PROJECT ABSTRACT

Master of Science in Applied Computer Science
E-Services Option

Adventist University of Africa

School of Post Graduate Studies

Title: DESIGN AND IMPLEMENTATION OF AN ON-LINE PLATFORM FOR INTEGRATION AND ANALYZING MULTIVARIATE MULTISOURCE MALARIA DATA

Researcher: Micah Asuke Ochola

Primary Advisor: Lossan Bonde, PhD

Second Reader: Collins Oduor, Phd

Date Completed: April 2023

One of the common public health problems reported by the World Health Organization (WHO) in the African Region is malaria, where the burden of the disease is highest globally. The greatest challenge experienced in the fight against Malaria is, surveillance, which leads to early detection and treatment, and is crucial for reducing transmission and preventing deaths. Malaria surveillance includes gathering, analyzing, and interpreting malaria-related data. Though there exist many facilities with Malaria data, the collection and integration of data from different sources has been a major challenge that needs to be addressed.

The proposed solution is aimed at the development of an online solution that can be used to collect malaria data from multiple sources including hospitals, drug stores and weather stations in various formats and aggregated into a format that can further be used in the prediction of malaria outbreak.

From the results, the system collects data from hospitals and drug stores, which is then integrated with weather data. The generated data was used to train a machine learning model, as a proof of concept to validate that it can it be used to predict malaria outbreaks.

This solution does not only solve the problem of data collection and integration but also ensures timely actions are taken in cases of out breaks. The implementation of this solution therefore significantly improves on the current practices by ensuring that hospital records and over the counter sale of drugs are reported electronically, daily and in real-time as opposed to manually and weekly. The solution also introduces the use of multi-source data in the analysis of malaria outbreaks rather than only focusing on hospital records as the only source of information for outbreak detection. Further to this, the project has the potential to contribute to the WHO Global Malaria Technical Strategy 2016-2030, as early detection and treatment of malaria are essential for reducing the burden of the disease. The methodology and system produced in this study can be used in other regions to improve malaria surveillance and outbreak prediction.

Adventist University of Africa

School of Postgraduate Studies

DESIGN AND IMPLEMENTATION OF AN ON-LINE PLATFORM FOR INTEGRATION AND ANALYZING MULTIVARIATE MULTISOURCE MALARIA DATA

A project

presented in partial fulfillment

of the requirements of the degree

Master of Science in Applied Computer Science

by

Ochola Micah Asuke

April 2023

DESIGN AND IMPLEMENTATION OF AN ON-LINE PLATFORM FOR INTEGRATION AND ANALYZING MULTIVARIATE, MULTISOURCE MALARIA DATA

A project

presented in partial fulfillment

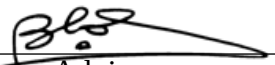
of the requirements of the degree

Master of Science in Applied Computer Science

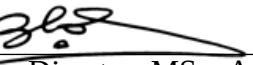
by

Micah Asuke Ochola

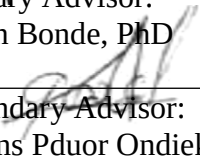
APPROVED BY:



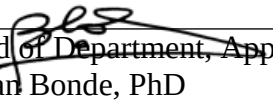
Primary Advisor:
Losan Bonde, PhD



Program Director, MSc. ACS
Losan Bonde, PhD



Secondary Advisor:
Collins Pduor Ondiek, PhD



Head of Department, Applied Sciences
Losan Bonde, PhD

:

Dean, School of Post Graduate Studies
Losan Bonde, PhD

AUA Main Campus

Date: _____

This work is as a result of countless and relentless sacrifices and is heartily dedicated to all the people who have been an inspiration throughout my course of study. From my family to classmates, lecturers and circle of friends who extended their help amid challenges while doing this project.

A special gratitude to my loving wife, Margaret Oirere whose words of encouragement and push for tenacity still rings in my ears. My Children Giovanni, Gianna, Gian and Gina you are the reason for what we do. May God bless you.

Above all, to our God Almighty who continually bestows His blessings in our everyday lives, especially for the strength, patience, wisdom, time, and guidance in realization of this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

eDW	Early Disease Electronic Warning System
IMSS	Integrated Malaria Surveillance System
MEP	Malaria Epidemic Prediction
NCAR	National Centre for Atmospheric Research
NN	Neural Network
SVM	Support Vector Machine
TP	True Positives
WHO	World Health Organization

ACKNOWLEDGMENTS

My most profound appreciation goes to Dr. Lossan Bonde and Dr. Collins Oduor, my advisors and mentors, for their time, effort, and understanding in helping me succeed in my studies. Their advice and wealth of experience have inspired me throughout my studies. In addition, I'd like to thank Mr. Samson Ooko for the technical assistance throughout my research. I'd also like to express my gratitude to the faculty of Post Graduate Studies. Thanks to their generosity and encouragement, my time spent studying at Adventist University of Africa has been truly rewarding. To conclude, I'd like to thank God, my family, my wife, and my children. It would have been impossible to finish my studies without their unwavering support over the past few years.

CHAPTER 1

INTRODUCTION

Background of the Study

Malaria is a serious disease that is curable but yet can be fatal. It is caused by the transmission of malaria parasites from person to person via the bite of an infected female Anopheles mosquito. According to data from 2020, there were approximately 241 million cases of malaria worldwide, resulting in over 627,000 deaths [1]. As displayed in figure 1 below, burden of malaria is particularly high in the African Region. According to the World Health Organization (WHO) in Africa, 95% of all malaria cases and 96% of all malaria-related deaths occurred in 2020. Shockingly, children under the age of five accounted for around 80% of all malaria-related fatalities in this region [2]. In Kenya, the situation is equally a concern, with an estimated 3.5 million new clinical cases and 10,700 deaths from malaria reported each year. Residents in western Kenya are more sensitive to infection [3]. An epidemic was recorded in Nandi County in July 2021 [4]. Unfortunately, the epidemic was not discovered quickly because of inadequate surveillance and data integration from several sources. As a result, the illness spread and may have caused more harm than if it had been recognized sooner. It is crucial to raise awareness about the seriousness of malaria and its impact on communities around the world. Improved surveillance and data integration can help detect outbreaks sooner, leading to more timely and effective responses. By working together and acting, the burden of malaria can be reduced, and prevent unnecessary suffering and loss of life.

The high prevalence of Malaria in Africa can be attributed to some of the following factors [1];

- The most common parasite species, Plasmodium falciparum, is also the most likely to cause severe cases of malaria and result in death.
- An effective mosquito species, the Anopheles gambiae complex, is responsible for the high transmission rate.
- Local climatic conditions usually permit year-round transmission. The spread and seasonality of malaria are influenced by climate, with adequate rainfall being necessary for mosquito survival and adequate temperature for parasite development [5].
- Malaria control efforts have been impeded by a lack of resources as well as socioeconomic insecurity.

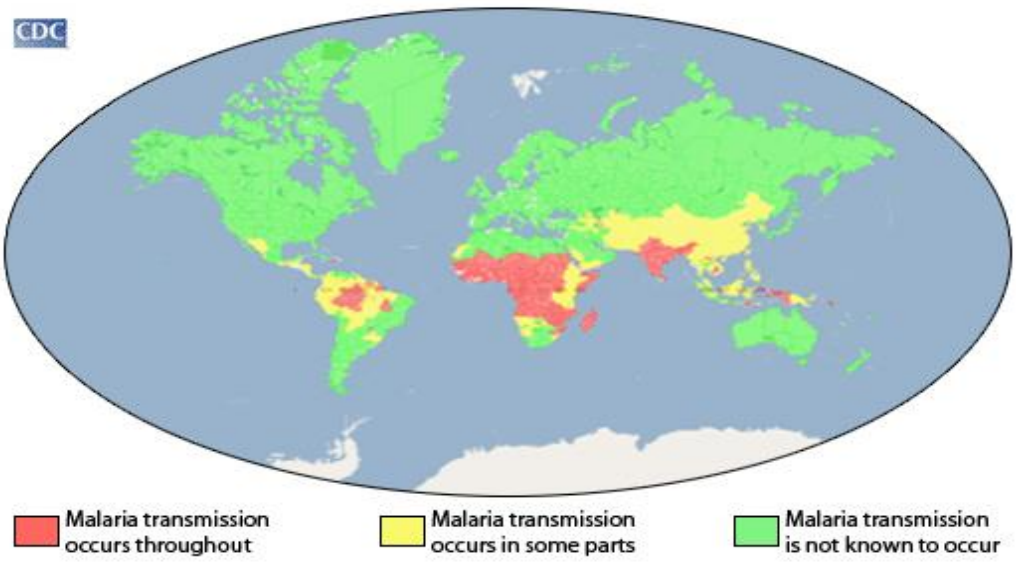


Figure 1. Global Malaria Transmission [1]

Early detection and treatment of malaria can significantly reduce illness, prevent fatalities, and aid in the prevention of disease transmission [2]. Malaria

surveillance involves the continuous and systematic gathering, analysis, and interpretation of malaria-related data, as well as its application in public health policy, implementation, and evaluation. Enhanced malaria surveillance of cases and fatalities supports health ministries in identifying the regions or demographic groups that are most affected, as well as aiding governments in tracking shifting disease trends. Robust malaria monitoring systems can assist countries in developing effective healthcare treatments and evaluating the effectiveness of their malaria control strategies [2]. In Kenya, the Canter for Disease Control (CDC) and Kenya Medical Research Institute (KEMRI) collaborate on health facility surveillance, which involves documenting illnesses identified during hospital or health center admissions [6].

It is stated in [7] that enhancing monitoring systems is critical to reducing malaria outbreaks. According to the findings of a study on malaria surveillance systems [8] common gaps across countries include a scarcity of surveillance in distant settlements or the private sector, insufficient health information architecture to collect high-quality case-based data, data from other sources, such as intervention information, are not well integrated, poor presentation of produced data, as well as its inability to make programmed decisions. In Kenya, data is acquired from several sources and formats, which takes time to combine, making the monitoring process challenging. Recent research [9] found that malaria monitoring is now accomplished by utilizing dynamic, networked systems that demand quick data sharing between diverse platforms. The occurrence of dynamic changes in one or more interacting components, which may result in inconsistencies and incompatibilities across infrastructure components, is a significant problem that these systems must overcome.

Data integration can also enable the prediction of future epidemics so actions are made in time.

The latest technologies have the potential to facilitate the development of improved analytics and timely prediction of malaria outbreaks. In Yemen, two systems and studies that aim to accomplish this are the Integrated Malaria Surveillance System (IMSS) and the Early Disease Electronic Warning System (eDEWS). Other approaches include developing a machine learning-based model to forecast malaria incidence based on climate variability [11], creating a machine learning model to predict malaria based on patient information from parasite case reports [12], and determining a suitable machine learning classifier technique for malaria incidence prediction [13]. Additionally, there have been efforts to predict malarial outbreaks using machine learning and deep learning approaches, such as the framework for Malaria Epidemic Prediction (MEP) in Ethiopia based on factors such as rainfall, relative humidity, mean temperature, elevation, and lag malaria cases [15], as well as other ongoing research in this field.

The need to integrate data from different sources and apply machine learning for better management of malaria has been emphasized. It is possible to extract new ground-breaking insights from historical or transactional data by applying machine learning and deep learning approaches, allowing us to make better-informed decisions and adopt the best tactics to cope with future events. Some research has been conducted to construct machine learning models to detect malaria using blood smear pictures; nevertheless, this technique has several disadvantages. In addition, the existing systems used for data integration cannot directly be used in Kenya and need modifications. Therefore, this study proposes the use of an online data integration

platform and a machine learning model for the prediction of malaria epidemics based on historically reported cases and weather information.

Implementing such a system would help to realize the WHO Global technical plan for Malaria 2016-2030 [16], updated in 2021 and offers a technical foundation for all malaria-endemic countries. It is meant to guide and support regional and national malaria control and eradication efforts when monitoring and case prediction are critical.

Statement of the Problem

There has been substantial progress in the creation of surveillance systems, with data now being gathered electronically by various healthcare providers, laboratories, and government organizations at the worldwide, national, state, and local levels [17]. Because of the availability of this data, public health authorities have been developing access, analysis, and reporting requirements that are met through in-house infectious disease information systems. However, building and deploying such technologies does not assure successful infectious data collection and exploitation in wider settings [18]. Additional technological and policy difficulties must be addressed. Some of these difficulties are as presented below:

To begin with, Figure 2 demonstrates that the majority of existing systems for disease control were developed independently, as described in [18]. This lack of integration between systems can lead to difficulties when sharing information between organizations, which may result in manual methods such as email attachments and data re-entry. These methods can be time-consuming and prone to errors, which may hinder the ability of disease-control organizations to respond to outbreaks effectively. Furthermore, most search and data analysis features within these systems are limited to corporate users only, which can restrict access to

important information for researchers, healthcare providers, and government entities. Real-time data sharing, particularly of databases, could significantly enhance scientific review and response times by utilizing input and action triggers provided by various government entities. By sharing data in real time, organizations can access up-to-date information and respond quickly to outbreaks, which is essential for effective disease control. In summary, the lack of integration between existing disease control systems can pose challenges to information sharing and hinder effective outbreak response. Real-time data sharing and access to search and analysis features can significantly enhance scientific review and response times, ultimately improving the effectiveness of disease control efforts. By addressing these challenges, the researchers can improve our ability to respond quickly and effectively to outbreaks, ultimately leading to better health outcomes for individuals and communities.

