ISA Journal of Medical Sciences (ISAJMS)

Homepage: https://isapublisher.com/isajms/
Email: office.isapublisher@gmail.com

Volume 2, Issue 6, Nov-Dec, 2025 ISSN: 3049-1746

Predicting and Preventing Malaria Outbreaks in Sokoto State, Northern Nigeria, using Machine Learning: Exploring the Relationship between Malaria Incidence and Climatic Data

Olayinka Oloyede

University of East London

Received: 01.12.2025 | **Accepted:** 11.12.2025 | **Published:** 15.12.2025

***Corresponding Author:** Olayinka Oloyede

DOI: *10.5281/zenodo.17934326*

Abstract **Original Research Article**

Malaria outbreaks pose a significant health threat globally, especially in Sub-Saharan Africa and Nigeria, which accounts for over 25% of the global burden of the disease. According to the Malaria Indicator Survey of 2021, Sokoto State is one of the states with a high malaria burden (between 31% and 40% prevalence) in Nigeria. This study evaluated the potential of predictive analytics using machine learning to predict and prevent malaria outbreaks by exploring the relationship between malaria incidence and climatic data specifically temperature and rainfall.

This study employed a retrospective observational design to analyze historical malaria incidence and climatic data. The monthly malaria incidence data, spanning January 2015 to December 2022, were obtained from the Nigeria District Health Information System (DHIS) while the Climatic data, including temperature and rainfall, was collected from the National Aeronautics and Space Administration (NASA) website for the same period. The collected data was cleaned to remove any inconsistencies, missing values, or outliers. Three supervised machine learning algorithms were chosen; Support Vector Machine (SVM), Random Forest Classifier, and K-Nearest Neighbors (KNN). The models were trained using the preprocessed data, with hyperparameter tuning to optimize performance.

The results of the models revealed that SVM achieved the highest accuracy of 76% in forecasting malaria outbreaks, significantly outperforming both Random Forest and KNN, which achieved 59% accuracy. The trained models were evaluated using appropriate metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.

This research demonstrates the effectiveness of machine learning, particularly SVM, in predicting malaria outbraks using climatic data in Sokoto State however, it is important to also consider other malaria-causing factors such as socioeconomic factors in subsequent studies. Overall, these models have the potential to equip people working to fight malaria with better information for preventing outbreaks and lessening their effects through early planning and actions.

Keywords: Malaria Outbreak Prediction, Machine Learning, Climatic Data Analysis, Support Vector Machine (SVM), Sokoto State Nigeria.

Introduction

Background of Study

The importance of data in the 21st century cannot be overemphasized. It has become the fabric upon which innovations, problem-solving, and strategic decision-making are hewn. It is being used to better understand the world, discover new patterns and insights, and improve decision-making and this is made possible due to the enormous volume of data being generated, collected, and stored digitally. According to Petroc Taylor (2023), the volume of data that is being created, captured, and consumed worldwide grew from 2 zettabytes in 2010 to 120 zettabytes in 2023 and it is projected to grow to about 181 zettabytes in 2025. The availability of such a volume of data has revolutionized several sectors such as the Tech sector where Google utilizes data to customize individual searches based on search history, politicians use data to customize political campaign messages to people more likely to support a particular candidate based on their web and search history in politics, financial institutions use historical data to identify potentially fraudulent transactions and public health practitioners employing data to predict possible epidemics in public health using historical data. (Murdoch and Detsky, 2013).

Machine learning (ML) was introduced by Authur Samuel in the 1950s with the intent of developing an algorithm that improves over time through the continuous learning of the model based on historical data (Nkiruka, Prasad, and Clement, 2021). Lying at the intersection between statistics and computer science, machine learning was designed for computers to improve automatically and make successful predictions using past experiences (Jordan and Mitchell, 2020). Improvement in several sectors of the economy of nations can be attributed to the advent of machine learning. In a study by Liakos et al. (2018), in agriculture, machine learning was applied to yield prediction, disease detection, and livestock management among others. Additionally, in public health, machine learning has been used in the estimation of disease incidence, and epidemic prediction (Haneef et al., 2021). Furthermore, ML has been used in customer relationship management and customer retention in finance, and several opportunities still exist and continue to open in the application of ML.

Sokoto, the capital of Sokoto State in north-western Nigeria, sits in a dry, sandy region with some hills. The rainy season is short, lasting from June to September (or sometimes October). It averages 550mm of rain annually, peaking in August. The hottest months are March and April, reaching up to 45°C. A cold, dry, and dusty wind called the Harmattan blows between November and February (Abdullahi et al., 2009).

Types of Machine Learning

ML can be divided into three categories: supervised learning, unsupervised learning, and reinforcement learning (Sarker, 2021).

- Supervised learning in which the machine is tasked with learning a function that maps an input to an output based on a sample input-output pair (Sarker, 2021). The machine is programmed to learn the mapping between the input variable and the output variable and apply the model to predict outputs for future data. It is used when specific targets are expected from a certain input dataset, and it is especially used in classification models, logistic regression, and K-nearest neighbor (Sarker et al., 2020). Supervised learning is one of the most common ML models and its application includes the categorization of loan seekers using their past credit history, patient categorization for a particular disease based on risk factors, and Spam filtering in emails among others (Dridi, 2022).
- Unsupervised learning unlike the supervised model uses unlabeled data to identify patterns that may not have been easily noticeable in a dataset and this is done without human interference (He et al., 2015). The unsupervised learning model focuses on unraveling interesting patterns and trends from the data instead of predicting a definite outcome as in the case of the supervised learning model. Because of a lack of definite possible outcomes, unsupervised learning is focused more on exploratory data analysis and is used in clustering data and association

pattern mining (James et al., 2023). Its application includes partitioning of Alzheimer's disease patients based on certain similarities (Alashwal et al., 2019), autism spectrum disorder identification and characterization (Parlett-Pelleriti et al., 2022).

- Reinforcement learning as the name implies is a machine learning model that allows machines to evaluate the best option in a particular environment to improve the efficiency of the model. This is usually done via trial and error using feedback from actions and experiences (Kaelbling, Littman, and Moore, 1996). Reinforcement learning is a very robust tool that can be used to improve the efficiency of automated tasks like autonomous driving tasks, automated gameplay, programming of robots, and supply chain logistics (Nkiruka, Prasad, and Clement, 2021). Reinforcement learning can be used to detect malicious cyber activities and botnet traffic and some of the most common ones include Monte Carlo learning, Q-learning, and Deep Q Networks (Kaelbling, Littman, and Moore, 1996).

