# A MALARIA EARLY WARNING ALERT APPLICATION FOR HEALTH OFFICIALS
**(*Imalaria*)**

By

CHRISTINE MWAMSIDU

096408

A Dissertation Submitted to the Faculty of Information in partial fulfillment of the requirements for the award of a degree in Master of Science in Computing and Information Systems.



**Strathmore University**

March 2024

## Declaration and Approval

I Christine Mwamsidu declare that this research has not been previously submitted to any other University for the award of a Degree in Master of Science in Computeing and Information Systems.

Sign: ________________________ Date: 29th March 2024

Student Name: Christine Mwamsidu

Registration Number: 096408

SUPERVISOR'S DECLARATION

This research proposal has been submitted for review with my approval as a university supervisor.

Sign: ________________________ Date: 29th March 2024

Supervisor's Name: Dr Joseph Orero

Faculty of Information Technology,

Strathmore University

## Table of Contents

## Abstract

Malaria is a critical public health issue that must not be overlooked since it affects majority of the world's population. The goal of the study was to determine how to use confirmed malaria case data, and current climatic conditions to forecast malaria epidemics and offer timely information to health officials. This research will help to inform and build a malaria early warning alert application that can offer public health experts with early warning signs and prediction of malaria outbreaks.

The first step was selection of the variables that has a significant impact on malaria transmission. From literature review, malaria incidences, rainfall/precipitation, humidity, and temperature were selected. Data collection was done where malaria data was collected from Kenya Health Information System (KHIS) and incidence rates calculated based on cases and population. Weather data was also collected from historical data API accessed through this link: https://open-meteo.com/. Hourly data collected was aggregated and averaged for the month derived. Selection of the best machine learning algorithm for the malaria early warning alert system was conducted and six models including Decision Tree Regressor, Linear Regression, Random Forest Regressor, Ada Boost Regressor, Gradient Boosting Regressor, SVR and decision tree were tested. Random forest Regressor which had the best score was selected.

Training of the model on the collected data to make predictions and provide alerts to health officials at facility level was done using 70% of the data. Evaluation of the performance of the model using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error RMSE was also conducted using 30% of the data.

Deployment of the model was done using TensorFlow's SavedModel format which was integrated into the React application to create the malaria early warning alert system. The UNHCR model for emergency response was adopted to calculate the alert/epidemic threshold. This model considers the threshold for a malaria outbreak to be 1.5 times the baseline over the previous three weeks. Alert (or 'epidemic threshold') and action thresholds were used to provide information to the public health staff. The rest of the data was labelled as normal threshold.

## Dedication

This research work is dedicated to all stakeholders involved in the elimination of Malaria a life-threatening disease. Special gratitude also goes to my husband Moses Emalu for his unwavering support and encouragement, and to all my mentors and lecturers at Strathmore University and colleagues at Amref Health Africa in Kenya, without their help and experience, this study would not have been possible.

## Acknowledgements

I am grateful to Amref Health Africa in Kenya for allowing me to study and increase my knowledge and skills in monitoring and evaluation. The drive for innovation and the use of technology inspired me to take up this course. My supervisor, Mr. Benson Ulo, and my colleagues have also supported me in accomplishing my goals. I sincerely thank my supervisor and lecturer, Dr. Joseph Orero. His mentorship and supervision, constant accessibility, availability, and dedication greatly inspired me. Lecturers in the Faculty of IT at Strathmore University revealed to me the true meaning of innovation using technology. I also acknowledge Moses Emalu, my husband for his moral and financial support. Finally, without the grace of the Almighty God, I would not have reached this point.

## Abbreviations/Acronyms

| | |
|---|---|
| ACT | Artemisinin-Based Combination Therapy |
| AFI | Acute Febrile Illness |
| API | Application Programming Interface |
| ARIMA | Autoregressive Integrated Moving Average |
| CDC | Centre for Disease Control |
| CHV | Community Health Volunteers |
| EMR | Electronic Medical Recording Systems |
| GEE | Google Earth Engine |
| HCW | Health Care Worker |
| HMIS | Health Management Information System |
| LIS | Laboratory Information Systems |
| NDVI | Normalized Difference Vegetation Index. |
| RDTs | Rapid Diagnostic Tests |
| WHO | World Health Organisation |
| IPTi | Intermittent Preventive Treatment of infants |
| IPTp | Intermittent Preventive Treatment pregnant women |
| SMC | Seasonal Malaria Chemoprevention |
| MDA | Mass Drug Administration |
| MAE | Mean Absolute Error |
| MSE | Mean Squared Error |
| RMSE | Root Mean Squared Error |

## List of Figures



## List of tables

# Chapter 1: Introduction

## 1.1 Background

Malaria, a disease of major public health importance, is caused by the parasitic protozoan and is transmitted by the Anopheles mosquito from one person to another (Cowman, Healer, Marapana, & Marsh, 2016). It affects majority of the world's population and India accounts for 80% of the global malaria burden. At present, more than 100 nations and regions are susceptible to malaria transmission, as reported by Starnes et al., (2018). Africa is disproportionately affected by this disease, accounting for 95% of all malaria cases and 96% of deaths worldwide in the year 2020, according to the World Health Organization, (2021).

It is estimated that in Kenya, between 13% to 15% of outpatient consultations in medical facilities are due to malaria. The likelihood of malaria transmission and infection is primarily influenced by altitude, rainfall patterns, and temperature. As a result, there are significant variations in the prevalence of malaria across geographic regions and seasons. Roughly 70% of the Kenyan population is at risk of contracting malaria, including 13 million individuals living in areas where malaria is endemic and another 19 million residing in highland areas that are prone to seasonal and epidemic transmission (Division of National Malaria Programme (DNMP) [Kenya] and ICF, 2021).

The Global Technical Strategy for Malaria 2016–2030 shows that surveillance involves tracking the disease and responses, then acting based on the received data. According to (Maharaj, 2017)), there is need for early warning for malaria epidemics to allow measures to be implemented to mitigate against it. This can enhance the National Malaria Programs ability to develop perfect early warning system that could predict epidemics up to six months or a year in advance (Merkord et al., 2017). Systems that have been automated can deliver a solution to these problems and also offer alternative solutions through an early detection mechanisms that helps to control the spread of the disease, which can save thousands of lives (Modu et al., 2017).

In view of this, this study aimed to develop a malaria outbreak detection alert for health officials using machine learning techniques. This will identify possible disease outbreaks and/or malaria hotspots. Such an application has the potential to reduce malaria incidences and improve the malaria surveillance system for a country.

## 1.2 Problem Statement

According to the World Health Organisation (2018), quite a lot countries with disproportionately high malaria burden have suboptimal surveillance systems which are are incapable to assess disease distribution and trends. In a majority of the countries that account for 85% of global malaria cases, trends are based on models rather than routine surveillance (WHO, 2022b) making it hard to enhance responses and respond to outbreaks. Moreover, responses at the facility level are sub optimal or unavailable. Therefore, it is essential to strengthen malaria surveillance systems to enable health officials to respond promptly and effectively in endemic areas, primarily to prevent outbreaks, monitor progress, and ensure accountability of governments and the global malaria community (World Health Organization, 2021).

To improve surveillance systems, a malaria outbreak alert system is important to a country. Health officials will be provided with key information needed to create interventions for curbing malaria infections early. If countries are capacitated, malaria which is a treatable disease can be eliminated through well-informed strategies.

## 1.3 Aim

The purpose of the study was to design and implement an application that will capacitate health officials with an early warning on malaria outbreaks. This will be through employing machine learning techniques using malaria incidences, metrological and geospatial data.

## 1.4 Specific Objectives

(i) To identify challenges affecting malaria outbreak or early warning detection.

(ii) To review related systems and models for predicting malaria outbreaks

(iii) To develop an application for malaria outbreak alert

(iv) To test the functionality of the application

## 1.5 Justification

Malaria eradication refers to the complete elimination of malaria infections worldwide caused by human malaria parasites through deliberate efforts. The pace of progress towards this goal varies depending on a country's health system strength, investment in malaria control, and other factors (Widdowson et al., 2003).

Effective surveillance is crucial to achieving goals towards malaria control and elimination (Bridges et al., 2012). Timely and accurate information about malaria transmission is essential and can be beneficial even in highland and semi-arid areas, where unstable malaria transmission and epidemics are linked to seasonal variations in temperature and rainfall. As populations in epidemic-prone areas lack immunity, malaria epidemics can result in severe morbidity and mortality. Early detection of such epidemics can allow for proactive interventions both preventive and curative (Baeza et al., 2014).

Resurgent malaria epidemics are a significant concern during malaria elimination efforts, and information on the location and timing of residual transmission hotspots is essential to facilitate rapid and effective public health responses. Therefore, innovative malaria information systems that enable data sharing among stakeholders and direct public health responses are urgently needed (Hemingway et al., 2016).

To prioritize malaria infection prevention, many countries have adopted WHO recommendations for key interventions. Despite significant progress, malaria remains a serious public health threat in Sub-Saharan Africa and other developing nations. Therefore, innovative prevention strategies are critical.

### 1.6 Scope and Limitation

The researcher will develop the system according to the Kenyan context. The country needs such a system hence the researcher will be able to provide a platform where this can be improved further. The researcher will not examine the broader question of how the data collected will influence public health decision making.

## Chapter 2: Literature Review

### 2.1 Introduction

This chapter will give an in-depth understanding of the research. The focus will be on facts about malaria, malaria situational analysis, tourism and malaria, challenges affecting malaria outbreak detection and a review of related systems and machine learning techniques for detecting malaria outbreaks.

### 2.2 Malaria

Malaria is an illness caused by a parasite that is transmitted through mosquito bites. Symptoms of malaria include fever, chills, and flu If it is not treated, people might develop severe complications and worse succumb from it. After being bitten by an Anopheles mosquito, there is a period known as the incubation period before symptoms start to appear which usually lasts between 7 to 30 days, with P. falciparum having shorter incubation periods compared to P. malariae, which tends to have longer ones (CDC, 2018).

Malaria is predominantly found in economically disadvantaged regions of tropical and subtropical areas around the world, where it remains a significant contributor to illness and mortality. Nevertheless, malaria is preventable (Jamison & World Bank, 2006). In regions where the transmission is high, the most disposed populations are young children with low immunity and pregnant women who also have low immunity suppressed by the pregnancy. The catastrophic costs to malaria really affect individuals, families, communities, and nations (CDC, 2018).

The World Report on Malaria 2021 by the World Health Organization highlights the key facts about the global malaria situation. In 2020, there were estimated 241 million cases of malaria worldwide with deaths standing at 627,000 cases. Inexplicably, WHO African Region is disproportionately affected by malaria, with a burden of 95% of reported malaria cases and 96% of deaths that are related to malaria in 2021. Children under 5 years old make up about 80% of all malaria deaths

In Kenya, malaria epidemiology of malaria mainly influenced by rainfall patterns, altitude and temperature. The country has four epidemiological zones for malaria, with the first one being the Endemic zone around Lake Victoria, western part of Kenya, and coastal regions. The spread of malaria in these areas is influenced by features such as rainfall, temperature, and humidity, which affect the vector life cycle. Malaria transmission is strong throughout the year in these regions, with high entomological inoculation rates recorded annually of 30 to 100 (Degefa et al., 2017).

The Seasonal malaria transmission zone are the arid and semi-arid areas of northern and southeastern Kenya, and there are short periods of severe malaria transmission during the rainy season in these areas. The formation of stagnant water pools due to elevated temperatures during these seasons serves as breeding grounds for mosquitoes, which contribute to the spread of malaria. In addition, seasonal climatic conditions such as flooding and El Niño can lead to malaria epidemics, with high morbidity rates due to low immunity levels in the population(Degefa et al., 2017).

