Scientific Research Publishing

Journal of Geographic Information System, 2024, 16, 115-135
https://www.scirp.org/journal/jgis
ISSN Online: 2151-1969
ISSN Print: 2151-1950

Advancing Malaria Prediction in Uganda through AI and Geospatial Analysis Models

Maria Assumpta Komugabe[1,2*], Richard Caballero[1], Itamar Shabtai[1], Simon Peter Musinguzi[3,4,5]

[1]Center for Information Systems & Technology, Claremont Graduate University, Claremont, California, USA
[2]Department of Information Technology, Faculty of Business and ICT, University of Kisubi, Entebbe, Uganda
[3]Faculty of Health Sciences, University of Kisubi, Entebbe, Uganda
[4]Department of Agriculture Production, Faculty of Agriculture, Kyambogo University, Kyambogo, Uganda
[5]Department of Agriculture and Natural Sciences, Faculty of Agriculture, Uganda Martyrs University, Kampala, Uganda
Email: *maria-assumpta.komugabe@cgu.edu, richard.caballero@cgu.edu, itamar.shabtai@cgu.edu, spmusinguzi@kyu.ac.ug

How to cite this paper: Komugabe, M.A., Caballero, R., Shabtai, I. and Musinguzi, S.P. (2024) Advancing Malaria Prediction in Uganda through AI and Geospatial Analysis Models. *Journal of Geographic Information System*, **16**, 115-135.
https://doi.org/10.4236/jgis.2024.162008

Received: February 6, 2024
Accepted: April 6, 2024
Published: April 9, 2024

Abstract

The resurgence of locally acquired malaria cases in the USA and the persistent global challenge of malaria transmission highlight the urgent need for research to prevent this disease. Despite significant eradication efforts, malaria remains a serious threat, particularly in regions like Africa. This study explores how integrating Gregor's Type IV theory with Geographic Information Systems (GIS) improves our understanding of disease dynamics, especially Malaria transmission patterns in Uganda. By combining data-driven algorithms, artificial intelligence, and geospatial analysis, the research aims to determine the most reliable predictors of Malaria incident rates and assess the impact of different factors on transmission. Using diverse predictive modeling techniques including Linear Regression, K-Nearest Neighbor, Neural Network, and Random Forest, the study found that; Random Forest model outperformed the others, demonstrating superior predictive accuracy with an R^2 of approximately 0.88 and a Mean Squared Error (MSE) of 0.0534, Antimalarial treatment was identified as the most influential factor, with mosquito net access associated with a significant reduction in incident rates, while higher temperatures correlated with increased rates. Our study concluded that the Random Forest model was effective in predicting malaria incident rates in Uganda and highlighted the significance of climate factors and preventive measures such as mosquito nets and antimalarial drugs. We recommended that districts with malaria hotspots lacking Indoor Residual Spraying (IRS) coverage prioritize its implementation to mitigate incident rates, while those with high malaria rates in 2020 require immediate attention. By advocating for the use of appropriate predictive models, our research emphasized the importance of evidence-based decision-making in malaria control strate-

gies, aiming to reduce transmission rates and save lives.

Keywords

Malaria, Predictive Modeling, Geospatial Analysis, Climate Factors, Preventive Measures

1. Introduction

In 2022, the World Health Organization (WHO) released a report revealing a significant malaria outbreak, affecting 249 million individuals globally and resulting in 608,000 fatalities. In Uganda specifically, more than 12.7 million people were infected, with 17,556 deaths recorded [1].