The application of machine learning to predict disease incidence using supervised learning has advanced over time and will be employed in this research to predict an impending malaria epidemic in Sokoto State, northern Nigeria, thereby helping to make advanced plans to avert high morbidity and mortality.

Problem Statement

Malaria is a serious and widespread infectious disease that affects millions of people in Nigeria, especially in the northern region. Parasites transmitted by female Anopheles mosquitoes cause malaria and can cause fever, headache, chills, vomiting, anaemia, and organ failure (CDC, 2022). The most common and deadly type of malaria parasite in Nigeria is Plasmodium falciparum. If not treated promptly, malaria can have a devastating outcome, especially in children. According to the World Health Organization (WHO, 2023a), an untreated malaria infection can lead to death within 24 hours in children.

The incidence and prevalence of malaria are two important indicators of the burden and transmission of the disease. Incidence refers to the number of new malaria cases in a given period, usually per year. Prevalence refers to the proportion of the population that is infected with malaria at a given point in time, usually measured by blood tests. Both incidence and prevalence can vary depending on the season, the region, the age group, and the intervention coverage.

According to the WHO report on malaria (WHO, 2023b), Nigeria had the most malaria cases and deaths in the world in 2021, accounting for more than half of the cases in West Africa. Additionally, the report revealed that the malaria incidence and mortality rates in Nigeria slightly changed from 2020 to 2021, with a small decrease in incidence and a small increase in mortality. Furthermore, the report indicates that malaria is not evenly distributed in Nigeria, but varies by region, season, age, and socioeconomic status. Sokoto state in the northern part of the country has the highest malaria risk, due to the favourable climate for mosquitoes.

According to the 2018 Nigeria Demographic and Health Survey (NDHS) report released by the National Population Commission (NPC), (NPC, 2018), the prevalence of malaria parasitemia in children under five years of age was 23%, a decrease from 27% in 2015 and 42% in 2010. However, there were significant regional, rural-urban, and socioeconomic differences in the prevalence of malaria. The prevalence of malaria ranged from 16% in the South and Southeast Zones to 34% in the Northwest Zone. The prevalence of malaria was 2.4 times higher in rural populations than in urban populations (31% vs. 13%). The prevalence of malaria was seven times higher among children in the lowest socioeconomic group than among children in the highest socioeconomic group (38% vs. 6%).

Malaria Prevention and Treatment

Though, the Nigerian government has pledged to alleviate the impact of malaria and achieve its elimination by 2030, and the National Malaria

Elimination Programme (NMEP) has formulated and executed various policies and strategies to prevent, diagnose, treat, and monitor malaria, with a focus on the distribution of Insecticide-treated nets (ITNs). These bed nets are treated with insecticides to kill or repel mosquitoes, proving to be a highly effective and economical method for malaria prevention. The NMEP has distributed millions of ITNs through mass campaigns and routine channels like antenatal care and immunization services. Per the 2018 National Demographic and Health Survey (NDHS), 69% of households possessed at least one ITN, with 50% of children under five years and 49% of pregnant women sleeping under an ITN the night before the survey.

In addition to ITNs, indoor residual spraying (IRS) is another strategy employed to curb malaria transmission. IRS involves applying insecticides on the walls and ceilings of houses where mosquitoes rest, thereby reducing mosquito density and lifespan and interrupting malaria transmission. The NMEP has implemented IRS in specific high-burden areas such as Zamfara, Nasarawa, Plateau, and Sokoto States. According to the 2021 World Malaria Report (WHO, 2023a), the IRS protected 2.3% of the population at risk in Nigeria in 2021. Moreover, the NMEP has implemented Intermittent Preventive Treatment in Pregnancy (IPTp), involving the administration of antimalarial drugs to pregnant women during routine antenatal care visits, irrespective of malaria symptoms. IPTp is effective in preventing maternal and foetal complications like anaemia, low birth weight, and stillbirth. Following the 2016 WHO antenatal care model, the NMEP recommends a minimum of eight pregnancy-related contacts and at least three doses of IPTp with Sulfadoxine-pyrimethamine (SP). However, as per the 2018 NDHS, only 31% of pregnant women received three or more doses of IPTp with SP during their last pregnancy. Case management, entailing prompt diagnosis and effective treatment of malaria cases, is another focus in the fight against malaria. Adhering to WHO guidelines, rapid diagnostic tests (RDTs) and artemisinin-based combination therapy (ACT) are recommended as the first-line treatment for uncomplicated malaria.

Research Motivation and Questions

While several other interventions are currently being implemented to reverse the negative trend of the malaria epidemic in northern Nigeria, most of these interventions are capital-intensive, coupled with the fact that most of the funds for malaria intervention programming are provided by donors. Additionally, Nigeria, being a developing country, is plagued by the paucity of funds for public health programming; hence, it is important to leverage the advancement in technologies, such as the use of machine learning, to improve the cost efficiency of the malaria intervention in the country. The motivation for this research stems from the successful deployment of machine learning models in the prediction of epidemic and disease outbreaks in some other nations (Mbunge et al., 2023), (Morang'a et al., 2020), (Ugwu, Onyejegbu, and Obagbuwa, 2010) and the persistent high malaria burden in Sokoto State, Nigeria, resulting in the loss of scarce financial resources, loss of productive hours, high mortality, and morbidity, which could have been prevented if an early warning system were in place.

Additionally, the WHO also recently suggested the adoption of digital technologies for universal health coverage. It is hoped that a machine learning model that can accurately predict malaria disease outbreaks can assist in developing early malaria warning systems, redesigning interventions, making informed decision-making, and subsequently strengthening malaria prevention and control measures in Sokoto State, Nigeria, thereby improving the population health and maximizing the scarce resources available. In this regard, the following questions are highlighted to guide the objectives and scope of this study:

- How accurately can machine learning models identify patterns and predictors of malaria outbreaks in Sokoto State, Nigeria, using climate data (specifically rainfall and temperature) and historical malaria prevalence data?
- Among the specified machine learning models (linear regression, K-means, support vector machine), which one performs best in

predicting malaria prevalence in Northern Nigeria?
- How do these models compare in terms of accuracy, sensitivity, specificity, and robustness?
- Can we develop a reliable prediction model that anticipates the onset of malaria epidemics up to three months in advance?

