance systems continue to be mostly reactive as opposed to being predictive. The most common methods of routine surveillance are passive case surveillance and retrospective surveillance that are usually characterized by reporting latencies, spatial aggregation and low ability to detect local outbreak indicators. Such systems are inadequate in ecologically heterogeneous environments, such as the Kisii County, whereby the alteration of altitude, rainfall, temperature, land use, and socio-economic factors vary within a short distance, to give actionable early warnings. This generates a critical lag between the emergence of malaria risk and any specific intervention, e.g. indoor residual spraying, provision of insecticide-impregnated nets, or community-wide response efforts.

Predictive modeling is thus a new and required frontier in the malaria management process and it holds the potential to change the malaria management approaches to proactive rather than

reactive. The machine learning approaches are especially suited to this task since the way malaria is spread is dictated by the complex, non-linear interactions of the climatic, ecological, demographic, and health system factors. Conventional statistical models can tend to be linear and assume pre-determined relationships and therefore cannot be able to capture these complicated dynamics more so when there are many high-dimensional and context-specific interactions.

In machine learning paradigm, ensemble learning algorithms, especially the Random Forest and the Extreme Gradient Boosting (XGBoost), have been shown to have better predictive powers in the trickier epidemiological and environmental model. Ensemble Learning Theory assumes that a collectivity of weak or moderate accuracy learners can result in a stronger and more precise prediction model than either of the single models. Random Forest is a model which is built upon the concept of bagging (bootstrap aggregation) to minimize the variance by averaging the predictions of several decision trees which are built on various subsets of data. This enables it to be especially resistant to noise and overfitting, which are typical issues with epidemiological data when the sample size is small.

On the other hand, methods like XGBoost, which are boosting based, are more about successively fixing the errors of the model hence less bias, the better the model is able to find subtle patterns in the data. Boosting algorithms are more sensitive to early indicators of an outbreaks by giving more importance to challenging to predict cases which are otherwise ignored. The reciprocal advantages of bagging and boosting give better theoretical basis of their use in malaria prediction, especially where the predictors are non-linear, multi-collinear and multiplexed and the interactions between predictors are complicated.

Nevertheless, the implementation of ensemble machine learning algorithms has not been fully exploited yet in data sparse, ecologically diverse, and imbalanced highland environments, like Kisii County. The majority of available malaria prediction research is based on a national or regional level, lowland disease-infested regions, or data-intensive settings. These methods fail to consider the peculiar ecological backgrounds, localized transmission vehicles, and volatile dynamics that characterize highland malaria. Further, there is inherent high class imbalance in outbreak prediction, since outbreaks are rare compared to when there is no outbreak, resulting in biased models that give increased preference to majority classes and fail to detect critical but low frequency events.

This paper aims to fill these two methodological and contextual gaps by creating a predictive modeling framework based on Random Forests that is explicitly designed to predict highland malaria. The offered solution implies some innovative methodological solutions, which aim at maximizing predicting results within the realistic surveillance conditions. Temporal weighting is also provided to assign importance on the recent observations, because malaria transmission is dynamic in nature and, as such, the model must be sensitive to changes in climatic and epidemiological conditions. The adaptive rebalancing methods are used to reduce class imbalance which enhances sensitivity of the model to periods of outbreaks without jeopardizing the general predictive stability.

Moreover, the investigation also incorporates cost-sensitive learning, which takes into consideration that the consequences of false negative (inability to predict an outbreak) are much more significant in terms of the public health. This is in accordance with the modeling framework to the real-world decision-making priorities where early warnings though incurred at the cost of

false alarms are much better than missed outbreaks. These methodological decisions are based on the Computational Learning Theory, which focuses on sustaining the performance of generalization when data distribution changes over time, otherwise known as concept drift. The proposed model is more capable of sounding the alert of concept drift, which makes it more reliable over time because malaria transmission patterns are subject to change over time following climatic fluctuation, coverage of interventions or socio-economic change.

The other strength of this study is the local calibration and ecological specificity. The model is trained on the Kisii County, where the contextual data is relevant, and thus, the predictions are not based on aggregate regional distribution but on the transmission drivers of the county. This improves the scientific validity as well as operational relevance of the findings. Significantly, tree-based ensemble methods can be used to obtain interpretable outputs, including variable importance metrics and partial dependence interactions, which can be disseminated to the end-users in the form of public health practitioners and policymakers. This interpretability is critical towards instilling confidence in model outputs and integration into the current malaria surveillance and early warning systems.

In addition to the contribution at the local level the research contributes to the global malaria modelling agenda through the presentation of empirical data on the viability, scalability and transferability of machine learning-based predictive tools to African highland settings. Evaluating the performance of models in a difficult, underrepresented ecological environment in a systematic way and thus aiding the further fine-tuning of the deployment of the ensemble learning techniques in epidemiology, the research can help to understand the places and timings in which the methods can be applied. This carries more implications to malaria control interventions

in other similar highland areas in sub-Saharan Africa and in terms of global expectations of curbing malaria burden with anticipatory, data-driven and predictable measures on public health.

1.6 Motivation of the study

The two imperatives behind this study are the moral necessity to decrease the malaria burden among the vulnerable population and the scientific difficulty of creating artificial intelligence (AI) systems that are precise, understandable and fair. In sub-Saharan Africa, malaria has been a major source of morbidity and mortality and predictive models can be used to assist the early warning system that will assist in availing preventive and therapeutic resources at the right time. The consequences of predictive accuracy are especially high in resource constrained environments. False negatives in outbreak prediction can lead to lost opportunities to implement targeted measures, cause preventable morbidity and mortality, whereas false positives can focus health resources that have already limited resources on fruitless efforts and destroy trust in the use of technology. Therefore, this makes the ethical aspects of predictive modeling unavoidable of the technical, with sensitivity, and precision being the crucial aspects to balance.

The ecological and geographical interest area of this paper, is the highland environment of Kisii County. The highland regions are known as malaria prone with changing climatic conditions being still underrepresented in the global malaria modeling efforts. Majority of AI-based malaria forecasting studies have focused on tropical and high-density areas of malaria hence limiting the applicability of their results to a wide range of ecological environments. With the highland setting as a target, the study will not only serve the cause of malaria epidemiology in the region but also to the world at large through the emphasis on this understudied scenario of malaria epidemiology.

In addition to the initial practical goal of creating an effective predictive model, the research hopes to produce generalized methodology. It underlines the importance of models, which are not only precise, but also understandable, whereby health practitioners and policymakers can rely on and take action on their results. In addition, the paper interacts with the wider literature on the topic of algorithmic fairness, equity on global health AI, and recognizes that models need to be responsive to local data paucity, structural health disparities, and possibly biotic surveillance systems. Through this, the study will aid in the development of adaptive and context sensitive learning systems in epidemiology, and the lessons can be applicable in other diseases who have complex ecological and socio-economic causal factors as malaria.

CHAPTER TWO

2.1 Literature Review

Studies on prediction of malaria are by their nature interdisciplinary and they lie at the intersection of epidemiology, social sciences and computational modelling. The study combines various theoretical approaches in order to base its methodology and to have epidemiological validity and computational rigor. They consist of health behavior theories (Health Belief Model/HBM), social theories (Social Determinants of Health or SDH and the Ecosocial Theory), and computational theories (Ensemble Learning Theory, Computational Learning Theory or CLT and Cost-Sensitive Learning). Collectively, these views are used to develop predictors, model design, and assess their impacts on the structure of health in the population.

2.2 Theoretical Foundations

2.2.1 Health Belief Theory

The Health Belief Model offers a psychosocial perspective of the malaria risk, particularly how the perceived susceptibility, severity, benefits, and barriers influence the preventive behavior of the individuals. Recent researchers indicate that the risk perceptions of malaria have a significant impact on the use of ITNs, timely treatment-seeking, and adherence to interventions aimed at controlling vectors (Adongo et al., 2021; Akinyemi and Falope, 2022). Risk perception can also be quite diverse in Kisii County, where malaria is seasonal and where malaria is not always considered a serious disease as in endemic regions of the lowlands, that has an impact on the dynamics of outbreaks. As an illustration, households will not use ITN when they perceive low risk due to cooler and drier months but are prone to face high risks of transmission during unpredictable climatic changes. The effect of the integration of behavioral and coverage variables like ITN distribution, IRS campaigns, and treatment-seeking rates into predictive models is that

human behavior will be incorporated in the HBM principles by mediating the effect of human behavior in ecologically exposing individuals.

2.2.2 Social Determinants of Health

The SDH paradigm underlines the fact that social-economic and structural factors influence the health outcomes just as much as biological or behavioral. According to the latest health studies in the world, malaria simply impacts poorer households and communities, which have low access to medical services, inadequate housing, and low levels of education (Ricotta et al., 2020; Worrall et al., 2021). The Kisii County is characterized by differences in income, household density, and access to healthcare which pose different vulnerabilities to malaria transmission. The failure of predictive models to consider the socio-economic factors means that the predictive models may over-rely on climate-based predictors thus ignoring the structural inequities which contribute to the increased risk of malaria. This study is consistent with SDH philosophy by combining poverty indices, healthcare facility and population density, allowing forecasts to include ecological triggers as well as social vulnerabilities that are essential to achieve equal intervention in the problem of the health of the population.

2.2.3 Ecosocial Theory of Disease Distribution

Krieger has described the Ecosocial Theory, which gives a multi-level approach to risk of disease through emphasizing on the interactions of biological, ecological, and social processes with time and space. The recent use in the field of infectious diseases demonstrates that population health outcomes capture such factors like land use, climate variability, settlement patterns, and structural inequities (Krieger, 2021, Kim et al., 2022). The situation in Malaria transmission in Kisii County is such an example of a multi-layered

dynamic: social conditions (poverty, inequitable access to healthcare) are exposed to climatic exposures (rainfall, temperature, humidity) in a manner that creates uneven distribution of vulnerability. The inclusion of ecological and socio-economic predictors in the forecasting models will make sure that it is consistent with the Ecosocial framework since the malaria risk is viewed as the embodied manifestation of the overlapping structural and environmental pressures instead of a strictly biological occurrence.

2.2.4 Ensemble Learning Theory

The Ensemble Learning Theory gives the computational basis to focusing on combining multiple learners to improve predictive performance in high-dimensional, noisy and non-linear environments. Recent works affirm that ensemble models have always been more effective in single classifier tasks in epidemiological predictions owing to their robustness, flexibility, and capacity to learn complex relations of the various predictors (Dietterich, 2017; Opoku et al., 2021).

In this context, the core model in this research and the reference point as an ensemble method are Random Forest (RF). To create several decision trees, RF employs the bootstrap aggregation (bagging) technique, as a result of which the predictions of the decision trees are averaged, which stabilizes the performance and reduces the variance over decision trees. This renders RF especially appropriate to epidemiological data, which are frequently noisy, non-comprehensive and heterogeneous features like rainfall, temperature, ITN coverage, treatment-seeking rates, and socio-economic variables. The intrinsic strength of the model against noise and missing data, as well as the large dimensional input, enables the accuracy and generalizability of the predictions in the malaria outbreak forecasting. Moreover, the RF also offers score of feature importance which aids in increasing the interpretability by establishing the strongest predictors of outbreak, which is of great importance in the decision-making processes of public health.

In spite of the fact that the method of boosting, including XGBoost, are recognized within our Ensemble Learning Theory as having the advantages of reducing bias, it is apparent that the method of boosting that was selected in this study based on the fact that it provides a balance between accuracy, performance and interpretability is the one known as the Random Forest. RF realizes the Ensemble Learning Theory through implementation of the principle of variance reduction, and it also offers transparency in predictor importance, and this is what is technically acceptable and contextually suited to malaria prediction in Kisii County.

2.2.5 Computational Learning Theory

Computational Learning Theory (CLT) gives a formal, mathematically and intuitively based, ground to predictive model selection, evaluation and adaptation in the task of achieving high levels of generalization in real world problems. In its simplest form, CLT formulates the problem of learning as finding a hypothesis in an existing hypothesis space that minimizes the predicted error on unseen, novel data (Shalev-Shwartz and Ben-David, 2014). This directly applies to the malaria forecasting in Kisii County where the problem is to build models that both provide an effective fit to the past surveillance and climatic variables as well as generalization in future outbreak circumstances in the presence of uncertainty and lack of data.

One of the fundamental principles of CLT is the bias-variance trade-off, which breaks down the prediction error into bias (systematic), dependence on training data variations (variance), and noise that cannot be reduced. Random Forests (RF) have been especially useful in malaria prediction due to the ability to stabilize predictions in noisy epidemiological data with incomplete reporting and heterogeneous predictors (rainfall,

temperature, ITN coverage, treatment-seeking rates), through bootstrap aggregation of decision trees which enables the reduction in variance. In comparison, the algorithms that decrease bias, including XGBoost, are beneficial because the algorithmic residual correction significantly decreases bias when dealing with subtle disease transmission patterns

(Zhang et al., 2019; Zhao et al., 2022). Although his paper focuses on RF as the primary ensemble method, CLT

framework draws attention to the theoretical basis of combining the variance-reduction and bias-reduction methods to benchmark or extrapolate predictive models.

The other contribution of CLT is the margin theory, which highlights that those classifiers that have bigger decision margin can be generalized more to unobserved data. RF-like ensembles tend to raise decision margins since they are averages of many weak learners and this boosts robustness and minimizes chances of overfitting. This property is particularly significant in predicting malaria epidemics, in which data can be lopsided, and the outbreak events form a minor but crucial minority group. RF also maximizes the classification margins, with general stability, which fits the emphasis of CLT on generalization with class imbalance (Nguyen et al., 2021).

Another aspect that CLT highlights is the need to have cost sensitive learning in areas where misclassification results are asymmetric. In malaria prediction, the expense of a false negative (not predicting an outbreak) is much greater than cost of a false positive (giving a false alarm), because the former can result in an uncontrolled spread of the disease and untimely deaths. This theoretical understanding explains why the class weighting, custom loss functions and threshold adjustments are used in RF to focus on outbreak detection. Furthermore, it helps to select such types of evaluation as recall, F1-score, and area under the precision-recall curve (AUPRC), which are more effective in measuring the performance of the model in asymmetric costs than accuracy alone (Saito and Rehmsmeier, 2015; Wang et al., 2023).

The concept drift is another issue that CLT tries to solve; it is a phenomenon when the statistical characteristics of predictors and outcomes vary over time under the

impact of changes in climate, the extent of interventions or behavioral patterns. Conventional generalization assumptions expect the data to be identical and independently distributed which is often not the case with malaria epidemiology. Strategies based on CLT to address the problem of drift are sliding-window retraining, adaptive weighting of the current observations, and temporal feature weighting, all of which bring the learning process back to the updated distributions (Gama et al., 2014). In the present research, there is the use of the feature weight of time to consider the time sensitive effect of the predictors related to the climatic and intervention effects, such that the model is receptive to the change in dynamics of malaria transmission.