Thirdly, the western highlands are categorized as malaria epidemic-prone areas, and the epidemics vary from year-to-year. Epidemics occur when there are sustained temperatures of around 18° C, which favors vector breeding and increases the intensity of transmission. The whole population is vulnerable during these outbreaks, and case fatality rates can be up to ten times greater than in regions where malaria is frequent (Degefa et al., 2017).

Finally, The central highlands, including Nairobi, are considered low-risk malaria areas. Normally, the temperatures are too low to allow the completion of the sporogonic cycle needed by the malaria parasite. However, due to climatic changes, increasing temperatures and changes in the hydrological cycle may make these areas suitable for malaria vector breeding, potentially leading to malaria transmission in areas where it previously did not exist (Degefa et al., 2017).

Early detection of a disease outbreak is important in its containment and finally finding the appropriate cure for a disease. However, that is surely not trivial given the various challenges experienced. Currently, several countries with a high burden of malaria have weak surveillance systems, making it challenging for them to assess disease distribution and trends and therefore optimize responses and respond to outbreaks. (Rani et al., 2014).

## 2.3 Challenges affecting malaria outbreak detection

Malaria early warning systems aim to prevent outbreaks and control the spread of the disease by providing timely information on the potential risk of malaria transmission. However, there are several challenges that affect their implementation and effectiveness.

**Suboptimal surveillance system:** Reliable data is crucial for effective malaria early warning systems. However, obtaining accurate data in a timely manner can be a challenge in many areas, particularly in low- and middle-income countries where data collection and management systems are often weak. Inaccurate data can be detrimental in ensuring that there are progressive efforts in detection of an abnormal trend of disease occurrence (Buehler et al., 2004). It is important that an outbreak reporting network to have quality data evaluated according to a minimum criterion agreed upon (Lawpoolsri et al., 2018). As reported by Guthmann et al. in 2007, malaria epidemics remain poorly documented in sub-Saharan Africa, moderately because they occur in remote and rural settings where data collection is inefficient.

**Climate variability and change:** Malaria transmission is strongly influenced by climate factors such as temperature, rainfall and humidity. There is need for a better early Warning System for Sub-Saharan Africa. Climate variability and change can make it difficult to accurately predict the risk of malaria transmission, which can affect the effectiveness of early warning systems. Changes in temperature and precipitation patterns can alter the breeding and survival rates of mosquitoes, leading to changes in the transmission of malaria (Ayanlade et al., 2022). Changes in the distribution of mosquitoes can lead to the spread of malaria to new regions, making early warning systems less effective in predicting and preventing outbreaks Shifts in vector populations and the emergence of new disease vectors could make current early warning systems outdated and ineffective (Nissan et al., 2021).

**Critical gaps in the diagnosis of malaria:** To direct malaria treatment, all cases should be confirmed with a diagnostic test, either RDT or light microscopy, even in low-transmission settings. These are the WHO guidelines. Lack of sufficient diagnostic facilities in the locality prohibited the collection of some specimens that could have been the most useful. Guracha et al., 2011). Wangdi & Clements (2018) also established that poor diagnostic capacity is a factor that affects reporting of malaria cases by HCW, principally in low transmission locations. Furthermore, there is an insufficient number of skilled HCWs, healthcare infrastructure, and laboratory diagnostic capacity to ensure that emerging diseases likely to cause epidemics are dealt with. Consequently, delays in notification of outbreaks can be a major barrier having an effective early warning and response systems (Rajatonirina et al., 2014).

**Minimal financial resources:** Implementing and maintaining effective malaria early warning systems can be expensive, particularly in low- and middle-income countries where funding is often limited. In a study conducted in a district in Zambia that had nine public health facilities, several challenges concerning a lack of resources were identified. Regarding logistical support, only one facility was powered by electricity and five had internet network access, while seven had either a mobile or telephone. (Haakonde et al., 2018). Furthermore, most of the evidence available is from the northern hemisphere, whereas low-income countries with inadequate resources and infrastructure are at the greatest risk of epidemic events (Steele et al., 2016). Governments in developing countries sometimes receive their first information about an outbreak from the media, and not from the official/national reporting system. In economically disadvantaged nations, there may be reluctance to publicly declare an outbreak and cause panic among the population due to the lack of adequate resources to effectively respond, as stated by the World Health Organization (WHO, 2005).

**Minimal end-user engagement:** Community engagement and education are important components of malaria early warning systems. However, engaging communities in these systems can be challenging, particularly in areas where there is low levels of literacy or limited access to information. Current efforts to develop, study, and disseminate digital health have been limited by minimal user engagement. The integration of user input can happen at any stage of the digital health development process. It is possible to involve users as equal partners with designers from the beginning of a project, when the technology is first conceived. Alternatively, user input may be sought only during the final stages of design, through usability testing. The design process can also consider the needs of those who may be indirectly impacted by the technology's use or who are responsible for funding it (Birnbaum et al., 2015). Lessons learned from digital health applied to the eradication program suggest that attaining a population-level impact is undermined by minimal end-user engagement (Guracha et al., 2011). In another study conducted by Toda et al. (2016), suboptimal on job training of HCWs who did not attend classroom trainings and a lack of post training follow up mechanisms including supportive supervision were weaknesses that affected implementation of user engagement strategies in place.

**Lack of coordination:** Another crucial aspect to consider is the involvement of relevant stakeholders. The first level, which is typically the community level, is essential for successful implementation. Additionally, coordination among various sectors is crucial when managing malaria cases. A study conducted in China found that joint coordination was effective in controlling malaria outbreaks and reducing malaria incidence rates. Malaria elimination cannot be achieved through the efforts of a few countries alone, but it must be a concerted international effort (Zhang et al., 2018). Collaboration is necessary when a person with malaria crosses the boundaries of multiple districts, provinces, or even countries to eliminate transmission (World Health Organization et al., 2012).

In conclusion, overcoming these challenges is essential for the effective implementation systems that will provide information early and reducing the impact of malaria. This requires investment in data collection and management systems, health systems, and community engagement and education, as well as the integration of climate information into early warning systems.

### 2.4 Existing malaria early warning systems

A malaria early warning system (MEWS) is a surveillance program designed to predict and prevent outbreaks of malaria by continuously monitoring and analysing data on disease transmission and environmental factors that affect malaria transmission. It uses a combination of epidemiological, entomological, and environmental data to predict changes in malaria transmission patterns and to inform public health interventions to prevent outbreaks (Maharaj, 2017).

This section outlines examples of applications or systems available for malaria outbreak detection alerts. In the literature review conducted, there are no applications that are specifically designed for detecting malaria outbreaks and providing alerts. Generally, data is pulled from the malaria surveillance applications and analysed to determine predictions. As discussed in this section, there are several malaria surveillance systems in place and machine learning techniques that have been used in malaria early warning detection.

### 2.4.1 Information systems to support malaria surveillance

There are several systems used to collect malaria incidence data for surveillance. "SMS for Life" is a system where health care workers would send malaria surveillance data via phone messages. Health workers used their mobile phones to send structured SMS text messages weekly. The study showed that it is feasible to use simple text messages sent via mobile phones to effectively transmit current surveillance data from local health facilities to higher levels. However, the precision of the data provided was not ideal. Further efforts should be directed toward enhancing the quality of surveillance data reported via SMS (Githinji et al., 2014).

MSOS, a web-based portal configured text messaging was also assessed. The system facilitated communication between HCWs and Heath officials or managers. It also monitored notification of diseases and response actions taken by the managers. The study revealed that this intervention expressively improved timely notifications. However, disease response was suboptimal, with response given to only one-fifth of detected cases. These findings are consistent with a study conducted in Tanzania that revealed significant improvements in vital registration coverage through SMS, but limitations in accurately reporting births and deaths events in the community (Toda et al., 2016).

SaTScan is also a free software used for detecting malaria outbreaks. It can analyze spatial, temporal, and space-time data using scan statistics for spatial, temporal, or space-time analysis. Its main objectives are to perform geographical disease surveillance, identify statistically significant spatial or space-time disease clusters, and assess the randomness of disease distribution. It is also used to conduct periodic disease surveillance to detect disease outbreaks early (Coleman et al., 2009).

DHIS2 is a global health information management system around the world. DHIS2 is a web-based platform that is open-source in nature and primarily utilized as a Health Management Information System (HMIS) (DHIS2, 2022). Kenya was the first country in Sub-Saharan Africa to deploy DHIS2 commonly known as KHIS (hiskenya.org). All the Kenyan sub counties and selected health facilities involved are connected to DHIS2 national server using the internet on computers (Open Health Marketplace, 2020).

The system is organized based on administrative units in Kenya. Each of the sub counties (299) have designated health facilities in all the 47 counties currently in existence. Aggregate reporting is done monthly by the health facilities on diseases, commodities and service. Due to limited internet connectivity and absence of infrastructure most rural health facilities send their reports to the sub county health records and information officers who enter the data into the web-based DHIS2 system (Karuri et al., 2014).

## 2.5 Machine learning techniques used in the prediction of malaria outbreaks

Early warning systems are critical in controlling the spread of malaria and saving lives. Machine learning techniques have been used in several research studies to develop malaria early warning systems.

### 2.5.1 Machine learning techniques used in outbreak detection/Early Warning

A study in Nigeria used five supervised machine learning methods to predict malaria outbreaks using meteorological and malaria incidence data from 2010 to 2020. The Naive Bayes model had the highest accuracy of 79.1%, while Support Vector Machine (SVM) was second with an average accuracy of 75.45%. K-Nearest Neighbor had an average accuracy of 70.8%, and Logistic Regression had an average accuracy of 68%. Linear Regression was not recommended due to its low average accuracy of 26.05% (Kazeem & Adebanji, 2021).

Random Forest is another machine learning technique that creates multiple decision trees and combines their results to make a final prediction. This approach can handle complex data and non-linear relationships, making it an effective tool for malaria early warning systems. According to Comert et al., 2020, decision trees provided an accuracy of 100% on test data used. This was similar to a study that showed, Random Forest was high with more than 90% in all evaluation metrics; the accuracy was 90.75%, F-score of 90.73%, Precision of 90.69% and Recall at 90.88%. With the exception of Support Vector Machine (SVM), other classifiers exhibited strong performance with accuracy rates exceeding 70% in all evaluation metrics. However, SVM lagged behind with an accuracy of approximately 60% across the evaluation metrics. This study recommended the use of Random Forest for malaria prediction (Dukuzumuremyi, 2020).

Artificial Neural Networks (ANNs) are another commonly used machine learning technique for predictive modeling. ANNs are designed to simulate the way the human brain works and have been applied to early warning systems for malaria. ANNs can handle large amounts of data, making them a suitable option for malaria early warning systems that need to process large amounts of data from multiple sources. This technique can correlate malaria cases and climate data. In a study conducted in Rwanda, results showed that Artificial Neural Network algorithm performed better with 93.9% and 88.2% of training and testing accuracy respectively in Bugesera district (Thakur & Dharavath, 2019)

Support Vector Machines (SVMs) are a machine learning method used for classification and regression. They have been employed in malaria early warning systems to predict malaria incidence. SVMs are effective at handling high-dimensional data and have been used in various studies to develop early warning systems for malaria. Additionally, SVMs have been used to identify risk factors associated with malaria transmission and to create predictive models for outbreaks (Ngai et al., 2011).