Objectives and Aims of Study

This study seeks to predict and manage malaria outbreaks in northern Nigeria proactively. It leverages a robust and data-driven model to forecast outbreaks by analyzing diverse data streams. This early warning system empowers stakeholders to proactively plan and implement interventions such as improved resource availability and strengthened healthcare infrastructure. Ultimately, these efforts aim to mitigate the public health impact of malaria outbreaks by reducing both mortality and morbidity rates in the region. The objectives of the study are listed below:

- To determine the precision of machine learning models in the identification of patterns and predictors of malaria outbreaks in northern Nigeria using climate (rainfall and temperature) and Malaria prevalence data.
- To compare the performance of different machine learning models, such as Random Forest, K-Nearest Neighbor, and Support Vector Machine, for predicting malaria prevalence in northern Nigeria using climatic data (rainfall and temperature) from weather stations and malaria prevalence data.
- To train and validate a prediction model for the onset of malaria epidemic up to three months in advance in northern Nigeria using Machine Learning

Significance of Study

Seasonality is a major factor in malaria transmission, with a high prevalence of malaria being recorded in months with certain climatic conditions. According to Mabaso et al. (2007), several studies have identified rainfall and temperature as major causes of malaria outbreaks. The development of a robust and accurate machine learning model that will accurately predict malaria outbreaks ahead of time using historical data will improve the malaria program intervention in northern Nigeria and will improve cost efficiency, whilst also reducing morbidity and mortality.

Delimitations, Limitations, and Assumptions

A few limitations have been identified in the study, especially as it relates to data availability and quality. Being a study that is conducted using secondary data (Data from the Nigeria Health Management Information System), historical malaria incidence data may contain some data quality issues and gaps due to reporting variations, misdiagnoses, or underreporting. Additionally, climatic data—especially at a local level—might suffer from sparse measurements or inaccuracies. Ensuring robust data collection and validation is essential for accurate model training and evaluation. Additionally, the quality of climate data from weather stations impacts the reliability of our predictions. Addressing these limitations requires careful curation and validation of datasets.

The second set of limitations pertains to spatial and temporal generalization. While our study focuses on Sokoto State in northern Nigeria, malaria dynamics can vary significantly within the region. Extrapolating findings to other regions or periods requires caution. Local variations in climate, vector behavior, and healthcare infrastructure may impact model performance. Furthermore, our models assume linear relationships between climatic variables and malaria incidence, which may oversimplify complex interactions. Non-climatic factors (e.g., socioeconomic status, population density) are not explicitly considered, potentially affecting the precision of our predictions. Balancing model complexity with generalization and addressing these assumptions is crucial for robust research outcomes. Acknowledging these limitations ensures transparency and informs future research directions.

Literature Review

This section reviews recent studies on the use of ML to predict disease outbreaks and epidemics and its merit in ensuring that proper preparations are made to avert potential epidemics and pandemics using this prediction model. This section also reviews the different models that have been employed in the prediction of these disease incidents, their robustness, accuracy, and limitations.

Relationship between Malaria Incidence and Climatic Data

In a study by Nkiruka, Prasad, and Clement (2021) on the prediction of malaria incidence using climate variability and machine learning in highly malaria-endemic countries of Burkina Faso, Mali, Niger Republic, Nigeria, Cameroon, and the Democratic Republic of Congo (DRC), climatic variabilities such as rainfall, humidity, surface radiation, air temperature, and atmospheric pressure were identified as key drivers of malaria transmission and prevalence. The study specifically identified a linear relationship between an increase in monthly minimum temperature and a corresponding increase in the likelihood of malaria transmission, leading to an increase in malaria cases. Additionally, a higher number of malaria cases was also recorded during the rainy season, further reinforcing the link between climatic data and malaria incidents. Having established the link between climatic data and malaria incidents, it is important to identify the possible ways of correctly predicting a malaria outbreak. In another study by Okunlola and Oyeyemi (2019), a significant association between malaria incidence and climatic factors, including temperature and rainfall, was also discovered. The authors utilized regression analysis to establish the relationship and highlighted the importance of incorporating climatic data in malaria prediction models.

Different Forecasting Techniques

Several strategies have been employed in predicting malaria cases using climatic data in recent times. One of the strategies employed in the prediction of malaria incidence using climatic data is statistical and mathematical models, such as time series. In a study by Makinde, Abiodun, and Ojo (2020) on the modeling of malaria incidence in Akure (south-west Nigeria), using a statistical model (the negative binomial approach) to predict malaria incidents due to climatic variability, the model (a discrete probability distribution model which models the number of failures in a sequence of independent and identically distributed Bernoulli trials before a denoted success occurs) was used. In this case, a positive malaria case was the denoted success, and the number of negative malaria cases was recorded for each positive malaria case. The negative binomial model was formulated for each of the inpatient, outpatient, and mortality counts, and it was found that temperature and relative humidity positively correlate with malaria incidents. Though the binomial approach offers a robust approach in forecasting malaria incidence and it can handle overdispersion and outliers, it is, however it is better used when explanatory variables are being factored into the analysis.

In another study by Kumar et al. (2014) on forecasting malaria cases using climatic factors in Delhi, India, where a time series was used for the forecasting of malaria using climatic factors, a significant relationship was identified between malaria incidence and various climatic factors such as rainfall, relative humidity, and median temperature. The study used the autoregressive average moving model (ARIMA) time series for the forecast. The ARIMA forecasting model, which has been used to forecast stock prices, health outcomes, epidemics, etc. (Benvenuto et al., 2020), (Ariyo, Adewumi, and Ayo, 2014) is a statistical model used to analyse and forecast time series data. It is usually used for data collected over regular intervals, like daily sales figures or monthly customer counts. The model predicts future values using past datasets. ARIMA models are described by a specific notation (p, d, q). These parameters indicate the specific types of past values considered (p), the degree of differencing used (d), and the number of past moving averages included (q). Though the ARIMA model is good for understanding patterns in time series data and making forecasts based on those patterns, it is required that the data is stationary, so there might be

a pre-processing step to make the data meet the specified requirement. ARIMA models, though powerful for analysing time series data and making forecasts based on past patterns has some limitations. One requirement is that the data exhibits stability over time, often achieved through pre-processing. Additionally, these models assume a continuation of past trends, which can be shaky ground for forecasting unexpected events. Furthermore, ARIMA models struggle with complex patterns and data that don't follow a normal distribution. In essence, they are well-suited for predictable scenarios but may not be the best choice for forecasting major disruptions or highly volatile data (www.sciencedirect.com, 2019).