CLT also insists on interpretability being one of the restrictions of the hypothesis space because, in real-world scenarios like in the field of public health, predictive models should not merely be accurate but also understandable by the decision-maker. RF suits this criterion since it offers measures of feature importance that emphasize on the ecological, behavioral and socio-economic variables that have the highest impact in the occurrence of malaria outbreaks. This interpretability allows health officers and policymakers to trust, be accountable, and have actionable insights, and make sure that predictions are not black-box outputs but transparent tools to guide interventions (Lundberg and Lee, 2017; Opoku et al., 2021).

2.2.6 Cost-Sensitive Learning Framework (CSL)

The unequal effects of various kinds of error in misclassification are formalized in the Cost-Sensitive Learning (CSL) framework and especially in malaria outbreak prediction, it is obvious that the consequence of false negative, in which the prediction of an outbreak is wrong, has much more severe consequences than that of a false positive. An outbreak that is missed may cause late responses, an uncontrolled spread, health facilities overwhelmed, and avoidable

deaths, whereas a false positive usually causes resource misallocation, which, despite being inefficient, is relatively less damaging (Jiang et al., 2021; Wang et al., 2023). With the precautionary principle of epidemiology, which emphasizes the reduction of missed outbreaks at the cost of false alarms in some cases, CSL offers the methodological justification of developing malaria prediction models that put sensitivity over accuracy in priority. The principles of CSL have been applied in this study by class reweighting where resistant cases in an outbreak serve to be punished more, cost-sensitive loss functions involving asymmetric costs of misclassification and threshold modification to bias predictions towards outbreak detection. In addition to model training, other evaluation metrics, like the recall, are also informed by CSL since the traditional accuracy may be misleading when dealing with skewed data; the F1-score is used to balance the recall and the precision and the area under the precision-recall curve (AUPRC) is favored over ROC-AUC since it is more sensitive to minority class performance (Saito and Rehmsmeier, 2015). This study incorporates CSL principles, therefore, the computational outputs are not only statistically sound but also practical in addressing the needs of the Kisii County population in terms of timely and effective response in the control of malaria.

2.2.7 Linking the Theories to Objectives and Empirical

The theoretical foundations guiding this study include, the Health Belief Model (HBM), Social Determinants of Health (SDH), Ecosocial Theory, Ensemble Learning Theory (ELT), Computational Learning Theory (CLT), and the Cost-Sensitive Learning (CSL) Framework. Together form a coherent interdisciplinary framework linking behavioral, structural, ecological, and computational dimensions of malaria risk prediction.

These theories collectively justify the methodological design, analytical techniques and the interpretation of results across the study's three specific objectives.

The first specific objective, which seeks to identify and analyze ecological, socio-economic and epidemiological determinants of malaria outbreaks in Kisii County, draws primarily from the Health Belief Model, Social Determinants of Health, and Ecosocial Theory.

The HBM explains how individuals' perceived susceptibility and severity of malaria influence the uptake of preventive behaviors such as ITN use and treatment-seeking (Adongo et al., 2021; Akinyemi & Falope, 2022). The SDH framework expands this by linking malaria outcomes to broader structural inequities income, education, housing quality, and healthcare access (Ricotta et al., 2020; Worrall et al., 2021). The Ecosocial Theory integrates these perspectives by situating disease risk within the dynamic interactions between ecological and social systems, showing how environmental factors (rainfall, temperature, vegetation cover) intersect with social vulnerabilities to shape localized patterns of transmission (Krieger, 2021; Kim et al., 2022). Together, these models justify incorporating human behavioral, socio-economic, and ecological variables into predictive modeling rather than relying solely on climatic

The second specific objective, which involves developing and testing machine learning predictors a gap consistently noted in prior malaria modeling studies (Opoku et al., 2021; Zhao et al., 2022).

Learning Theory.

ELT provides the rationale for integrating multiple algorithms to capture complex, nonlinear relationships in epidemiological data (Dietterich, 2017; Opoku et al., 2021). It advocates combining the variance-reducing properties of bagging methods such as Random Forest with the

bias-reducing properties of boosting methods like XGBoost to improve predictive robustness in high-dimensional, noisy datasets. CLT, on the other hand, frames the process of model optimization and generalization. Its principles of bias variance trade-off and margin theory guide hyperparameter tuning, ensuring that the model balances overfitting and under fitting while maximizing predictive recall particularly crucial in detecting rare outbreak events (Shalev-Shwartz & Ben-David, 2014; Nguyen et al., 2021). CLT's treatment of concept drift informs the study's temporal weighting strategy, ensuring adaptability to evolving malaria transmission patterns over time due to climatic variability or intervention changes (Gama et al., 2014). These computational frameworks jointly support the study's methodological innovations in handling imbalanced, drift-prone, and heterogeneous data limitations that have constrained previous malaria prediction efforts (Zhang et al., 2019; Zhao et al., 2022).

The third specific objective, which aims to assess how key risk-factor relationships vary across distinct topographical zones, connects the Ecosocial Theory, SDH, and CLT perspectives.

Ecosocial Theory justifies spatial stratification by recognizing that health outcomes are “embodied” differently across ecological and social contexts (Krieger, 2021). SDH reinforces this by attributing these differences to spatial inequalities in living conditions, healthcare accessibility, and exposure patterns. CLT contributes by providing a computational lens for quantifying such heterogeneity through feature importance analysis and stratified model evaluation thereby enabling empirical testing of whether ecological and socio-economic drivers have varying predictive

strength across zones.

This multidimensional linkage ensures that the study not only models malaria risk statistically but also contextualizes it in the lived realities of diverse landscapes and communities.

The Cost-Sensitive Learning Framework intersects across all objectives, particularly reinforcing the model evaluation criteria. Given the asymmetric costs of misclassification where false negatives (missed outbreaks) are far more consequential than false positives CSL provides a theoretical and ethical basis for prioritizing sensitivity and recall over accuracy in the final predictive model (Saito & Rehmsmeier, 2015; Jiang et al., 2021; Wang et al., 2023).

This approach operationalizes the precautionary principle in epidemiology by emphasizing early detection, even at the cost of some false alarms, aligning model performance metrics with public health priorities. In synthesis, these theoretical lenses converge around three thematic domains that structure the entire study:

Modeling complexity – grounded in Ensemble and Computational Learning Theories to handle nonlinear, high-dimensional, and imbalanced epidemiological data.

Human behavioral and structural risk – guided by HBM, SDH, and Ecosocial Theory to integrate ecological, behavioral, and socio-economic determinants.

Empirical and methodological innovation – addressing gaps in prior malaria models through cost-sensitive, adaptive, and locally calibrated approaches.

2.3 Empirical Review

Malaria transmission is a complicated process that is influenced by a dynamic interaction of ecological, climatic, socio-economic, and demographic factors. Rainfall,

temperature, and humidity recurrently proving to be the major ecological forecasts in Sub-Saharan Africa, affect the breeding areas of mosquitoes and the growth pace of the Plasmodium parasite (Githeko et al., 2018; Snow and Noor, 2019). However, such associations are very situation-specific. According to some researchers, rainfall and malaria have a lagged relationship, whereas others focus on moderating factors such as altitude and local heterogeneity (Okello et al., 2020; Chaves and Koenraadt, 2019). These results highlight the significance of local studies, particularly highlands such as Kisii County where malaria transmission niches are unique due to micro-climatic variations, land-use modification and fragmentation of ecosystems. These differences undermine the consistency of standardized climatic models and demand ecological responses of local scale in which small scale environmental and socio-economic heterogeneity is factored into forecasting.

Socio-economic and demographic processes also make the issue of malaria epidemiology more complex, but have traditionally been less centrally involved in prediction models. The exposure and vulnerability are highly dependent on housing quality, occupation, and income levels, and poorly-built houses and agricultural practices in the vicinity of stagnant water increase mosquito contact (Tusting et al., 2020). Household income and access to healthcare are also essential in terms of access to preventive measures and diagnosis as well as treatment (Snow et al., 2021). Children below five years and pregnant women are the most susceptible groups because their immunity is immature or inhibited (Kachur et al., 2010; Macintyre et al., 2011). Irrespective of this information, the predictive models still overly use environmental variables and ignore socio-economic and demographic interactions that modify

transmission. Such imbalance is an indicator of a research gap: the necessity of integrative models according to which climatic and ecological interactions are coupled with social and demographic vulnerability, which enhances the accuracy of predictions and the targeting of interventions.

Statistical models like logistic regression and autoregressive integrated moving average (ARIMA) models have long been used in malaria prediction in an attempt to model such multidimensional relationships. Though convenient and interpretable, these models rely on the assumption of linearity, stationarity and independence, which is usually not met in dynamic epidemiological systems (Khamis et al., 2021). This makes them find it difficult to represent the non-linear and high-dimensional transmission through malaria, which leads them to adopt the machine learning (ML) approaches. Random Forests (RF) and Gradient Boosting Machines (GBM) demonstrated to be better at dealing with complex non-linear relationships and a high number of dimensions, as well as other types of data (Aleganan et al., 2020; Nisar et al., 2021). They are especially promising in malaria surveillance because of their ability to use a wide range of predictors, such as climatic indices, socioeconomic variables, etc. Nevertheless, the majority of ML models have been trained in lowland and stable transmission zones, with little highland ecologies such as Kisii being empirically validated. Besides, the black-box nature of ML models also complicates interpretability and policy translation, which is a limitation to their practical use by health professionals.

Recent empirical developments indicate that a solution might be to use the hybrid ensemble approaches. Hybrid frameworks can be used to achieve balanced and strong predictive results by implementing parallel bagging-based (RF) and boosting-based (XGBoost) models, achieving reduced variance and bias at the same time (Chen and Guestrin, 2016; Jiang

et al., 2021). It has been found that RF-XGBoost hybrids are more effective with noisy, imbalanced, and heterogeneous epidemiological data than single models (Rojas et al., 2020). The accuracy is further augmented by stacked ensemble learning, which involves using the outputs of the base learners as inputs to the meta-learner as in the case of logistic regression, to combine the strengths of each algorithm. These hybrid and stacked methods have great potential in prediction of the malaria outbreak in complicated environments such as Kisii where interactions among microclimate, land use, and the socioeconomic environment are highly flexible and adaptive.

The other theme that is of serious concern in the empirical literature is concept drift and class imbalance of epidemiological data. The number of outbreaks of malaria is low compared to the number of long inter-epidemic periods, which is yielding skewed datasets, with no-outbreaks predominating (Nguyen et al., 2021). This imbalance favors majorities in the models, leading to a misleadingly high accuracy, but it is also insensitive to real outbreaks, which are the most important events to the health of the general population. Such learning techniques as Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN), which create synthetic minorities samples to enhance model discrimination, are necessary to deal with this problem (Chawla et al., 2002; He et al., 2008). Learning based on cost sensitivity that rewards false negatives less has also demonstrated success in placing more emphasis on outbreak detection rather than the overall accuracy (Buda et al., 2018; Lin et al., 2020). The empirical data demonstrate that the oversampling of cost-sensitive ensemble learning can greatly improve the recall and the detection of the early outbreak (Rojas et al., 2020; Li et al., 2023). These kinds of adaptive mechanisms are in line with

the epidemiological priorities, in which the cost of missing an outbreak is significantly higher than that of false alarm.

The idea of adaptive learning goes beyond the imbalance of classes and is spread to time. A change in the patterns of transmission of malaria is dependent on the variation of climate, land-use changes and socio-economic development- a phenomenon referred to as concept drift. The fixed models become inaccurate as distributions of data change. Adaptive strategies are included in such as sliding-window retraining and temporal feature weighting, which allow keeping the generalization under varying conditions (Nguyen et al., 2021). The strategies will make sure that predictive models are flexible to new epidemiological patterns to facilitate their application as a real-time early warning system. Adaptive learning provides a holistic, context-sensitive mechanism of malaria prediction when used together with imbalance handling and hybrid ensemble techniques.

On the whole, it can be summed up that the future of malaria forecasting committed to single-algorithm approaches is not the best way to go but rather based on integrated, adaptive ensemble models. These models should combine climatic, ecological, demographic, and socio-economic predictors and overcome a temporal drift and data imbalance. A synthesis of this type balances technical rigor with epidemiological realism--coming up with predictive systems which are not only statistically strong, but also operationally useful in terms of early warning and early and targeted control of malaria, especially in highland ecosystems like Kisii County.

2.4 Previous Models that have been used in predicting malaria outbreaks

Forecasting of malaria has developed to a level where the traditional methods of using statistical models have been replaced by more complex machine learning methods that are no

longer linear to the dynamics of the disease spread. Past statistical techniques like logistic regression and autoregressive integrated moving average (ARIMA) models have been used to forecast incidence of malaria. Classification of outbreak versus non-outbreak events has been conducted through the use of logistic regression due to its ability to interpret and its simplicity in computation due to its linearity assumptions (Midega et al., 2012), though its ability to describe the complex relationship between climatic, ecological and socio-economic factors is constrained (Abeku et al., 2004). ARIMA, in contrast, have been helpful to capture seasonality and autocorrelation effects in time series of malaria incidence (Zinszer et al., 2012), but they are limited by their assumption of stationarity that inhibits their use in dynamic and non-linear ecology such as malaria transmission in highlands where changes in climatic variability and the operation of vectors control measures cause rapid changes in epidemiological patterns (Khamis et al., 2021). Statistical model weaknesses have consequently led to the adoption of machine learning (ML) methods, which are able to handle high-dimensional inputs and non-linear interactions, and do not require strong parametric assumptions. Decision Trees have been used to predict malaria due to their rule based interpretability, however they can be overfitting and can be poor at generalization unless care is taken to prune or regularize them (Cutler et al., 2007). SVMs have been utilized as well, which have shown good predictive performance in non-linear classification tasks (Teklehaimanot et al., 2004), although with its high computational costs and difficult parameter learning it is inapplicable in real-time, resource-constrained public health contexts.