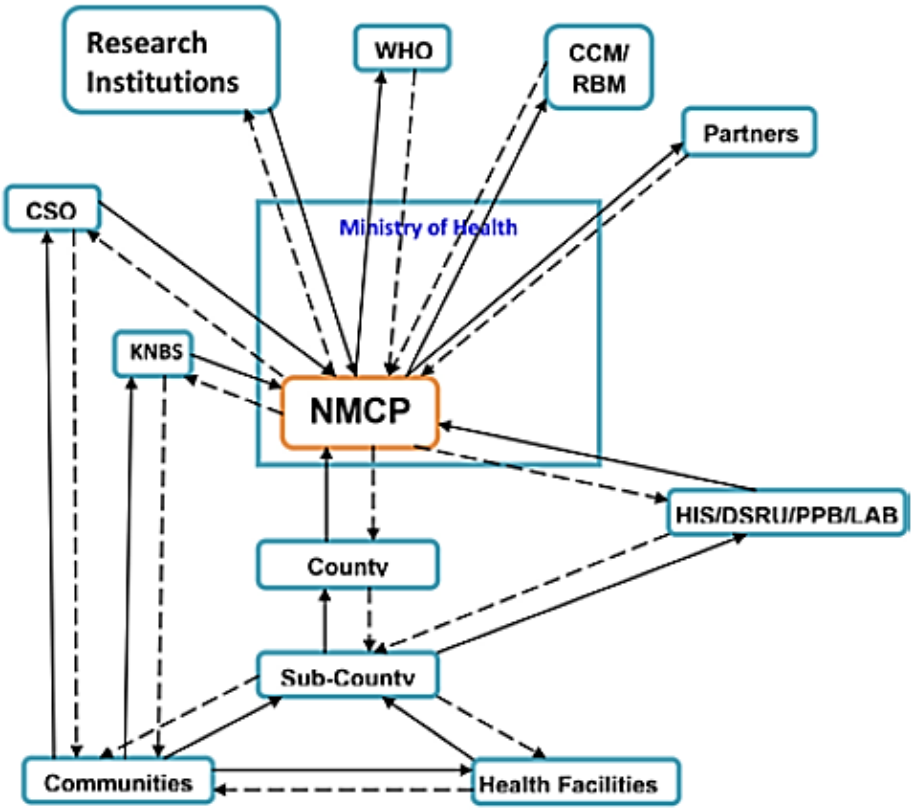


Figure 2. Data Sources for Malaria M & E Plan [18]

In addition, the emergence of digital systems has led to a significant increase in the amount of information that can be collected and analyzed for every public health epidemic. However, despite statistical packages for analysis and health information systems for data collection, current surveillance systems lack accessible multivariate data integration for data analysis and visualization. This creates a significant gap in the ability to process information promptly and limits support for professionals analyzing data and developing predictive models. Therefore, there is a pressing need for an integrated data environment platform that can capture multivariate data and use big data analysis techniques for predictive modeling and visualization. An information system that integrates data from collection to modeling and visualization-facilitated analysis will result in higher quality data input, more timely analysis, better epidemic prediction analysis, and possibly improved disaster event management.

Moreover, the current reporting and alerting method often relies on manual processes and human involvement, which can lead to delays and errors. This can be particularly problematic when swift action is needed to prevent the spread of infectious diseases and to protect public health. Therefore, there is a need for more efficient and automated reporting and alerting systems that can quickly and accurately transmit critical information to relevant stakeholders. Such systems should be designed to integrate seamlessly with existing public health information systems and be capable of providing real-time alerts and notifications. By leveraging advanced technologies such as artificial intelligence and machine learning, these systems can enhance public health response capabilities and help prevent the spread of infectious diseases.

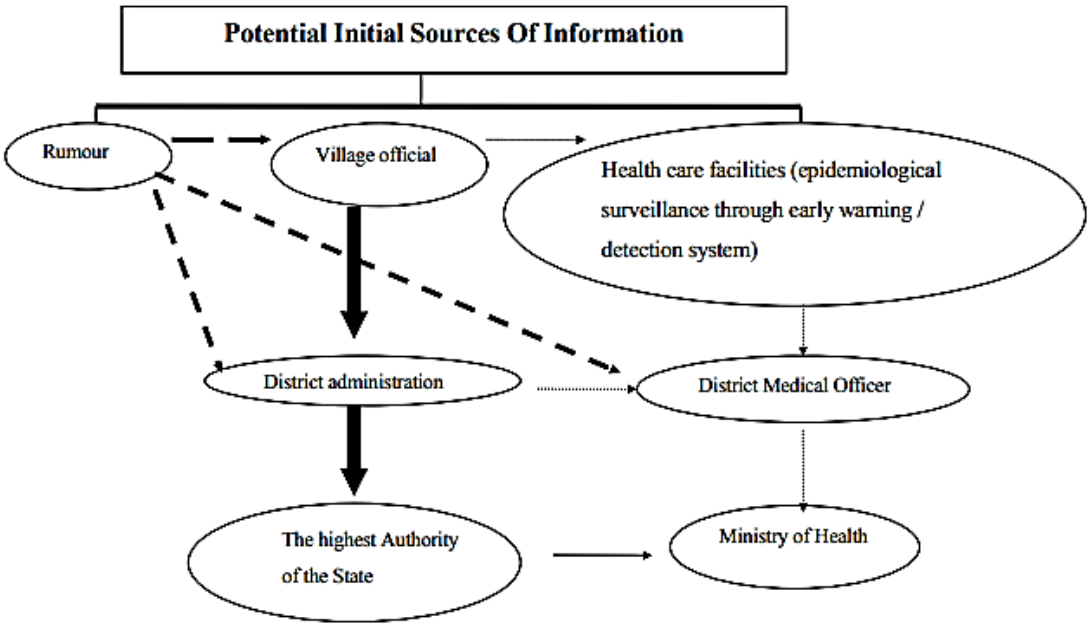


Figure 3. Current Ways of Reporting Any Unusual Events from Communities/ Villages

Scope and Limitations of the Study

The general objective of the research was to design and model an online platform that will capture and integrate multivariate data that can be used for malaria surveillance and improve outbreak detection and notification. With the federation of data from various sources, the platform will then generate aggregated data that can be used to train a machine learning model to predict outbreak of malaria from time to time. The research assumes that the techniques used to design the platform are scalable for the surveillance of other diseases.

For this study data was collected from Nandi country in Kenya with the weather data being collected from online platforms. The data integration system is focused on the integration of data of confirmed cases from different hospitals, data from drug stores and the prevailing weather data. A machine learning model was also trained using the aggregated data and tested based on reported cases and the prevailing weather conditions.

The main limitation of this study was the reluctance of healthcare providers to give the needed information this was overcome by giving assurances to ensure the privacy and confidentiality of the collected information.

Significance of the Study

Having a well-developed disease surveillance and outbreak detection platform will enable:

1. Forecasting on the trends, or increase of outbreak weeks or months in advance to better respond to Malaria threats.
2. Integration big data in malaria surveillance of multivariate data improving where traditional surveillance has largely been ineffective and expensive.

The benefits of more integrated surveillance and response systems are numerous and far-reaching. By reducing illness cases and deaths, such systems can contribute to a healthier population and a more productive workforce. In addition, they can help to minimize the economic impact of epidemics, which can affect individuals, businesses, and entire countries. Moreover, integrated surveillance systems can improve communication and coordination among government agencies, healthcare providers, and public health organizations, leading to more effective response operations and better resource use. Furthermore, by providing consistent and uniform data, these systems can provide disease pattern comparisons across different nations and periods, allowing for more accurate global disease monitoring and prevention.

Operational Definition of Terms

Accuracy: The degree to which a projected value matches the actual value.

Confidentiality: The safeguarding of sensitive data against unauthorized access, disclosure, or usage.

Data Integration: Data integration is the process of merging and harmonizing data from several sources into a single format suitable for analysis.

Deep Learning: Deep learning is a type of machine learning that use multiple-layer neural networks to uncover complex patterns and connections in data.

Machine Learning: Machine learning is a subset of artificial intelligence in which computers learn from data and improve their performance on a specific activity without being explicitly directed to do so.

Malaria surveillance: Malaria surveillance is the systematic and continual collection, analysis, and interpretation of data on malaria cases and their distribution to inform public health policies.

Malaria: A parasitic illness caused by the Plasmodium parasite and spread by Anopheles mosquito bites.

Multivariate data: Information including two or more interdependent variables with a cause-and-effect connection.

Online platform: A software program accessible via a web browser and accessible via the internet.

Outbreak detection: The discovery and communication of an elevated occurrence of a disease in a specific region that exceeds predicted values.

Prediction: The process of estimating future outcomes or occurrences using historical data and statistical models.

Prevailing weather conditions: The present and projected weather trends in a certain location, such as temperature, humidity, and rainfall.

Privacy: Individuals' personal information, including their health state, should not be divulged or shared without their agreement.

Prototype: Prototyping is a software development approach that entails creating a functional model of a system to enhance and verify its design before completely implementing it.

Reported Cases: Malaria cases that have been confirmed and reported by healthcare practitioners are referred to as reported cases.

Synthetic Data: Synthetic data generation is the process of creating artificial data by employing statistical models and algorithms to produce data that mimics real-world data.

Visualization dashboard: A visualization dashboard is a graphical user interface that displays data in a visual manner, such as charts, graphs, and maps, to allow users to easily assess? information.

CHAPTER 2

METHODOLOGY

For the creation of the multivariate data-gathering information system, the software prototyping methodology was chosen as the software development method. The choice of this architecture was influenced by the need to understand user needs and develop system requirements at an early stage. Furthermore, user input was obtained, allowing the researcher to grasp what is anticipated from the solution. Some of the benefits of software prototyping

- The prototype is a depiction of the proposed solution with fewer features.
- Before installation, users can assess the system's performance and provide ideas for improvement.
- The system developer is given the chance to comprehend needs that may have been considered in the early phases.
- accelerates the system development process
- It is possible to get a clear and complete grasp of the needs.

Overview of the Approach

The system prototype process was divided into four main stages; functional selection, system creation, system assessment, and system usage. The initial stage involved choosing the functionalities which functionalities which were to be prototyped based on the objective and scope of the projects. These functionalities were mainly, an interface for data entry from various sources, an interface to create

accounts and set parameters, a capability to integrate the multivariate source data and generate an aggregated data that can be used in training a model to predict malaria outbreak. This was followed by designing and building the prototype, the online system was then reviewed in the requirements reengineered and refined to meet the system needs. The system is currently undergoing testing for future improvements even though it meets the initial user requirements.

The prototyping process was an iterative process and in cooperated reviews on behalf of different stakeholders as the design was improved. The third and fourth phases were repeated until the system performed to acceptable levels. The steps can be summarized as presented in figure 4 with the following main components

a) Initial statement of requirements.

b) Prototype creation.

c) The prototype is built, tested, and put to use.

d) Prototype revision and refinement.

Several iterations are carried out during the process, with the third and fourth phases being repeated until the system is accepted by the user, as seen in Figure 4

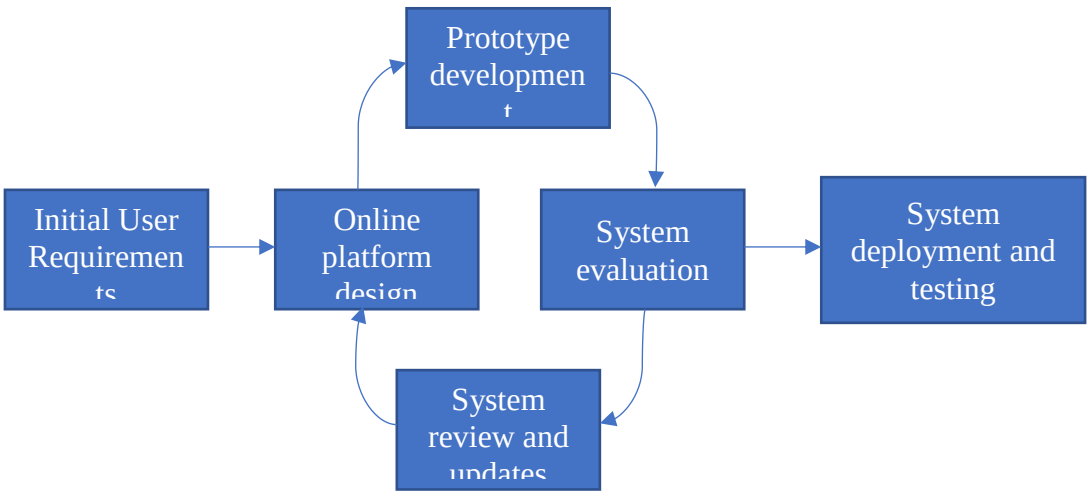


Figure 4. System Prototyping Steps

Online Platform

General System Flow

The system has user accounts for the hospitals and drug store to enter the reported malaria cases and drug store. In addition, an API is used to be able to pull weather data from an approved weather site. All the data from multi source locations and in multivariate forms are stored in the system. The system further processes the collected data and aggregates the same to produce a processed version which can be used to train or inference to predict an outbreak of malaria or not. Different users should be able to view the collected and processed information. Figure 5 show the unified platform data collection and integration process

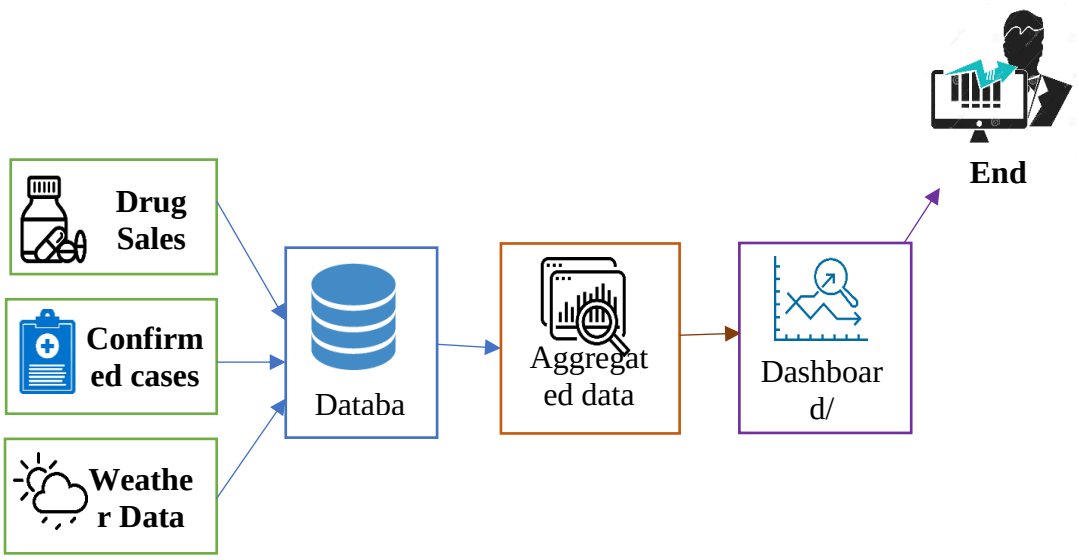


Figure 5. General System Flow

System Tools

The following tools were used in the development of the system;

- Server: Apache Tomcat was used as the web server for the system. It is a popular open-source web server and servlet container for deploying and

running Java-based web applications. Because of its scalability and dependability, it is frequently utilized in industry.