While other statistical and mathematical methods have been successfully applied in the forecast of malaria incidence, the application of Machine Learning Models to the same problem in recent times has brought better predictions due to the ability of machine learning models to uncover hidden insights, capture and manage non-linear patterns, and their capacity to handle very large and complex datasets. In a study by Mbunge et al. (2022) on the application of machine learning models to predict malaria incidence using malaria cases and environmental risk factors in sub-Saharan Africa, models such as logistic regression, support vector machine, decision tree, and random forest classifier were used. The logistic regression model and random forest classifier recorded 83% prediction accuracy; however, the random forest classifier achieved a higher precision of 87% more than other methods, while the decision tree had the least precision of 66%. The study reiterated the importance of machine learning models in the prediction of malaria incidents with a very high degree of accuracy, helping stakeholders and players effectively allocate resources such as human, financial, and physical resources while also helping stakeholders develop early warning systems to monitor the possible spread of the disease and subsequently strengthen control measures.

In another study by Aliyu Adamu and Singh (2021) on malaria prediction model using machine learning algorithms where six machine learning models (Support Vector Machine, Naïve Bayes, Decision Tree, Artificial Neural Network (ANN), Random Forest, Logistic Regression, and Gradient Extreme Boost Algorithm) were evaluated to determine the most accurate forecast model. The models used a dataset downloaded from the WHO web portal on malaria incidence per 1000 population and other malaria-causing factors such as availability of basic drinking water, average temperature, and average rainfall. Like in the earlier evaluated study, the random forest model achieved the highest percentage of accuracy among the six deployed models with an accuracy of 97.72%, precision of 100%, F1 score of 98%, and error rate of 2.27% beating the closest rival, logistic regression, with 95.45% accuracy, 95% precision, and 95% F1 score.

Choosing the right ML Model

Identifying and choosing the right ML model to use for malaria forecasting is pivotal to the success of malaria prediction modeling. According to Ileperuma et al. (2023), the prediction task must first be evaluated as either a classification, regression, or clustering problem, and since malaria outbreak is a binary classification (being either a malaria outbreak or not), the use of classification algorithms such as Random Forest, Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) are top choices for the forecasting model. Additionally, selecting the right machine learning model requires careful consideration of the data at hand. Large datasets and numerous data points benefit from algorithms like Gradient Boosting and Random Forest. High dimensionality, where data has many features, might necessitate techniques to reduce the feature space. Imbalanced data, where one category dominates, can be tackled with specific algorithms (Random Forest, SVM) or adjustments like oversampling or under sampling. Understanding which features hold the most weight in predictions is crucial, and Random Forest offers built-in features for this. However, a trade-off exists between a model's interpretability and its performance. While decision trees and Random Forest are clear to understand, deep learning models often prioritize high performance at the expense of being like a black box. Finally, the presence of noisy data or outliers necessitates the use

of robust algorithms like Random Forest and SVM. To this end, Random Forest, SVM, and K-Nearest Neighbour will be used for this research work.

Machine Learning Models

Random Forest

Random forest is a powerful machine-learning technique that utilizes an ensemble of decision trees. Each tree is built using a random subset of data points (bootstrap sample) from the training data. Additionally, randomness is introduced by considering only a random subset of features (feature bagging) when splitting each node in the trees (Pal, 2005). This approach helps reduce correlation among trees and prevents overfitting. Random forest is versatile and can handle both regression and classification tasks. It is popular due to its high accuracy and ability to work with missing data. Furthermore, the algorithm allows for easy evaluation of feature importance through metrics like Gini importance and mean decrease in impurity. While powerful, random forest comes with certain drawbacks. Training the model can be computationally expensive due to the large number of decision trees involved. This also translates to requiring more resources for data storage. Additionally, interpreting the predictions from a random forest can be complex compared to a single decision tree (Parmar, Katariya and Patel, 2018).

Support Vector Machine (SVM)

Support Vector Machines (SVMs) are powerful supervised learning algorithms designed for classification tasks. Unlike other algorithms that merely draw a separation line, SVMs excel at finding the optimal hyperplane in high-dimensional spaces (Jakkula, 2006). This hyperplane acts as a clear decision boundary, effectively separating data points belonging to different classes. The key strength of SVMs lies in their ability to maximize the margin between the hyperplane and the closest data points from each class (Huang et al., 2018). These critical data points, known as support vectors, play a crucial role in defining the hyperplane's position. A wider margin results in a more robust separation between classes, leading to a model that generalizes well on unseen data.

However, real-world data is rarely perfectly separable in its original form. To address this challenge, SVMs utilize kernel functions. These functions project data points into a higher-dimensional space, allowing the creation of a hyperplane that effectively separates classes, even when they are indistinguishable in lower dimensions. Once the optimal hyperplane is established, classifying new data points becomes straightforward. SVMs analyze the position of the new point relative to the hyperplane. Points on one side are assigned to a specific class, while those on the other belong to the opposing class. SVMs achieve robust classification by identifying a hyperplane that maximizes the margin between the decision boundary and the most influential data points (support vectors) (Jakkula, 2006). This data-driven approach enables the creation of models that generalize effectively to unseen data, making SVMs valuable tools for various classification problems.

K-Nearest Neighbour

K-Nearest Neighbours (KNN) is a straightforward yet powerful machine-learning algorithm for both classification and regression tasks. Unlike complex models that require intricate training procedures, KNN operates on the fundamental principle of similarity.

At its core, KNN assumes that data points close together in the feature space likely share similar characteristics. This translates to similar classifications or values. To predict a new data point, KNN identifies its k nearest neighbours within the existing training data. Here, k represents a user-defined parameter specifying the number of neighbors to consider.

For classification, KNN assigns the new data point the most frequent class label amongst its k nearest neighbors (Wang et al., 2003). Imagine classifying an image based on color. If the k closest images to the unknown image are predominantly classified as "cat," the KNN algorithm would likely classify the unknown image as a "cat" as well.