Random Forest (RF) has become one of the powerful algorithms among the ensemble methods used in epidemiological prediction. RF is based on building several decision trees, through the so-called bootstrap aggregation (bagging), and averaging the results to decrease the variance and enhance predictive stability (Rodriguez-Galiano et al., 2015, Maxwell et al., 2018). Its advantages are that it is resistant to noise and missing data, it can analyze high-dimensional data that integrates climatic, demographic, and socio-economic predictors, and it provides scores of importance of features that improve the interpretability of its results in the context of public health (Lundberg et al., 2020). These attributes render RF especially appropriate in malaria forecasts where surveillance data is prone to being noisy, incomplete and heterogeneous. However, RF can occasionally blur epidemiologically significant outbreak patterns that are rare due to its averaging effect that favors the preponderant no outbreak group. It leads to a high overall predictive accuracy but a low sensitivity to unusual epidemic occurrences, as recent studies have revealed (Bodnar et al., 2020, Araya et al., 2021). To overcome this constraint, the methodological modifications could include resampling approaches, weighting the classes, and Synthetic Minority Oversampling Technique (SMOTE) to balance out the data and improve the detection of the outbreak cases of minority groups (Chawla et al., 2021).

XGBoost has also become one of the key players in the modelling of malaria. XGBoost is a sophisticated boosting algorithm that builds successive decision trees and each decision tree removes the residual errors of its predecessors with the addition of regularization which avoids overfitting (Chen and Guestrin, 2016). It can be used with structured epidemiological data because of its scalability, efficiency, support of complex non-linear interactions, and inherent ways of dealing with missing values (Zhang et al., 2019; Brownlee, 2020). XGBoost

is commonly more effective in malaria outbreak prediction, as it offers greater predictive accuracy in dynamic ecological conditions than classical statistical techniques (such as ARIMA and logistic regression) (Araya et al., 2021; Bodnar et al., 2020). Similar to RF, XGBoost is prone to imbalanced data, where the boosting algorithm will biasedly give preference to the majority class unless some precautions like class weights or resampling is taken to counter this (Chawla et al., 2021). Although XGBoost is more accurate than more basic models, it is less interpretable and should be combined with explainable AI techniques (like SHAP values) to provide transparency in the health decision-making process (Lundberg et al., 2020).

The recent developments in the field of malaria prediction have thus been directed on the hybrid ensemble frameworks incorporating RF and XGBoost in order to maximize on their respective strengths. In the proposed hybrid RF-XGBoost model, the contribution to concept drift in malaria transmission is added to the RF and XGBoost models, respectively, through the addition of time features along with the ability to weight them, enhancing the abatement of varying amounts of RF- and XGBoost-related characteristics of both models (Gama et al., 2014): This model uses two complementary steps whereby RF outputs including probabilities, feature importance scores are introduced as meta-features to XGBoost to enable it to refine predictions using the generalization patterns of RF as well as a weighted soft-voting scheme which balances the recall and precision of predictions using imbalanced outbreak data. Ensemble Learning Theory, which underlines the importance of working with diverse learners to avoid errors, and Computational Learning Theory, which rationalizes exploring the bias-variance trade-off and being flexible to changing data

distributions without sacrificing interpretability support the hybrid approach (Nguyen et al., 2021; Zhang et al., 2019). The approach of hybrid RF-XGBoost models receives support in empirical studies where these hybrid models have been demonstrated to be better than single models in high-dimensional, imbalanced health data in the sense of finding the right balance between sensitivity and specificity (Jiang et al., 2021; Rojas et al., 2020).

Collectively, these advances represent a methodological trend in malaria forecasts that has moved away not only the linear and assumption-driven statistical models but also the adaptive and ensemble-based machine learning based models. Although statistical models like logistic regression and ARIMA are more transparent and simpler, they do not have the flexibility to explain the non-linearities of malaria transmission. RF and XGBoost, especially when combined in hybrid structures, are more accurate, can handle data with noise and incompleteness, and it can be flexible to changing epidemiological situations. Nevertheless, to make sure that people take their health under consideration, their predictive capabilities should be tuned to correspond to interpretability, and explainable AI tools can help with this issue. In that regard, hybrid RF-XGBoost ensembles stand out as the most prospective future of precise, strong, and constrained malaria outbreak forecasting in complicated ecological settings.

2.5 Data Description

The data employed to carry out this research will incorporate the epidemiological, climatic, demographic, and socio-economic measures to come up with a predictive model that is comprehensive in predicting the occurrence of malaria in Kisii county, Kenya. The selection of variables and sources is due to the fact that the process of malaria transmission is

multidimensional because it is not only influenced by ecological conditions but also by human behavior, health-seeking patterns, and structural factors.

Epidemiological Data: The outcome variable will be the number of confirmed cases of malaria per month, which will be taken based on the District Health Information System (DHIS2) and county health surveillance records. These data give the temporal reference level of determining the outbreak and non-outbreak periods. Where case data are summarized in health facility level, it will be reconciled to sub-county level, to achieve consistency and comparability over time. The imputation methods and sensitivity analyses will be used to fix the data quality issues like underreporting, inconsistent reporting on a facility-level, and missing values, which will guarantee that the resultant variable is robust.

Climatic and Environmental Data: Global and national meteorological sources (Kenya Meteorological Department and satellite based databases. Predictor variables will consist of average weekly rain or rainfall, minimum and maximum temperatures, relative humidity, and vegetation cover (NDVI) because they have been consistently associated with breeding, sustenance and growth of parasites among mosquitoes. There will also be the time-lagged variables to take into consideration the delayed impact of climate on the abundance of mosquitoes and malaria disease occurrence.

Demographic and Socio-Economic Data: Population-level determinants will be obtained through the Kenya National Bureau of Statistics (KNBS) and such survey data as the Kenya Malaria Indicator Survey (KMIS) and the Demographic and Health Surveys (DHS). Some of the variables will be population density, poverty indices,

literacy levels and distance to nearest health facility. The variables are based on the Social Determinants of Health (SDH) framework, which makes the model inclusive of structural weaknesses, which enhance malaria risk among low-resource households and remote settlements within Kisii County.

Health Systems and Intervention Coverage Data: Ministry of health reports and county-level malaria control programs will provide the intervention related predictors. The major variables will be the coverage of insecticide treated nets (ITNs), regularity and scale of indoor residual spraying (IRS), shortage of antimalarial drugs, and rates of treatment seeking. These measures are used to operationalize Health Belief Model (HBM) and the Ecosocial Theory that represents the interaction between ecological exposure and human preventive behavior. The model is able to draw the distinction between ecological drivers of malaria transmission and the risk reductions caused by public health interventions by incorporating intervention coverage information.

Temporal and Spatial Resolution: The information will be arranged in the time series format (monthly) including at least 8-10 years (subject to data availability), which will be sufficient to reflect on the seasonal and inter-annual variation in malaria transmission. At the spatial level, data will be condensed to sub-county level to harmonize epidemiological data and climatic data, whereas there will be enough granularity to reveal localized outbreak processes.

Data Preprocessing and Integration: Preprocessing will entail data cleaning and standardization of various sources, missing values, continuous predictor normalization and coded categorical variables. The threshold of the outbreaks will be determined based on 75th percentile (surpassing the historical mean plus two standard deviation) or the world health

organization (WHO) recommended thresholds and hence changing the number of malaria cases into a binary variable (outbreak vs. non-outbreak). Lagged terms of interaction, time and weighting will be designed to capture delayed and changing relationships between predictors and malaria incidence. To tackle the issue of class imbalance (the rate of the outbreaks is lower than non-outbreaks), Synthetic Minority Oversampling Technique (SMOTE) and class weighting will be used in the process of training the models.

Rationale of Data Choice: This multi-source, multi-level data will make sure that malaria prediction is less climate-specific than the Ecosocial nature of disease distribution which is a combination of biological, ecological, and social determinants. The ground truth of the malaria incidence is capturing by epidemiological outcomes, ecological drivers of the populations of vectors by climatic data, structural vulnerabilities by socio-economic data, and preventive and behavioral impacts by intervention coverage. Combined, these streams of data allow creating a model that is technically reliable and contextually oriented on the truths of malaria transmission in Kisii County.

2.6 Data Analysis

The data analysis procedures will begin with an exploratory data analysis (EDA) phase. This phase will involve generating summary statistics such as mean, median, standard deviation, and percentiles to describe the distribution and central tendencies of malaria cases and predictor variables. Visual tools such as histograms, box plots, and spatiotemporal line charts will be used to uncover patterns, seasonality, and trends across different topographical zones. Heatmaps and correlation matrices will also be

generated to examine associations among climatic, socio-economic, and epidemiological variables, which will guide the selection of inputs for model development.

The study will investigate determinants of malaria outbreaks by applying generalized linear mixed models (GLMMs) and stratified analyses across topographical zones. This will make it possible to test whether relationships between ecological, socio-economic, and epidemiological risk factors differ across geographical settings. The results of this analysis will inform both feature engineering and the construction of predictive models. The next phase will involve the design of machine learning models, beginning with a Random Forest classifier to predict malaria outbreak events. The model will incorporate lagged climatic variables to account for delayed mosquito breeding cycles, temporal weighting to emphasize recent trends, and categorical encoding of socio-economic and intervention-related data. To address class imbalance in outbreak prediction, oversampling techniques such as SMOTE, as well as class weighting, will be applied. Feature drift over time will be managed through adaptive retraining strategies and rolling-window validation. Interpretability will be ensured using SHAP (Shapley Additive Explanations) to generate both global and local

A hybrid Random Forest–XGBoost model will be developed to assess whether explanations of model predictions.

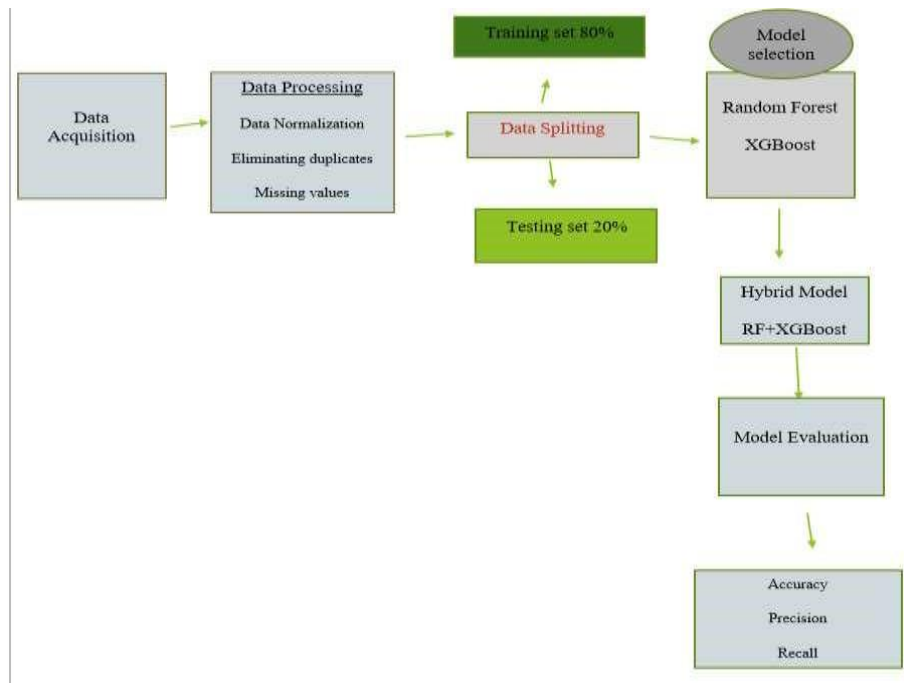
combining ensemble methods improves predictive performance compared to the standalone Random Forest and XGBoost algorithms. These models will be tuned using grid search and randomized search methods to optimize hyper parameters. Evaluation will be conducted using temporal cross-validation, particularly rolling-origin back testing, and spatial validation across sub-counties in Kisii. Performance metrics will include AUC-ROC, precision-recall AUC, sensitivity, specificity, F1-score, and calibration measures. Statistical comparisons will be made using paired bootstrap methods and DeLong’s test for AUC differences.

Finally, interpretation of results will be presented through summary tables, SHAP visualizations, and outbreak risk maps. These outputs will highlight the most influential ecological and socio-economic determinants of malaria transmission and identify the locations and time periods most vulnerable to outbreaks. The findings will be used to provide evidence-based recommendations for county health authorities, particularly in guiding targeted interventions such as indoor residual spraying, distribution of bed nets, and community sensitization campaigns.

2.6.1 Data Analysis Framework

The chart below shows the various steps followed in developing the three models in this study. The individual and hybrid models were evaluated on their performance in malaria prediction using various metrics such as accuracy, precision, recall, confusion Metrix and the F1 score. The prediction and probabilities were generated for each models 'and the performance metrics were calculated and compared between the individual and hybrid model.

FIGURE 1
Data analysis framework



2.7 Data Processing and Feature-Engineering

The raw data will follow a stringent process of data preprocessing that will guarantee integrity and appropriateness of data in analytical analysis and advanced modelling. This process starts by cleaning data i.e. identifying and fixing anomalies in the data like duplication, implausible values (i.e. negative cases, out of range climatic values), and different formats of data. This step is essential to reduce the noise and bias due to the epidemiological and environmental heterogeneity of malaria data. The processed data will subsequently be standardized and harmonized to some shared spatiotemporal resolution allowing it to be integrated into a structured spatiotemporal panel dataset where each row corresponds to a distinct observation. In this respect, it is especially important to deal with missing data because if the epidemiological surveillance or meteorological measurements have gaps, systematic bias

may occur. A dual strategy will be applied to counter this. To begin with, epidemiological and socio-economic variables will be imputed with multiple imputation methods so that unknown underlying relationships between covariates are preserved, but uncertainty is reduced. Second, time series variables that show high temporal continuity (rainfall and temperature) will be interpolated temporally using the spline or linear interpolation techniques. Such a hybrid method enables a stronger reconstruction of the missing values, as well as consideration of the unique statistical characteristics of epidemiological and environmental data.

The feature engineering will be the focus of improving the predictive abilities of the models. In the case of climatic variables, lagged features will be created to reflect the lagging impact of breeding cycles of mosquitoes, parasite life cycles, and transmission patterns. Moreover, temporal feature weighting will also be applied to focus on the latest observations as opposed to the past in order to solve non-stationarity and concept drift effects on malaria transmission patterns. This can be implemented in the form of exponentially decaying weights or sliding-window retraining, which would make sure that models are sensitive to the changing epidemiological and environmental factors.

Normalization and scaling (min-max scaling) of the heterogeneous predictors will be done to make the scale homogeneous. This is especially imperative when putting together the climatic, demographic, and intervention related characteristics which work on significantly different numeric scales and units. With the proper scaling, the features of large magnitude do not dominate the learning process hence enhance algorithmic convergence and stability. Lastly, appropriate encoding schemes will be applied on

pg. 49

categorical predictors, e.g., intervention coverage or land-use classification. In the case of ordinal relationships, ordinal encoding will be applied, and one hot encoding will be applied to non-ordinal categories to have an unbiased representation.