- Back end: For the development of the back-end Java Spring was used. It is a feature-rich framework for developing Java web applications. Its characteristics, which include inversion of control, dependency injection, and aspect-oriented programming, assist developers in creating scalable and maintainable programs.
- Front End: Both Thymeleaf and Bootstrap were used in combination in developing the front end. Thymeleaf is a server-side Java template engine that allows developers to generate HTML templates for server-side rendering. It is frequently used in Spring Framework-based web applications. Bootstrap is a free and open-source front-end framework for building responsive web pages and online apps. It offers pre-built UI components like buttons, forms, and navigation bars that may be changed to meet the requirements of a given project.
- Database: PostgreSQL was used as the relational database. It is a well-known open-source relational database management system that is noted for its resilience, dependability, and support for sophisticated SQL capabilities. It's widely used in online applications that need a scalable and efficient database backend, such as those used to manage malaria data.

Data Analysis Process

Big data analytics can revolutionize healthcare by enabling healthcare providers to leverage specialized tools to extract insights from clinical and other data repositories and achieve specific outcomes. The five key phases involved in big data healthcare analytics are data acquisition, data storage, data management, data

analysis, and data visualization and reporting. These phases are depicted in Figure 6, which illustrates the process of big data analysis in healthcare management.

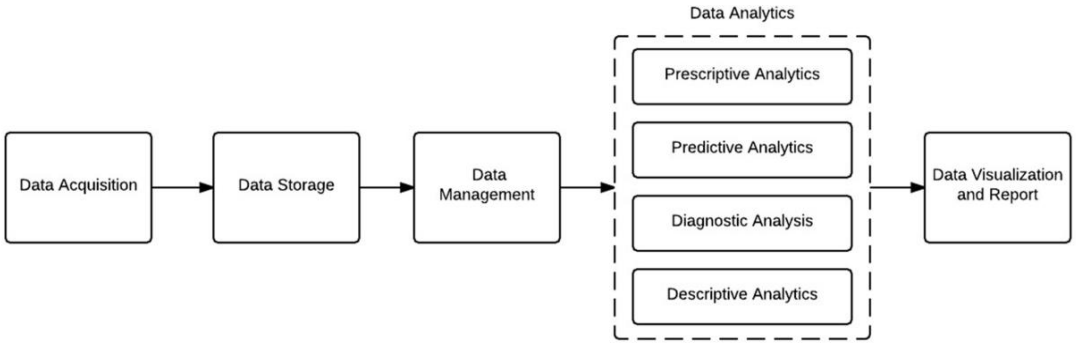


Figure 6. Data Analysis Steps

Machine Learning

Machine learning was used was used to validate the implementation of the aggregate data and the tentative results. The objective of the study was not to implement the machine learning, but a moel was trained as a proof of concept that data collected can be run through the system can be used to train and inference to predict cases of malaria outbreaks.

Machine Learning Steps

The main objective of machine learning is to extract valuable insights from data. Therefore, data plays a crucial role in unlocking the potential of machine learning, the designed system will help collect even more of such data to enable more robust models. The machine learning model training was done in five different steps as presented in Figure 7, each of the steps was focused on leveraging the power of data. The steps are a move from the traditional seven steps as the steps relating to data

have been combine and data collection and management which will involve, collection of data, data cleaning and data validation.

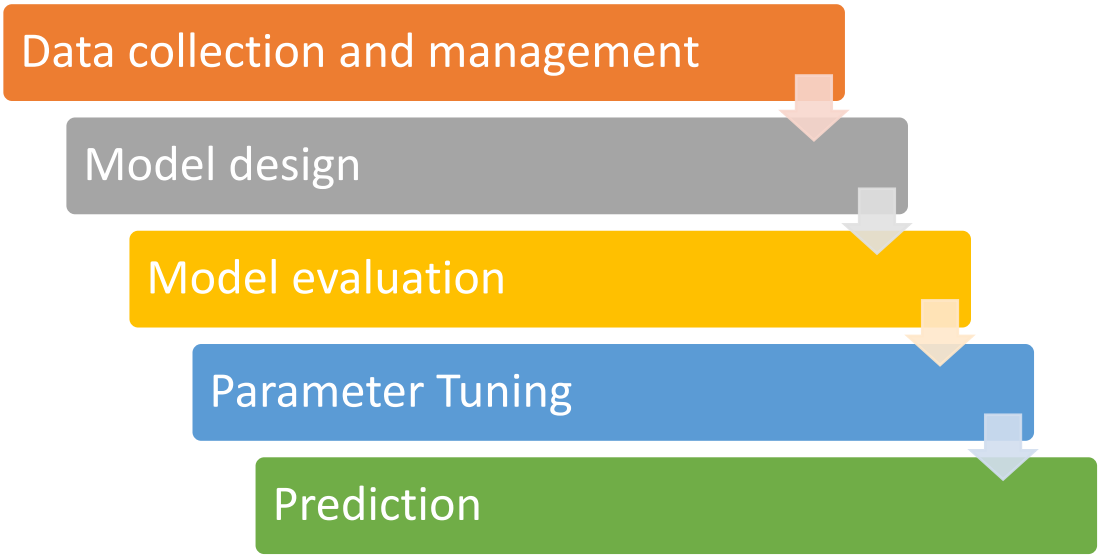


Figure 7. Machine Learning Steps

The process of machine learning was centered around data, to extract meaning from it. The machine learning process began with data collectoion, which included gathering meteorological data and monthly reported malaria cases from 2016 through 2021. This was followed by data preparation, which involved cleaning, validation, and looking for trends and outliers in the obtained data. The data collection and management process was followed by the model design which involved the selection of the optimal model to utilize, taking into consideration aspects such as data type, model purpose, preparation time, scalability, and accuracy. After selecting the model the model training step began, during which the model was developed using labelled sample data and gradually modified to generate better predictions. After training, the model was tested against an unused control dataset to determine its performance in real-world applications. The parameter tuning step included evaluating and modifying the original parameters to improve the model's accuracy. Finally,

predictions were done using the model after going through all of the preceding processes of data collection, preparation, selection, training, and testing.

Machine Learning Tools

The study utilized two tools for machine learning: COLAB and Edge Impulse. COLAB is a Google Research product that allows users to write and run Python code in the browser, making it a suitable tool for machine learning, data analysis, and education. On the other hand, Edge Impulse is a popular embedded machine learning development platform used by over 1,000 organizations and over 10,000 machine learning projects globally.

CHAPTER 3

PROPOSED SOLUTION

In this section the proposed solution for malaria data collection and integration platform is presented in details

Data Collection and Integration Platform

System Requirements

The project's main goal is to gather data on malaria cases from hospitals, monitor anti-malarial medicine sales from registered shops, incorporate meteorological data, and aggregate the data from various sources making it possible to apply a machine learning model to predict outbreaks. In addition, a data visualization dashboard with period filtering is required. The requirements listed below offer a full description of what is required for the system, including whether or not the function is a project minimum need.

- The system should provide a registration form for the hospital/health officers to sign up as users, which requires them to enter their staff ID, name, hospital, and role.
- The system should allow registered users to submit reported cases each day through an online form, which includes details like the number of cases, the type of disease, location, and date.
- The system should provide a dashboard that displays the data submitted by the hospital/health officers in a graphical format.

- The system should generate outbreak alerts if it detects a sudden increase in the number of reported cases in a particular location or disease.
- The system should provide a user management feature that allows the admin to add or remove system users, hospitals, or stores.
- The system admin should be able to update the system settings or configurations.
- The system should have an option to enter weather information or use an API to automatically fetch the data.
- The system should provide a registration form for drug sellers to sign up as users, which requires them to enter their staff ID, name, store, and role.
- The system should allow registered users to submit reported sales each day through an online form, which includes details like the type and quantity of the drug sold, location, and date.
- The system should provide a dashboard that displays the sales data submitted by the drug sellers in a graphical format.
- The system should aggregate the data collected from the hospitals/health officers and drug sellers.
- The system should send alerts and update the dashboard if it detects any potential outbreaks or unusual trends in the data.

Non-Functional Requirements

Non-functional requirements are features of quality that specify how a system should behave. These are some examples:

- Availability: The system's capabilities and services should be available for use with all operations 99.99 percent of the time.

- Usability: The system should be easy to use.
- The system should be dependable for at least ten years.
- Scalability: The system must be able to expand without compromising its performance.
- Data Integrity: The system must be capable of protecting users' access to personal data.
- Performance: The system should always respond to various user activities.
- Recoverability: In the event of a breakdown, the system should feature a self-recovery backup solution.
- Flexibility: A service-based architecture that is adaptable will be extremely desirable for future expansion.
- Security: guarantee that the program is safeguarded against unwanted access to the system and its data.

It is critical to review the requirements and capture data to develop a thorough list of functional and non-functional demands while designing and building the system. This analysis is essential throughout the system design stage because it provides a thorough grasp of the project's aims and objectives. Designers may discover the fundamental functions that the system must accomplish, as well as any limits or limitations that must be considered, by assessing the requirements to capture data. This exhaustive set of functional and non-functional criteria may then be used to steer the system's development, ensuring that it satisfies the project's objectives and specifications.

System Overview

The system is made up of three main conceptual pillars as displayed in figure 8, the data collection pillar, Data integration pillar and the data visualization pillar. The pillars will be able to work together to achieve the overall system goal by use of the laid down infrastructures.

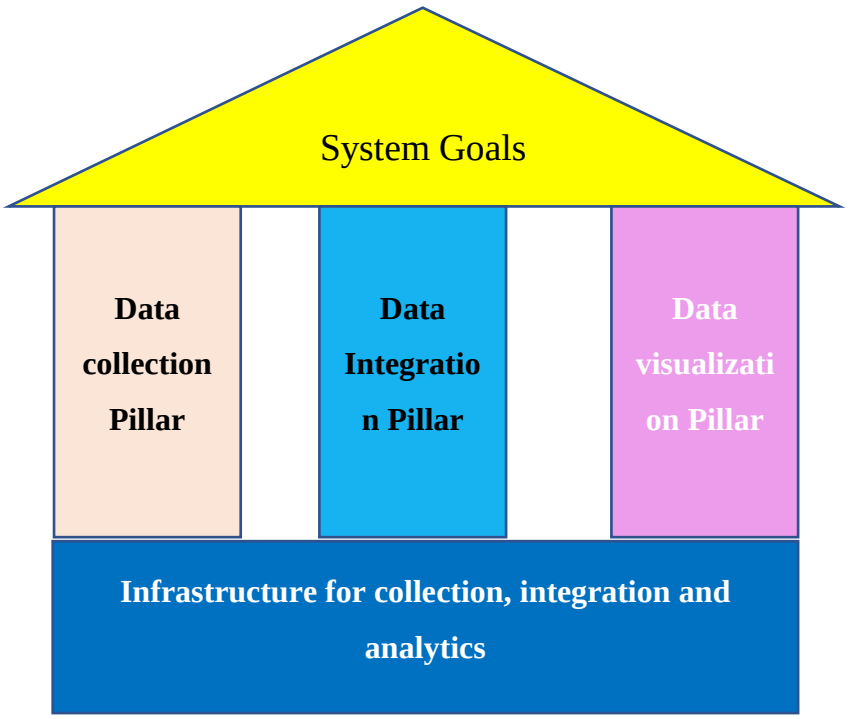


Figure 8. System Overview

Conceptual Design

To accomplish its missions, conceptually the system has three major subsystems as presented in figure 9. The system has two main separate content repositories namely, the data management repository and integrated data repository. So as to enhance security the two systems are only accessible to authorized users within defined networks. The data management repository supports the daily collection of data, such as accepting entry of daily sales and drug sale and also weather updates from an API. The integrated data repository stores the integrated data

that has been processed from those in the data management repository, the two repositories communicated when needed but have independent storage. The access is in the data integration zone for access to the daily trends and prediction of malaria outbreaks.

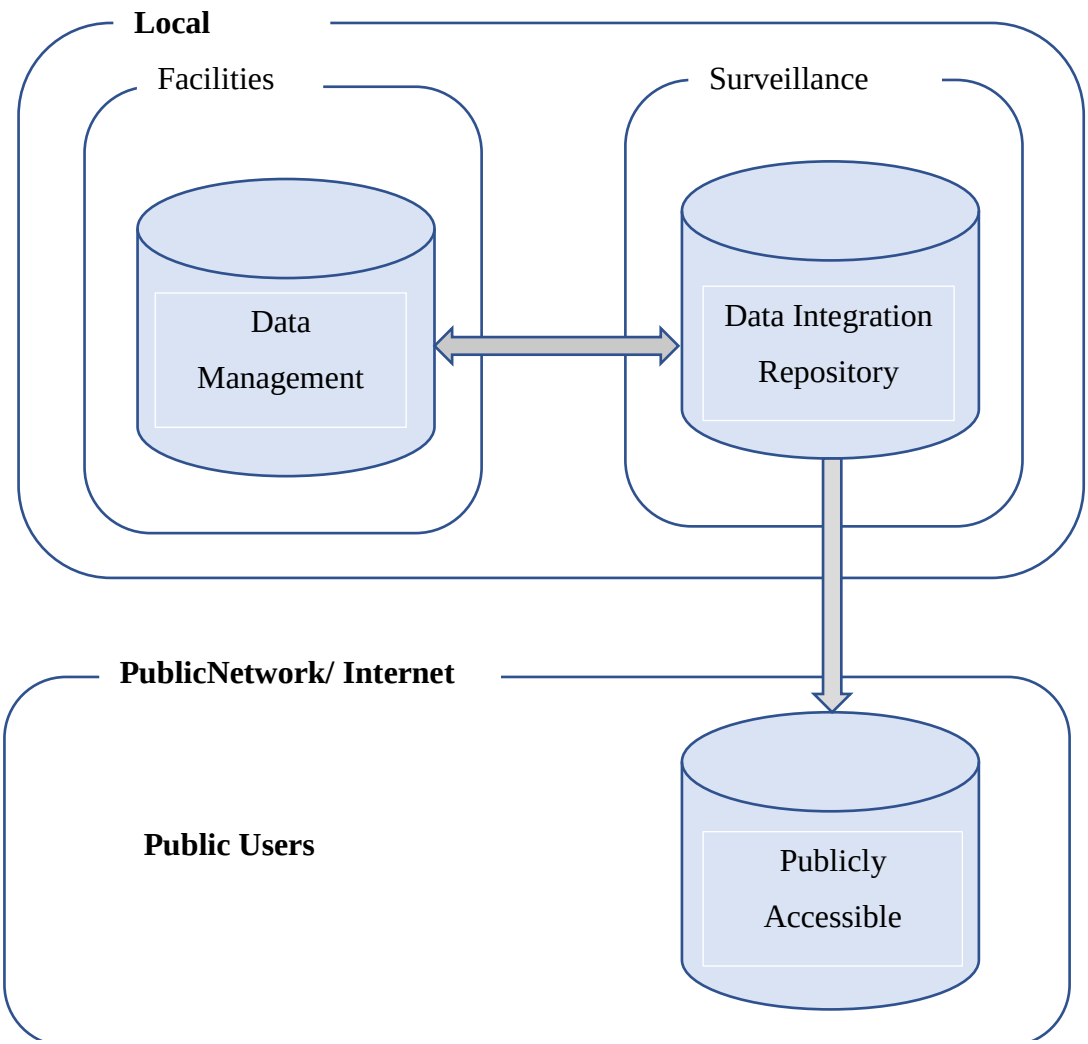


Figure 9. Conceptual Design

User Characteristics

The data collection and integration platform support two categories of users: authorized users and public users. The data management and data integration subsystems are only accessible to the authorized users. Authorized users are further

categorized to functional specialists and system administrators or managers. The following lists the specialists and managers that are supported.