In regression problems, KNN predicts the value of a new data point by averaging the values of its k nearest neighbors. This approach leverages the assumption that neighboring data points share similar values, allowing for a reasonable estimation of the unknown value.

While KNN offers a clear advantage in terms of its simplicity and interpretability, it comes with certain limitations. The algorithm's performance heavily relies on the chosen value of k. Selecting a very low k might lead to overfitting, where the model is overly sensitive to specific data points in the training set. Conversely, a large k value can result in underfitting, where the model fails to capture the intricacies of the data. Additionally, KNN can be computationally expensive for large datasets as it involves calculating distances between the new data point and all data points in the training set (Guo et al., 2003).

Overall, KNN serves as a versatile tool for various machine-learning applications due to its ease of understanding and implementation. However, careful consideration of the k parameter and potential computational cost is crucial for achieving optimal performance.

Methodology

Malaria Incidence Data

The malaria incidence data for Sokoto state were sourced from the Nigeria Health Management Information System (HMIS) powered by the District Health Information System (DHIS) platform on https://dhis2nigeria.org.ng/. The Nigeria HMIS platform is the central storage of Nigeria's health information data for all disease areas. Data for disease areas such as Malaria, HIV/AIDS, Maternal and Neonatal Child Health (MNCH), and routine immunization are hosted on the platform. Malaria incidence data for 2 indicators were downloaded, and they are:

1. Number of confirmed malaria cases with MRDT
2. Number of confirmed malaria cases with microscopy

Monthly data was downloaded from January 2015 to December 2022. The data for both indicators were summed together in Microsoft Excel to determine the total malaria incidence for each month.

Climate Data

The climatic data was sourced from the National Aeronautics and Space Administration (NASA) Data Access Viewer Website (https://power.larc.nasa.gov/data-access-viewer/). Data was sourced for Sokoto state by month from January 2015 to December 2022. The monthly maximum temperature and average precipitation data were collected and aggregated with the malaria incidence data. The data was merged using Microsoft Excel and converted to comma-separated value format (CSV) for processing using the Python programming language.

Data Processing and Transformation

A preprocessing of the data was done before the modelling. This included the translation and encoding of the data so it could be parsed by the computer. The data cleaning process includes data collection, feature selection, and data partitioning. The data was prepared for modelling using MS Excel. In preparing the data for modelling, the World Health Organisation's (WHO) definition of an epidemic, where the case numbers rise above the five-year average plus two standard deviations, was used (Harvey, Valkenburg and Amara, 2021).

The data from 2015 to 2020 was used to calculate the 5-year average and standard deviation. The monthly data were then categorized in a calculated column to show "High" (High malaria Incidence) for a malaria incidence ratio above the calculated epidemic level and "Low" (Low malaria Incidence) for a malaria incidence ratio that is lower than the epidemic level.

Data Modelling

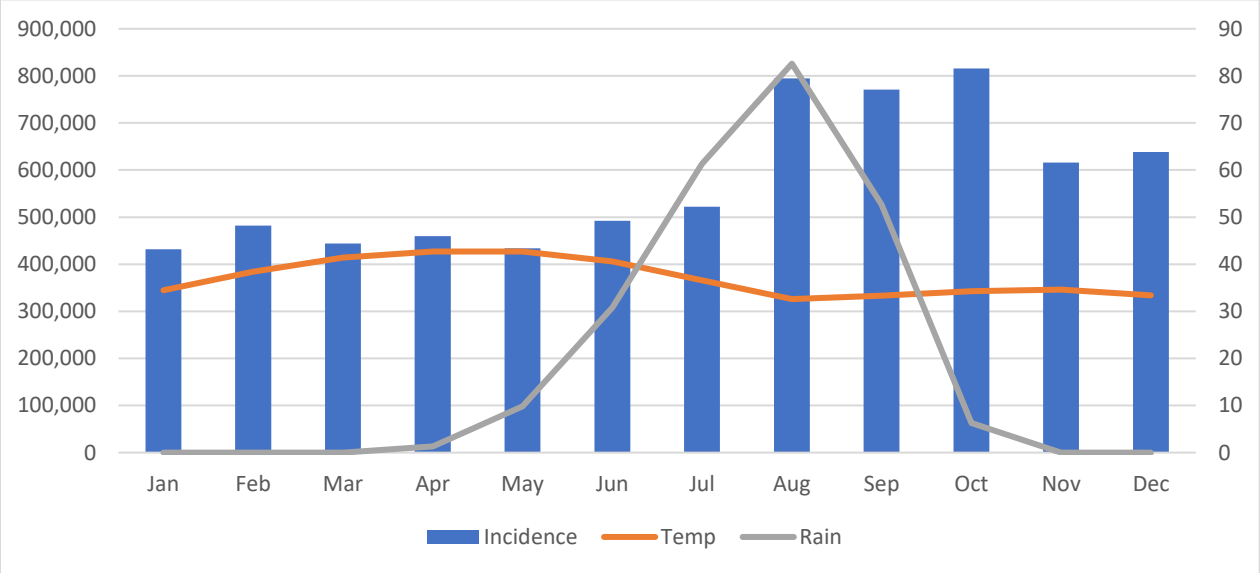
The data processing and modelling were done using the Python 3 programming language using Google Colab. Python is a popular language for machine learning and data science. It supports statistical analysis, machine learning, data visualization, and graphical plotting. It offers extensive libraries like

Scikit-learn, Pandas, and NumPy, along with user-friendly data structures for classification, clustering, and desktop applications in machine learning and data mining (Pratiyush Guleria and Sood, 2018). The dataset was split into two parts; 30% of the data was used as training data, while the remaining 70% was used for the modelling. The three selected models, namely Random Forest Classifier, Support Vector Machine, and K-Nearest Neigbour were trained and used for the modelling and the model report was printed and compared.

Malaria Incidence and Climatic Data

As assumed for this study and established by several other studies, a linear relationship between malaria incidence, temperature, and precipitation has been experienced severally. Figure 1 below shows a plot of the malaria incidence category against the recorded maximum temperature. A high malaria incidence was recorded when the monthly temperature dropped because of increased precipitation. According to Odongo-Aginya et al. (2005) in a study on Malaria incidence and rainfall in Uganda, a direct relationship was observed between malaria transmission and monthly rainfall in Entebbe Municipality, where the study was carried out, further reinforcing the pattern discovered in this study. Additionally, Mafwele and Lee (2022) also reinforced this thought, having discovered through their study that increased precipitation and temperature fluctuation are linearly related to malaria incidence.