Naive Bayes is a simple machine learning algorithm used for classification. This algorithm has been used in malaria early warning systems to predict the occurrence of malaria outbreaks. Naive Bayes uses probability theory to make predictions, making it a quick and efficient option for malaria early warning systems (Kazeem & Adebanji, 2021).

K-Nearest Neighbors (KNN) is a non-parametric classification method utilized in malaria early warning systems for predicting malaria incidence. KNN functions by identifying the k closest neighbors to a particular data point and producing a prediction based on their values. Due to its simplicity and ease of implementation, KNN is a viable choice for malaria early warning systems (Nkiruka et al., 2021).

Time series analysis is also popular machine learning technique that can be used to predict outbreaks or in early warning systems. This technique can be used to analyze and forecast the spread of malaria over time based on historical data. In studies conducted in South Eastern Iran and Afghanistan, time-series models were effective in predicting malaria incidences with satisfactory accuracy in a malaria early warning system (Ostovar et al., 2016). Both time series and cross-sectional data was used in the model. The dependent variable used was total malaria cases out of local transmission in the desired time and the independent variables were the total rainfall measured in millilitres, minimum and maximum mean temperature (Nkiruka et al., 2021).

In the Netherlands, the ISIS system used a simple linear regression model that was modified to account for factors like seasonal changes, past outbreaks, and long-term trends. This approach required minimal adjustments or model checks. To predict the total number of cases for the current week, a regression line was created by plotting the totals from the previous five years for the nine weeks surrounding the same week. For example, to predict the value for week 20, values from weeks 16 to 24 of the past five years were used to create a regression line (Watkins et al., 2006).

This was also backed by a study aimed at identification of population predictors of malaria outbreaks in endemic municipalities in Colombia to develop an early warning system for malaria outbreaks. The measurement of variables was based on historic registries, and logistic regression was performed to analyze the data. Attitude above sea level (OR 3.65, 95% CI 1.34–9.98), rainfall variability (OR 1.85, 95% CI 1.40–2.44), and the proportion of inhabitants over 45 years old (OR 0.17, 95% CI 0.08–0.38) were factors associated with malaria outbreaks in Colombian municipalities (Mateus & Carrasquilla, 2011).

There are other early warning systems for other diseases like Ebola, Dengue Fever, Lime disease, rift valley fever among others (National Research Council (US) Committee on Climate, 2001). According to (WHO, 2022a) Other early warning systems for diseases include WHO's Early Warning, Alert and Response System (EWARS). The system is specifically designed to enhance the detection of disease outbreaks in emergency settings, such as countries experiencing conflict or recovering from a natural disaster. It is not intended as a routine tool, but rather as a supplementary resource to the national disease surveillance system during emergencies. Congo, for instance, deployed this system after an Ebola outbreak, incurring a cost of USD 438 per detected Ebola case and USD 1.8 per alert received.

A study was conducted to evaluate various methods for detecting outbreaks using public health surveillance data. The study found that there is considerable diversity in the methodologies employed to evaluate outbreak detection techniques. No single approach was found to be superior, and the approaches used were generally Descriptive, Derived, Epidemiological, and Simulation approaches. It is necessary to establish a basic framework for evaluating outbreak detection methods and recommend the use of multiple evaluation approaches to provide a comprehensive and contextualized description of outbreak detection performance (Watkins et al., 2006).

An effective operational early warning system relies on several critical components, including climate forecasts, ongoing epidemiological surveillance or surveillance systems, and environmental observations. These components can serve as inputs for predicting potential outbreaks or other health risks. To strengthen, it is important to consider vulnerability assessments assessing affected populations and risk analysis to know the foreseen extent of the disease impact. Finally, using information from the early warning system, a response public communication strategy can be developed. The population should work together with the health officials to prevent or curb any further damage in case of an outbreak (National Research Council (US) Committee on Climate, 2001).

There are several machine learning techniques that have been used in research studies to develop malaria early warning systems. The choice of machine learning technique will depend on the specific requirements and characteristics of the data being used for the malaria early warning system. The use of machine learning techniques in malaria early warning systems has the potential to improve disease control and save lives.

### 2.5.2 Examples of surveillance data and threshold used



When there is a malaria outbreak, a certain number of cases in a specific location within a certain timeframe must be surpassed, which is referred to as a threshold. There are different types of thresholds, including alert and action/epidemic thresholds, which indicate when further investigation or action is necessary. The UNHCR, (2023), considers the threshold for a malaria outbreak to be 1.5 times the baseline over the previous three weeks. In Kenya, two types of thresholds are used: alert and action/epidemic thresholds. The alert threshold signals to health workers that more investigations are necessary, while the action/epidemic threshold is reached when there is a consistent increase above the alert threshold, indicating that further action or response is necessary. The alert threshold is calculated by finding the third quartile of the number of cases per week over at least five years, and the action/epidemic threshold is calculated by finding the mean plus 1.5 standard deviations of the number of cases per week over at least five years (Division of National Malaria Program, 2020).

.Several studies have shown that malaria parasites and vectors are very sensitive to changes in environmental factors such as rainfall patterns, water temperatures, humidity, and air. Precipitation has the most significant impact, as it creates breeding sites for mosquitoes. Temperature is also important, with a range of 13-35oC being conducive to the breeding of Anopheles mosquitoes. Long-term analysis is necessary to accurately forecast high-risk periods for malaria outbreaks (Ototo et al., 2022).

Similarly, it was confirmed by another study that precipitation has the most significant impact, followed by temperature, which can affect the breeding of Anopheles mosquitoes. To forecast high-risk periods for malaria, it is crucial to consider all relevant parameters and conduct long-term analyses, as relying on one year of data may lead to errors in predicting high-risk periods. Time series data analysis can help identify high-risk periods within a year by extracting relevant data such as NDVI, rainfall, temperature, and the number of malaria cases from previous months (Nyasa et al., 2022).

Google Earth Engine (GEE) provides a platform for collecting climatic data, including surface temperature, climate, atmospheric, and weather data. Using GEE, researchers can analyze time-series remote sensing data quickly and easily by providing fast and accessible processing space and easy access to free remote sensing data (Google, 2022).

Overall, understanding the relationship between environmental factors and malaria outbreaks is essential in preventing and controlling malaria. By implementing thresholds and monitoring climatic conditions, health workers can identify and respond to malaria outbreaks promptly, saving lives and resources (Bashar & Tuno, 2014).

**2.8 Conceptual Model**

The application under development will provide malaria outbreak/early warning alerts and key information on malaria. Long-term data both in the past and future on malaria cases and climate variables including temperature, rainfall and humidity will be pulled to the application through APIs. Data will be modelled through machine learning techniques with thresholds that will allow for the prediction of malaria outbreaks. The users will mainly be health officials

Data to be used in the system will be new malaria case data from surveillance systems counties in Kenya. Climate data will be retrieved from the weather forecasts, GEE among others. Geospatial information will also be used to allow the application to issue alerts to specific health officials in an area up to the facility level.

The conceptual framework is as below:

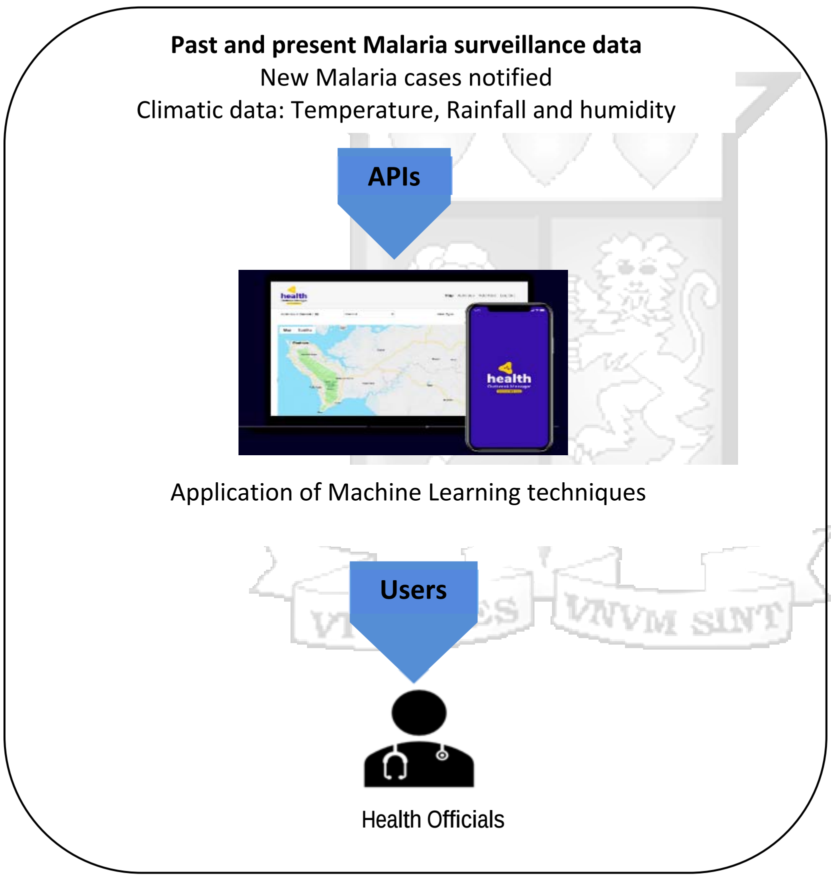



**Figure 1: The conceptual model for the malaria outbreak detection application**

# Chapter 3: Research Methodology

## 3.1 Introduction

This section explains the methodology that will be used to conduct this study and the system design that will be used.

## 3.2 Research Design

A positivistic approach was applied because the research is based on facts that will be measured using the application. A case study from sampled counties will be used to gain in-depth information on malaria detection practices. In this study, qualitative as well as quantitative data collection methods will be applied.

## 3.3 Location of the Study

Due to the researcher's convenience, the study will use secondary data from all the 47 counties that are divided into Endemic, Seasonal malaria, Malaria epidemic and low malaria zones.

## 3.4 Target Population

The target population included National Malaria Control Program staff, public health experts, Health workers and the community living in malaria endemic zones. They will be reached out to provide feedback once the application is developed.

## 3.5 Sampling

Purposive sampling was be applied in the selection of the study sites used to test the application. This is because the researcher was required to test the system where longitudinal data is available. Cluster sampling was be used to select respondents. This is because of the various managerial and technical staff involved.

## 3.6 Data Collection

The project used secondary data from KHIS for malaria cases and population data from Kenya bureau of statistics used to calculate incidence rates per 1,000 population. Weather data was derived from historical weather data API using this link: https://open-meteo.com/.

### 3.7 Data Analysis

Data collected was analysed using excel. This was before utilization of the data for the development of the system.

### 3.8 System Development Methodology

Waterfall methodology was used in this study as a formal approach to implementing System Development Life Cycle SDLC. It identifies system requirements early and minimizes changes as the project progresses. Each phase must be fully completed before the next, providing certainty before development.

#### 3.8.1 Requirement Gathering and Analysis

The requirements, system requirements, and functional requirements are documented in chapter 4.

#### 3.8.1 System Analysis

The researcher acquired the system requirements as discussed in chapter 4 and created use cases, activity diagrams and data flow diagrams.

#### 3.8.2 System Design

Inputs used were malaria cases from the KHIS systems and climate data (rainfall, humidity and temperature) from historical data API https://open-meteo.com/. The outputs are the logical and physical design. In logical design, database, interfaces and reports was developed. In physical design, the researcher selected the appropriate specifications for both hardware and software that defined the system architecture.

a. The application was built using various frameworks. The programming language that will be used is python and JavaScript. At the front end, React js was used and at the server side, Node js. Various Application Programming Interfaces (APIs) to localize malaria cases and climate data in the application was used. To predict the outbreak which is the main component of the system, several machine learning techniques were tested to get the one that has the pest precision and accuracy levels. In addition, malaria outbreak thresholds for both malaria cases and climatic data were be applied to define malaria outbreaks. Climatic data to be used includes rainfall, humidity and temperature. Geospatial data which included data on the location of confirmed cases of malaria, as well as information on population density was used to localize the alerts.