- Cloud systems provider
- Health specialist
- Pharmacist
- Drug sellers
- Hospital Records specialists
- Health officers
- Public Users
- System Administrators

Each authorized user will be assigned responsibilities that govern what the user may do after successfully authenticating with the system. The data collection management subsystem is also restricted to approved users with data collection rights. The access subsystem is available to the general public without authentication. Public users will have the possibility to personalize the appearance of their accounts.

Use Cases

A Use Case diagram as displayed in figure 10 can help communicate ideas to users and stakeholders. When completed, it also gives a high-level knowledge of the system's functional needs. The Use Case diagram assists in defining the system's scope and bounds by identifying actors, use cases, and their interactions. It also serves as a foundation for developing system requirements, creating test cases, and verifying that the system satisfies the demands of the users. Overall, the Use Case diagram is critical during the system design stage, assisting in ensuring that the system is developed to satisfy the needs of the users.

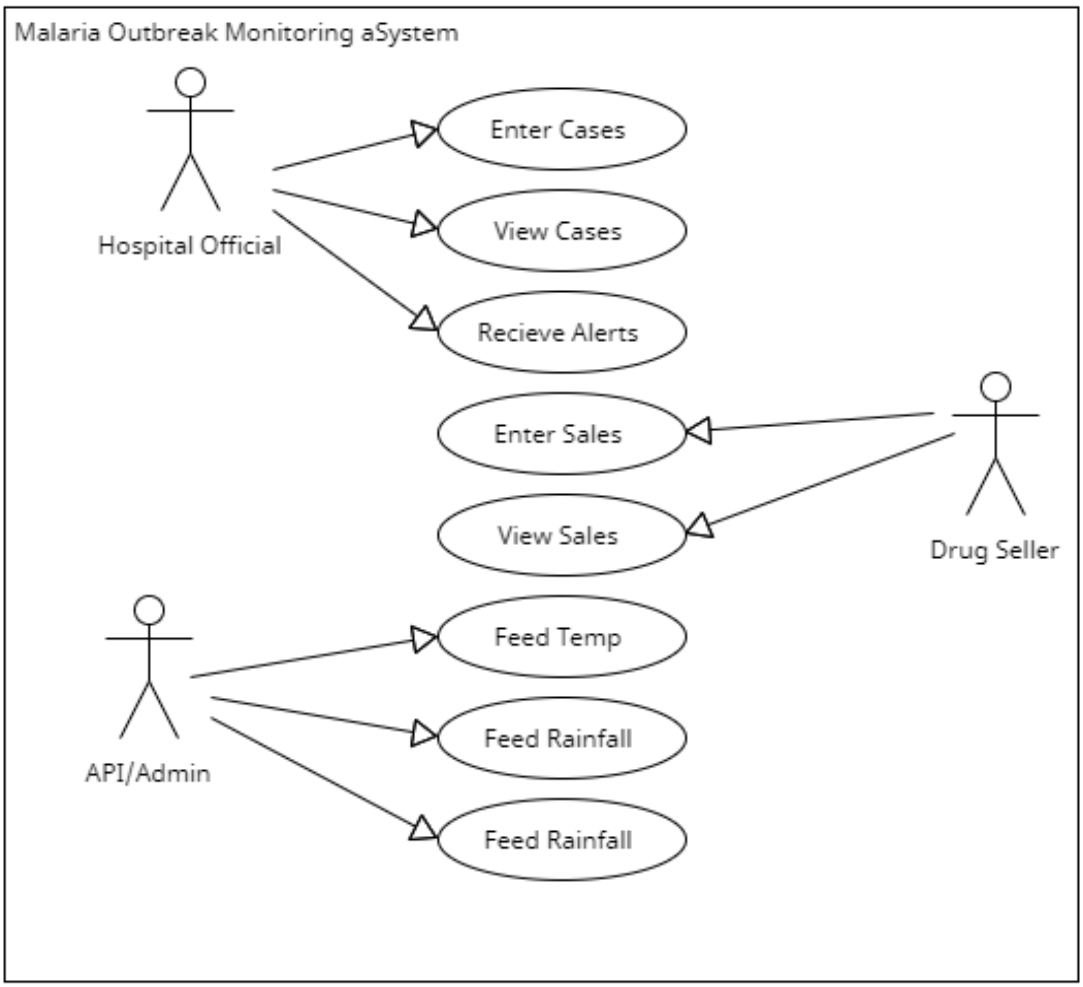


Figure 10. Use Case Diagram

Process Flow

A process flow diagram is a visual representation of the steps involved in a particular process or workflow. In the context of the malaria outbreak prediction system, a process flow diagram would outline the specific steps involved in collecting, processing, and analyzing data, as well as the interactions between different components of the system. A process flow diagram may be useful in a variety of ways. It gives a clear and comprehensive summary of the whole process, which makes it easier to grasp and identify possible areas for development or optimization. It can also aid in the identification of possible bottlenecks or inefficiencies in the process, allowing for targeted changes.

A process flow diagram can also assist in identifying relationships between system components. For example, the data-gathering process may be contingent on the availability of specific resources or data sources, which may influence the overall accuracy of the system's predictions. Finally, a process flow diagram may be utilized as a tool for communication and cooperation among project stakeholders. It may enable conversation and feedback by giving a visual depiction of the process, as well as assuring that everyone engaged has a clear grasp of the system's design and operation. Figure 11 displays the process flow for authentication of users who will be using the system.

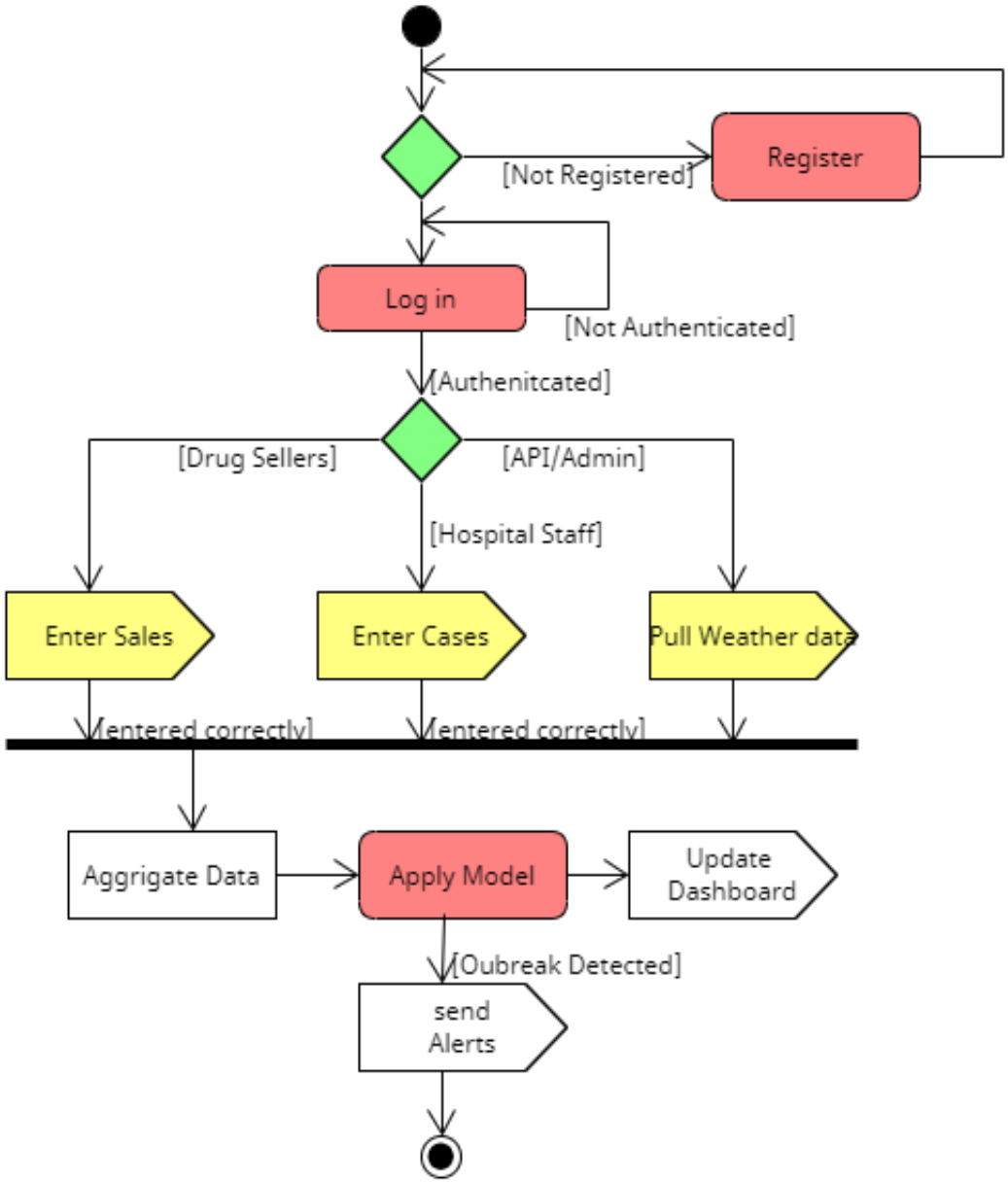


Figure 11. System Process Flow

Component Diagram

The component diagram shows the various components of the system that work together to achieve the overall system goals. The main components are; the hospitals where cases are recorded, the drug stores where the sales information is recorded, the health officer components to view different aspects of the system, the data integration and visualization components. In addition these components interact

with the database thought the security and persistence components s shown in figure 12.