Figure 1 – Malaria Incidence versus Temperature and Precipitation



Predictive Model Result

As seen in tables 2, 3, and 4, which are the results of the models (Random Forest, K-Nearest Neighbour, and Support Vector Machine, respectively), it is seen that machine learning can become a useful tool in the prediction of a possible malaria epidemic. Based on the results of the models, Support Vector Machine (SVM) achieved the highest prediction accuracy of 76% meaning the model was able to accurately predict 76% of high malaria cases correctly, while both KNN and Random Forest achieved 59% accuracy, showing that they were only able to predict about 59% of high malaria cases correctly.

Table 2 – Random Forest Classifier Report

Random Forest	Precision	Recall	F1-Score	Support
High	0.61	0.69	0.65	12
Low	0.55	0.46	0.5	17
Accuracy			0.59	29
Macro Avg	0.58	0.57	0.57	29
Weighted Avg	0.58	0.59	0.58	29

Table 3 – K-Nearest Neigbour Report

KNN	Precision	Recall	F1-Score	Support
High	0.62	0.62	0.62	12
Low	0.54	0.54	0.54	17
Accuracy			0.59	29
Macro Avg	0.58	0.58	0.58	29
Weighted Avg	0.59	0.59	0.59	29

Table 4 – Support Vector Machine Report

SVM	Precision	Recall	F1-Score	Support
High	0.76	0.81	0.79	12
Low	0.75	0.69	0.72	12
Accuracy			0.76	29
Macro Avg	0.76	0.75	0.75	29
Weighted Avg	0.76	0.76	0.76	29

Additionally, SVM also recorded the highest F1 score of 79% showing that the model can make correct predictions of both high and low malaria cases with about 79% accuracy, while Random Forest followed with a 65% F1 score, and KNN achieved the lowest F1 score of 62%. Furthermore, SVM also achieved the highest Recall percentage of the 3 models, with 81% showing the model correctly identified 81% of actual positives, which is followed by Random Forest at 69% and KNN at 62%. The result shows that SVC is the best model for malaria incidence prediction based on the dataset used for this research work. Additionally, it is not advisable to use either KNN or Random Forest due to the low accuracy recorded.

Discussion

This study employed three machine learning models: Random Forest (RF), K-Nearest Neighbour (KNN), and Support Vector Machine (SVM) to predict malaria outbreaks in Sokoto State, Nigeria, based on climatic data. The findings revealed that KNN demonstrated a higher prediction accuracy of 76% compared to 59% for the other two models. KNN's better performance can be attributed to its simplicity and effectiveness in handling non-linear relationships within the data. KNN's effectiveness in classifying data points based on their closeness to other points in the feature space likely enabled it to better capture the intricate relationships between climatic variables and malaria incidence. On the

other hand, the lower accuracy of RF and SVM could be due to issues such as sensitivity to data quality, overfitting in RF, or less-than-ideal kernel selection and data scaling in SVM. Enhancing their performance might require additional hyperparameter tuning and feature engineering. Additionally, several factors can be responsible for model performance, and they include data quality, algorithm complexity, hyperparameter tuning, and data volume, and this is echoed by other studies (Battineni et al., 2020).

The result also showed that climatic data have a significant relationship with malaria incidence and outbreaks in Sokoto State, with temperature, humidity, and rainfall emerging as critical predicting factors. High temperatures and humidity levels create favourable conditions for mosquito breeding, while rainfall patterns influence the availability of breeding sites. These findings align with existing literature on the relationship between climate and malaria transmission. The predictive model developed in this study has important implications for public health strategies in Sokoto State. By identifying climatic conditions that precede malaria outbreaks, health officials can allocate resources more efficiently and implement preventive measures proactively. For instance, targeted distribution of insecticide-treated nets and indoor residual spraying can be intensified during periods of high risk, potentially reducing the incidence of malaria (Ahmed Abubakar Jajere et al., 2023). Additionally, implementing real-time monitoring systems and early warning systems can facilitate timely interventions and resource allocation. Ethical implications, such as data privacy, bias, and fairness, should be carefully considered throughout the development and deployment of machine learning models for public health applications.

While this study provides valuable insights in the prediction of malaria outbreaks, the accuracy and reliability of the models can be improved using high quality, comprehensive granular data. The use of a more diverse source of data which includes socio-economic, demographic and healthcare datasets can help to improve predictive capacity of the models. Additionally, combining multiple models through ensemble methods can enhance predictive accuracy and robustness (Ahmed Abubakar Jajere et al., 2023).

Conclusion and Recommendation

Malaria remains a significant global health burden, particularly in resource-limited settings. The World Health Organization (WHO) estimates that in 2022 (WHO, 2022), there were 249 with Nigeria accounting for over 66 million of the cases, which is attributable to the malaria-endemic northern Nigeria. Early detection of malaria, especially in children and pregnant women, leading to prompt intervention, is important in the prevention of severe illness and possible death from the illness. Predictive analytics using machine learning offers a good option for predicting malaria incidence and early detection of possible epidemics, potentially enabling the development of preventative measures and allocation of resources by the relevant stakeholders to the areas with identified risk.

This work explored the possibility of using machine learning algorithms to predict possible malaria incidence and possible incoming epidemics based on two climatic factors, namely rainfall and temperature. The research investigated the performance of three machine learning models, namely, Random Forests, Support Vector Machines (SVMs), and K-Nearest Neighbors (KNN). The results demonstrated that the SVM model achieved the highest prediction accuracy of 76% in predicting, highlighting its potential as a valuable tool for informing malaria public health interventions. However, several limitations exist, and it's important to acknowledge that model performance can be influenced by several factors, such as the quality and completeness of the training data, the specific geographical location, and the local variations in climatic patterns and mosquito populations.

It is suggested that future research directions should focus on addressing these identified limitations and further exploring the potential of machine learning for malaria prediction. Additional factors affecting malaria transmission and prevalence, such as

socioeconomic factors (poverty, education level, lack of access to clean water and sanitation, and inadequate housing) all contribute to malaria risk, and human mobility data, that is, the population movement patterns, which have also been identified to influence malaria transmission. Integrating data on travel and migration patterns could improve the models' ability to predict outbreaks in areas with high population turnover (Gbenga Adegbite et al., 2023), (Debnath et al., 2024).