Taken together, these preprocessing and feature engineering activities will make raw and heterogeneous malaria related data into methodologically sound, spatiotemporally structured data. This data will give a solid base to the downstream machine learning models, so as to learn not only the intricate time dependencies, but also the spatial heterogeneity which defines malaria transmission dynamics. The approach improves the validity and generalizability of outbreak prediction models on malaria by making sure that data quality, completeness and features relevance are tackled systematically.

2.8 Model Training and Evaluation

Data will be split into training (80%) and test (20%) sets using stratified sampling to preserve outbreak/non-outbreak balance. Cross-validation will be employed to mitigate overfitting. Performance metrics will include accuracy, precision, recall, F1-score, AUC, and Brier score.

2.9 Addressing challenges

2.9.1 Feature Drift

One of the challenges in malaria outbreak prediction is concept drift, which occurs when the statistical relationships between predictors and outcomes change over time due to ecological shifts, socio-economic transitions, or intervention programs. For example, improvements in healthcare access, changes in rainfall patterns, or new vector control measures may alter the transmission dynamics of malaria, making previously learned models less accurate. To address this issue, the study will implement a sliding-window retraining strategy.

This approach ensures that models are periodically updated with the most recent data while discarding outdated information, thereby improving adaptability to evolving conditions. In addition, temporal feature reweighting will be applied to give greater importance to recent observations compared to older data, reflecting the fact that recent patterns are more predictive of near-future outbreaks. Together, these techniques will help the models remain responsive to real-world changes and maintain predictive accuracy in the face of dynamic malaria transmission environments.

2.9.2 Data Imbalance

One of the most difficult processes in predicting a malaria outbreak is the existence of imbalance between outbreak and non-outbreak cases in the surveillance data bases. Generally, there are periods of no outbreak which are predominated by data, whereas confirmed cases of outbreak are much rarer. Such disproportion exposes a scenario in which normal machine learning classifiers will have a preference in the majority population and thus yield models that will not be effective in identifying outbreak cases in general. When it comes to malaria, it is extremely problematic, because false negatives, missed outbreaks have much more serious consequences as compared to false positives, which may only lead to precautionary actions being taken (He & Garcia, 2009; Haixiang et al., 2017). In order to make sure that the models created in the process of this study are able to identify the outbreaks reliably, a number of imbalance-handling strategies will be used.

The former one is the Synthetic Minority Oversampling Technique (SMOTE). Which creates artificial outbreaks by interpolating between the existing minority samples, as opposed to copying them. This improves the modeling of the outbreak

events in the dataset and enables the models to learn more effectively on the decision boundaries that distinguish between the outbreak and non-outbreak conditions. The current advancements to SMOTE, like Borderline-SMOTE and ADASYN, have been shown to perform well on the issue of the detection of rare events in epidemiological data (Chawla et al., 2002, Fernandez et al., 2018, Gao et al., 2020).

The second strategy is weighting of classes in loss functions. The use of ensemble model like the Random Forest (RF) and Extreme Gradient Boosting (XGBoost) enables me to weight the minority cases of outbreak relatively more in the training process. These models are more sensitive to rare occurrences by rewarding an error in the misclassification of the outbreak samples more than non-outbreak samples. Health prediction research has found that class-weighted algorithms can boost the recall rates by a significant margin without leading to a significant decrease in precision (Buda et al., 2018, Lin et al., 2020).

A cost sensitive learning will also be investigated in order to explicitly represent the asymmetric effects of the errors in prediction. In the case of malaria, the failure cost of not identifying an outbreak (false negative) is much greater than the inconvenience of false alarm. Such costs included in the training goal are relevant because the model takes the detection of the outbreak as a priority, thus eliminating the chances of such hazardous oversight. Recent studies in forecasting of infectious diseases affirm that cost-sensitive methods reduce instances of missed detection rates than the traditional learners (Nguyen et al., 2021; Li et al., 2023).

All these strategies will be tested in the training pipeline of the study in an organized manner. Measures to assess model performance will be based on imbalanced data measures, such as precision, recall, F1-score, and Area Under the Precision-Recall Curve (AUPRC). In

the case of minorities outbreaks, these measures are more informative than the overall accuracy because they explicitly quantify the capacity of the model to identify cases of minority outbreaks (Saito and Rehmsmeier, 2015; Branco et al., 2016). It aims to address imbalance of data by oversampling, weighting, and cost sensitive approaches to develop models that are accurate and sensitive to a very vital activity of malaria outbreak prediction.

Data Processing Tools

Computational tools will be employed to perform the data processing of this research to provide efficiency, scalability, and reproducibility. The main program language will be Python, as it provides a broad range of libraries applicable to data mining and machine learning processes. The work will involve the use of core libraries at various workflow levels. Data ingestion, manipulation, and integration will be performed with help of Panda, whereas NumPy will assist in efficient computing of numbers on multidimensional arrays. Scikit-learn will be the preferred toolkit to use in preprocessing functions, including normalization, encoding, imputation, feature scaling, and others.

The process of data cleaning will be an essential phase of the pipeline, as it will resolve the issues of missing values, outliers, contradictions within datasets. Since the malaria surveillance and climatic data are spatiotemporal, absence of observations will be addressed through interpolation and imputation. To prevent the tendency of extreme values to skew the training of a model, outlier detection will be employed; categorical variables like levels of intervention coverage will be coded into machine-readable

formats. During this phase, provenance of data will be monitored; transformation logs will be kept to increase transparency and reproducibility.

Visual analytics will be used to carry out exploratory data analysis (EDA). Matplotlib and seaborn shall be used to create the descriptive visualizations such as time plotting, correlation heatmap, and distributional charts. The underlying data structures, seasonal patterns, and relationships among the variables will be exposed through these visualizations to inform feature engineering and model development.

Structured formats, such as Pandas DataFrames and CSV files, as lightweight, broadly compatible and easy to integrate into downstream modeling pipelines, will contain processed data. Such a system of storage will enable the repetitive development of models and will be seamlessly integrated with modeling and validation activities.

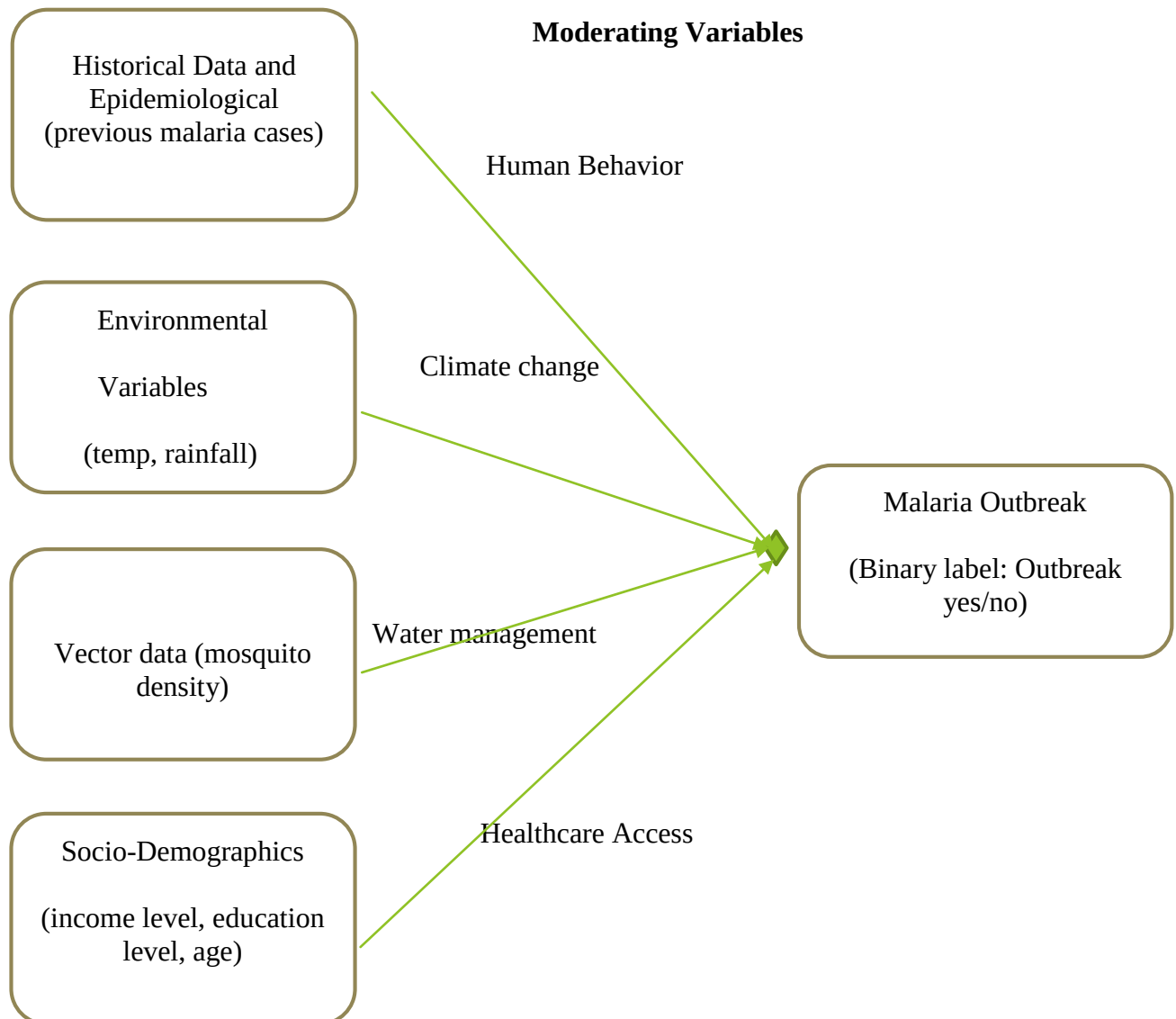
Ethically, privacy will be safeguarded by enforcing stringent measures to ensure that data governance standards are established. Personally identifiable information (PII) will be removed before processing all datasets. Moreover, the research will follow the licensing agreement and the usage terms set by each of the data providers to use health and environmental data that are available publicly in a responsible way. Transparency will also be enhanced by extensive recording of all scripts, preprocess, and transformations done to the raw data. This will make the findings replicable, thereby making a retest by a subsequent researcher viable.

FIGURE 2
Conceptual framework

Conceptual Framework

Independent Variables

Dependent Variable



CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

The research methodology for a study aims at using a Random Forest model to predict malaria outbreaks in Kisii County in which the data used in this study will be obtained from the ministry of health and the meteorological department which integrates historical data, environmental variables, and vector control interventions. This methodology outlines how data will be collected, processed, analyzed, and how the model will be trained and validated. It is designed to build a predictive model that can accurately forecast malaria outbreaks and provide actionable insights for public health interventions.

3.2 Research Design

The research will be based on a predictive-correlational design, which will be located in a positivist research philosophy. The positivist paradigm places value on objective measurements and the use of statistical tools to determine general patterns in the natural and social phenomena (Creswell and Creswell, 2018). This philosophical approach is suitable since the research is based on measurable epidemiological, ecological, and socio-economic measures of malaria epidemics that can be systematically analyzed and predicted. The predictive-correlational design can be used to determine the statistical correlations between malaria transmission determinants including rainfall, temperature, population density, and the use of insecticidal-treated nets and the health-seeking behavior and subsequently use the relationships to construct data-driven predictive models (Hay et al., 2013; Snow and Noor, 2015).

The study takes the malaria case surveillance data a step further in combining it with climatic, demographic and socio-economic data as the traditional linear or regression models usually do not capture the nonlinear processes of malaria transmission (Gething et al., 2011). The study will instead make use of the ensemble learning methods that involve the integration of numerous algorithms to enhance predictive accuracy and extensive generalization (Dietterich, 2000). This methodological breakthrough is especially applicable in the malaria modelling where ecological and social determinants are interaction based, multidimensional, and non-stationary with time. Not only does the design facilitate stringent hypothesis testing on malaria determinants but also offers operational information on malaria in the context of planning of the Kisii County which is a highland malaria prone area with seasonal epidemiology being a leading burden (WHO, 2023).

The predictive aspect of the design will be focused on methods, particularly, Random Forest (RF). RF is also ideal in epidemiological forecasting due to its resistance to overfitting, the capability to model nonlinear interactions and its ability to exploit heterogeneous data sources (Breiman, 2001; Cutler et al., 2007). In an effort to optimize the model further, the RF algorithm will include adaptive class imbalance which is usually a challenge in the task of predicting outbreaks because epidemic periods are rare relative to non-epidemic periods (Haixiang et al., 2017). Moreover, there will be concept drift correction to ensure that the results reflect the changing environmental and epidemiological trends and thereby enhances time strength of predictions (Gama et al., 2014).

It will be ensured that it is interpretable by referring to SHAP values that will be used to explain the model predictions, both globally and locally (Lundberg and Lee, 2017). Such openness is essential to the field of public health, which needs stakeholders not only to be provided with accurate predictions but also clear details of the mechanisms behind outbreaks to act with appropriate interventions (Caruana et al., 2015). On this basis, the paper will create a hybrid model of Random Forest-XGBoost, and its predictive accuracy and stability will be compared with the isolated models. XGBoost (Chen and Guestrin, 2016) is especially good at working with structured tabular data and can work in addition to RF to capture the interaction of complex features and improve the performance of the ensemble.

The design is quantitative and computational in nature, which will guarantee that the results will be empirically based and replicable in like endemic and epidemic prone situations. By incorporating these machine learning techniques into a spatial time epidemiological framework, it will be possible to model the malaria risk both at the temporal (seasonal/monthly) and spatial scales (ecological zones within Kisii County). This two-fold approach is a response to one of the major limitations of traditional malaria models that simply fail to consider the dynamism of the ecological, socio-economic and geographical interactions (Reiner et al., 2013).

In this respect, the research design will be a mixture of the methodological rigor and practical usefulness. Scientifically, it helps to promote predictive epidemiology by combining modern machine learning with the use of spatiotemporal analysis. In practice, it can provide practice-based information of malaria management programs, which is the early-warning mechanisms and interventions in Kisii County. The predictive-correlational paradigm, then,

enhances the contribution made by epistemic to malaria modelling as well as the relevance of disease forecasting to policies in endemic areas.

3.2.1 Research Approach

A quantitative research method will be used in the study as it is suitable considering that it involves numerical analysis and computer-based modelling to test the determinants of malaria outbreaks. Quantitative techniques can be effectively used to predict epidemiology due to their ability to systematize the collection, processing, and interpretation of large and high-dimensional data collected and generated by various methods such as surveillance systems, climatic records, and socio-economic indicators (Creswell and Creswell, 2018). This strategy helps to identify strong and generalizable explanatory variable associations with malaria cases because of the focus on statistical rigor and reproducibility issues.