### 3.8.3 System Implementation

The actual writing of the codes based on the designs discussed above was done. The system was developed in small programs. Each unit developed and tested for its functionality, also known as unit testing.

### 3.7.4 System Testing

After performing unit testing, the researcher combined all modules created during the implementation phase and integrated them into the system.

## 3.9 Research Quality

The results obtained will be replicated in other trials. Solutions developed closely replicate the real application that the Ministry of Health might adopt.

## 3.10 Ethical Considerations

Measures were put in place to protect intellectual property in areas where the research was undertaken. During the study, privacy and confidentiality of data was maintained. Ethical approval and certificate from National Commission for Science, Technology & Innovation (NACOSTI) were also acquired.

# Chapter 4: System Analysis, Design and Architecture

## 4.1 Introduction

This describes how the systems requirements were gathered, and the various methodologies used to design and develop the malaria outbreak detection model.

### 4.1.1 Requirement Gathering and Analysis

Through literature review, the following requirements were documented.

### 4.1.2 User Requirements

- Real-time data monitoring: Application provides real-time monitoring of malaria cases and climatic data that will allow health officials to stay updated on the current situation and take immediate action if necessary.
- Geospatial Mapping: The application should have a geospatial mapping feature that displays the geographical distribution of malaria cases. This helps health officials identify high-risk areas and allocate resources accordingly.
- Collaboration and Communication: The application should facilitate collaboration and communication among health officials, allowing them to share information, exchange best practices, and coordinate response efforts.
- Integration with Existing Systems: The application should be compatible with existing health information systems or databases to ensure seamless data integration and avoid duplication of efforts. This enables health officials to access and consolidate data from various sources for comprehensive analysis.
- User-Friendly Interface: The application should have an intuitive and user-friendly interface that is easy to navigate and understand. Health officials with varying levels of technical expertise should be able to use the application without extensive training.
- Mobile Accessibility: The application should be accessible on mobile devices, allowing health officials to access information and receive alerts on the go. This is particularly important for field workers who may need to report cases or collect data from remote areas.
- Data Security and Privacy: The application should adhere to strict data security and privacy standards to protect sensitive health information. It should implement encryption, user

authentication, and access control mechanisms to ensure that only authorized personnel can access and modify the data. requirements

### 4.1.3 System requirements

- Platform Compatibility: The application should be compatible with multiple platforms, including desktop computers, laptops, and mobile devices (iOS and Android), to ensure widespread accessibility for health officials.
- Scalability: The application should be designed to handle a large volume of data and accommodate potential growth in the number of users and data sources. It should be able to scale up its infrastructure and processing capabilities as needed.
- Data Integration: The application should support seamless integration with various data sources, such as health databases, climate data repositories, and other relevant sources. It should have the ability to aggregate and synchronize data from different systems into a unified platform.
- Data Storage and Management: The application should provide robust data storage and management capabilities to handle the influx of real-time and historical data. It should ensure data integrity, security, and efficient retrieval for analysis and reporting purposes.
- Data Processing and Analytics: The application should have powerful data processing and analytics capabilities to handle complex algorithms and predictive modelling. It should be able to perform real-time data analysis, generate meaningful insights, and produce visualizations for decision-making.
- Alerting and Notification System: The application should have a reliable and timely alerting system to notify health officials about potential malaria outbreaks, changes in risk levels, or other critical information. It should support multiple communication channels, such as email, SMS, push notifications, and in-app alerts.
- Robust Connectivity: The application should be designed to function in areas with varying levels of connectivity, including low-bandwidth or intermittent internet access. It should have offline capabilities to collect and store data locally, and synchronize when an internet connection becomes available.
- Security and Privacy: The application should incorporate robust security measures to protect sensitive data from unauthorized access, manipulation, or breaches. It should

employ encryption, authentication mechanisms, access controls, and secure transmission protocols to ensure data security and privacy compliance.

- Backup and Disaster Recovery: The application should have a robust backup and disaster recovery mechanism to prevent data loss and ensure business continuity in case of system failures or disasters. It should regularly backup data and have procedures in place for quick system restoration.
- System Performance and Reliability: The application should be designed to deliver high performance and reliability, ensuring minimal downtime or delays. It should be able to handle concurrent user access, process data efficiently, and respond promptly to user requests.
- Compliance with Standards and Regulations: The application should adhere to relevant standards, regulations, and guidelines related to healthcare data management, privacy, and security.

#### 4.1.4 Functional Requirements

- User Registration and Authentication: The application should provide a user registration process and support authentication mechanisms to ensure that only authorized health officials can access the system.
- Data Collection and Reporting: The application should allow health officials to input and report malaria-related data, including cases, patient demographics, geographical information, and severity levels. It should provide structured forms or templates for data collection.
- Real-time Data Visualization: The application should offer interactive and visual representations of data, such as charts, graphs, and maps, to facilitate data analysis and decision-making by health officials.
- Alert Generation and Distribution: The application should generate alerts based on predefined thresholds and criteria.
- Geospatial Mapping and Heatmaps: The application should provide geospatial mapping functionality to visualize malaria cases and risk levels based on geographical location. It should generate heatmaps to identify high-risk areas and support targeted interventions.
- Early Warning System: The application should implement algorithms and models to predict potential malaria outbreaks based on historical data, climate information, and other relevant

factors. It should generate early warning notifications for health officials to take proactive measures.

#### 4.1.5 Non-Functional Requirements:

- Performance: The application should have a fast and responsive user interface, ensuring quick data retrieval, processing, and analysis. It should handle concurrent user access and maintain optimal performance even during peak usage periods.
- Reliability and Availability: The application should be reliable, with minimal downtime and disruptions. It should have mechanisms in place to ensure high availability, such as redundancy, failover, and backup systems.
- Usability: The application should have an intuitive and user-friendly interface, making it easy for health officials to navigate, input data, and interpret visualizations. It should require minimal training for users to understand and operate effectively.

### 4.2 Systems Design

System design entail the use of various diagrams that include a sequence diagram, use case diagram, entity relationship diagrams, data flow diagram and the application wireframe.

#### 4.2.1 Data Flow Diagrams (DFD)

A data flow diagram is a tool used to visualize the structure and behavior of a system. In the proposed system, the data flows from the surveillance systems and weather forecast APIs to the application, which uses machine learning techniques to model the data and generate malaria outbreak predictions. The predictions are then sent to specific health officials through the application, using geo-spatial information to target alerts to specific areas.

#### 4.2.2 Input data:

1. Malaria incidences surveillance systems in counties in Kenya
2. Average climate data per month including temperature, rainfall/precipitation, and humidity.
3. Geo-spatial information

#### 4.2.3 Data processing:

- Data is modelled using machine learning techniques to predict malaria outbreaks
- Thresholds are set for malaria outbreak predictions

- The application uses geo-spatial information to target alerts to specific areas and health officials

### 4.2.4 Output data:

1. Malaria outbreak/early warning alerts sent to health officials

DFD's offer different levels of abstraction represented by 0, 1 and 2. We shall show a representation up to the 2 level of abstraction.

**Detailed data flow:**


Figure 2:0-Level DFD

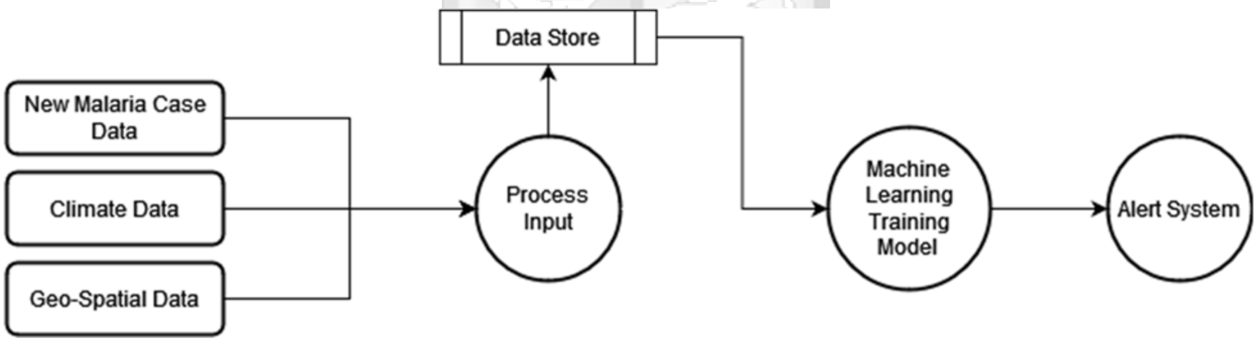



Figure 3: 1-Level DFD

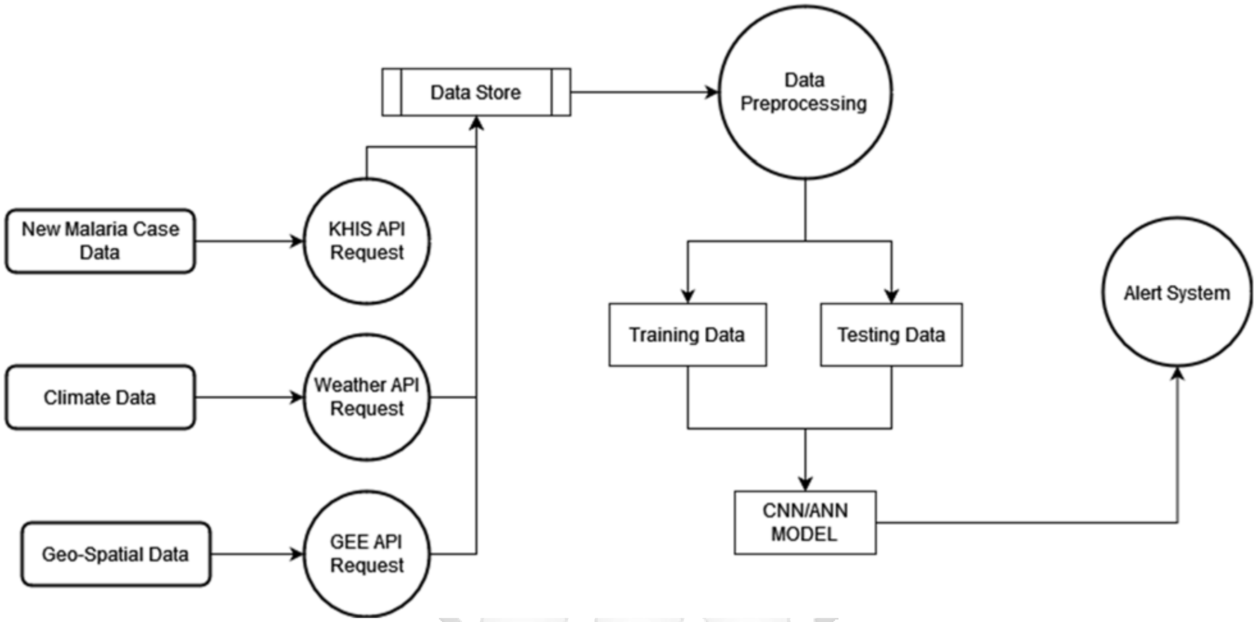



Figure 4: 2-Level DFD

### 4.2.5 Use Case Diagrams

In this system, a health official registers in the system. After registration, they are able to login and access the system. The health official can then interact with the system and get alerts on potential outbreaks across the country.