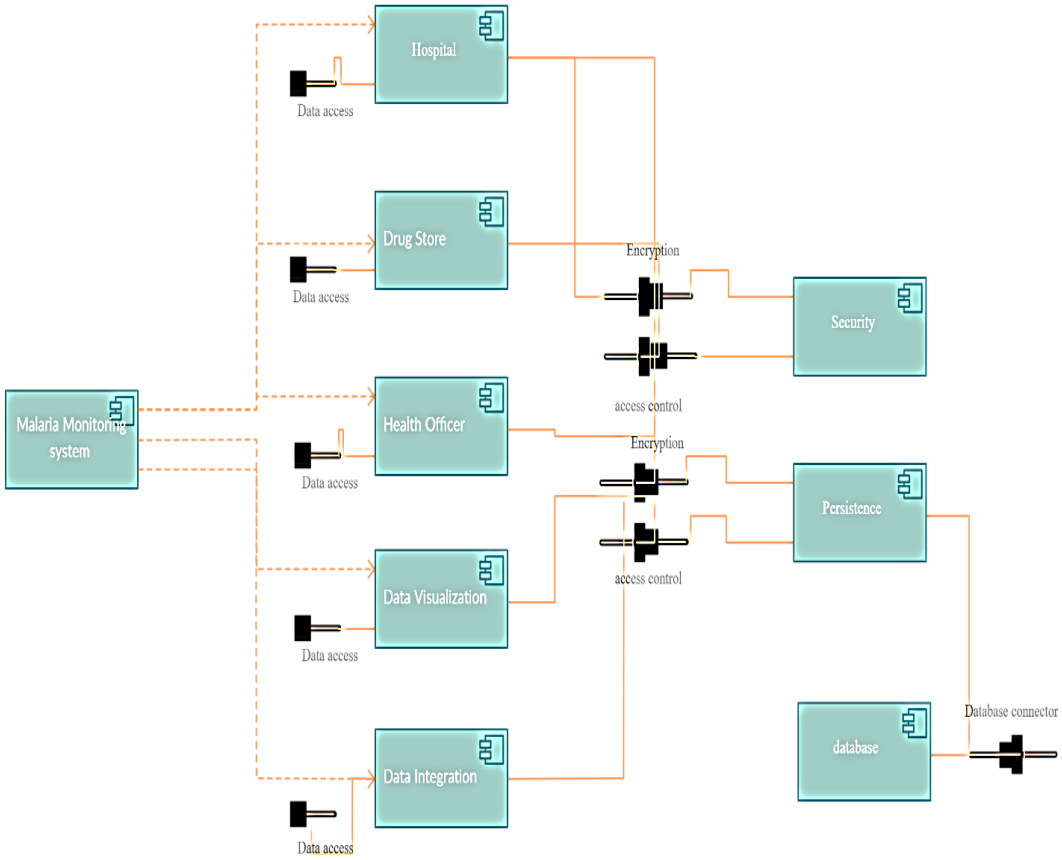


Figure 12 Component Diagram

Log in Sequence Diagram

Figure 13 show the log in and authentication sequence for the system

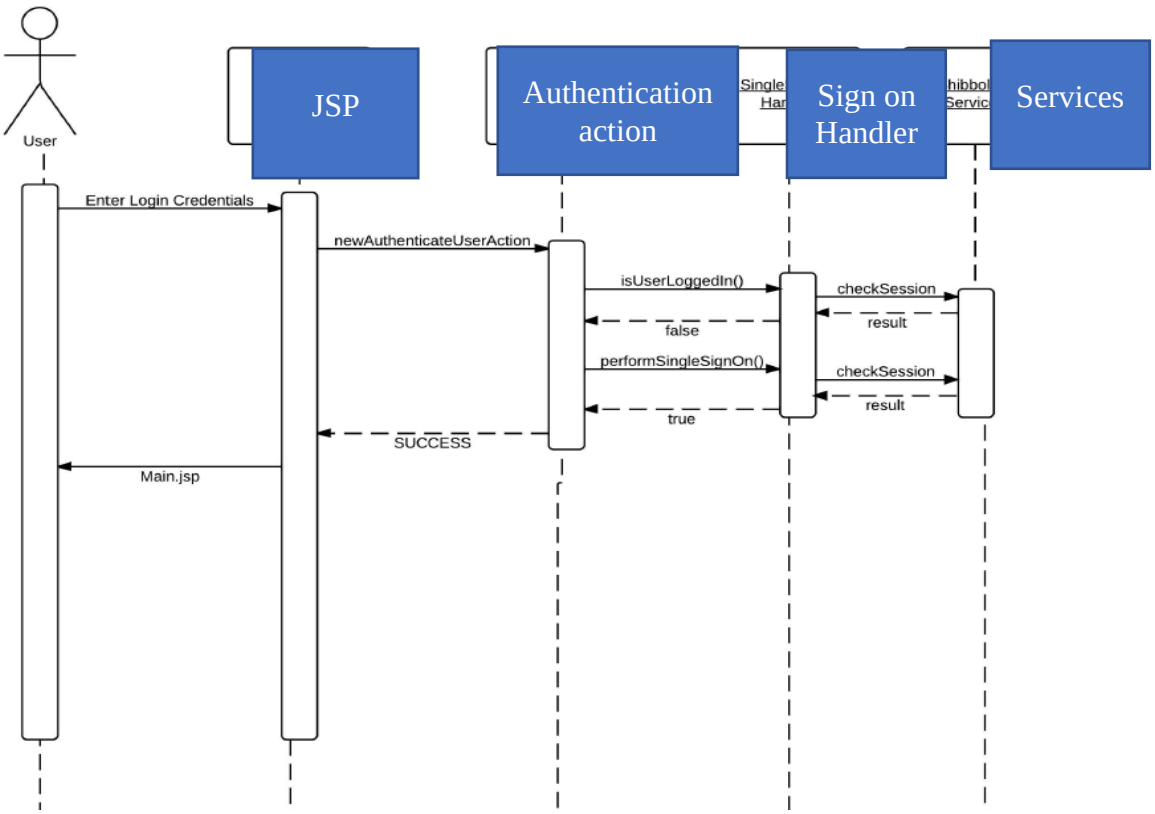


Figure 13 Login Sequence Diagram

System Implementation

A relational database was used in implementing the solution. The system is made up of several tables namely;

- User Table- Contains the users' details (UserID, user name, email, Mobile number, FacilityID)
- Authentication table- Contain user credentials (UserID, Password)
- Facilities Table-contain information about the facilities (Fields: Facility ID, Facility name, Facility Type, Facility level, email address, Mobile Number)

Table 1. Sample Facilities Table

Code	Facility	Type	Level	Email	Mobile	Location	County
178593	Alexandria General Hospital-Kapsab	Hospital	Level 5	ocholam@g	+254722134156	Kapsabet	Nandi
136078	Alpha Hill Medical Center Ltd	Hospital	Level 5	ocholam@g	+254722134156	Kapsabet	Nandi
131156	Apex Medical Services	Hospital	Level 5	ocholam@g	+254722134156	Nandi Hills	Nandi
13212	Baraton Jeremic Community Medica	Hospital	Level 5	ocholam@g	+254722134156	Kapsabet	Nandi
131157	Bethesda Medical Clinic	Clinic	Level 5	ocholam@g	+254722134156	Nandi Hills	Nandi

- Cases Table – Contains reported cases per day (FacilityID, Cases recorded, Description, Date)

Table 2. Sample Reported Cases Table

Date	Cases	Facility ID	Facility	Description
01-Jan-19	115	88324	Apex Medical Services	Ok
02-Jan-19	115	88324	Apex Medical Services	Ok
03-Jan-19	115	88324	Apex Medical Services	Ok
04-Jan-19	115	88324	Apex Medical Services	Ok
05-Jan-19	115	88324	Apex Medical Services	Ok
06-Jan-19	115	88324	Apex Medical Services	Ok

- Drug Table – Contains reported sales per day (FacilityID, Quantity sold, Description, Date)

Table 3. Sample Over The Counter Sales Table

Date	Facility ID	Quantity Sold	Facility	Description
1-Jan-2019	9939	115	Kamobo Chemists	Malaria Drugs
2-Jan-2019	9939	115	Kamobo Chemists	Malaria Drugs
3-Jan-2019	9939	115	Kamobo Chemists	Malaria Drugs
4-Jan-2019	9939	115	Kamobo Chemists	Malaria Drugs

- Climate Table – Contain the climatic data (Date, Rainfall, Temperature, Humidity)

Table 4. Sample Climatic Data Table

Date	Temprature	Humidity	Rainfall
1-Jan-2019	24	56	1.0
2-Jan-2019	23	56	0.1
3-Jan-2019	25	56	0.5
4-Jan-2019	24	63	2.5
5-Jan-2019	24	60	2.2
6-Jan-2019	23	60	2.2
7-Jan-2019	24	47	0.8

- Aggregation Table- Contains the aggregated data (Date, Drug sold, Recorded cases, Rainfall, Temperature, Humidity)

Table 5. Sample Aggregated Data Table

Date	Quantity Sold	Drugs Sold	Temprature	Humidity	Rainfall	Confirmed Cases
1-Jan-2019	115	367	24	56	1	345
2-Jan-2019	115	489	23	56	0.1	356
3-Jan-2019	115	379	25	56	0.5	345
4-Jan-2019	115	279	24	63	2.5	435

Figure 14 gives the layout of the main user interface

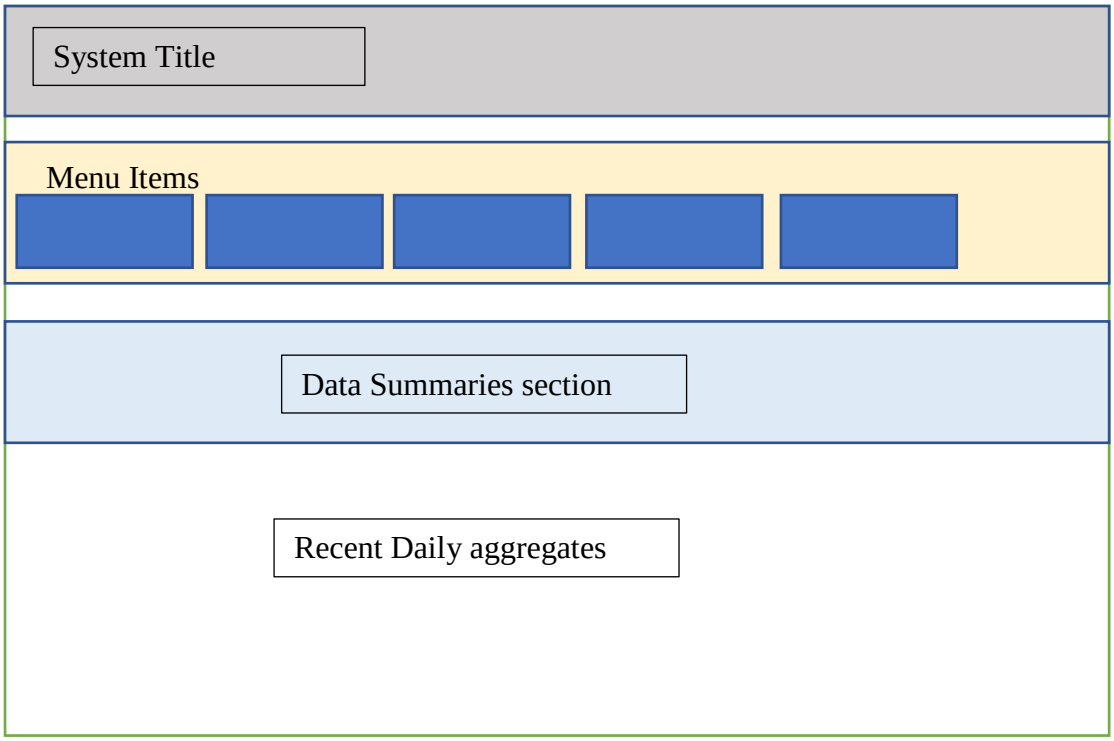


Figure 14. Interface Layout

System Deployment

The system was deployed as a prototype and can be accessed using the links provided:

URL: http://38.242.238.38:8080/login

Admin UserId: admin@gmail.com

Admin Password: datamanager

Machine Learning Model

Input: Raw Data

The officials in charge of health in a selected county of Kenya provided data on confirmed malaria incidence rates for all sub counties from 2016 to 2021. The dataset contained a normalized yearly figure of confirmed malaria cases per 1000 population, calculated by dividing confirmed cases by the corresponding population size. Confirmed malaria occurrences were cases that had been verified and documented by various hospitals and healthcare institutions and then reported to the

county's national health information system. Figure 15 displayed the county's yearly malaria incidence and drug sales over the past five years, ending in 2021.

Climate data were obtained from the National Centre for Atmospheric Research (NCAR) archive. The NCAR dataset consisted of observational data from 2016 to 2021, including daily data on temperature, rainfall, and relative humidity. Since monthly records of malaria incidence were not available in the selected county, this study only analyzed monthly records of both meteorological factors and malaria incidence reports.

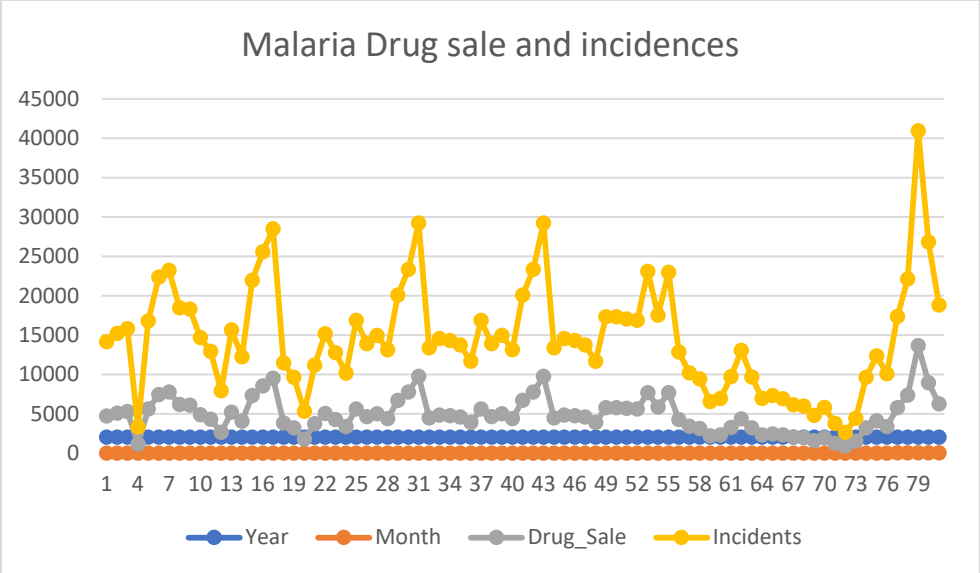


Figure 15. Drug Sales and Malaria Incidences

Data Pre-Processing

The dataset was standardized using the min-max scaler to bring them to the same scale. Then, one of the WHO-recommended approaches for setting malaria incidence thresholds was used to convert the target variable from continuous to categorical variables. The method involved taking the mean of the previous five years

(n=5) and adding two times the standard deviation (SD).

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

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

Malaria incidence threshold = x + 2(SD)

Where: *x = population mean xi = annual incidence report n = total number of years*

SD = Standard Deviation

The thresholds attained were 18362. As a result, anytime the number of malaria incidences exceeds these criteria, it is considered a high incidence, and vice versa. Outbreak and No Outbreak were the two outcome types for the target variable.

Table 6. Sample Monthly Average Data Table

Year	Month	Temprature	Rainfall	Drug_Sale	Incidents
2016	January	23	248.13	4719	14158
2016	February	25	42.77	5072	15217
2016	March	26	76.12	5274	15822
2016	April	22	320.28	1114	3341
2016	May	20	260.34	5592	16777
2016	June	20	135.6	7447	22341
2016	July	20	74.08	7741	23223
2016	August	22	104.24	6157	18471
2016	Septembe	22	159.64	6092	18275

Model Training

In this project, the data sets were preformatted using a custom Python script before being submitted to an online platform called Edge Impulse. Edge Impulse is a platform that allows users to develop and deploy machine learning models for embedded systems. The preformatted data sets consisted of three groups: one with outbreak season samples, one with data from no outbreak seasons, and a test set

containing 20% of raw data from each group. The purpose of having separate training and test sets is to evaluate the performance of the machine learning model on data it has not seen before. To split the data into these groups, the holdout strategy was used, as shown in Fig. 16. The holdout strategy involves splitting the data into training and test sets randomly. This ensures that the distribution of the data is maintained in both sets and prevents overfitting, which occurs when a machine learning model performs well on the training set but poorly on the test set. After the data sets were created, they were forwarded to Edge Impulse for additional processing and model training. Edge Impulse includes a wide range of machine learning models and data processing and analysis capabilities, making it a perfect platform for designing and executing machine learning models on embedded devices.