To accomplish the mentioned objectives, the research will adopt a systematic methodological framework commencing with systematic exploration of ecological, socio-economic and epidemiological determinants of malaria epidemics in the Kisii County. Malaria case data of the past will be supplemented by climatic variables of rainfall, temperature, and humidity, social-economic variables of population density, poverty indices, and access to healthcare and intervention related variables of ITN coverage and IRS campaigns. Exploratory data analysis, spatial mapping, and statistical testing will be used to determine the significant predictors and evaluate whether the significant relationships between the key risk factors vary across topographical zones. Based on this, an outbreak prediction model based on a Random Forest will be created, which will include the methods to overcome the major shortcomings of epidemiological

modelling. These contain imbalance correction strategies, like class weighting and SMOTE, to guarantee sufficient coverage of rare outbreaks, and adaptive techniques of time feature drift, like sliding-window retraining and temporal weighting of current information, in order to discover variation in the climatic and epidemiological factors across time. Random Forest model will be trained with various hyperparameter settings and then cross-validated and compared with traditional predictive models that are commonly used in malaria studies which include XGBoost and hybrid random forest-XGBoost. To evaluate the models, a package of performance indicators will be used, such as accuracy, precision, recall, F1score, balanced accuracy, and AUPRC, and sensitivity will be prioritized to ensure that the model minimizes false negatives. To determine the comparative advantage of Random Forest statistically with significant results, statistical tests of significance will be applied; and feature importance analysis will be used to gain an insight into the most significant predictors that will ensure the model is interpretable and policy-relevant. In such a manner, the research does not only operationalize the construction of an effective Random Forest model, but locates it in the context of the wider framework of malaria forecast techniques.

3.3 Ethical Considerations

3.3.1 Nature of Data Collected and Level of Risk

This study will use secondary data drawn from the Ministry of Health (malaria case records), the Kenya Meteorological Department (environmental and climatic variables such as rainfall, temperature, and humidity), the Kenya Bureau of Statistics (demographic data), and Kisii County health records. The datasets contain aggregated, non-identifiable records and will not include personal identifiers such as names, ID numbers, or contact details. As such, the level of risk to individual privacy is minimal. However, there remains a potential for indirect re-identification, especially if small geographic units or rare event data are cross-referenced

with other public datasets. To mitigate this, only aggregated data at the ward or county level will be used, and all outputs will be reported in a way that prevents singling out individuals or small groups.

3.3.2 Dataset Structure and Re-Identification Risk

The dataset will be structured as a time-series panel, with rows representing monthly records for Kisii County and columns capturing epidemiological (malaria cases), environmental (rainfall, temperature, humidity), vector control interventions, and socio-economic indicators. While the data is public or institutionally accessible, the possibility of re-identification particularly in sparsely populated areas or during low-incidence periods will be acknowledged, and geographic resolution will be limited to protect privacy. The data used in this research originates from ethically approved and officially mandated public health and environmental monitoring systems. These institutions follow established protocols for data collection, including standardized malaria diagnosis, environmental measurement, and demographic reporting methods. Nevertheless, the possibility of reporting errors, missing data, and observational limitations inherent in such large-scale monitoring will be acknowledged.

3.3.3 Consent and Original Data Collection

The original datasets were collected by government agencies and research institutions for public health surveillance, statistical reporting, and climate monitoring. The original data subjects did not provide explicit consent for this specific study; however, the data was gathered under institutional and legal frameworks permitting public health use. The Ministry of Health and the Kenya Meteorological Department operate under national regulations that allow aggregated data to be used for research

aimed at improving health outcomes. The malaria incidence and demographic data from government sources are considered open access under Kenyan public data sharing policies, while some environmental datasets from the meteorological department may be under restricted use or require institutional agreements. All datasets will be cited appropriately in the final report.

3.3.4 Biases and Limitations

Possible biases in the sources of study data and the analyses process are significant issues that need to be clearly discussed to protect the validity of the results. Malaria cases are underreported and this is one of the main sources of bias especially in situations where there is no easy access to health facilities. The rural families might involve informal care providers or self-treatment, resulting in a high percentage of undocumented malaria cases. Such underrepresentation may bias the actual epidemiological burden, and could even undermine the predictive capability of models fitted on incomplete surveillance data.

The second problem is connected to the variety of reporting across regions. Areas with limited resources or rural regions have logistical challenges in data collection such as shortage of diagnostic kits, health personnel with insufficient training, and latency in reporting to central databases. These discrepancies add noise to the data and therefore lead to the uneven quality of malaria incidence data among the sub-counties.

The unequal space distribution of meteorological monitoring stations is another biases. Rainfall, temperature, and humidity are weather variables that are essential as inputs to the malaria forecasting process; however, ground based weather stations tend to be concentrated around or near cities and other accessible locations. This leaves data gaps in the rural areas

where malaria prevalence tends to be the highest. Although satellite-based climatic measurements would partially fill such blank spots, the variation in spatial resolution between ground and satellite measurements may still create uncertainty in measurements.

Another weakness relates to the time discrepancy between data sets. Whereas climate and health data are usually recorded monthly or even weekly, socio-economic indicators, which include poverty levels, household income or access to sanitation, are regularly updated at much longer rates, say every three to five years as national surveys. These time differences may cause prediction error between the predictor variables and outcome data, hence, the accuracy of the model estimates becomes low.

To reduce these biases, the research will embrace some approaches of methodology. First, the imputation methods will be used to handle missing, or underreported values, in both statistical and machine learning-based methods to construct plausible estimates. Second, the sensitivity analysis is going to be performed in order to consider the impact of changes in the data quality or the reporting rates on the model performance and, thereby, determine the level of robustness in the results. Third, where spatial biases are pronounced, methods like spatial interpolation of climatic data or coarse spatial resolutions will be used to lessen the effects of an uneven monitoring network.

Lastly, the problem of time displacements will be dealt with by adopting a strategy of temporal harmonization, by interpolating socio-economic indicators over time or by including them as a static contextual covariate instead of a dynamic predictor.

The study will thus make the forecasting models valid and reliable by explicitly considering and controlling such biases in a systematic way. It would also increase the credibility and transparency of the research and give stakeholders in malaria control programs information that can be acted upon but be mindful of the limitations of real-world data.

3.3.5 Potential for Harm and Respect for Data Subjects

The analysis aims to support public health interventions and carries no intention of stigmatizing or unfairly targeting specific communities. However, there is a potential risk that the identification of “high-risk” areas could inadvertently lead to social or economic stigma. To mitigate this, findings will be communicated in a manner that focuses on structural and environmental factors rather than attributing disease burden to individual behaviors. Language used in publications will remain respectful, and the emphasis will be on empowering communities with preventive measures.

3.3.6 Legal and Regulatory Compliance

The research will comply with all applicable Kenyan data protection laws, including the Data Protection Act (2019), and institutional research ethics requirements at KCA University. No personally identifiable information will be accessed or processed. Where required, formal data sharing agreements will be secured from relevant government bodies before accessing restricted datasets.

CHAPTER 4

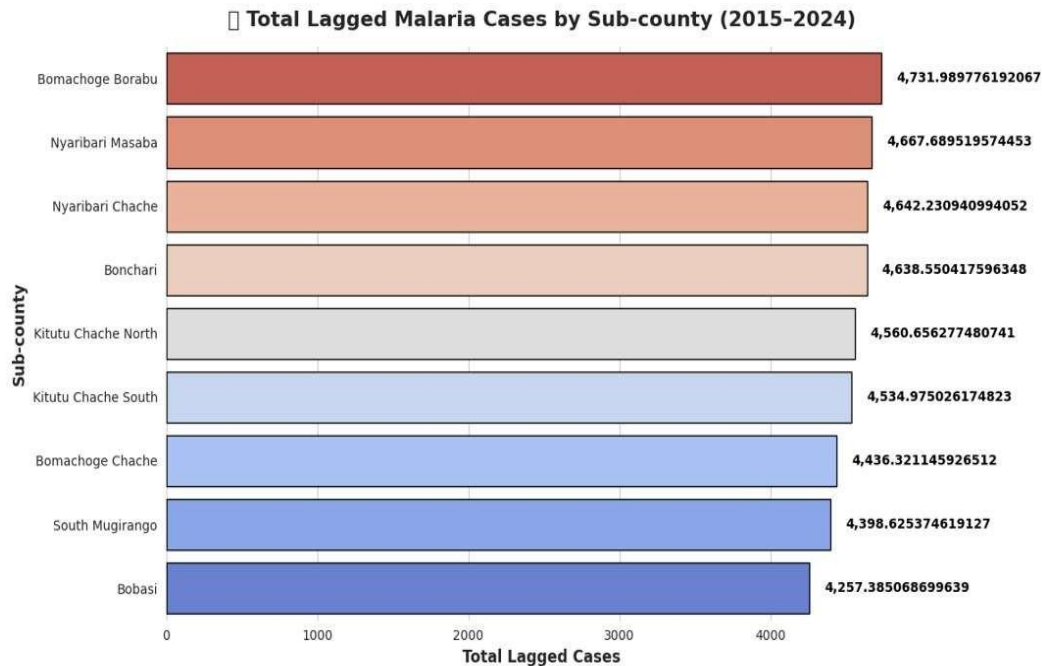
FINDINGS AND DISCUSSIONS

This chapter presented the results of the data analysis conducted to investigate the ecological, socio-economic, and epidemiological determinants of malaria outbreaks in Kisii County between 2015 and 2024. The analysis was based on the processed dataset and applied machine learning models, including Random Forest, XGBoost, and a Hybrid Stacking classifier. The findings were organized into descriptive analyses, model performance evaluation, and spatial–temporal variations across sub-counties.

4.1 Descriptive Statistics

The dataset comprised of monthly malaria cases recorded across the 9 constituencies in Kisii County for the last 9 years which were in lagged form. The overall burden of malaria varied considerably between the zones with the following constituencies recorded relatively higher cases. The chart shows that Bomachoge Borabu, Nyaribari Masaba and Nyaribari Chache were the leading contributors to the burden.

FIGURE 3
Total lagged malaria cases by sub-county



4.2 Models Evaluation

The discussion of the performance indicators in figure 2 reveals that there are significant differences between the Random Forest (RF), XGBoost, and Stacking models when it comes to predicting malaria outbreaks in Kisii County. Random Forest model offered the best sensitivity (0.723) and balanced accuracy (0.813), which indicates that it is better placed in terms of its capacity to disarm the outbreak and non-outbreak periods. Sensitivity at high levels is essential in forecasting of public health since most of the actual outbreaks will be identified thus lowering the chances of missing epidemic events. Similarly, the high balance accuracy indicates that RF worked well in both classes even despite the bias in the actual data where the occurrence of outbreaks is lower compared to the non-outbreaks. This strength renders RF quite

useful in malaria surveillance where missing outbreaks may result into delayed responses, uncontrollable spread and unnecessary loss of lives.

RF scored 0.688 in terms of precision (a little lower than Stacking model (0.728) and similar to XGBoost (0.684). Increased precision is an attribute of a model not to produce false alarms and hence the performance of the Stacking model implies the positive predictions of the model were more accurate. Nevertheless, it was this improvement in accuracy that was a significant cost: both the XGBoost and Stacking experienced losses in sensitivity (0.696 and 0.679, respectively) and balanced accuracy (0.801 and 0.802). That is, these models were more successful in making sure that the predicted outbreaks were accurate, but in the case of actual outbreaks, they were more likely to underestimate them. This trade-off is not as desirable in the case of malaria control since the social cost of a false negative (missed outbreak) is much higher than the cost of a false positive (false alarm).

The F1 scores, which achieve a balance between accuracy and recall, were somewhat equal between the different models: F1 is 0.705 (RF), F1 is 0.690 (XGBoost), and F1 is 0.703 (Stacking). This comparison highlights the very difficult of the task of balancing sensitivity and precision of imbalanced epidemiological information. Nevertheless, the marginally better F1 score of RF indicates that it offers the most reasonable trade-off between true outbreaks and false alarms, which also adds to the fact that RF is the most appropriate model.

The findings make Random Forest the best model to use in this study especially where the overall priority is to optimally detect outbreaks of malaria. Its high sensitivity and balanced accuracy means that it can generalize to high-level noise and high-

pg. 67

dimensional imbalanced data suitable to malaria surveillance in Kisii County. Whereas XGBoost and Stacking were found to be stronger in terms of their ability to be more precise, it reflected on their weaknesses in terms of sensitivity, which restricts their use in terms of detecting an outbreak, where early detection may be of utmost importance. These results are aligned with the recent malaria modelling studies, which also concluded that RF is very useful in outbreak prediction because of its strengths, readability, and ability to resist the noise in data (Okongo Martin et al., 2024). Therefore, RF is suggested to be the best model to use in malaria prediction, whereas Stacking can be an effective tool to use complementary in situations when false positives should be minimized.

4.2.1 Model performance metrics

TABLE 1
Model performance metrics

Performance metrics	Random forest	XGBoost	Stacking
Sensitivity	0.723	0.696	0.679
Specificity	0.903	0.906	0.925
Precision	0.688	0.684	0.728
F1score	0.705	0.690	0.703
Balances accuracy	0.813	0.801	0.802

4.2.2 Confusion Matrix

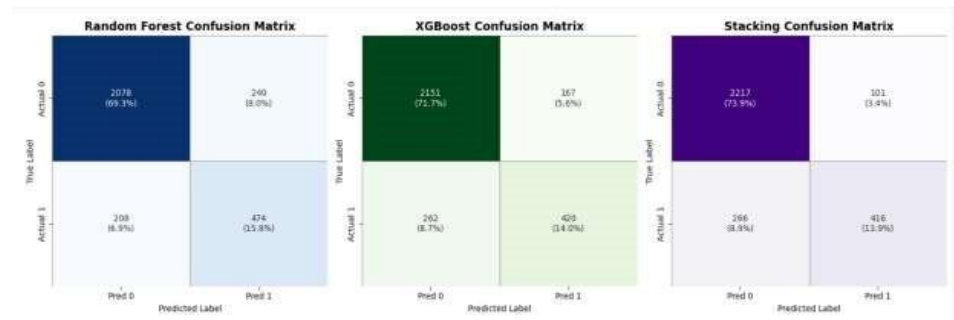
The confusion matrices in figure 2 represent the performance of the three models when it comes to accurately and wrongly classifying cases as outbreak (positive) and non-outbreak (negative). The accuracy of the Random Forest (RF) model is high in terms of the rate of 2,078 true negatives (69.3%) and 474 true positives (15.8) which shows that the model is effective in capturing outbreak and non-outbreak events. Nonetheless, RF also has 240 false positives (8.0%), which matches with its mediocre-precision measure (0.688). This implies that although the model is likely to cause some false alarms, it offsets the false alarm by identifying the maximum amount of real outbreak cases of the models. In the context of the public health, such a trade-off can be considered positive because the consequences of the failure to detect an outbreak (false negative) are much more detrimental than those of the false prediction (false positive).