**Use case description**

Table 1: Use case description on loading new data

| | |
|---|---|
| **Use Case 1** | Load new data into application |
| **Primary Actor** | Health Official |
| **Brief Description** | This use case defines how the health official loads new data into the application for regression |
| **Pre-condition** | Health official is identified and authenticated<br><br>Regression model Random Forest Regression Trained |
| Success Scenario | |
| **Actor Action** | **System Responsibility** |
| Health official loads new climate data | |

| | |
|---|---|
| | System converts data values into csv format as input to the Random Forest Regression Model via the Kaggle public API |
| | System performs regression |
| | System returns regression/predicted result |
| User views the percentage and accuracy level of the predicted cases per 1000 of the population | |
| | If threshold exceeded, system alerts of possible epidemic |
| **Extensions** | |
| The system periodically captures new data via the weather API | |

Table 2: Use case description on alert notification

| | |
|---|---|
| **Use Case 2** | Alert Notification |
| **Primary Actors** | Health Official |
| **Brief Description** | This use case describes how the health official receives an alert on possible malaria epidemics |
| **Pre-condition** | Successful prediction of malaria incidence cases superseding the threshold set |
| **Post-Condition** | Recommended course of action |
| **Main Success Scenario** | |
| **Actor Action** | **System Responsibility** |
| Health Worker accesses the system | |
| | Alerts User of possible epidemic |
| View alert and recommended course of action | |
| | Displays level of accuracy |
| Exit the system | |

# Chapter 5: System Implementation and Testing

## 5.1 Introduction

This section describes in detail the process undertaken to conduct system implementation and testing. Below is a summary of the process that was undertaken.

a. Feature selection: Selection of the most important variables that has a significant impact on malaria transmission.
b. Data collection: Collection of relevant data on malaria incidence and climatic data (Rainfall/precipitation, humidity and temperature).
c. Model selection: Selection of the best machine learning algorithm for the malaria early warning alert system, Artificial Neural Networks (ANNs).
d. Model training: Training of the model on the collected data to make predictions and provide alerts to health officials at facility level.
e. Model evaluation: Evaluation of the performance of the model using metrics such as accuracy, precision, and recall.
f. Deployment: Deploy the model in a real-world setting, incorporating feedback from stakeholders and making necessary updates to improve performance.

## 5.2 Feature Selection

In the feature selection process, we identified the most relevant variables that contribute to malaria outbreaks. These variables were determined through a scientific literature review. The selected features included temperature, rainfall/precipitation, humidity, and disease prevalence as here referred to as malaria incidence. These features were deemed crucial for accurately predicting and forecasting malaria outbreaks in the given area.

## 5.3 Data collection

This research identified weather variables such as temperature, rainfall, and humidity as important indicators for predicting malaria outbreaks. The weather variables were obtained from the weather API https://open-meteo.com/ in JSON format and seamlessly integrated into the application. In addition, Malaria data was derived from the Kenya Health Information System (KHIS) and integrated into the application. The data variables represent the inputs used in malaria predictions as explained below:

- **Malaria incidence data**: Monthly aggregated data from Jan 2018 to May 2023 as reported by health facilities in KHIS and population data from Kenya Bureau of statistics was used to calculate the malaria incidence per 1,000 people at risk. Formula for calculating malaria incidence: *confirmed Cases / population At Risk ×1000*
- **Rainfall/precipitation**: Data on average rainfall in millilitres per month was used from Jan 2018 to May 2023.
- **Relative Humidity**: Data on average humidity in percentage per month from Jan 2018 to May 2023 was used.
- **Temperature**: Data on average temperature in degrees Celsius per month. from Jan 2018 to May 2023.

The data was processed using `python` and stored in a CSV file format for easy accessibility and analysis. The collection of weather data was done for each county in Kenya using Google Colab to process and aggregate the collected information. The code below for retrieving the data was executed on Google Colab.

**Installing packages in Google Colab**

```
[2] pip install openmeteo-requests

pip install requests-cache retry-requests numpy pandas

[4] import openmeteo_requests

import requests_cache
import pandas as pd
from retry_requests import retry

# Setup the Open-Meteo API client with cache and retry on error
cache_session = requests_cache.CachedSession('.cache', expire_after = -1)
retry_session = retry(cache_session, retries = 5, backoff_factor = 0.2)
openmeteo = openmeteo_requests.Client(session = retry_session)
```

Figure 5: Installing packages in Google Colab

```python
# Make sure all required weather variables are listed here
# The order of variables in hourly or daily is important to assign them correctly below
url = "https://archive-api.open-meteo.com/v1/archive"
params = {
  "latitude": 1.2333,
  "longitude": 35.1167,
  "start_date": "2018-01-01",
  "end_date": "2023-05-31",
  "hourly": ["temperature_2m", "relative_humidity_2m", "rain"],
  "timezone": "Africa/Cairo"
}
responses = openmeteo.weather_api(url, params=params)

# Process first location. Add a for-loop for multiple locations or weather models
response = responses[0]
print(f"Coordinates {response.Latitude()}°E {response.Longitude()}°N")
print(f"Elevation {response.Elevation()} m asl")
print(f"Timezone {response.Timezone()} {response.TimezoneAbbreviation()}")
print(f"Timezone difference to GMT+0 {response.UtcOffsetSeconds()} s")

# Process hourly data. The order of variables needs to be the same as requested.
hourly = response.Hourly()
hourly_temperature_2m = hourly.Variables(0).ValuesAsNumpy()
hourly_relative_humidity_2m = hourly.Variables(1).ValuesAsNumpy()
hourly_rain = hourly.Variables(2).ValuesAsNumpy()

hourly_data = {"date": pd.date_range(
  start = pd.to_datetime(hourly.Time(), unit = "s"),
  end = pd.to_datetime(hourly.TimeEnd(), unit = "s"),
  freq = pd.Timedelta(seconds = hourly.Interval()),
  inclusive = "left"
)}
hourly_data["temperature_2m"] = hourly_temperature_2m
hourly_data["relative_humidity_2m"] = hourly_relative_humidity_2m
hourly_data["rain"] = hourly_rain

hourly_dataframe = pd.DataFrame(data = hourly_data)
# print(hourly_dataframe)
```

Figure 6: Code from retrieving data from the historical data API



## 5.4 Model selection

Under this step, machine learning techniques including Decision Tree Regressor, Linear Regression, Random Forest Regressor, Ada Boost Regressor, Gradient Boosting Regressor, SVR and decision tree were tested to select the best model for malaria early warning. This process involved the use of statistical correlation analysis to determine the degree of the relationship between the relative movements of two variables (Zhao et al., 2020). Kaggle, a platform used for data science was used in selecting of the appropriate models.

The process started with importation of libraries to support the testing. A sample data of few variables was loaded into Kaggle and the data features verified to be okay. Data cleaning was also conducted which was okay. The modelling libraries mentioned above were loaded and feature engineering involving conversion of the data into numerical was done.

Data was split into features and output data. Six models including Decision Tree Regressor, Linear Regression, Random Forest Regressor, Ada Boost Regressor, Gradient Boosting Regressor, SVR and decision tree were tested. The results were featured in 3 parts as explained below.

```
# Regression problem - DecisionTreeRegressor

# Instatiate our model
dt_regression = DecisionTreeRegressor()

# Train Model
dt_regression.fit(x_train_r, y_train_r)

# Predict on our Model
train_predictions = dt_regression.predict(x_train_r)
test_predictions = dt_regression.predict(x_test_r)

print("Mean Absolute Error:", mean_absolute_error(y_test_r, test_predictions))
print("Mean Squared Error:", mean_squared_error(y_test_r, test_predictions))
print("R2 Score:", r2_score(y_test_r, test_predictions))
```

```
Mean Absolute Error: 9.403333333333333
Mean Squared Error: 191.99243333333342
R2 Score: -4.9362905791340825
```

Figure 7: Sample model evaluation using kaggle

Table 3: Interpretation of the three parameters used to determine the best model

| Parameter | Interpretation |
|---|---|
| **Mean Absolute Error (MAE):** | • It explains how far off, on average, the predictions are from the actual values.<br>• The lower the MAE, the better the model performance. **A value of 0 would indicate a perfect model.** |
| **Mean Squared Error (MSE):** | • The Mean Squared Error measures the average of the squares of the errors between the predicted values and the actual values.<br>• The lower values of MSE indicate better model performance. **A value of 0 would indicate a perfect model.** |

| | |
|---|---|
| **R^2 Score (Coefficient of Determination):** | • Measures the proportion of the variance in the dependent variable that is predictable from the independent variables.<br>• A negative R^2 score indicates that the model is performing worse than a horizontal line fitting the data.<br>• Higher R^2 scores are desirable, indicating that a larger proportion of the variance is explained by the model. |

Based on the explanation above, when interpreting these results, a lower MAE and MSE values, and a higher R^2 score, indicating better performance of the model in terms of accuracy and explanatory power were considered.

### 5.4.2 Interpretation

The table below summarises the results of the various models that were tested and their results.

The model with the best results was **Random Forest Regression.** It was therefore selected as the most suitable model for predicting malaria outbreaks.

Table 4: Analysis of the tested machine learning models

| | Machine Learning Selection | Mean Absolute Error | Mean Squared Error | R2 Score |
|---|---|---|---|---|
| 1 | Decision Tree Regressor | 9.4 | 192.0 | -4 |
| 2 | Linear Regression | 7.3 | 83.0 | -1.6 |
| 3 | Random Forest Regressor | 2.54 | 15.1 | 0.5 |
| 4 | Ada Boost Regressor | 5.9 | 62.9 | -0.9 |
| 5 | Gradient Boosting Regressor | 5.0 | 47.8 | -0.4 |
| 6 | SVR | 5.9 | 54.9 | -0.7 |

Random Forest Regression is a popular machine learning method used for regression tasks. It is a machine learning technique used for solving regression problems. It is an ensemble learning method that combines the predictions of multiple decision trees to obtain more accurate results. Each decision tree is constructed using a random subset of the training data and a random subset of the features.

### 5.5 Model training

In the context of artificial intelligence (AI), model training refers to the process of teaching a machine learning model to make predictions or decisions based on input data. This process involves presenting the model with a dataset containing examples of input data along with the **corresponding correct output (labels or targets)**. The model then learns patterns and relationships within the data to make accurate predictions or classifications when presented with new, unseen data.

The training was achieved by use of TensorFlow Random Forest (Decision Forests) on a dataset containing numerical features. The process below was applied in training the model.

#### Installing TensorFlow Decision Forests

Installation of TF-DF - a library to train, run and interpret decision forest models (e.g., Random Forests, Gradient Boosted Trees) in TensorFlow.

```
!pip install tensorflow tensorflow_decision_forests 'tensorflowjs>=4.4.0'
```

#### Importing libraries

```
# Prepare and load the model with TensorFlow

import tensorflow as tf
import tensorflow_decision_forests as tfdf
import tensorflowjs as tfjs
from google.colab import files

import os
import io
import numpy as np
import pandas as pd
import math

# from tensorflow import keras
# import tensorflowjs as tfjs
```

**Note**: The hidden code cell limits the output height in colab.

```
#@title

from IPython.core.magic import register_line_magic
from IPython.display import Javascript
from IPython.display import display as ipy_display

# Some of the model training logs can cover the full
```

```
# screen if not compressed to a smaller viewport.
# This magic allows setting a max height for a cell.
@register_line_magic
def set_cell_height(size):
  ipy_display(
      Javascript("google.colab.output.setIframeHeight(0, true, {maxHeight: "
+
                   str(size) + "})"))

# Check the version of TensorFlow Decision Forests
print("Found TensorFlow Decision Forests v" + tfdf.__version__)

Found TensorFlow Decision Forests v1.8.1
```

**Training a Random Forest Regression model**

Training the regression model on the i_malaria_kenya dataset. The objective of this is to predict malaria_incidence.