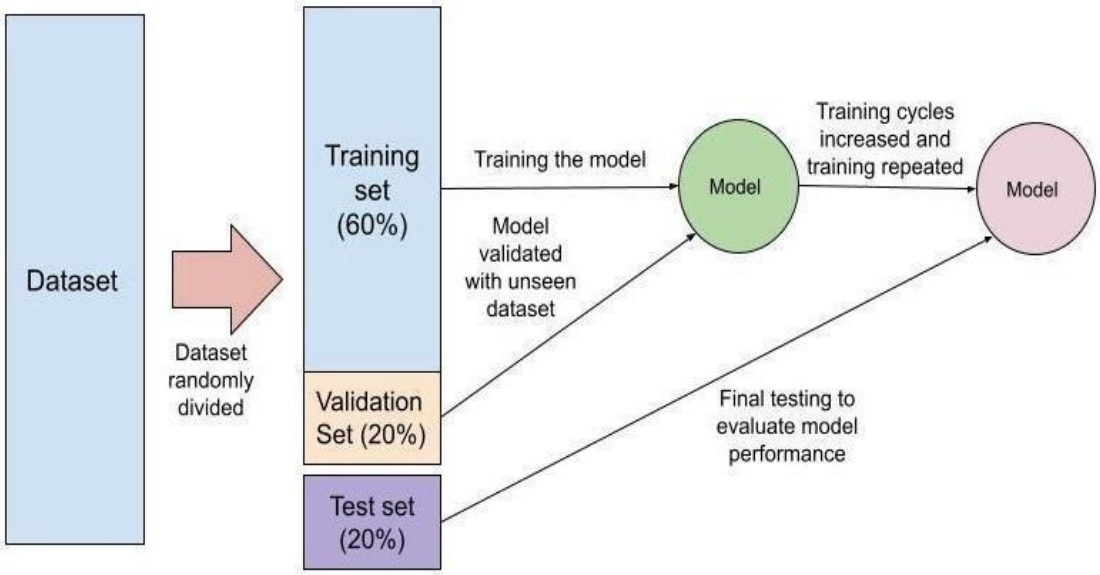


Figure 16. Hold Out Method

A window size of 1000ms was chosen to train this model, which implies that the data was divided into 1000ms pieces. The sample rate was set to 2ms, which means that data was collected every 2ms within each window. The increment for

window size was likewise set to 1000ms, which meant that a new window was produced every 1000ms.

To handle the raw data, a block with input axes from all four parameters was established. This signifies that the model was fed data from all four parameters. The parameters were most likely chosen based on their applicability to the disease prediction task.

A Neural Network (NN) was employed as the machine learning model, utilizing raw data as input and two output characteristics, Outbreak, and no outbreak. A neural network is a machine learning model inspired by the structure and function of the human brain. It is made up of linked layers of nodes that may learn patterns and correlations between input and output data.

A sample plot of the raw data is depicted in Figure 17, which likely shows a graphical representation of the data used for training the model. The plot can be used to visualize the data patterns and help with understanding the input data and model output.

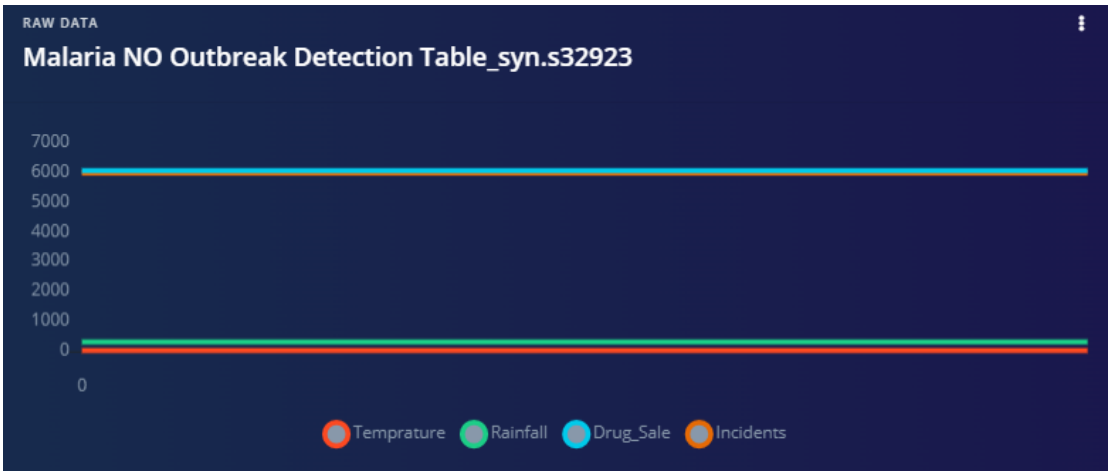


Figure 17. Raw Data Sample

When developing machine learning models for embedded systems, it is critical to keep computing needs for feature extraction to a minimum. This is because

embedded systems frequently have limited resources such as processing power, memory, and energy. One method for reducing computing costs is to extract features before model training. Identifying and picking the most significant information from raw data and transforming it into a more comprehensible form is what feature extraction entails. This reduces the amount of data that needs to be processed, thereby saving computational resources.

To reduce computational requirements, the features were extracted before training the models in this project. Figure 18 depicts an example of the processed features, which is most likely a graphical representation of the extracted features. These features might comprise numerical or categorical data that has been preprocessed to offer useful information to the machine learning algorithm. Instead of needing to analyze the raw data, the machine learning models may focus on understanding the patterns and correlations between the features by extracting the features beforehand. In embedded systems with limited processing resources, this can lead to faster and more accurate predictions. Feature extraction is an essential approach for decreasing processing needs in embedded machine learning models. It enables quicker model training and efficient data processing, which is critical for applications that demand real-time predictions.

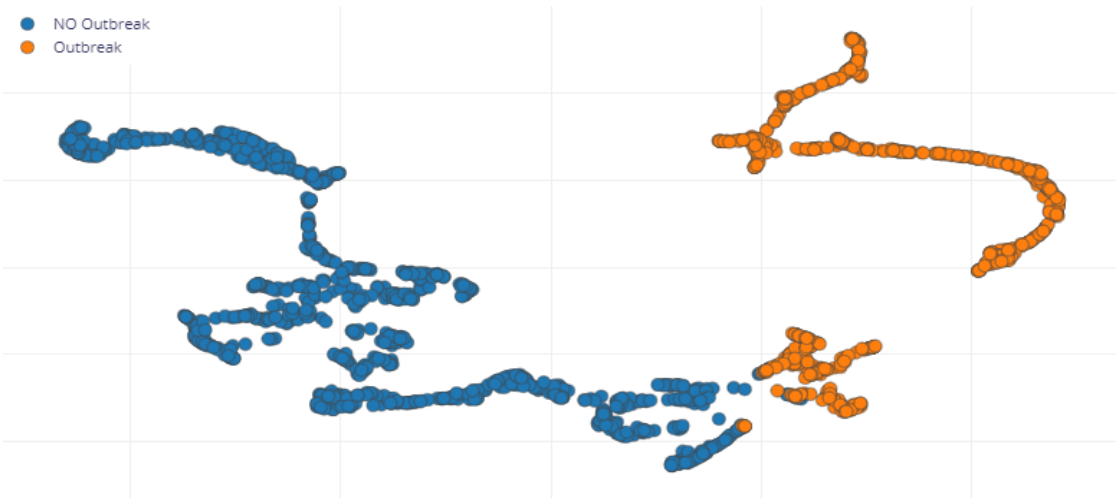


Figure 18. Generated Data Features

The process of training a neural network involves feeding it with a dataset, called the training set, and adjusting its parameters to minimize the difference between the predicted and actual outputs. In the case of this project, the neural network models were trained using 100 epochs, which means the dataset was iterated over 100 times. The learning rate was set at 0.0005, which determines the size of the steps taken by the algorithm to optimize the model parameters.

Additionally, a minimum confidence threshold of 0.80 was set, which means that only predictions with a confidence level greater than 0.80 were accepted. This threshold value helps to ensure that the model's predictions are accurate enough for the intended application.

This project's neural network design consists of four layers: an input layer, two dense layers, and an output layer. The variables utilized to predict the outcome were represented by four features in the input layer. The two dense layers each had ten and twenty neurons, referring to the number of nodes in the layer. The more nodes in a layer, the more complicated the model may be, but it also increases the risk of overfitting. Finally, two features in the output layer represented the projected output.

CHAPTER 4

RESULTS AND BENCHMARKING

System Results

The developed system is a complete platform that integrates numerous sources of data to anticipate malaria outbreaks in a specific location. Several components of the system work together to gather, process, and analyze data to deliver actionable insights to healthcare practitioners and policymakers. The system is a complete platform that incorporates many data sources to forecast malaria outbreaks in a specific location. The system is intended to be user-friendly, adaptable, and secure. A dashboard, system setup page, facilities administration, drug sale management, climatic data, and user authentication are all included. By combining these components, the system gives actionable data to healthcare practitioners and policymakers to aid in the prevention and management of malaria epidemics.

A dashboard in the system summarizes reported malaria cases, medicine sales, and meteorological data in a specific location. The dashboard is user-friendly and can be configured by period to allow for easy trend analysis. Figure 19 give a screen shot of the system dashboard.

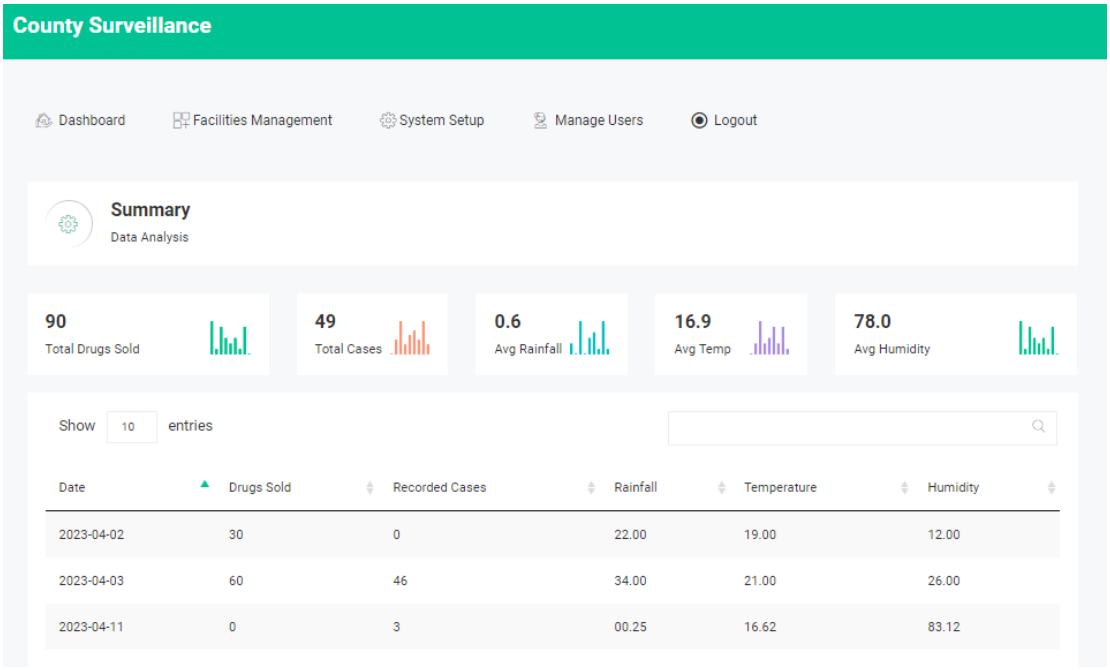


Figure 19. System Dashboard

A system setup page is available as shown in figure 20 to configure parameters such as hospitals, users, pharmacy store categories, and amenities. This enables the system to be easily customized to meet the demands of diverse healthcare professionals in various places. The system's facilities management component enables the input of malaria sales and reported cases from hospitals and pharmacy stores. This information is analyzed and evaluated using machine learning techniques to forecast malaria outbreaks in the region.

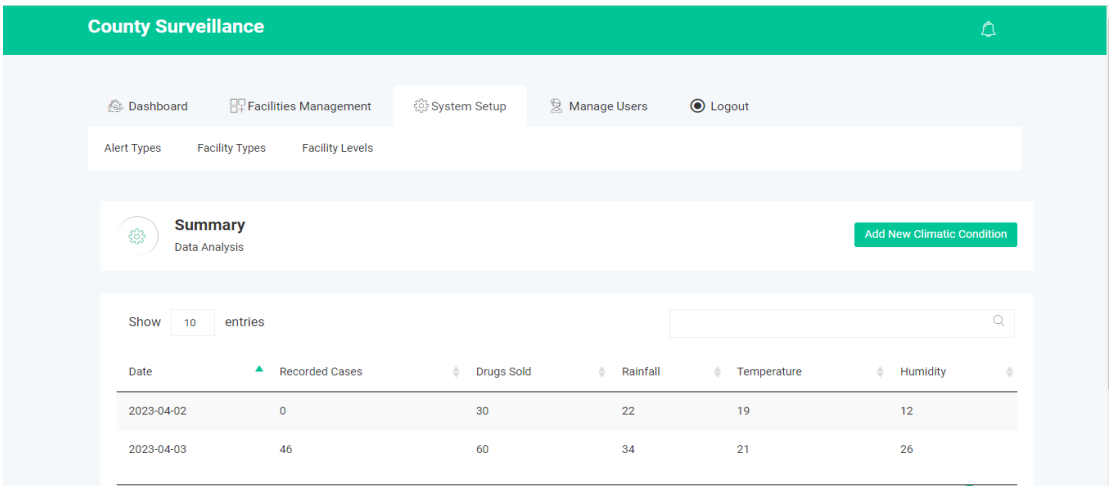


Figure 20. Setup Page

The system also includes a drug sale management component shown by figure 21, which tracks malaria treatment sales in registered sites. This data is critical for anticipating malaria outbreaks since it shows areas where there are no hospital visits but over the counter sales of malaria drugs are rapidly increasing, signalling a possible epidemic.