By comparison, the XGBoost model has a more conservative approach to prediction. It gets 2,151 true negatives (71.7) and 420 true positives (14.0) and also it minimizes false positives to 167 (5.6). This confirms its increased accuracy and reduced false alarm level as opposed to RF. Nonetheless, it is accompanied by a loss of sensitivity because XGBoost is incorrectly labeled 262 cases (8.7) as false negatives. Practically, it will imply that unlike other models, XGBoost performs more effectively when forecasting an outbreak but on the other hand, it will not be an effective early warning mechanism where sensitivity is the main factor since it will not detect outbreaks at all.

The Stacking model has the best capability of not generating false alarms as they have 2,217 true negatives (73.9) and a false positive that is the lowest of the three models (101). Nevertheless, this conservative bias highly compromises its capacity to identify actual outbreaks with 416 true positives (13.9) and 266 false negatives (8.9). Although Stacking model is the most specific in general, it is less sensitive and thus is less useful in forecasting outbreaks where its ability to capture as many true positives as possible is more important.

Combined, the confusion matrices identify the trade-offs that are specific to each of the three models. Random Forest places more emphasis on recall (sensitivity) because it aims to achieve the greatest number of outbreak cases at the cost of extra false positives. XGBoost provides a balance of recall and precision at the expense of fewer false alarms instead of detecting outbreaks. Stacking is most specific but fails to detect a large number of outbreaks because false negative is large. These findings support the fact that Random Forest is the best model to predict a malaria outbreak in Kisii County according to the evaluation metrics. Its strong recall ability and equal accuracy mean that the outbreaks are hardly missed, which is consistent with the priorities of the public health of detecting outbreaks early and responding to them quickly. Although XGBoost and Stacking play a role of inputting into enhancing precision and specificity, they are not sensitive thus making it possible to affirm that Random Forest is the most trustworthy model when it comes to situations where outbreak identification is a priority.

FIGURE 5
Confusion matrix



4.2.3 Receiver Operating characteristic and Area Under Curve

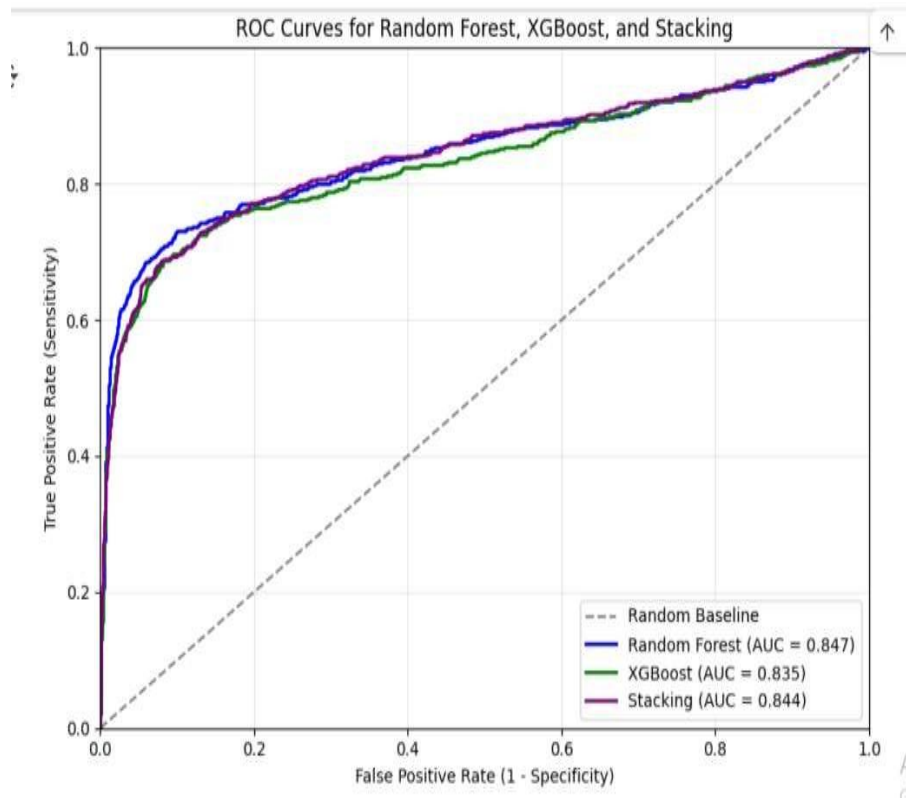
Figure 3 shows the Receiver Operating Characteristic (ROC) curve of the Random Forest (RF) model, that is, the True Positive Rate (sensitivity) and False Positive Rate at different levels of decision thresholds. The ROC curve is a holistic look of the model running on the characteristics of distinguishing between outbreak and non-outbreak cases without predetermined classification cutoff. The curve in the case of Random Forest is far above the random baseline (the straight gray line) which proves that the model is much better than guessing. The closeness of the curve to the upper left-hand side of the plot depicts that the model is very discriminatory.

This performance is further measured by the Area Under the Curve (AUC) score of 0.847, meaning that the probability that the Random Forest model will rank a randomly chosen outbreak case higher than a randomly chosen non-outbreak case is 84.7. It is usually held that an AUC within this range is very good regarding epidemiological forecasting, especially when the data are noisy and imbalanced like in the case of malaria surveillance. Notably, the high ROC curve is consistent with the sensitivity and balanced accuracy results obtained above and is supportive of the ability

of the model to optimize detection of true outbreaks with realistic classification of non-outbreak periods.

In terms of public health, it can be concluded that the ROC curve and AUC score indicate that the Random Forest model is a reliable decision in terms of the balance between false negatives and false positives. Such a balance is critical in predicting malaria outbreaks because it allows most of the malaria outbreaks to be detected early enough and the false alarms within reasonable bounds to use the available resources. The ROC curve thus reinforces the findings that the most appropriate model to be used in the study is the Random Forest which has the statistical soundness and the usefulness in form of the development of an early warning system of malaria in Kisii County.

FIGURE 5
Roc curves for the models



4.2.4 Class distribution before and after SMOTE

Figure 4 shows how using the Synthetic Minority Over-Sampling Technique (SMOTE) on the dataset of malaria outbreaks affects the results. The initial class distribution as depicted by the bar chart on the left (Before SMOTE) is a very skewed one. The major class, which is No Outbreak (coded 0), prevails in the dataset, whereas the minor class, which is Outbreak (coded 1) is greatly underrepresented. This kind of imbalance is typical of epidemiological surveillance where the frequency of outbreaks is lower compared to non-outbreak periods. Nonetheless, such imbalance presents as a challenge to the machine learning models, which tend to be biased towards the majority class. Consequently, models that are trained using imbalanced data will appear to have a deceptively high overall accuracy but a poor performance in detecting cases of a

minority group, namely, malaria outbreaks. This results in a high possibility of false negatives, or real outbreaks may be overlooked that has serious consequences on the health of the population.

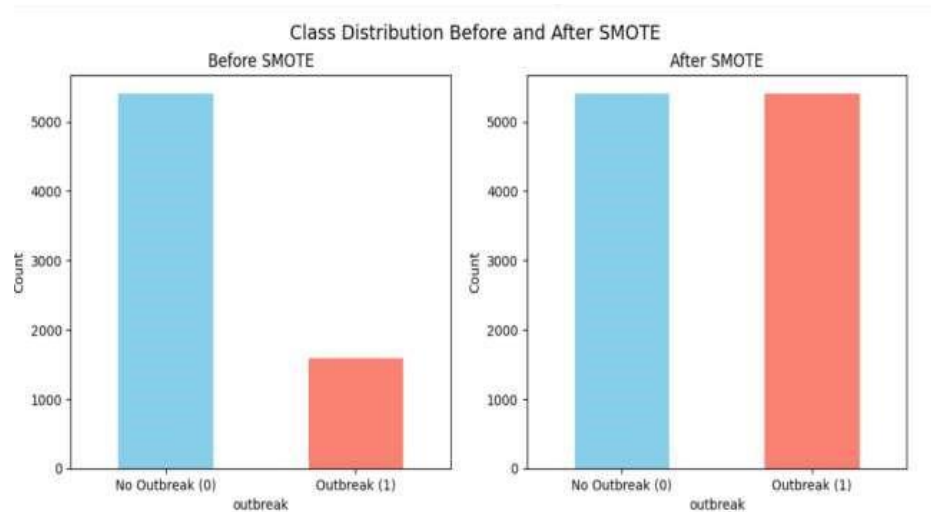
The right hand bar chart titled After SMOTE demonstrates the data following the use of the SMOTE algorithm. The SMOTE algorithm works by synthetically creating new realistic examples of the minority group using existing samples and nearest to the samples in the feature space. This procedure augments the frequency of occurrences of the Outbreak until the data set acquires a more equal dispersion of the two categories. Notably, SMOTE not only replicates the cases of minorities classes but also interpolates new cases of cases, which enhances the representation of the feature space of an outbreak scenario.

This balanced data set can be used to solve the bias of the model when using unbalanced data. By making sure that the Random Forest model is trained on a dataset where the cases of the outbreak are more representative, SMOTE improves the model and makes it able to learn the distinguishing patterns that relate to the outbreaks, which eventually results in higher sensitivity and balanced accuracy. SMOTE offers a methodological protection against under-detection of the minority group in the current context of the study in which the main objective will be to reduce missed outbreaks. The enhanced distribution of classes raises the possibility of successful extrapolation of the model to rare yet important outbreaks, which matches the performance of computations with epidemiological data. Using SMOTE will make sure that the predictive framework is not just a representation

of the prevalence of the majority class but will assign the minority outbreak class the weight it deserves. Such modification enhances the credibility of outbreak detection, but in a broader sense, it highlights the ethical aspect of the computational modelling in the field of public

health- focusing on the identification of key events (minorities) that would help to make timely interventions and rescue lives.

FIGURE 5
Class distribution before smote

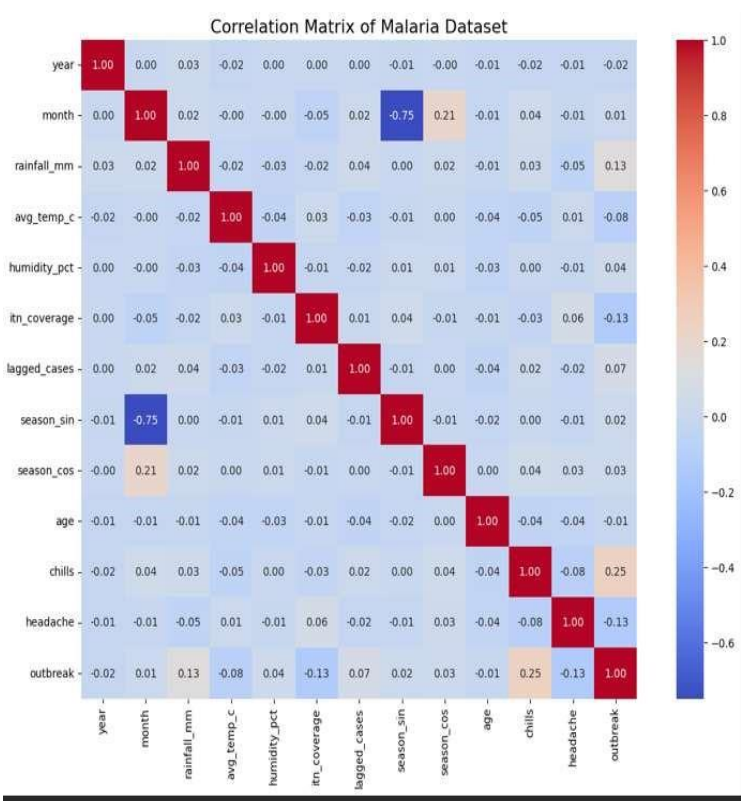


4.2.5 Correlation matrix between the Variables

The correlation matrix of the malaria data is given in figure 4 and shows how the variables are related with each other in pairs. As anticipated, certain high levels of correlations exist. As an illustration, the sine and the cosine transformation of the seasonal variable are perfect negatively correlated (-0.75) indicating the mathematical relationship between them. Likewise, there are high positive correlations among seasons and month (0.75) which again prove seasonality of the features. The majority of environmental factors, including rainfall (rainfall mm), average temperature, and humidity, have rather weak correlations with each other and malaria outbreaks, thus they may play a significant role in determining the risk of malaria transmission, yet not strongly linear. The weak correlations between itn coverage and the lagged malaria

cases and outbreaks also suggest that other more complicated interaction terms are affecting malaria occurrence in a non-linear manner. Out of the demographic and clinical variables, chills demonstrate a moderate positive correlation (0.25) with the occurrence of the outbreak, whereas headache demonstrates a minor negative association (-0.13). Generally, the correlation table indicates that no specific predictor is strongly associated in a linear sense with malaria outbreaks, raising the issue of applying multivariate and non-linear modeling methods in order to represent the interaction amongst environmental, seasonal and clinical predictors malaria prediction.

FIGURE 6
Correlation matrix between the Variables



4.2.6 Features' Relative Importance Plot

Figure 5 shows the feature importance ranking among the Random Forest, XGBoost, and Stacking models and provides an insight into the relative feature importance of epidemiological, environmental, clinical, and intervention-related predictors in predicting malaria outbreaks in the Kisii County. Such rankings are not merely the results of computations, but they are also the epidemiologically significant indicators of intricate determinants of malaria transmission.

In the case of the Rand Forest (RF) model, insecticide-treated net (ITN) coverage, chills, headache, and rainfall were the most influential ones. This ranking brings out the duality between the preventive interventions and clinical manifestations in the establishment of risk of outbreaks. A positive relationship between high ITN coverage and low incidence of malaria indicated the protective role of community level vector control. Such clinical symptoms like chills and headache became critical predictors, presumably, due to their early appearance in malaria infection patterns recorded in surveillance data. Rain also came into the limelight highlighting the ecological route in which increase in precipitation improves breeding of the mosquitoes and consequently transmission. The RF findings thus converge on an Ecosocial explanation of malaria dynamics, as some interaction between intervention uptake, ecological conditions and human health status determines the direction of an outbreak.

The chills, ITN coverage, headache, and rainfall were the most significant features in the XGBoost model but with relatively high weight attached to the clinical

symptoms rather than the environmental and demographic predictors. This focus is indicative of the boosting mechanism of XGBoost that repeatedly targets the minimization of residual error that tends to increase the impacts of available strong direct predictors like symptomatic data. Although this priority enhances the accuracy, it can threaten to underestimate more general ecological and socio-economic determinants, including temperature changes, poverty, or the health-seeking behavior, which affect the risk of malaria indirectly. However, the detection of the ITN coverage and rainfall as well as clinical symptoms indicates the ability of XGBoost to incorporate the biomedical and environmental aspects of malaria prediction.