Current machine learning applications in malaria control demonstrate promise but hold the potential for further refinement. The implementation of advanced deep learning models could significantly enhance predictive accuracy. By addressing existing limitations and pursuing ongoing research efforts, machine learning can evolve into a robust tool for optimizing malaria control strategies (Ahmed Abubakar Jajere et al., 2023). This advancement can ultimately contribute to a diminished disease burden, improved public health outcomes, and a significant stride towards achieving the World Health Organization's objective of malaria elimination.

References

Abdullahi, K., Abubakar, U., Adamu, T., Daneji, A.I., Aliyu, R.U., Jiya, N., Ibraheem, M.T.O. and Nata'ala, S.U. (2009). Malaria in Sokoto, North Western Nigeria. *African Journal of Biotechnology*, [online] 8(24). doi:https://doi.org/10.4314/ajb.v8i24.68803.

Ahmed Abubakar Jajere, Muhammed Bukar Ngamdu, Ibrahim, U. and Ahmed, N. (2023). Analysis Of Impact Of Rainfall And Temperature Variability On Malaria Incidence In Yamaltu/Deba Local Government Of Gombe State, Nigeria. *Deleted Journal*, 1(2), pp.75–80. doi:https://doi.org/10.26480/magg.02.2023.75.80.

Alashwal, H., El Halaby, M., Crouse, J.J., Abdalla, A. and Moustafa, A.A. (2019). The Application of Unsupervised Clustering Methods to Alzheimer's Disease. *Frontiers in Computational Neuroscience*, 13(31). doi:https://doi.org/10.3389/fncom.2019.00031.

Aliyu Adamu, Y. and Singh, J. (2021). Malaria Prediction Model Using Machine Learning Algorithms. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(10), pp.7488–7496. doi:https://doi.org/10.17762/turcomat.v12i10.5655.

Ariyo, A.A., Adewumi, A.O. and Ayo, C.K. (2014). Stock Price Prediction Using the ARIMA Model. *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*. [online] doi:https://doi.org/10.1109/uksim.2014.67.

Battineni, G., Sagaro, G.G., Chinatalapudi, N. and Amenta, F. (2020). Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis. *Journal of Personalized Medicine*, [online] 10(2). doi:https://doi.org/10.3390/jpm10020021.

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S. and Ciccozzi, M. (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data in Brief*, 29, p.105340. doi:https://doi.org/10.1016/j.dib.2020.105340.

Center for Disease Control, CDC (2022). *About the Analytics and Modeling Track | CDC*. [online] www.cdc.gov. Available at: https://www.cdc.gov/pef/analytics-and-modeling/index.html.

Debnath, A., Anirban Tarafdar, A. Poojitha Reddy and Bhattacharya, P. (2024). ROVM integrated advanced machine learning-based malaria prediction strategy in Tripura. *˜The œJournal of supercomputing/Journal of supercomputing*. doi:https://doi.org/10.1007/s11227-024-06094-w.

Dridi, S. (2022). Supervised Learning - A Systematic Literature Review. *www.researchgate.net*. doi:https://doi.org/10.31219/osf.io/tysr4.

Gbenga Adegbite, S.O. Edeki, Itunuoluwa Isewon, Emmanuel, J., Dokunmu, T.M., Rotimi, S.O., Jelili Oyelade and Adebiyi, E. (2023). Mathematical modeling of malaria transmission dynamics in

humans with mobility and control states. *Infectious Disease Modelling*, [online] 8(4), pp.1015–1031. doi:https://doi.org/10.1016/j.idm.2023.08.005.

Guo, G., Wang, H., Bell, D., Bi, Y. and Greer, K. (2003). KNN Model-Based Approach in Classification. *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, 2888, pp.986–996. doi:https://doi.org/10.1007/978-3-540-39964-3_62.

Haneef, R., Kab, S., Hrzic, R., Fuentes, S., Fosse-Edorh, S., Cosson, E. and Gallay, A. (2021). Use of artificial intelligence for public health surveillance: a case study to develop a machine Learning-algorithm to estimate the incidence of diabetes mellitus in France. *Archives of Public Health*, 79(1). doi:https://doi.org/10.1186/s13690-021-00687-0.

Harvey, D., Valkenburg, W. and Amara, A. (2021). Predicting malaria epidemics in Burkina Faso with machine learning. *PLOS ONE*, 16(6), p.e0253302. doi:https://doi.org/10.1371/journal.pone.0253302.

He, K., Zhang, X., Ren, S. and Sun, J. (2015). Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9), pp.1904–1916. doi:https://doi.org/10.1109/tpami.2015.2389824.

Huang, S., Cai, N., Pacheco, P.P., Narrandes, S., Wang, Y. and Xu, W. (2018). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics & Proteomics*, [online] 15(1), pp.41–51. Available at: https://cgp.iiarjournals.org/content/15/1/41.short.

Ileperuma, K., Jampani, M., Sellahewa, U., Panjwani, S. and Amarnath, G. (2023). *Predicting Malaria Prevalence with Machine Learning Models Using December 2023 Colombo, Sri Lanka*. [online] Available at: https://cgspace.cgiar.org/server/api/core/bitstreams/f8d9192b-4407-4122-af43-73437550cbdc/content [Accessed 13 Mar. 2024].

Jakkula, V. (2006). *Tutorial on Support Vector Machine (SVM)*. [online] Available at: https://course.khoury.northeastern.edu/cs5100f11/resources/jakkula.pdf.

James, G., Witten, D., Hastie, T., Tibshirani, R. and Taylor, J. (2023). Unsupervised Learning. pp.503–556. doi:https://doi.org/10.1007/978-3-031-38747-0_12.

Jordan, M.I. and Mitchell, T.M. (2020). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), pp.255–260. doi:https://doi.org/10.1126/science.aaa8415.

Kaelbling, L.P., Littman, M.L. and Moore, A.W. (1996). Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4, pp.237–285. doi:https://doi.org/10.1613/jair.301.