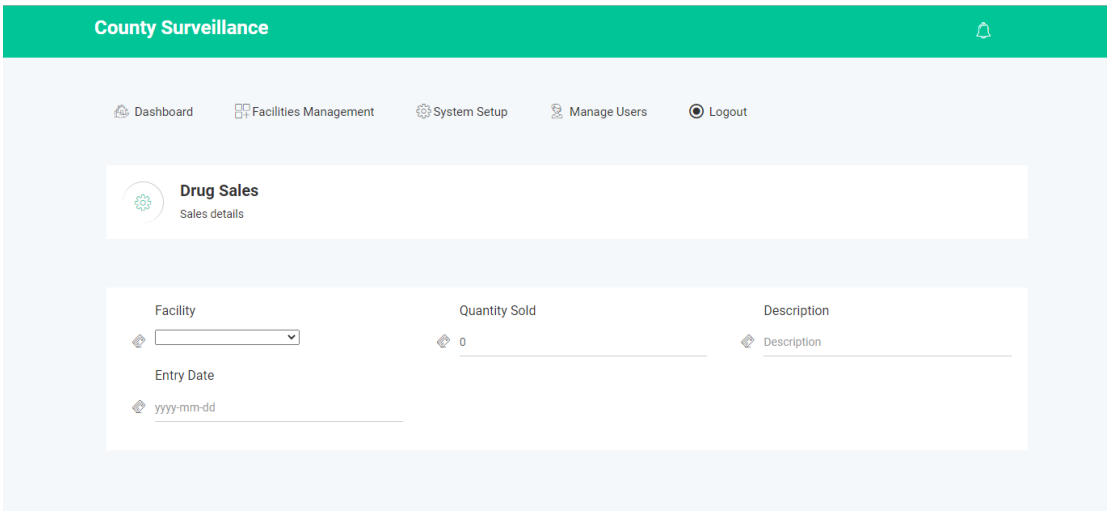


Figure 21. Over the Counter Sale of Drugs Page

The system's climate data component gives information on weather trends in the region. Climate change greatly influences the El Niño cycle that is known to be associated with increased risks of some diseases transmitted by mosquitoes, such as malaria, dengue, and Rift Valley fever By Capturing this information we are ablet to use it to forecast malaria outbreaks in conjunction with other data sources. See figure 22.

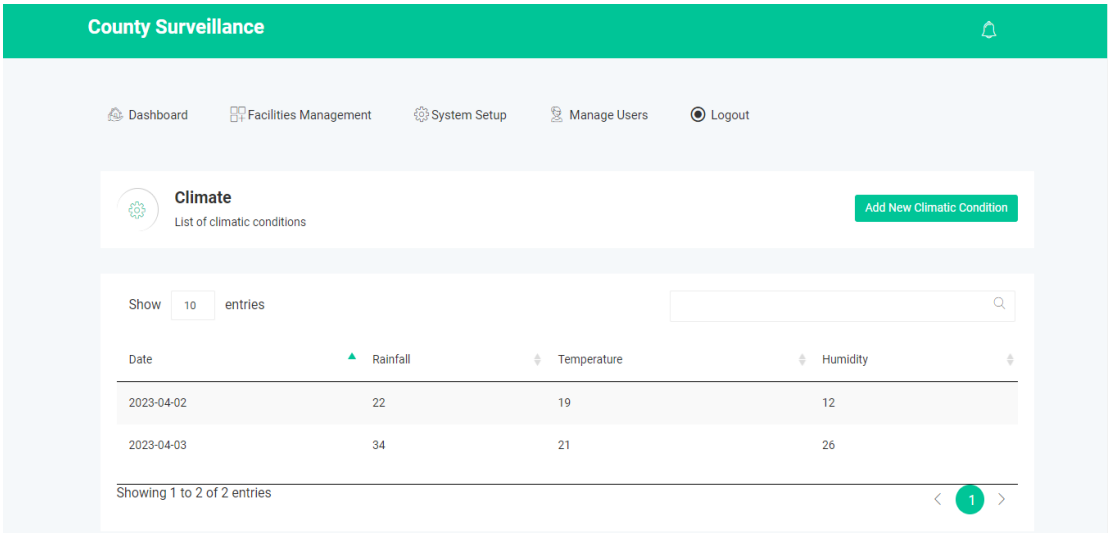


Figure 22. Climate Data Page

Figure 23 shows that users and facilities must be registered, and authentication is used to verify that only authorized users have access to the system. Depending on their function in the healthcare system, various users have varied access permissions. This protects sensitive data and restricts access to only authorized individuals.

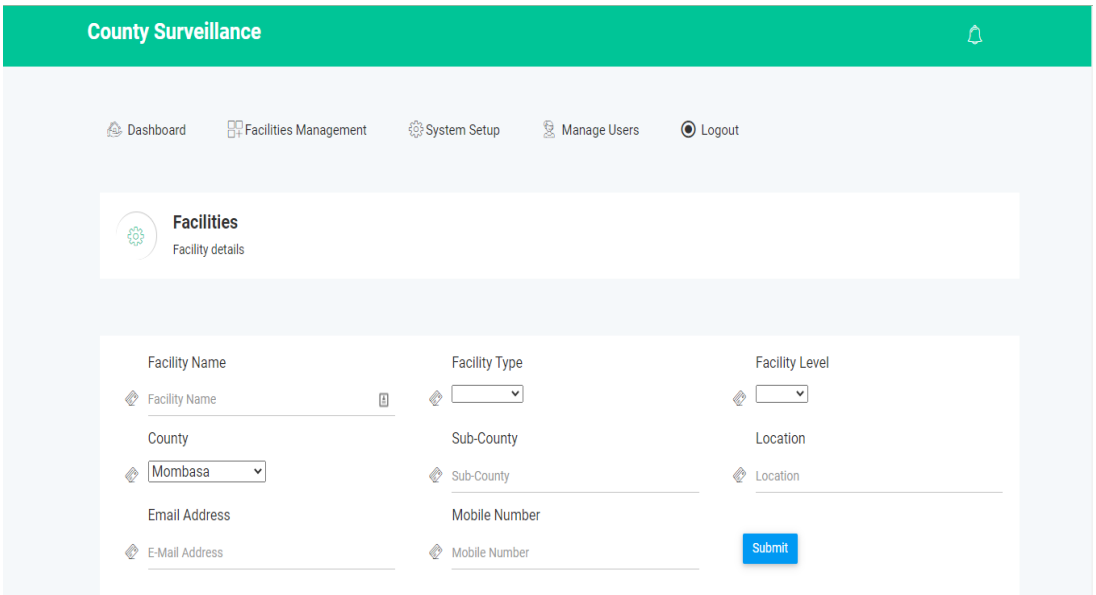


Figure 23. Facility Management Page

Model Training Results

A machine learning model's validation accuracy is a measure of how well it works on data that it has never seen before. It is critical to assess the model's performance on a validation set since it determines if the model is overfitting or underfitting the data. In this situation, the model's validation accuracy was 90.0 percent. This signifies that the model categorized 90.0 percent of the examples in the validation set correctly. The model's loss on the validation set was 0.25 percent, which indicates how well the model minimizes mistakes.

Visualizing the model's performance on the validation set is done through a confusion matrix. This matrix shows the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for each class in the validation set. A true positive is an instance correctly classified as belonging to the positive class, while a false positive is an instance incorrectly classified as belonging to the positive class. A true negative is an instance correctly classified as belonging to the negative class, while a false negative is an instance incorrectly classified as belonging to the negative class.

Upon examining the confusion matrix, it was determined that the model accurately classified 96.7% of instances as belonging to the positive class (Outbreak) and 85.9% of instances as belonging to the negative class (No Outbreak). However, it incorrectly classified 14% of instances as belonging to the positive class when they belonged to the negative class and 3.3% of instances as belonging to the negative class when they belonged to the positive class. These metrics provide insight into the performance of the model and can help in further optimization and improvement.

Table 7. Model Confusion Matrix Table

Machine learning model Training Performance on validation set **Accuracy 90.0% Loss 0.25**		
	No Outbreak	Outbreak
No Outbreak	85.9%	14.1%
Outbreak	3.3%	96.7%

Table 8. Model F1 Scores

Machine learning model Training Performance on validation set		
	No Outbreak	Outbreak
F1 Score	0.91	0.88

CHAPTER 5

CONCLUSION

Summary of Work

Malaria is a life-threatening illness that mostly affects impoverished countries in Sub-Saharan Africa. It spreads by the bite of infected female Anopheles mosquitos and kills a significant number of people in this region. Malaria's high prevalence is exacerbated by factors such as a lack of healthcare infrastructure, poverty, and insufficient public health activities.

To address the high incidence of malaria, the suggested solution entailed the construction of an online malaria data integration platform that collects multivariate malaria data from various sources including hospitals where the cases are diagnosed and drug store. The system also collects meteorological information which is an important factor in malaria spread. The platform is then able to integrate the data for further analysis and decision making. The platform was developed following the prototype system development process, which stresses quick development and ongoing feedback to guarantee that the final product satisfies the needs of the users.

As a proof of concept to show how the integrated data could be useful, a machine learning model was developed using historical data gathered in a Kenyan county over five years. The best machine learning model attained an accuracy of 90% in predicting weather there is an outbreak of malaria or not based on the aggregated information collected daily. Before making the prediction, the algorithm collects data from hospitals and drug sales, which is then combined with meteorological data. The

system gathers information on the number of malaria cases reported in the county, and hospitals give daily totals. This information is then merged with meteorological data such as temperature, precipitation, and humidity. The data is then fed into the machine learning algorithm, which predicts the risk of a malaria breakout in the county.

The results are displayed in a visualization dashboard with an easy-to-use interface for viewing predicted epidemic statistics. Using the dashboard, users may check information by hospitals, counties, and pharmacy stores. The dashboard also provides information on the meteorological conditions most likely to create a malaria pandemic, allowing users to take precautionary measures to avert an outbreak.

Finally, the proposed method provides a practical option to tackle the high prevalence of malaria in Sub-Saharan Africa. The aggregation of data and the prediction of malaria outbreaks based on historical data and meteorological information have the potential to save lives by allowing public health experts to take preventative measures before epidemics begin. The platform's user-friendly design makes it straightforward for users to access and grasp expected epidemic data, making it a critical tool for Sub-Saharan African public health experts.

Discussion

Malaria is a major public health problem in Sub-Saharan Africa, and anticipating outbreaks is essential for effective prevention and control. The proposed solution addresses this difficulty by building an online malaria data integration platform and a machine learning model trained on historical reported cases and meteorological data. The use of the prototype system development approach enables the speedy and efficient creation of a functional system. This method is great for projects with ambiguous requirements since it allows for rapid iteration and feedback.

Furthermore, it allows for the early diagnosis of potential issues, thereby saving time and money in the long run.

The use of machine learning to predict malaria outbreaks is a realistic option since it can manage massive amounts of data and uncover tendencies that people may not see right away. The model's training data was collected in a Kenyan county during five years. This data is likely to be reflective of local conditions, but keep in mind that the model's accuracy in other locations may be modified by a variety of variables. As a result, further testing and validation in other places are necessary to ensure the model's accuracy. The data was created using permitted online data synthesis techniques to solve the issue of poor data availability. The top model achieved an accuracy of 90%, which is a promising result. However, the limitations of synthetic data must be recognized, since they may not accurately reflect the underlying data's distribution. More study and validation are needed to verify the model's accuracy.

To anticipate malaria outbreaks, the system collects data from hospitals and pharmacies, which is then integrated with climatic information. The merging of many data sources provides a more full view of the factors that contribute to malaria spread. However, it is critical to consider data constraints such as underreporting or missing data, which may affect the model's accuracy. To maintain the model accurately, it is necessary to monitor and improve data quality regularly. The findings are shown on a visualization dashboard, which effectively communicates the forecasts to public health experts and other stakeholders. The dashboard aids in data interpretation and decision-making processes. It is vital, however, to ensure that the dashboard is simple to use and accessible to a wide range of users, including those with little technical experience.

In terms of future advances, the proposed technique might be expanded to more sites to evaluate the model's accuracy under different conditions. Furthermore, the system might be improved by including additional data sources such as satellite photographs or social media data. In addition, the system might be modified to encompass additional diseases such as dengue fever and the Zika virus.

Finally, the proposed approach might be used to anticipate malaria outbreaks in Sub-Saharan Africa. However, to ensure the system's accuracy and durability, it is necessary to assess its restrictions as well as potential future upgrades. The combination of machine learning and several data sources provides a more full picture of the factors that contribute to malaria transmission, and the visualization dashboard aids in data comprehension.

Limitations

While the proposed solution is a promising approach to reducing malaria's high incidence in Sub-Saharan Africa, numerous project obstacles must be addressed. The following are some of the constraints:

- Data Availability: The initiative is dependent on the availability of accurate and timely data from county hospitals and pharmacy retailers. This data may not be easily available in some places, limiting the model's efficacy.
- Generalizability: The model was trained using data from a single Kenyan county, which may or may not be typical of the whole area. As a result, the model's capacity to forecast malaria outbreaks in other areas may be restricted.

- Weather Data: The model's accuracy is dependent on the availability of accurate and trustworthy weather data. However, meteorological data may not be easily available in some places, affecting the model's accuracy.
- Access to Technology: The proposed solution necessitates the use of technology such as cell phones, internet access, and laptops, which may not be available in all places. This may limit the platform's accessibility to public health professionals in some places.
- The use of synthetic data-generating methods raises ethical questions about patient data privacy and confidentiality. It is critical to guarantee that all data gathered and processed respect patient privacy and confidentiality.
- Maintenance and sustainability: To stay successful in predicting malaria outbreaks, the system requires frequent maintenance and upgrades. Furthermore, the platform's viability needs partners to commit to providing the required resources for maintenance and upgrades.

To summarize, while the proposed solution represents a promising approach to combating the high prevalence of malaria in Sub-Saharan Africa, these limitations must be addressed for the platform to be effective in predicting malaria outbreaks and accessible to regional public health officials.

Perspectives and Future Work

Future work might be done in several areas to improve the suggested solution:

- Expansion to other regions: The model should be tested in other Sub-Saharan African regions to determine its efficacy and accuracy in forecasting malaria outbreaks under different conditions. Data from multiple sources, such as hospitals, pharmacies, and weather stations, would need to be collected and merged to train the model.

- More data sources, such as satellite data, mosquito population data, and socioeconomic data, might be integrated: The model may be improved by incorporating additional data sources, such as satellite data, mosquito population data, and socioeconomic data. This might help find new factors that contribute to the spread of malaria and improve the model's accuracy.
- Real-time data collection and analysis: The platform might be enhanced by collecting and analyzing real-time data to provide more accurate and rapid forecasts of malaria outbreaks. This would involve the development of a data collection and analysis system capable of processing and analyzing massive amounts of data in real-time.
- Integration with other healthcare systems: The platform may be combined with other healthcare systems to provide a more complete view of the population's health. This might entail connecting the platform to electronic medical records systems to produce a more accurate picture of the incidence of malaria and other illnesses in the region.
- Deployment on mobile devices: The platform might be applied to mobile devices to enable real-time access to data and forecasts for public health officials. This would necessitate the creation of a mobile application that could run on low-cost devices and function in locations with restricted internet access.
- Collaboration with local communities: The platform's effectiveness is dependent on local communities' active engagement in data collecting and reporting. To ensure the platform's success and sustainability, future work should focus on creating trust and collaborating with local communities.