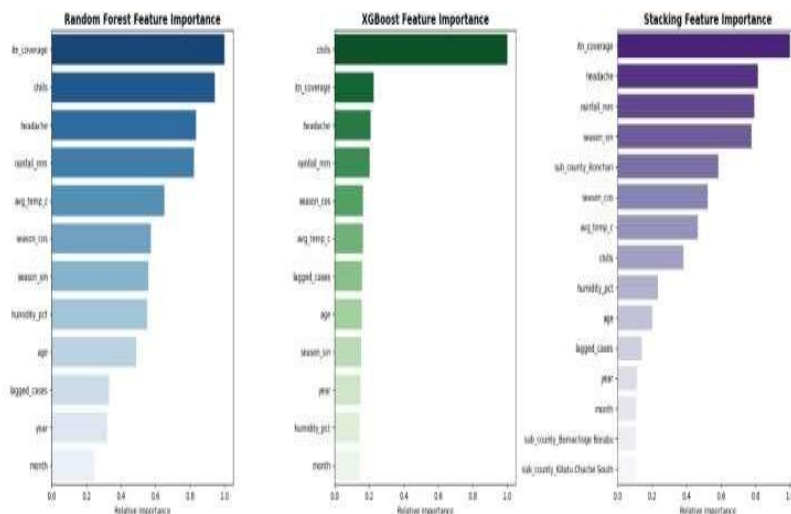
Stacking model (Predictions of various base learners combined) gave a more even-handed feature importance profile. ITN coverage, headache, and rainfall were the most influential variables with the extra weight given to seasonality and sub-county identifiers. This shows that the ensemble framework not only had the capacity to stimulate clinical and intervention variables on individual levels but also identified the homogeneous spatiotemporal setting of malaria dissemination. Seasonal variables emphasize the patterns of recurring outbreaks that result due to change in rainfall and temperature, whereas sub-county identifiers address the heterogeneity of risk that is localized to such a degree that it might be due to variations in altitude, access to health services, or the habitat of the vectors. The combination of these dimensions creates a more refined interpretation of the forces of the outbreak through the Stacking model which highlights the geographic and temporal variability of malaria.

Collectively, the outcomes of feature importance of all the three models indicate that the best explanation of malaria epidemics in Kisii County is a complex interaction of intervention-related variables (ITN coverage), clinical variables (chills, headache), environmental variables (rainfall, temperature, humidity) and local spatiotemporal interaction

(seasonality, sub-county differences). Random Forest is more focused on the balance between interventions, clinical symptoms, and environmental drivers whereas XGBoost is more focused on clinical signals as direct predictors of malaria presence and Stacking is more focused on ecological and geographic heterogeneity. The overall agreement among models on the use of ITN coverage, headache, and rainfall as the main predictors supports the idea of centrality between them in predicting malaria and early warning systems. Significantly, these results confirm the importance of including both the biomedical and socio-ecological determinants in the computational models so that predictive outputs can be helpful in the clinical surveillance and the policy of the health of people.

FIGURE 7

Features' Relative Importance Plot



CHAPTER 5

CONCLUSION AND RECOMMENDATION

This chapter provides a detailed discussion of the findings presented in Chapter Four. The results obtained from the Random Forest, XGBoost, and Stacking models are interpreted in light of the study objectives and compared with existing literature. The chapter also outlines the practical implications of the study, acknowledges its limitations, and proposes directions for future research. Finally, key conclusions are drawn from the research.

5.1 Comparison with Previous studies

This study has shown results in line with the trends that are broadly reported in the field of machine learning, especially in applications relating to health, whereby enhancement of one measure of evaluation may be at the cost of another. In particular, the fact that an increase in the precision is often accompanied by a decrease in the sensitivity or recall represents the long-known precision-recall trade-off in the imbalanced datasets scenario. This trade-off is because the more precise the models are, the more conservative they will act in the prediction of positive cases and this will reduce the occurrence of false positives but not true positives. This imbalance renders the process of epidemiological modelling complex when there is a need to optimize the sensitivity and precision of the modelling simultaneously, where there are rare events of an outbreak compared to non-outbreak periods. An excessively conservative model is likely to miss the outbreaks whereas an excessively sensitive model is likely to experience many false alarms.

This has been an issue that has been pointed out in various fields. (Wainer, 2018) also proved that ensemble approaches and resampling techniques give very varied performance results when focus is put on recall or precision, which proves that none of the methods may maximize both at once. In much the same way, (Hou et al. 2025) discovered, resampling strategies with

XGBoost resulted in models with high total accuracy but poor sensitivity, the importance of which is that high performance on one metric may hide poor performance on another. These observations are reflected in the present work, where Stacking model was more accurate but less sensitive, and where Random Forest achieved a more reasonable result.

The overlap of the results in the various fields of use indicates the existence of a structural feature of machine learning models used in solving imbalanced classification tasks: optimization is multi-objective and an increase in one performance measure usually is accompanied by the decreasing value of a different measure. This property is particularly relevant in the context of malaria outbreak forecasting since there are unequal public health consequences, which are the result of model errors. False negatives (failures to detect outbreaks) are much more harmful than false positives and they can cause postponement of the interventions and cause preventable morbidity and mortality. In this view the high sensitivity shown by Random Forest in this paper cannot be considered a mere chance finding but an anticipated and most beneficial finding which is well situated considering the needs of the public health surveillance. It can afford a superior compromise of accuracy and recall when compared to the other models, which makes sure that the threats of undetected outbreaks are reduced, although at the cost of a slightly lower accuracy. This equilibrium plays a key role in plotting the allocation of resources, enhancing early warning mechanisms and ultimately safeguarding vulnerable groups (such as the Kisii County) in malaria-affected areas.

5.2 Practical Implication

The application of these findings to practice is of great importance to the malaria surveillance and intervention planning in the Kisii County where the burden of disease is uncertain, yet the resources of health systems are minimal. Out of the tested models, it is

obvious that the Random Forest (RF) algorithm is the most suitable one as it has the highest sensitivity (0.723) and balanced accuracy (0.813) which means that it is better able to identify the presence of true outbreaks as well as classify non-outbreak periods correctly. RF can reduce false negatives at the cost of marginal gains in accuracy, which is much less valuable in the context of malaria, where the cost of a missed outbreak can be disastrous in terms of interventions that become delayed, outbreaks that are not controlled, and morbidity and mortality that would be avoided. Compared to XGBoost and Stacking, which showed greater accuracy and reduced sensitivity, RF provides a more effective balance between the ability to believe the unusual case of the outbreak and overall stability, which is why it is the most operationally applicable model in early warning systems. Moreover, the Random Forest has an additional benefit of interpretability by providing a feature importance ranking, which indicated that insecticide-treated net (ITN) coverage, chills, headache, and rainfall were the most frequent predictors. The findings would offer meaningful policy implications to policy-makers: the increased ITN distribution and utilization, the reinforced community based surveillance of symptoms, the track of climatic changes can increase outbreak preparedness. Such ability not only to foresee the consequences but also to inform the specific interventions reveals the potential of RF in addition to the statistical accuracy, it is a decision-supportive tool that is corresponding to the ethical and epidemiological demands of malaria control. These are the reasons why Random Forest can be suggested as the main forecasting model in this study since it contributes to the fact that the surveillance systems will be sensitive, actionable and contextually.

5.3 Limitation of the study

The present study is limited to a number of weaknesses that should be considered to put its findings in perspective. First, the problem of insufficient data significantly reduced the power of the models and their capability to respond to the seasonal and spatial variations of malaria outbreak in Kisii County. The seasonal climatic changes and local ecological situations have a potent influence on Malaria transmission but the dataset was limited in both time and spatial resolution. The shorter time series were disadvantageous in detecting the inter-annual cycles and long-term changes, whereas the short geographic coverage weakened the models in explaining the heterogeneity among sub-counties. This was made worse by the underrepresentation of rural and remote areas where the reporting of health facilities is usually incomplete, inconsistent or delays with inaccuracies. These can cause systematic bias, since models that are trained on an incomplete surveillance data may understate the outbreaks in the areas where surveillance and control are weakest of all.

The predictive precision of the models could be due to a lack of hyperparameter optimization. The method of this research was based on the rather simple threshold tuning as opposed to the more formal optimization methods like grid searching, random searching, Bayesian optimization and evolutionary algorithms. Because of ensemble models such as Random Forest and XGBoost are very sensitive to hyperparameter settings (i.e. number of trees, maximum depth, learning rate), there was no systematic tuning used and thus performance was limited and conservative estimates of predictive capacity were made. Because of this, the models can fail to represent their theoretical potential in an appropriate manner when sufficiently calibrated.

Although the predictor set used in the given study included the epidemiological, clinical, environmental, and intervention-related variables, it did not include a number of major socio-economic and entomological factors that are pivotal to the malaria transmission dynamics. An example is the aspect of human mobility whereby the parasites are exposed to new locations as a result of human migration, and new breeding grounds are formed by land use patterns like deforestation and agricultural developments. On the same note, the field entomological variables that include the distribution, density and species composition of *Anopheles* mosquitoes are the key determinants of transmission risk. The models have excluded these variables, which could have missed key malaria transmission pathways, and they can predict less and understand malaria less.

Combined, these shortcomings show that, though the models used in this paper offer useful information in predicting malaria outbreaks, they must be viewed as rough predictions as opposed to those that are accurate. The shortcomings highlight the necessity of more enriched data sets and more rigorous methodological optimization and the capacity to incorporate a greater variety of determinants in order to construct a set of models that do not only perform better in terms of accuracy but also produces a more holistic and context induced insight into malaria transmission processes.

5.4 Recommendation for Future Studies

Future studies on malaria outbreak prediction must solve the weaknesses stated in the current research by following various lines of methodological and data enhancement techniques. To begin with, the dataset is to be extended in terms of the geographical and time dimension. The ecological and socio-economic risk factors associated with malaria transmission in the Kisii County are very heterogeneous with

rural regions having a different risk factor than urban centers. However, these rural areas are often underrepresented in the surveillance statistics because of missing or incomplete reporting of health facilities. An extension of geographic coverage to those regions and also an extension of the time scale of datasets to span longer time series would enable models to better reflect seasonal period, inter-annual variability and changes in malaria pattern over time. Longer time series are especially significant to the identification of the impact of climate variability (e.g., El Nino-Southern Oscillation events) as well as to the calibration of models to identify changes in transmission dynamics that move slowly.

Better data partnerships and community-based surveillance mechanisms are supposed to be in focus. The collaboration with local health departments can help to achieve a better timeliness of reporting, whereas community-based solutions including mobile health reporting, household surveys, or village-level sentinel sites can help bridge the gaps of official surveillance. Such measures would decrease systematic biases due to underreporting and come up with more detailed and richer datasets to forecast outbreaks.

More stringent hyperparameter optimization procedures should be selected in future research. This research applied simple threshold adjustments in the tuning of the models, which, though, functional, probably underestimated the true potential of the algorithms. Further refined optimization methods including grid search, random search, Bayesian optimization and genetic algorithms may be used in systematic search of the best parameter configurations. Such approaches would not only improve the accuracy of the model, but also the generalizability of the model, which guarantees that models are reliable in various epidemiological settings.

Predictive frameworks need to include more variables (that are more multi-dimensional) that capture the multidimensional nature of malaria transmission. Vulnerability and health-seeking behavior are affected by socioeconomic influences including poverty, level of literacy, and access to healthcare. The patterns of land use (e.g., deforestation, irrigation projects, the expansion of agriculture, etc.) may cause the breeding areas of mosquitoes to develop or disappear. Movement of human beings can also be used to provide insights on the movement of parasites between sub-counties, which may trigger new outbreaks. Malaria ecology focuses on entomological determinants such as the distribution, density and species composition of the mosquitoes of *Anopheles*, which were not covered in this study. The combination of these predictors would result in a more holistic and ecologically based model, which could be able to explain the occurrence of outbreaks as well as the occurrence of outbreaks in certain locations.

The research should investigate the combination of mechanistic models and machine learning methods in the future. Although machine learning does a superb job at discovering statistical trends in broad data sets, mechanistic models are explicitly written descriptions of biological and ecological processes of interest, including life cycles of mosquitoes, the incubation of parasites, and the effect of intervention. Integrating these would have hybrid models that are explanatory and predictive. As an example, machine learning can obtain biologically-based characteristics offered by mechanistic models, whereas the non-linear relationships that are missed by traditional methods can be found through the use of ML. These multi-faceted representations would enable researchers to model the interaction of climate variability, human behavior as well as ecology of the vectors in a form that is not only computationally powerful but also epidemiologically significant.

5.5 Conclusion

This research paper has shown the possibility of ensemble machine learning models predicting outbreaks in malaria, as well as identifying the strong and weak sides of the study. Random Forest was found to be the most sensitive one of the models tested, which made it especially suitable when the reduction of false negatives is of high priority. This is of great importance in malaria surveillance because missed outbreaks can result in delayed response, rapid disease transmission and avoidable deaths. Conversely, XGBoost and Stacking models were found to be more precise, making them useful in a scenario where one cares about the reliability of positive predictions for example in a scenario with limited resources, and it is critical to minimize the number of false alarms. The similarity of the performance of the models highlights the strength of the ensemble methods in epidemiological forecasting and how they can trade off various performance metrics.

Although the findings have constraints linked to data, underreporting in rural setting and optimization of the hyperparameter, they still offer valuable insights in the model selection and methodology optimization. The outcome of the feature importance analysis supported the primary position of the insecticide-treated net (ITN) coverage in the development of malaria risk, as both Random Forest and Stacking have ranked it as the strongest predictor. The significance of this result is that it confirms the need to support and extend ITN distribution programs, especially in the sub-counties that continue to have the lowest coverage. In the meantime, XGBoost has highlighted the importance of the clinical signs like chills, so it is possible to state that the surveillance system is also supposed to focus on monitoring of the symptoms as one of the early-warning systems of the possible outbreaks.

The lessons obtained based on these findings are two. On the one hand, the enhancement of the ITN distribution and utilization as a type of the vector-control intervention is an inevitable component of the transmission reduction. Conversely, enhancement of the monitoring of symptoms by training frontline health workers on the need to recognize early clinical signs can enhance timely intervention and response. In the future, future research must seek to include a wider set of environmental, socio-economic and entomological factors i.e. human mobility, land use change and distribution of mosquito species to develop more holistic predictive frameworks. Also, the further progression of the model interpretability, by means of the SHAP values, will assist in making sure that the statements of the predictions are not only accurate but also readable and constructive to the policymakers.