**Loading the dataset and converting it in a tf.Dataset**

We use pandas to load the dataset

**Upload dataset from local drive**

During the data preprocessing step, the collected environmental data including temperature, rainfall, humidity, and malaria incidence were cleaned and standardized to ensure consistency and accuracy.

```
# Upload the dataset
from google.colab import files
uploaded = files.upload()

# Wait for it to be 100% uploaded

<IPython.core.display.HTML object>

Saving malaria_data_update_2202024.csv to malaria_data_update_2202024.csv

# Load the dataset into a Pandas Dataframe.
dataset_df = pd.read_csv(io.BytesIO(uploaded['malaria_data_update_2202024.csv
']))

# Dataset is now stored in a Pandas Dataframe

# Display the first 3 examples.
print(dataset_df.head(10))
```

```
   temperature  relative_humidity  precipitation  malaria_incidence
0    19.029011          44.595102       0.002554             12.323
1    19.367176          47.662115       0.018750             10.601
2    16.995274          73.773335       0.293414              6.653
3    16.324750          83.044013       0.463750              5.059
4    16.553473          80.310311       0.370296              8.528
5    15.553361          84.975755       0.367361             10.171
6    15.548836          76.526145       0.201210             13.362
7    16.003205          74.299108       0.292339              9.727
8    17.552875          59.908556       0.060972              7.780
9    17.785059          63.468779       0.095833              8.620
```

Next split the dataset into training and testing:

```
# Split the dataset into a training and testing dataset.
def split_dataset(dataset, test_ratio=0.30):
  """Splits a panda dataframe in two."""
  test_indices = np.random.rand(len(dataset)) < test_ratio
  return dataset[~test_indices], dataset[test_indices]

train_ds_pd, test_ds_pd = split_dataset(dataset_df)
print("{} examples in training, {} examples for testing.".format(
    len(train_ds_pd), len(test_ds_pd)))

2134 examples in training, 921 examples for testing.
```

And finally, convert the pandas dataframe (pd.Dataframe) into tensorflow datasets (tf.data.Dataset):

Verify column names

```
print(train_ds_pd.columns)

Index(['temperature', 'relative_humidity', 'precipitation',
       'malaria_incidence'],
      dtype='object')

train_ds = tfdf.keras.pd_dataframe_to_tf_dataset(train_ds_pd, label="malaria_
incidence", task=tfdf.keras.Task.REGRESSION)
test_ds = tfdf.keras.pd_dataframe_to_tf_dataset(test_ds_pd, label="malaria_in
cidence", task=tfdf.keras.Task.REGRESSION)
```

**Training the model**

Next, the dataset was split into training and testing sets in a ration of 70:30, with a portion of the data reserved for model training and the rest for evaluating the model's performance. The random forest regression model was trained using the training dataset, and its performance was evaluated on the testing dataset using evaluation metrics such as mean squared error,  R-squared value.

```
%set_cell_height 300

# Configure the model.
model_7 = tfdf.keras.RandomForestModel(task = tfdf.keras.Task.REGRESSION)

# Train the model.
model_7.fit(train_ds)

<IPython.core.display.Javascript object>

Use /tmp/tmpnxib3m47 as temporary training directory
Reading training dataset...
Training dataset read in 0:00:05.726985. Found 2134 examples.
Training model...
Model trained in 0:00:01.654937
Compiling model...
Model compiled.

<keras.src.callbacks.History at 0x7876278e8d90>
```

### 5.4 Evaluate the model

The evaluation of a model is a critical step in assessing its performance and determining its effectiveness in solving a specific task. Evaluation helps measure how well the model generalizes to new, unseen data and provides insights into areas where the model may need improvement. After training the model, it was evaluated using the test dataset. Performance metrics such as mean squared error (MSE), mean absolute error (MAE) and were calculated to assess how well the model generalized to unseen data.

**Results on Evaluation of the model on the test dataset.**

```
[ ] model_7.compile(metrics=["mse", "mae"])
    evaluation = model_7.evaluate(test_ds, return_dict=True)

    print(evaluation)
    print()
    print(f"MSE: {evaluation['mse']}")
    print(f"RMSE: {math.sqrt(evaluation['mse'])}")
    print(f"MAE: {evaluation['mae']}")
```

```
1/1 [==============================] - 1s 537ms/step - loss: 0.0000e+00 - mse: 154.6718 - mae: 5.3104
{'loss': 0.0, 'mse': 154.67178344726562, 'mae': 5.310389995574951}

MSE: 154.67178344726562
RMSE: 12.436711118590221
MAE: 5.310389995574951
```

**Remarks**

The MSE value of 154.6718 indicates that, on average, the squared difference between the actual and predicted values is approximately 154.67.

The RMSE value of approximately 12.4367 suggests that, on average, the predictions deviate from the actual values by approximately 12.44 units.

The MAE value of 5.3104 indicates that, on average, the absolute difference between the actual and predicted values is approximately 5.31.

Based on the above results, the model's performance seems reasonable, as indicated by relatively low values of MSE, RMSE, and MAE. These results align to the findings during selection of the model that led to the selection of random Forrest regression.



This data was then exported to a saved model as per the code below.

```
# Export the model to a SavedModel.
model_7.save("/tmp/tfdf_model_latest")

!tensorflowjs_converter \
    --input_format=tf_saved_model \
    /tmp/tfdf_model_latest \
    /tmp/tfdf_model_v3

# Download the converted TFJS model
!zip -r tfdf_model_v3.zip /tmp/tfdf_model_v3/
files.download("tfdf_model_v3.zip")

  adding: tmp/tfdf_model_v3/ (stored 0%)
  adding: tmp/tfdf_model_v3/assets.zip (stored 0%)
  adding: tmp/tfdf_model_v3/group1-shard1of1.bin (deflated 6%)
  adding: tmp/tfdf_model_v3/model.json (deflated 87%)
```

### 5.6 Deployment

Deployment of a model in artificial intelligence (AI) refers to the process of making the trained model available for use in real-world applications. After the model training and evaluation stage, the next step in system implementation was to create a user-friendly interface for health officials to access the malaria early warning alert application. This interface would allow for the input of new weather data and relay of real-time predictions and alerts regarding potential malaria outbreaks.

Here's an overview of the deployment process:

#### Model Serialization

Before deployment, the trained model needs to be serialized or saved to disk in a format that can be easily loaded and used by other systems or applications. Common serialization formats include pickle (for Python-based models), ONNX (Open Neural Network Exchange), or TensorFlow's SavedModel format. This application makes use of TensorFlow's SavedModel format for model serialization and deployment to ensure compatibility and ease of integration with the ReactJS front-end of the application.

```
[ ] # Export the model to a SavedModel.
    model_7.save("/tmp/tfdf_model_latest")
```

```
[ ] !tensorflowjs_converter \
        --input_format=tf_saved_model \
        /tmp/tfdf_model_latest \
        /tmp/tfdf_model_v3
```

```
[ ] # Download the converted TFJS model
    !zip -r tfdf_model_v3.zip /tmp/tfdf_model_v3/
    files.download("tfdf_model_v3.zip")
```

*A snippet of the serialization/conversion of the model to a SavedModel*

#### Model Containerization

Models are often deployed within containers (e.g., Docker containers) to encapsulate all dependencies, including the model itself, libraries, and runtime environments. Containerization ensures that the model can be deployed consistently across different environments and platforms. However, for our application, we do not require containerization as we will be deploying the model on a cloud-based platform.

**Cloud-based deployment**

This offers several advantages for deploying the malaria early warning alert application. It allows for scalability and flexibility, as cloud platforms provide resources that can easily handle varying levels of workload and traffic. Cloud platforms such as Vercel which was adopted for use in this project allow for easier and faster integration with the front-end of the application and provide seamless deployment and hosting capabilities. Cloud-based deployment ensures high availability and reliability, as the application can be accessed from anywhere with an internet connection.

The integration of the model into the React application was achieved by pulling in the necessary TensorFlow packages from a CDN. This allows the application to interact with and utilize the trained model for making predictions or decisions based on input data.

```
<script src="https://cdn.jsdelivr.net/npm/@tensorflow/tfjs@4.5.0/dist/tf.min.js"></script>
<script src="https://cdn.jsdelivr.net/npm/@tensorflow/tfjs-tfdf/dist/tf-tfdf.min.js"></script>
    You, 2 days ago • Uncommitted changes
```

The next step involved loading the serialized model into memory, making it ready for inference. Once the model is loaded into memory, it can be used to make predictions based on input data.

```javascript
const tf = window.tf;
const tfdf = window.tfdf;

// Malaria incidence prediction
const predictiveModelInference = async (weatherData) => {
  // Load the model
  const model = await tfdf.loadTFDFModel("http://127.0.0.1:3000/tfdf_model/model.json");

  // Perform an inference
  const result = await model.executeAsync({
    precipitation: tf.tensor([weatherData[2]]),
    relative_humidity: tf.tensor([weatherData[1]]),
    temperature: tf.tensor([weatherData[0]]),
  });

  return result.dataSync()[1];
};
```

The function 'predictiveModelInference' accepts input data and performs inference using the loaded model. The results of the inference are then returned by the function, which can be further utilized by the React application for displaying alerts or generating visualizations based on the predictions.

The predicted value, herein referred to as `currentIncidence`, is then used in conjunction with other variables and thresholds to determine the level of malaria risk and issue appropriate warnings or alerts. These variables include past malaria incidence data, climate data (such as temperature and rainfall), population density, and historical patterns of malaria outbreaks.

An 'alert threshold' (or 'epidemic threshold') indicates the level of incidence above which a disease requires an urgent response. For malaria, an alert threshold occurs when confirmed malaria cases reach 1.5 times the baseline level. This threshold serves as a trigger for the early warning system to issue alerts or take proactive measures to mitigate the risk of a malaria outbreak.