REFERENCES

[1] CDC, “Impact of Malaria Worldwide,” *Centers Dis. Control Prev.*, 2020, Accessed: Feb. 18, 2022. [Online]. Available: https://www.cdc.gov/malaria/malaria_worldwide/impact.html.

[2] WHO, “Malaria,” 2021. https://www.who.int/news-room/fact-sheets/detail/malaria (accessed Feb. 18, 2022).

[3] C.-C. for D. C. and Prevention, “CDC - Malaria - Malaria Worldwide - CDC’s Global Malaria Activities - President’s Malaria Initiative (PMI),” 2021.

[4] CitizenTV, “Nandi County records high number of malaria cases - YouTube,” 2021. https://www.youtube.com/watch?v=8O9YFgkppjY (accessed Feb. 18, 2022).

[5] Centers for Disease Control and Prevention (CDC), “CDC - Malaria - About Malaria - Biology,” *Centers for Disease Control and Prevention*. pp. 1–2, 2019, Accessed: Feb. 20, 2022. [Online]. Available: https://www.cdc.gov/malaria/about/biology/index.html.

[6] C. for D. C. and Prevention, “CDC - Malaria - Malaria Worldwide - CDC’s Global Malaria Activities - Kenya,” 2019. https://www.cdc.gov/malaria/malaria_worldwide/cdc_activities/kenya.html (accessed Feb. 20, 2022).

[7] C. Nsanzabana, “Strengthening Surveillance Systems for Malaria Elimination by Integrating Molecular and Genomic Data,” *Trop. Med. Infect. Dis. 2019, Vol. 4, Page 139*, vol. 4, no. 4, p. 139, Dec. 2019, doi: 10.3390/TROPICALMED4040139.

[8] C. Lourenço *et al.*, “Strengthening surveillance systems for malaria elimination: A global landscaping of system performance, 2015-2017,” *Malar. J.*, vol. 18, no. 1, pp. 1–11, Sep. 2019, doi: 10.1186/S12936-019-2960-2/FIGURES/2.

[9] J. H. Brenas, M. S. Al-Manir, C. J. O. Baker, and A. Shaban-Nejad, “A Malaria Analytics Framework to Support Evolution and Interoperability of Global Health Surveillance Systems,” *IEEE Access*, vol. 5, pp. 21605–21619, Oct. 2017, doi: 10.1109/ACCESS.2017.2761232.

[10] L. S. Anam, M. M. Badi, M. A. Assada, and A. A. Al Serouri, “Evaluation of Two Malaria Surveillance Systems in Yemen Using Updated CDC Guidelines: Lessons Learned and Future Perspectives,” *Inq. (United States)*, vol. 56, pp. 1–8, Oct. 2019, doi: 10.1177/0046958019880736.

[11] O. Nkiruka, R. Prasad, and O. Clement, “Prediction of malaria incidence using climate variability and machine learning,” *Informatics Med. Unlocked*, vol. 22, p. 100508, Jan. 2021, doi: 10.1016/J.IMU.2020.100508.

[12] Y. W. Lee, J. W. Choi, and E. H. Shin, “Machine learning model for predicting malaria using clinical information,” *Comput. Biol. Med.*, vol. 129, p. 104151, Feb. 2021, doi: 10.1016/J.COMPBIOMED.2020.104151.

[13] P. Mohapatra, N. K. Tripathi, I. Pal, and S. Shrestha, “Determining suitable machine learning classifier technique for prediction of malaria incidents attributed to climate of Odisha,” *https://doi.org/10.1080/09603123.2021.1905782*, 2021, doi: 10.1080/09603123.2021.1905782.

[14] G. Kalipe, V. Gautham, and R. K. Behera, “Predicting Malarial Outbreak using Machine Learning and Deep Learning Approach: A Review and Analysis,” *Proc. - 2018 Int. Conf. Inf. Technol. ICIT 2018*, pp. 33–38, Dec. 2018, doi: 10.1109/ICIT.2018.00019.

[15] B. E. Chekol and H. Hagras, “Employing Machine Learning Techniques for the Malaria Epidemic Prediction in Ethiopia,” *2018 10th Comput. Sci. Electron. Eng. Conf. CEEC 2018 - Proc.*, pp. 89–94, Mar. 2019, doi: 10.1109/CEEC.2018.8674210.

[16] W. Health Organization, “GLOBAL TECHNICAL STRATEGY FOR MALARIA 2016-2030 Global Technical Strategy for Malaria 2016-2030 Global Malaria Programme World Health Organization.”

[17] R. W. Pinner, C. A. Rebmann, A. Schuchat, and J. M. Hughes, “Disease surveillance and the academic, clinical, and public health communities,” *Emerg. Infect. Dis.*, vol. 9, no. 7, pp. 781–787, Jul. 2003, doi: 10.3201/EID0907.030083.

[18] B. A. Kay, R. J. Timperi, S. S. Morse, D. Forslund, J. J. McGowan, and T. O’Brien, “Innovative information-sharing strategies.,” *Emerg. Infect. Dis.*, vol. 4, no. 3, p. 465, 1998, doi: 10.3201/eid0403.980334.

[19] M. Edelstein, L. M. Lee, A. Herten-Crabb, D. L. Heymann, and D. R. Harper, “Strengthening global public health surveillance through data and benefit sharing,” *Emerg. Infect. Dis.*, vol. 24, no. 7, pp. 1324–1330, Jul. 2018, doi: 10.3201/EID2407.151830.

[20] World Health Organization, *World Malaria Report: 20 years of global progress and challenges*, vol. WHO/HTM/GM, no. December. 2020.

[21] N. Maurice *et al.*, “Malaria Epidemic Prediction Model by Using Twitter Data and Precipitation Volume in Nigeria,” *Multimed. Soc.*, vol. 22, no. 5, pp. 588–600, 2019.

[22] K. E. Mace, N. W. Lucchi, and K. R. Tan, “Malaria Surveillance — United States, 2017,” *MMWR Surveill. Summ.*, vol. 70, no. 2, pp. 1–35, 2021, doi:

10.15585/MMWR.SS7002A1.

[23] E. Mbunge, R. C. Millham, M. N. Sibiya, and S. Takavarasha, “Diverging Mobile Technology’s Cognitive Techniques into Tackling Malaria in Sub-Saharan Africa: A Review,” *Lect. Notes Networks Syst.*, vol. 232 LNNS, pp. 679–699, Oct. 2021, doi: 10.1007/978-3-030-90318-3_54.

[24] A. Ismail, A. Shehab, and I. M. El-Henawy, *Healthcare Analysis in Smart Big Data Analytics: Reviews, Challenges and Recommendations*. Springer International Publishing, 2019.

[25] S. SA, “Big Data in Healthcare Management: A Review of Literature,” *Am. J. Theor. Appl. Bus.*, vol. 4, no. 2, p. 57, 2018, doi: 10.11648/j.ajtab.20180402.14.

[26] A. Meri *et al.*, “Modelling the utilization of cloud health information systems in the Iraqi public healthcare sector,” *Telemat. Informatics*, vol. 36, no. December, pp. 132–146, 2019, doi: 10.1016/j.tele.2018.12.001.

[27] G. L. Tortorella, T. A. Saurin, F. S. Fogliatto, V. M. Rosa, L. M. Tonetto, and F. Magrabi, “Impacts of Healthcare 4.0 digital technologies on the resilience of hospitals,” *Technol. Forecast. Soc. Change*, vol. 166, no. December 2020, p. 120666, 2021, doi: 10.1016/j.techfore.2021.120666.

[28] N. J. Ravindran and P. Gopalakrishnan, “Predictive Analysis for Healthcare Sector Using Big data Technology,” *Proc. 2nd Int. Conf. Green Comput. Internet Things, ICGCIoT 2018*, pp. 326–331, Aug. 2018, doi: 10.1109/ICGCIOT.2018.8753090.

[29] R. Agarwal, “Predictive Analysis in Health Care System Using AI,” pp. 117–131, 2022, doi: 10.1007/978-981-16-6265-2_8.

[30] F. Gonçalves, R. Pereira, J. Ferreira, J. B. Vasconcelos, F. Melo, and I. Velez, “Predictive Analysis in Healthcare: Emergency Wait Time Prediction,” *Adv. Intell. Syst. Comput.*, vol. 806, pp. 138–145, Jun. 2018, doi: 10.1007/978-3-030-01746-0_16.

[31] V. K. Daliya, T. K. Ramesh, and A. Shashikanth, “A Machine Learning based Ensemble Approach for Predictive Analysis of Healthcare Data,” *2020 2nd PhD Colloq. Ethically Driven Innov. Technol. Soc. PhD Ed. 2020*, Nov. 2020, doi: 10.1109/PHDEDITS51180.2020.9315300.

[32] A. Acharya, J. Patel, and J. Patel, “Predictive Analysis in Healthcare Using Feature Selection,” *Biomed. Data Min. Inf. Retr.*, pp. 53–101, Aug. 2021, doi: 10.1002/9781119711278.CH3.

[33] A. Menon, M. S. Aishwarya, A. Maria Joykutty, A. Y. Av, and A. Y. Av, “Data Visualization and Predictive Analysis for Smart Healthcare: Tool for a Hospital,” *TENSYMP 2021 - 2021 IEEE Reg. 10 Symp.*, Aug. 2021, doi: 10.1109/TENSYMP52854.2021.9550822.

[34] P. Phoobane, M. Masinde, and T. Mabhaudhi, “Predicting Infectious Diseases: A Bibliometric Review on Africa,” *Int. J. Environ. Res. Public Heal. 2022*,

Vol. 19, Page 1893, vol. 19, no. 3, p. 1893, Feb. 2022, doi: 10.3390/IJERPH19031893.

[35] P. M. Brock *et al.*, "Predictive analysis across spatial scales links zoonotic malaria to deforestation," *Proc. R. Soc. B*, vol. 286, no. 1894, Jan. 2019, doi: 10.1098/RSPB.2018.2351.

[36] D. Chaya Jagtap and N. Usha Rani, "Cuckoo search based ensemble classifier for predictive analysis of malaria infection scope on thin blood smears," *Indian J. Public Heal. Res. Dev.*, vol. 10, no. 5, pp. 1019–1031, 2019, doi: 10.5958/0976-5506.2019.01209.9.

[37] S. Thakur and R. Dharavath, "Artificial neural network based prediction of malaria abundances using big data: A knowledge capturing approach," *Clin. Epidemiol. Glob. Heal.*, vol. 7, no. 1, pp. 121–126, Mar. 2019, doi: 10.1016/J.CEGH.2018.03.001.

[38] J. B. Awotunde, R. G. Jimoh, I. D. Oladipo, and M. Abdulraheem, "Prediction of Malaria Fever Using Long-Short-Term Memory and Big Data," *Commun. Comput. Inf. Sci.*, vol. 1350, pp. 41–53, Nov. 2020, doi: 10.1007/978-3-030-69143-1_4.

[39] J. S. Kimuyu, "Comparative spatial–temporal analysis and predictive modeling of climate change-induced malaria vectors' invasion in new hotspots in Kenya," *SN Appl. Sci.*, vol. 3, no. 8, pp. 1–15, Aug. 2021, doi: 10.1007/S42452-021-04722-1/FIGURES/8.

[40] C. J. Standley, E. Graeden, J. Kerr, E. M. Sorrell, and R. Katz, "Decision support for evidence-based integration of disease control: A proof of concept for malaria and schistosomiasis," *PLoS Negl. Trop. Dis.*, vol. 12, no. 4, p. e0006328, Apr. 2018, doi: 10.1371/JOURNAL.PNTD.0006328.

[41] M. Verma, K. Kishore, M. Kumar, A. R. Sondh, G. Aggarwal, and S. Kathirvel, "Google search trends predicting disease outbreaks: An analysis from India," *Healthc. Inform. Res.*, vol. 24, no. 4, pp. 300–308, 2018, doi: 10.4258/hir.2018.24.4.300.

CURRICULUM VITAE

MICAH A. OCHOLA
+254721969726
cholam@gmail.com
P.O2500–30100ELDORET-KENYA

PROFILE

Skilled Training Specialist and professional presenter with over 10 years career experience in IT Training environments. Highly adept in team building and leadership, coaching and mentoring, and change management. Possesses excellent organizational and time management abilities.

Vastly involved in the areas of Distance Education, Open Learning and Communications skills, design of communications strategy, course development and production, course delivery, tutorial support, students support systems, Research and evaluation Management of distance education systems

EDUCATION

- Msc. Applied Computer Science E-Service
  Adventist University of Africa, 2017 – Date

- Bsc. Software Engineering
  University of E.A Baraton, 2001 – 2006

CERTIFICATION

- Certified Software Developers Proffessional (CSDP)
- Certified Cisco Networks Associate (CCNA) CHIPUKA
- International Computer Driving License (ICDL)

SKILLS

- Project Management Communication Skills Technical Writing Abilities. Adult Learning Documentation Skills Computer Programming

MEMBERSHIP

- IEEE

- ACM
- CSK

OTHER TRAINING CERTIFICATES

- Kenya Accountants Technicians Course (KASNEB)
- Evaluation in Humanitarian Settings (UNICEF)
- Financial Reporting(Last Mile Learning)
- Managing in the Humanitarian Sector (UNHCR)
- Results Based Management (UNHCR)

PROFFESSIONAL EXPERIENCE

SPECIALIST TRAINING
International Rescue Committee |Nairobi |2016 – Present

SOFTWARE DEVELOPMENT CONSULTANT
Advantech |Nairobi | Dec 2015 – June 2016

SOFTWARE APPLICATIONS DEVELOPMENT TRAINER
Code Pamoja & Dew CIS Solutions | Nairobi |January 2016 – June 2016

LEAD IMPLEMENTATION DEVELOPER
Emerald Flamingo Beach Resort and Spa Hotel | Mombasa |January 2014 – Dec 2015

TEACHING ASSISTANT
University of Eastern Africa, Baraton | Eldoret |2007 - 2015

SOFTWARE DEVELOPER
Fintech (K) | Nairobi | Jun 2006 – Dec 2006