Kumar, V., Mangal, A., Panesar, S., Yadav, G., Talwar, R., Raut, D. and Singh, S. (2014). Forecasting Malaria Cases Using Climatic Factors in Delhi, India: A Time Series Analysis. *Malaria Research and Treatment*, 2014, pp.1–6. doi:https://doi.org/10.1155/2014/482851.

Liakos, K.G., Busato, P., Moshou, D., Pearson, S. and Bochtis, D. (2018). Machine Learning in Agriculture: A Review. *Sensors (Basel, Switzerland)*, [online] 18(8), p.2674. doi:https://doi.org/10.3390/s18082674.

Mabaso, M.L.H., Craig, M., Ross, A. and Smith, T. (2007). Environmental Predictors Of The Seasonality Of Malaria Transmission In Africa: The Challenge. *The American Journal of Tropical Medicine and Hygiene*, [online] 76(1), pp.33–38. doi:https://doi.org/10.4269/ajtmh.2007.76.33.

Mafwele, B.J. and Lee, J.W. (2022). Relationships between transmission of malaria in Africa and climate factors. *Scientific Reports*, 12(1). doi:https://doi.org/10.1038/s41598-022-18782-9.

Makinde, O.S., Abiodun, G.J. and Ojo, O.T. (2020). Modelling of malaria incidence in Akure, Nigeria: negative binomial approach. *GeoJournal*, 86(3), pp.1327–1336. doi:https://doi.org/10.1007/s10708-019-10134-x.

Mbunge, E., Milham, R.C., Maureen Nokuthula Sibiya and Takavarasha, S. (2023). Machine Learning Techniques for Predicting Malaria: Unpacking Emerging Challenges and Opportunities for Tackling Malaria in Sub-saharan Africa. pp.327–344. doi:https://doi.org/10.1007/978-3-031-35314-7_30.

Mbunge, E., Millham, R.C., Sibiya, M.N. and Takavarasha, S. (2022). Application of machine learning models to predict malaria using malaria cases and environmental risk factors. *2022 Conference on Information Communications Technology and Society (ICTAS)*. doi:https://doi.org/10.1109/ictas53252.2022.9744657.

Morang'a, C.M., Amenga–Etego, L., Bah, S.Y., Appiah, V., Amuzu, D.S.Y., Amoako, N., Abugri, J., Oduro, A.R., Cunnington, A.J., Awandare, G.A. and Otto, T.D. (2020). Machine learning approaches classify clinical malaria outcomes based on haematological parameters. *BMC Medicine*, [online] 18(1). doi:https://doi.org/10.1186/s12916-020-01823-3.

Murdoch, T.B. and Detsky, A.S. (2013). The Inevitable Application of Big Data to Health Care. *JAMA*, [online] 309(13), p.1351. doi:https://doi.org/10.1001/jama.2013.393.

Nigeria Population Commission and ICF International (2018). Nigeria Demographic And Health Survey 2018. *Ngfrepository.org.ng*. [online] doi:http://ngfrepository.org.ng:8080/jspui/handle/123456789/3145.

Nkiruka, O., Prasad, R. and Clement, O. (2021). Prediction of malaria incidence using climate variability and machine learning. *Informatics in Medicine Unlocked*, 22, p.100508. doi:https://doi.org/10.1016/j.imu.2020.100508.

Odongo-Aginya, E., Ssegwanyi, G., Kategere, P. and Vuzi, P.C. (2005). Relationship between malaria infection intensity and rainfall pattern in Entebbe peninsula, Uganda. *African Health Sciences*, [online] 5(3), pp.238–245. Available at: https://www.ajol.info/index.php/ahs/article/view/7023.

Okunlola, O.A. and Oyeyemi, O.T. (2019). Spatio-temporal analysis of association between incidence of malaria and environmental predictors of malaria transmission in Nigeria. *Scientific Reports*, [online] 9(1), p.17500. doi:https://doi.org/10.1038/s41598-019-53814-x.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1), pp.217–222. doi:https://doi.org/10.1080/01431160412331269698.

Parlett-Pelleriti, C.M., Stevens, E., Dixon, D. and Linstead, E.J. (2022). Applications of Unsupervised Machine Learning in Autism Spectrum Disorder Research: a Review. *Review Journal of Autism and Developmental Disorders*. doi:https://doi.org/10.1007/s40489-021-00299-y.

Parmar, A., Katariya, R. and Patel, V. (2018). A Review on Random Forest: An Ensemble Classifier. *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*, 26, pp.758–763. doi:https://doi.org/10.1007/978-3-030-03146-6_86.

Petroc Taylor (2023). *Data Created Worldwide 2010-2025* . [online] Statista. Available at: https://www.statista.com/statistics/871513/worldwide-data-created/.

Pratiyush Guleria and Sood, M. (2018). Predictive Data Modeling: Educational Data Classification and Comparative Analysis of Classifiers Using Python. doi:https://doi.org/10.1109/pdgc.2018.8745727.

Sarker, I.H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, [online] 2(3), pp.1–21. doi:https://doi.org/10.1007/s42979-021-00592-x.

Sarker, I.H., Kayes, A.S.M., Badsha, S., Alqahtani, H., Watters, P. and Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. *Journal of Big Data*, [online] 7(1). doi:https://doi.org/10.1186/s40537-020-00318-5.

Ugwu, C., Onyejegbu, N.L. and Obagbuwa, I.C. (2010). The Application of Machine Learning Technique for Malaria Diagnosis. *International Journal of Green Computing*, 1(1), pp.68–77. doi:https://doi.org/10.4018/jgc.2010010107.

Wang, H., Düntsch, I., Gediga, G., University, B., Wang@ulst, H., Uk, I., Düntsch, G. and Gediga (2003). *Nearest Neighbours without k: A Classification Formalism Based on Probability Nearest Neighbours without : A Classification Formalism based on Probability (Extended Abstract)*. [online] Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=0de171728b1bc379b759fb9a96346ed0d5f2a044.

World Health Organization, WHO (2023a). *Malaria*. [online] World Health Organization. Available at: https://www.who.int/news-room/fact-sheets/detail/malaria.

World Health Organization, WHO (2023b). *World malaria report 2023*. [online] www.who.int. Available at: https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2023.

www.sciencedirect.com. (2019). *Autoregressive Integrated Moving Average - an overview | ScienceDirect Topics*. [online] Available at: https://www.sciencedirect.com/topics/mathematics/autoregressive-integrated-moving-average.