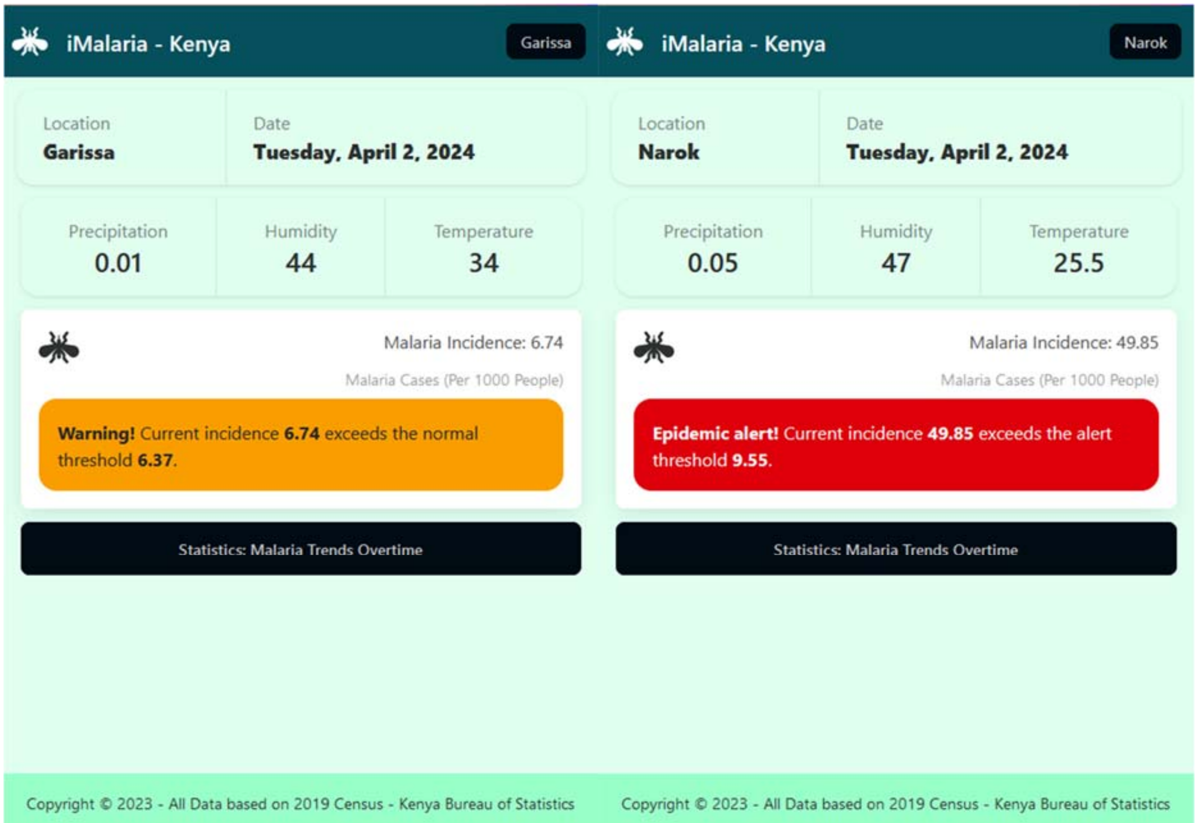
The baseline was determined by calculating the average historical incidence of malaria cases seen in the previous 3 weeks. An alert/epidemic threshold is then achieved by multiplying the baseline level by 1.5. A normal threshold is set as the average historical incidence of malaria cases seen in the previous 3 weeks. A warning threshold occurs when the number of malaria cases exceeds the normal threshold but has not reached the alert threshold.

```javascript
function calculateMalariaThresholds(currentIncidence) {
  // Calculate the average historical incidence for the last 3 weeks
  const recentData = historicalData.slice(-3); // Get last 3 elements
  const averageIncidence = recentData.reduce((sum, value) => sum + value, 0) / recentData.length;

  // Calculate thresholds based on average
  const normalThreshold = averageIncidence;
  const warningThreshold = currentIncidence > averageIncidence || currentIncidence < 1.5 *
  averageIncidence ? currentIncidence : null;
  const alertThreshold = 1.5 * averageIncidence;

  // Analyze current incidence
  if (currentIncidence > alertThreshold) {
    return (···
    );
  } else if (currentIncidence <= warningThreshold) {
    return (···
    );
  } else {
    return (···
    );
  }
}
```

The result of the predicted value is thereafter displayed to the user in the React application, providing them with relevant information about the current malaria risk level based on the input data and model predictions.

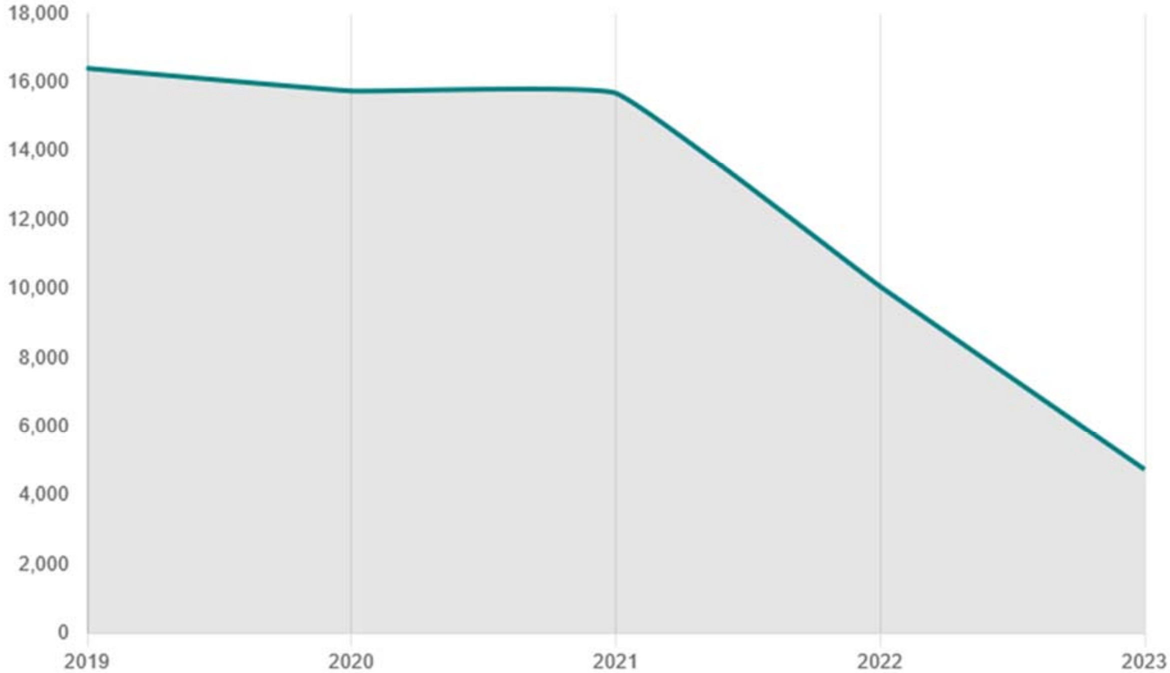



A graphical representation of the predicted malaria incidence over time can also be included in the React application, allowing users to visualize the trends and patterns in malaria transmission.

Total Malaria Cases Per Year Over The Last 5 Years For Narok County

| Year | Total Cases |
|---|---|
| 2019 | **16,414** |
| 2020 | **15,757** |
| 2021 | **15,691** |
| 2022 | **10,072** |
| 2023 | **4,757** |

Malaria Cases Over Time



# Chapter 6: Discussions

## 6.1 Introduction

This chapter outlines show how the objectives, were met, benefits of the system and limitations.

## 6.2 Review of Research Objectives

On the first objective which was to identify challenges affecting malaria outbreak or early warning detection. Several challenges were outlined in the literature review. In summary, the challenges were; suboptimal surveillance system, climate variability and change, critical gaps in the diagnosis of malaria, minimal financial resources, minimal end-user engagement and lack of coordination. The application will provide solutions in relation to the challenges that will improve early warning of malaria outbreaks and surveillance in general.

The second objective was to review related systems and models for predicting malaria outbreaks. Through literature review, various systems were analysed that were based on text messaging, statistical measures and machine learning. The developed system under the research was providing a simpler tool that does not require very technical staff to conduct and can be used in the roll out of primary health care.

On the third objective, on development of an application for malaria outbreak alert, this was met by this research in the form of a prototype. The system can be rolled out to be used by the ministry of health to support in planning and forecasting and also the rollout of Primary Health Care.

The last objective was to test the functionality of the application. Testing and compatibility were done to test the functionality of the system on diverse web browsers and user testing done to identify user friendliness, navigation and user acceptance to the system.

## 6.3 Benefits of the imalaria system

The benefits of the system are as outlined below:

**Early Warning System:** By leveraging predictive analytics, the system can potentially forecast malaria outbreaks before they occur, allowing for timely interventions and resource allocation.

**Improved Resource Allocation:** With accurate predictions, healthcare authorities can allocate resources such as medications, bed nets, and personnel more efficiently to areas at higher risk of malaria outbreaks, thereby maximizing the impact of interventions.

**Reduced Disease Burden:** Anticipating outbreaks allows for proactive measures such as vector control, community education, and targeted vaccination campaigns, ultimately leading to a reduction in the incidence and severity of malaria cases.

**Data-Driven Decision Making:** The system relies on data analytics and machine learning algorithms to process vast amounts of historical and real-time data, providing decision-makers with actionable insights to guide their response strategies.

**Customized Interventions:** Tailored interventions can be designed based on the specific risk factors identified by the system, leading to more effective and targeted public health interventions.

**Enhanced Surveillance:** The system can complement traditional surveillance methods by identifying emerging trends and patterns in malaria transmission, thereby strengthening overall disease surveillance efforts.

### 6.4 Limitations of the developed system

Limitations of the system have been summarised as below.

**Data Availability and Quality**: The accuracy of predictions heavily relies on the availability and quality of data. Limited or unreliable data inputs can lead to inaccurate forecasts.

**Model Uncertainty**: Predictive models inherently involve uncertainty, and there may be limitations in accurately predicting complex, dynamic systems such as malaria transmission. Variability in environmental factors, human behavior, and other unknown variables can affect the reliability of predictions.

**Sensitivity to Assumptions:** The model's performance may be sensitive to the assumptions made during its development, including feature selection, algorithm choice, and parameter tuning. Deviations from these assumptions could impact the model's accuracy.

**Resource Constraints:** Implementing interventions based on predictions requires adequate resources and infrastructure, which may be lacking in resource-constrained settings where malaria burden is often the highest.

**Ethical Considerations:** Predictive models raise ethical concerns related to privacy, data security, and potential stigmatization of communities identified as high-risk. Safeguards must be in place to address these ethical considerations.

**Limited Generalizability:** Predictive models developed in one geographic region or time period may not generalize well to other regions or time periods due to differences in environmental factors, population characteristics, and disease dynamics.

# Chapter 7: Conclusions and Recommendations

### 7.1 Conclusions

The malaria early warning alert application works by pulling long-term data on malaria cases and climate variables, such as temperature, rainfall, and humidity, through APIs. This data is then modelled using machine learning techniques with thresholds that allow for the prediction of malaria outbreaks. The application uses geospatial information to issue alerts to specific health officials in an area up to the facility level. Health officials will view the data via their electronic gadgets, including their mobile phones, which will be color-coded to indicate the likelihood of an outbreak (yellow), an outbreak (red), or no outbreak (green). This will enable health officials to order commodities in time, target specific areas with interventions, and issue alerts to community members to take precautions.

### 7.2 Recommendations

The following are recommendations to be applied i the rollout of the imalaria system.

**Continuous Evaluation and Improvement**: Regular evaluation of the performance of the prototype against historical outbreak data and real-time surveillance reports is key in the rollout of this model. This evaluation should include metrics such as accuracy, sensitivity, specificity, and predictive value. Feedback from end-users, including public health officials, healthcare workers, and community members, to identify areas for improvement and refine the system's algorithms and predictive models is also very key. Continuously update of the system's data inputs, including environmental, demographic, and epidemiological data, to ensure relevance and accuracy in predicting malaria outbreaks.

**Integration with Existing Public Health Infrastructure**: Integration of the "imalaria system with existing public health infrastructure, including disease surveillance systems, health information systems, and decision support tools, to enhance its usability and impact. Provision training and capacity-building support to healthcare workers and public health officials on how to interpret and use the system's output effectively in decision-making processes. Fostering collaboration and partnerships with local, national, and international stakeholders, including government agencies, non-governmental organizations, academic institutions, and technology providers, to promote the adoption and sustainability of the system.

**Community Engagement and Empowerment**: Engagement of local communities in the design, implementation, and evaluation of the "imalaria" system to ensure its relevance, acceptability, and effectiveness within the context of their socio-cultural and environmental settings. Empower community members to actively participate in malaria prevention and control efforts by providing them with access to relevant information, resources, and tools generated by the system. Promote community-led initiatives, such as vector control activities, environmental sanitation campaigns, and health education programs, supported by insights generated by the system to complement traditional public health interventions.