The emergence of locally acquired malaria cases in the USA from May to August 2023 has once again highlighted the ongoing threat posed by this disease. While the number of malaria cases in the USA is comparatively low compared to regions such as Africa, South America, and Asia, these occurrences offer important insights into disease patterns and control methods. The efforts to eliminate malaria in the USA, including mosquito control measures and the use of effective anti-malarial medications, emphasize the need for continued research and vigilance in the fight against the disease [2]. Meanwhile, Malaria transmission remains a persistent global challenge, especially through Anopheles mosquitos carrying Plasmodium parasites. Despite extensive global efforts, malaria continues to persist, underscoring the crucial role of vector control programs [3]. In 2022, global malaria cases reached 249 million, with 95% concentrated in 29 countries. Nigeria, the Democratic Republic of the Congo, Uganda, and Mozambique collectively contributed half of the total cases. Malaria, transmitted by mosquitoes, is preventable and treatable, but severe cases require urgent care. Uganda, ranking fifth in African malaria cases, faces an annual incidence rate of 478 cases per 1,000 people, impacting outpatient visits, hospital admissions, and fatalities [4] [5].

Geographic Information Systems (GIS) have emerged as invaluable tools in the realm of public health, particularly in disease control and surveillance efforts. By integrating spatial data with health-related information, GIS facilitates the visualization, analysis, and interpretation of disease patterns, ultimately aiding in effective decision-making and resource allocation [6]. The case of Zambia demonstrates the transformative potential of GIS in malaria eradication efforts, highlighting its importance in addressing pressing public health challenges. As technology continues to evolve and global health priorities evolve, the integration of GIS into disease control strategies will remain paramount in safeguarding public health and promoting well-being worldwide. The success of this initiative underscores the critical role of GIS in malaria control efforts and serves as a model for other countries facing similar health challenges like Uganda [7].

The complexity of understanding and predicting disease dynamics has long been a challenge in public health research. To tackle this issue, researchers have turned to Gregor's Type IV theory, which emphasizes the comprehensive exploration of phenomena by addressing fundamental questions of what, how, why, when, where, and what will be. This theory not only offers predictions but also provides testable propositions and causal explanations, thereby enhancing our understanding of disease mechanisms and behaviors. Central to Gregor's theory is the intertwined pursuit of explanation and prediction. Explanation serves as a crucial tool in fostering human comprehension, aiming to induce subjective understanding among individuals. Philosophical examinations into the nature of explanations suggest that a thorough explanation involves logically deriving phenomena from premises that encompass a covering law. This approach lays the groundwork for robust and insightful predictions in disease research, guiding researchers towards effective interventions and strategies [8].

In parallel with theoretical advancements, Geographic Information Systems (GIS) have emerged as indispensable tools in disease research. GIS possesses the unique capability to integrate diverse datasets, including disease data and environmental information, within a unified geographical framework. This integration facilitates the analysis of spatial patterns and variations, offering valuable insights into disease distribution and differentiation based on geographic location. Moreover, GIS significantly amplifies its explanatory capacity, particularly when diseases are associated with environmental risk factors. By visualizing spatial relationships and correlations, GIS enables researchers to uncover hidden patterns and identify potential causal mechanisms underlying disease transmission. This spatial perspective enhances our understanding of the complex interplay between environmental factors and disease outcomes, thereby informing targeted interventions and preventive measures [9].

Therefore, integration of Gregor's Type IV theory and GIS in disease research represents a powerful approach to understanding and predicting disease dynamics. By combining theoretical insights with spatial analysis, researchers can unravel the intricate complexities of disease transmission and behavior, ultimately paving the way for more effective public health interventions and strategies. As advancements in theory and technology continue to evolve, this interdisciplinary approach holds immense promise in addressing pressing health challenges and improving global health outcomes.

In addressing malaria in Uganda, this research aimed to utilize AI predictive modeling. Machine learning algorithms analyzed extensive datasets, incorporating geospatial and historical information for pattern recognition and correlation analysis. The results demonstrated AI's capability to forecast future malaria incidence rates, providing valuable insights for targeted interventions.

Climatic and geographical factors, particularly rainfall and temperature, significantly influence the biology, manifestation, and distribution of mosquitoes. This impact extends to the seasonal transmission of malaria, with temperature

and rainfall contributing to its occurrence with a lag time before and after the rainy season [10]. Over several years, malaria has persisted as a leading global cause of death, notably prevalent in Africa. The influence of climate on malaria proliferation, wet and warm environments foster the breeding of mosquitoes responsible for transmission [11]. This research focused on assessing the effectiveness of climatic indicators, including precipitation and mean temperature, in predicting fluctuations in malaria incidents. Multiple classification algorithms were tested to identify the most suitable one for predicting outcomes in this specific scenario.