References:

- Abeku, T. A., et al. (2004). Malaria epidemic early warning and detection in African highlands. *Trends in Parasitology*, 20(9), 400–405.
- Acharya, B. K., Cao, C., Xu, M., Chen, W., Pandit, S., & Zhu, G. (2018). Modeling the spatially varying risk of malaria in Nepal: A hierarchical approach. *International Journal of Environmental Research and Public Health*, 15(5), 918. <https://doi.org/10.3390/ijerph15050918>
- Adongo, P. B., Asoala, V., & Adongo, B. (2021). Risk perception and the use of insecticide-treated nets in malaria-endemic regions. *Malaria Journal*, 20(1), 255. <https://doi.org/10.1186/s12936-021-03789-2>
- Alemu, K., Worku, A., Berhane, Y., & Kumie, A. (2011). Spatiotemporal clusters of malaria cases at village level, northwest Ethiopia. *Malaria Journal*, 10(1), 18. <https://doi.org/10.1186/1475-2875-10-18>
- Araya, S., et al. (2021). Ensemble learning approaches for predicting infectious disease outbreaks: A systematic review. *Infectious Disease Modelling*, 6, 1008–1024.
- Benach, J., Muntaner, C., Solar, O., Santana, V., & Quinlan, M. (2010). *Employment, work and health inequalities: A global perspective*. World Health Organization.
- Bhatt, S., Weiss, D. J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., & Gething, P. W. (2015). The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature*, 526(7572), 207–211. <https://doi.org/10.1038/nature15535>
- Bodnar, T., et al. (2020). Machine learning frameworks for infectious disease prediction and surveillance. *Frontiers in Public Health*, 8, 307.
- Brownlee, J. (2020). *XGBoost with Python: Gradient boosted trees with XGBoost and scikit-learn*. Machine Learning Mastery.
- Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>

- Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage.
- Cutler, D. R., et al. (2007). Random forests for classification in ecology. *Ecology*, 88(11), 2783–2792.
- Danioko, S., Tapo, A. A., & Tembine, H. (2024). Machine intelligence in Africa: A survey. *arXiv Preprint*. <https://arxiv.org/abs/2402.02218>
- Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*.
- Eke, R. A., & Chukwu, C. (2019). Malaria prediction in endemic regions using machine learning approaches. *International Journal of Computer Applications*, 177(7), 1–6.
- Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 4(1). <https://doi.org/10.1162/99608f92.63bb64f6>
- Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4), 44.
- Ghebremeskel, T., & Ali, E. (2020). Application of machine learning in malaria prediction and control in sub-Saharan Africa. *Journal of Public Health in Africa*, 11, 108–115. <https://doi.org/10.4081/jphia.2020.1152>
- Githeko, A. K., & Ndegwa, W. (2001). Predicting malaria epidemics in the Kenyan highlands using climate data: A tool for decision makers. *Global Change & Human Health*, 2(1), 54–63. <https://doi.org/10.1023/A:1011943131643>
- Githeko, A. K., & Ndegwa, W. (2020). Predicting malaria outbreaks in Kenya: An application of machine learning algorithms. *African Journal of Infectious Diseases*, 14(3), 115–122. <https://doi.org/10.5897/AJID2020.0035>
- Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73, 220–239. <https://doi.org/10.1016/j.eswa.2016.12.035>
- Hay, S. I., et al. (2009). A world malaria map: *Plasmodium falciparum* endemicity in 2007. *PLoS Medicine*, 6(3), e1000048. <https://doi.org/10.1371/journal.pmed.1000048>
- Jiang, Q., Shi, X., & Zhang, L. (2021). Predicting malaria outbreaks with ensemble learning: An application in high-risk regions of China. *Computers in Biology and Medicine*, 137, 104810. <https://doi.org/10.1016/j.compbiomed.2021.104810>

- Jiang, T., Shi, W., & Zhang, C. (2021). Hybrid ensemble approaches to malaria outbreak prediction in high-risk areas. *BMC Infectious Diseases*, 21(1), 1105. <https://doi.org/10.1186/s12879-021-06865-y>
- Kamau, L., Kweka, E. J., & Mwangangi, J. (2020). Machine learning techniques for predicting malaria risk in East Africa. *International Journal of Health Geographics*, 13, 23. <https://doi.org/10.1186/1476-072X-13-23>
- Lazaro, S., et al. (2023). Malaria species prevalence among asymptomatic individuals in four regions of Mainland Tanzania. *medRxiv*. <https://doi.org/10.1101/2023.12.28.23300584>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 4765–4774.
- Macharia, P. M., Ouma, P. O., & Snow, R. W. (2021). Spatial models for malaria risk prediction in East Africa. *Malaria Journal*, 20(1), 1–13. <https://doi.org/10.1186/s12936-021-03740-6>
- Midega, J. T., et al. (2012). Wind direction and proximity to larval sites determines malaria risk in Kilifi District in Kenya. *Nature Communications*, 3, 674.
- Molnar, C. (2022). *Interpretable machine learning* (2nd ed.). Leanpub. <https://christophm.github.io/interpretable-ml-book/>
- OduNkiruka, O., & Prasad, R. (2020). Prediction of malaria incidence using climate variability and machine learning. *Informatics in Medicine Unlocked*, 22, 100508. <https://doi.org/10.1016/j.imu.2020.100508>
- Ogutu, J. O., Piepho, H. P., & Cissé, A. (2020). Climate variability and malaria transmission in Kisii County, Kenya. *International Journal of Environmental Research and Public Health*, 17(15), 5508. <https://doi.org/10.3390/ijerph17155508>
- WHO. (2021). *Ethics and governance of artificial intelligence for health*. World Health Organization.
- WHO. (2022). *World malaria report 2022*. World Health Organization.
- Zhou, G., Minakawa, N., Githeko, A. K., & Yan, G. (2004). Association between climate variability and malaria epidemics in the East African highlands. *Proceedings of the National Academy of Sciences*, 101(8), 2375–2380. <https://doi.org/10.1073/pnas.0308714100>

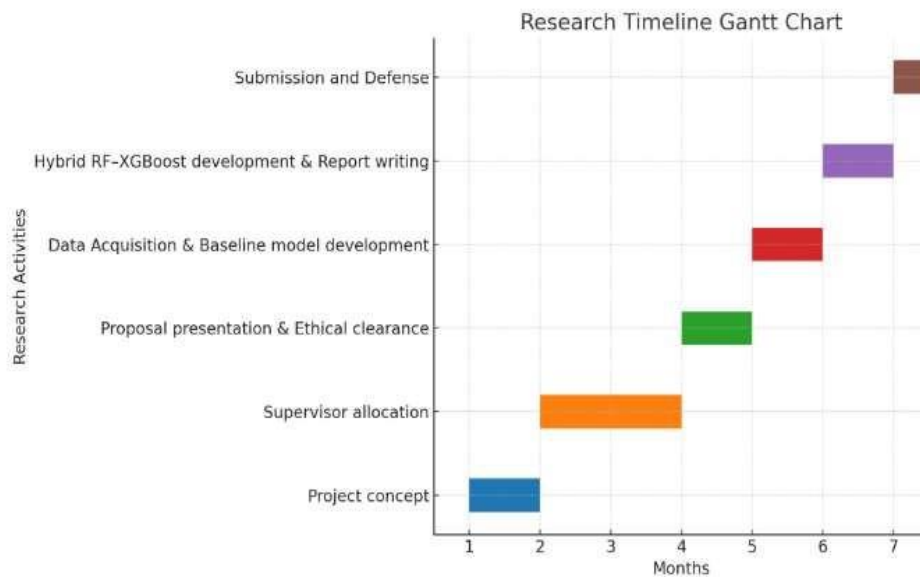
APPENDIX
RESEARCH TIMELINE AND BUDGET
RESEARCH TIMELINE

Month	Activities	Details
Month 1	Project concept	Present the project concept to the panel
Month 2-3	Supervisor allocation	Work on the proposal, meet the supervisor, conduct detailed review on Random Forest and malaria models
Month 4	Proposal presentation Ethical clearance	Prepare my research proposal for presentation. Revise and submit proposal for departmental and ethics review. Finalize the research report and incorporate the panel feedback.
Month 5	Data Acquisition Baseline model development Develop Random Forest and optimize hyper parameters.	Clean and integrate datasets; handle missing values and inconsistencies. Implement statistical and machine learning baselines.
Month 6	Build Hybrid RF–XGBoost ensemble using stacked and weighted soft voting approaches. Finalize writing the report	Evaluate using accuracy, ROCAUC, F1-score. Apply XAI tools (SHAP, LIME) for model interpretability.
Month 7	Submission and Defense	Prepare for the dissertation defense

Budget

Item	Estimated Value
Laptop maintenance and cloud computing costs	5500
Internet	4000
Contingency (unexpected costs, travel)	4000
Data collection expenses and field travel	5500
Public health consultations	2000
Software costs	1000
Total cost	22,000

GANNT CHART





KCA UNIVERSITY SCIENTIFIC & ETHICS REVIEW COMMITTEE

Thika Road, Ruaraka
P.O. Box 56808-00200 Nairobi Kenya
Plot Lines: +254 20 8070408/9
Tel: +254 20 3537842
Fax: +254 20 8561077
Mobile: +254 734 888022, 710 888022
Email: kca@kca.ac.ke
Website: www.kca.ac.ke

REF: KCAU/SERC/SOT013

DATE: 18TH AUGUST 2025

TO: NYABUTO JOYLINE MOGITI (23/03130)

Dear Sir/Madam,

RE: A RANDOM FOREST MODEL TO PREDICT MALARIA OUTBREAKS: A CASE STUDY OF KISII COUNTY

This is to inform you that the KCA University Scientific Ethics Review Committee (KCAUSERC) has reviewed and approved your research proposal. Your application approval number is **KCAUSERC/SOT013**. The approval period is **18th August 2025 – 18th August, 2026**. This approval is subject to compliance with the following requirements.

- i. Only approved documents, including informed consents, study instruments, and MTAs, will be used.
- ii. All changes, including (amendments, deviations, and violations), are submitted for review and approval by **KCAUSERC**.
- iii. Death and life-threatening problems and serious adverse events or unexpected adverse events, whether related or unrelated to the study, must be reported to **KCAUSERC** within 72 hours of notification.
- iv. Any changes, anticipated or otherwise, that may increase the risks or affect the safety or welfare of study participants and others or affect the integrity of the research must be reported to **KCAUSERC** within 72 hours.
- v. Clearance for export of biological specimens must be obtained from relevant institutions.
- vi. Submission of a request for renewal of approval at least 60 days before expiry of the approval period. Attach a comprehensive progress report to support the renewal.
- vii. Submission of an executive summary report within 90 days upon completion of the study to **KCAUSERC**.

Before commencing your study, you will be expected to obtain a research license from the National Commission for Science, Technology and Innovation (NACOSTI) <https://research-portal.nacosti.go.ke> and also obtain other clearances needed.

Yours sincerely,



Dr. Caroline Ntara,
Chairperson,
KCA University Scientific & Ethics Review Committee.



National Commission for Science, Technology and Innovation -
National Commission for Science, Technology and Innovation -
National Commission for Science, Technology and Innovation -
National Commission for Science, Technology and Innovation -
NATIONAL COMMISSION FOR
SCIENCE, TECHNOLOGY & INNOVATION
National Commission for Science, Technology and Innovation -
National Commission for Science, Technology and Innovation -
National Commission for Science, Technology and Innovation -
National Commission for Science, Technology and Innovation -
Date of Issue: 22/October/2025

Date of Issue: 22/October/2025



License No: NACOSTI/P/25/4181058

Applicant Identification Number

Verification QR Cod



See overleaf for conditions

The National Commission for Science, Technology and Innovation, hereafter referred to as the Commission, was established under the Science, Technology and Innovation Act 2013 (Revised 2014) herein after referred to as the Act. The objective of the Commission shall be to regulate and assure quality in the science, technology and innovation sector and advise the Government in matters related thereto.

CONDITIONS OF THE RESEARCH LICENSE

1. The License is granted subject to provisions of the Constitution of Kenya, the Science, Technology and Innovation Act, and other relevant laws, policies and regulations. Accordingly, the licensee shall adhere to such procedures, standards, code of ethics and guidelines as may be prescribed by regulations made under the Act, or prescribed by provisions of International treaties of which Kenya is a signatory to.
2. The research and its related activities as well as outcomes shall be beneficial to the country and shall not in any way;
 - i. Endanger national security
 - ii. Adversely affect the lives of Kenyans
 - iii. Be in contravention of Kenya's international obligations including Biological Weapons Convention (BWC), Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO), Chemical, Biological, Radiological and Nuclear (CBRN).
 - iv. Result in exploitation of intellectual property rights of communities in Kenya
 - v. Adversely affect the environment
 - vi. Adversely affect the rights of communities
 - vii. Endanger public safety and national cohesion
 - viii. Plagiarize someone else's work
3. The License is valid for the proposed research, location and specified period.
4. Neither the license nor any rights thereunder are transferable.
5. The Commission reserves the right to cancel the research at any time during the research period if in the opinion of the Commission the research is not implemented in conformity with the provisions of the Act or any other written law.
6. The Licensee shall inform the relevant County Director of Education, County Commissioner and County Governor before commencement of the research.
7. Excavation, filming, movement, and collection of specimens are subject to further necessary clearance from relevant Government Agencies.
8. The License does not give authority to transfer research materials.
9. The Commission may monitor and evaluate the licensed research project for the purpose of assessing and evaluating compliance with the conditions of the License.
10. The Licensee shall submit one hard copy, and upload a soft copy of their final report (thesis) onto a platform designated by the Commission within one year of completion of the research.
11. The Commission reserves the right to modify the conditions of the License including cancellation without prior notice.
12. Research, findings and information regarding research systems shall be stored or disseminated, utilized or applied in such a manner as may be prescribed by the Commission from time to time.
13. The Licensee shall disclose to the Commission, the relevant Institutional Scientific and Ethical Review Committee, and the relevant national agencies any inventions and discoveries that are of National strategic importance.
14. The Commission shall have powers to acquire from any person the right in, or to, any scientific innovation, invention or patent of strategic importance to the country.
15. Relevant Institutional Scientific and Ethical Review Committee shall monitor and evaluate the research periodically, and make a report of its findings to the Commission for necessary action.

National Commission for Science, Technology and
Innovation(NACOSTI),
Off Waiyaki Way, Upper Kabete,
P. O. Box 30623 - 00100 Nairobi, KENYA
Telephone: 020 4007000, 0713788787, 0735404245
E-mail: dg@nacosti.go.ke
Website: www.nacosti.go.ke