By implementing these recommendations, stakeholders can maximize the utility and impact of the "imalaria system as an early warning system for malaria outbreaks, ultimately contributing to the reduction of malaria-related morbidity and mortality in endemic regions.

### 7.3 Future developments

Overall, while the "imalaria" system holds promise for improving malaria control efforts, it is essential to recognize and address its limitations to maximize its effectiveness and mitigate potential risks. Ongoing monitoring, evaluation, and refinement of the system are crucial for ensuring its relevance and reliability in supporting public health decision-making.

## References

Ayanlade, A., Sergi, C. M., Sakdapolrak, P., Ayanlade, O. S., Di Carlo, P., Babatimehin, O. I., Weldemariam, L. F., & Jegede, M. O. (2022). Climate change engenders a better Early Warning System development across Sub-Saharan Africa: The malaria case. *Resources, Environment and Sustainability*, *10*, 100080. https://doi.org/10.1016/j.resenv.2022.100080

Baeza, A., Bouma, M. J., Dhiman, R., & Pascual, M. (2014). Malaria control under unstable dynamics: Reactive vs. climate-based strategies. *Acta Tropica*, *129*, 42–51. https://doi.org/10.1016/j.actatropica.2013.04.001

Bashar, K., & Tuno, N. (2014). Seasonal abundance of Anopheles mosquitoes and their association with meteorological factors and malaria incidence in Bangladesh. *Parasites & Vectors*, *7*(1), 442. https://doi.org/10.1186/1756-3305-7-442

Bridges, D. J., Winters, A. M., & Hamer, D. H. (2012). Malaria elimination: Surveillance and response. *Pathogens and Global Health*, *106*(4), 224–231. https://doi.org/10.1179/2047773212Y.0000000035

Degefa, T., Yewhalaw, D., Zhou, G., Lee, M., Atieli, H., Githeko, A. K., & Yan, G. (2017). Indoor and outdoor malaria vector surveillance in western Kenya: Implications for better understanding of residual transmission. *Malaria Journal*, *16*(1), 443. https://doi.org/10.1186/s12936-017-2098-z

Division of National Malaria Program. (2020, March). *Guidelines for Malaria Epidemic Preparedness and Response in Kenya. 2nd Edition.* https://www.measureevaluation.org/countries/kenya/kenya-malaria-resources-1/Kenya%20Malaria%20EPR%20Guidelines_Final%20Signed%20-1.pdf/at_download/file

Division of National Malaria Programme (DNMP) [Kenya] and ICF. (2021, September). *Kenya Malaria Indicator Survey 2020*. https://dhsprogram.com/pubs/pdf/MIS36/MIS36.pdf

Google. (2022). *Google Earth Engine*. https://www.google.com/earth/education/tools/google-earth-engine/

Hemingway, J., Shretta, R., Wells, T. N. C., Bell, D., Djimdé, A. A., Achee, N., & Qi, G. (2016). Tools and Strategies for Malaria Control and Elimination: What Do We Need to Achieve a Grand Convergence in Malaria? *PLoS Biology*, *14*(3), e1002380. https://doi.org/10.1371/journal.pbio.1002380

Karuri, J., Waiganjo, P., Orwa, D., & Manya, A. (2014). DHIS2: The Tool to Improve Health Data Demand and Use in Kenya. *Journal of Health Informatics in Developing Countries*, *8*(1), Article 1. https://jhidc.org/index.php/jhidc/article/view/113

Kazeem, I., & Adebanji, S. (2021). A MODEL FOR PREDICTING MALARIA OUTBREAK USING MACHINE LEARNING TECHNIQUE. *Scientific Annals of Computer Science*, *19*, 9–15.

Maharaj, R. (2017). Early warning systems for the detection of malaria outbreaks. *The Indian Journal of Medical Research*, *146*(5), 560–562. https://doi.org/10.4103/ijmr.IJMR_933_17

McKelvie, W. R., Haghdoost, A. A., & Raeisi, A. (2012). Defining and detecting malaria epidemics in south-east Iran. *Malaria Journal*, *11*(1), 81. https://doi.org/10.1186/1475-2875-11-81

Modu, B., Polovina, N., Lan, Y., Konur, S., Asyhari, A., & Peng, Y. (2017). Towards a Predictive Analytics-Based Intelligent Malaria Outbreak Warning System. *Applied Sciences*, *7*(8), 836. https://doi.org/10.3390/app7080836

National Research Council (US) Committee on Climate, E. (2001). Toward the Development of Disease Early Warning Systems. In *Under the Weather: Climate, Ecosystems, and Infectious Disease*. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK222241/

Nissan, H., Ukawuba, I., & Thomson, M. (2021). Climate-proofing a malaria eradication strategy. *Malaria Journal*, *20*(1), 190. https://doi.org/10.1186/s12936-021-03718-x

Nkiruka, O., Prasad, R., & Clement, O. (2021). 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

Nyasa, R. B., Awatboh, F., Kwenti, T. E., Titanji, V. P. K., & Ayamba, N. L. M. (2022). The effect of climatic factors on the number of malaria cases in an inland and a coastal setting from 2011 to 2017 in the equatorial rain forest of Cameroon. *BMC Infectious Diseases*, *22*(1), 461. https://doi.org/10.1186/s12879-022-07445-9

Open Health Marketplace. (2020). *District Health Information System 2 (DHIS2) | Open Health News*. https://www.openhealthnews.com/resources/district-health-information-system-2-dhis2

Ostovar, A., Haghdoost, A. A., Rahimiforoushani, A., Raeisi, A., & Majdzadeh, R. (2016). Time Series Analysis of Meteorological Factors Influencing Malaria in South Eastern Iran. *Journal of Arthropod-Borne Diseases*, *10*(2), 222–236.

Ototo, E. N., Ogutu, J. O., Githeko, A., Said, M. Y., Kamau, L., Namanya, D., Simiyu, S., & Mutimba, S. (2022). Forecasting the Potential Effects of Climate Change on Malaria in the Lake Victoria Basin Using Regionalized Climate Projections. *Acta Parasitologica*, *67*(4), 1535–1563. https://doi.org/10.1007/s11686-022-00588-4

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

UNHCR. (2023). *Disease surveillance thresholds—UNHCR|Emergency Handbook*. https://emergency.unhcr.org/entry/38802/disease-surveillance-thresholds

Watkins, R. E., Eagleson, S., Hall, R. G., Dailey, L., & Plant, A. J. (2006). Approaches to the evaluation of outbreak detection methods. *BMC Public Health*, *6*(1), 263. https://doi.org/10.1186/1471-2458-6-263

WHO. (2022a, March). *Early Warning, Alert and Response System (EWARS)*. https://www.who.int/emergencies/surveillance/early-warning-alert-and-response-system-ewars

WHO. (2022b). *Malaria surveillance assessment toolkit: Implementation reference guide*. https://www.who.int/teams/global-malaria-programme/surveillance/surveillance-system-assessments

Widdowson, M.-A., Bosman, A., van Straten, E., Tinga, M., Chaves, S., van Eerden, L., & van Pelt, W. (2003). Automated, Laboratory-based System Using the Internet for Disease Outbreak Detection, the Netherlands. *Emerging Infectious Diseases*, *9*(9), 1046–1052. https://doi.org/10.3201/eid0909.020450

World Health Organization. (2021). *World malaria report 2021*. World Health Organization. https://apps.who.int/iris/handle/10665/350147

World Health Organization, World Health Organization, & Global Malaria Programme. (2012). *Disease surveillance for malaria control: An operational manual*. World Health Organization. http://apps.who.int/iris/bitstream/10665/44851/1/9789241503341_eng.pdf

Youssefi, F., Zoej, M. J. V., Hanafi-Bojd, A. A., Dariane, A. B., Khaki, M., Safdarinezhad, A., & Ghaderpour, E. (2022). Temporal Monitoring and Predicting of the Abundance of Malaria Vectors Using Time Series Analysis of Remote Sensing Data through Google Earth Engine. *Sensors (Basel, Switzerland)*, *22*(5), 1942. https://doi.org/10.3390/s22051942

Zhang, S., Zhang, L., Feng, J., Yin, J., Feng, X., Xia, Z., Frutos, R., Manguin, S., & Zhou, S. (2018). Malaria Elimination in the People's Republic of China: Current Progress, Challenges, and Prospects. In S. Manguin & V. Dev (Eds.), *Towards Malaria Elimination—A Leap Forward*. InTech. https://doi.org/10.5772/intechopen.77282

## Annex 1 Sample User Acceptance Testing tool

**User Acceptance testing survey**

User Acceptance Testing (UAT) is a crucial stage in the software development process where you as the end-users, test the system to verify if it meets their requirements and expectations. This testing phase is performed to ensure that the developed system is fit for purpose, usable and meets the user's needs before it is deployed for live use. The purpose of this questionnaire is to gather information about your experience in use of the application. The questionnaire should take approximately five (5) minutes to complete. All responses will be kept confidential. Thank you in advance for your time and contribution to our research.

**Questionnaire**

How frequently did you receive alerts or notifications from the application?

a) Rarely [ ]
b) Sometimes [ ]
c) Often [ ]
d) Almost always [ ]
e) Never [ ]

To what extent did the system accurately detect and alert users of potential malaria outbreaks in your area?

a) Very accurate [ ]
b) Accurate [ ]
c) Somewhat accurate [ ]
d) Inaccurate [ ]
e) I don't know [ ]

Were the alerts or notifications received timely and relevant to your needs?

a) Yes, always [ ]
b) Most of the time [ ]
c) Sometimes [ ]
d) Rarely [ ]
e) Never [ ]

Did the early warning system provide enough information for you to take appropriate action to prevent or control a malaria outbreak?

a) Yes, always [ ]
b) Most of the time [ ]
c) Sometimes [ ]
d) Rarely [ ]
e) Never [ ]

Was the system user-friendly and easy to navigate?

a) Yes [ ]
b) No [ ]
c) I don't Know [ ]

Did the system provide enough information for you to make informed decisions about malaria prevention and control measures in your area?

a) Yes [ ]
b) No [ ]
c) I don't Know [ ]

Were there any technical difficulties or glitches you encountered while using the system?

a) Yes [ ]
b) No [ ]
c) I don't Know [ ]

If Yes , what were the challenges in accessing and using the application?

1. Slow load times [ ]
2. Technical difficulties with the application [ ]
3. Incomplete or inaccurate data [ ]
4. Lack of training or understanding of the system [ ]
5. Insufficient data bundles [ ]
6. Others Specify ………………………………………………………………………..

Did the application provide timely and relevant updates on changes in malaria risk and transmission in your area?

a) Yes [ ]
b) No [ ]
c) I don't Know [ ]

Did the system integrate well with other relevant health and disease surveillance systems you use?

a) Yes [ ]
b) No [ ]
c) I don't Know [ ]

Did you feel adequately trained on how to use the system and understand the information it provides?

a) Very adequately
b) Adequately
c) Fairly adequately
d) Not adequate

How satisfied were you with the overall performance and usefulness of the application in detecting and responding to malaria outbreaks?

a) Very satisfied
b) Satisfied
c) Neutral
d) Fairly satisfied
e) Not satisfied

What improvements would you suggest of the application to better address challenges affecting malaria outbreak or early warning detection?

a) Improved data collection and analysis
b) Better user training and support
c) Increased funding and resources
d) Integration with other systems or tools
e) Others Specify ………………………………………………………………………..