Cluster detection involves statistically analyzing the spatial or spatio-temporal distribution of diseases to identify specific areas and periods of concentration. This approach is crucial for controlling infectious diseases by informing targeted interventions and ensures cost-efficient resource use in high disease burden clusters. The application of spatial clustering and mapping techniques has proven effective in studying mosquito-borne infections, especially in the spatial and spatio-temporal analysis of malaria. Additionally, it can reveal areas lacking access to malaria care [12].

Several scholarly inquiries have been undertaken to examine the distribution of malaria epidemics across various geographical regions within Uganda. Nevertheless, a substantial proportion of these investigations has exhibited limitations by omitting critical factors that exert influence on both the incidence rates of malaria and the efficacy of preventive measures. Consequently, there exists a gap in the existing research landscape, warranting a scholarly exploration into the utilization of Machine Learning (ML), Artificial Intelligence (AI) methodologies and geospatial analysis. This avenue holds the potential to discern complex patterns and enhance our comprehension of the intricate dynamics inherent in malaria distribution and prevention in Uganda [13] [14].

Employing artificial intelligence methods and clustering analysis tools, the goal was to identify recurring features contributing to malaria outbreaks in Uganda. In addressing the impact of malaria on populations, a focus on preventative measures and early treatment is crucial. Anticipating outbreaks enables officials to forewarn at-risk populations, implement additional mosquito-control measures, and allocate resources to local clinics in advance. Through geographic analysis techniques, we pinpointed areas requiring attention, and our predictive model aids in understanding the most impactful malaria preventative measures for Uganda.

Artificial Intelligent Models in Malaria Prediction

Machine learning, a subset of artificial intelligence, demonstrates systems' capacity to learn from past experiences and enhance performance without explicit programming. It involves the exploration and development of algorithms aimed at making data-driven predictions. Machine learning algorithms are broadly categorized into supervised and unsupervised learning. Supervised learning algo-

rithms aim to model relationships between input features and target predictions using labeled datasets. Examples include Support Vector Machine, K-Nearest Neighbor, Naive Bayes, Logistic Regression, and Linear Regression algorithms [15].

Artificial intelligence (AI) learning techniques, spanning from conventional machine learning to deep learning, play a crucial role in predictive modeling of vector-borne diseases. These methods employ diverse numerical, probabilistic, and optimization techniques, enabling computers to analyze complex datasets effectively. Over the past few decades, machine learning and deep learning methods have demonstrated their significance, particularly in the realm of vector-borne illnesses. Their importance extends beyond technological advancements to personalized healthcare, emphasizing their potential in disease prediction and management [16].

Neural network classifiers are computational models consisting of interconnected neurons organized in layers. They process input data through nonlinear functions, adjusting weights during training to optimize performance. These models find widespread use in various fields, including pattern recognition and medical diagnosis. Specifically, in predicting diseases like malaria, neural networks can analyze verbal inputs from patients to provide accurate assessments based on learned patterns. Overall, neural network classifiers offer a versatile and powerful approach to data analysis and pattern recognition, promising solutions to diverse challenges, including disease prediction [17].

Despite global efforts using tools like insecticide-treated nets, indoor spraying, and preventive therapies, Uganda's malaria burden persists. Machine learning proves effective in handling complex datasets, aiding policymakers in early warning system establishment and strengthening prevention measures [13] [14].

While scholars focus on machine learning for predictive studies, its implementation in Uganda and sub-Saharan Africa is in its early stages. Challenges include limited, inaccurate data due to inconsistencies in collection methods. Nonetheless, machine learning addresses malaria challenges, including detection, diagnosis, mosquito identification, outbreak prediction, and transmission forecasting. Machine learning, a subset of artificial intelligence, extracts patterns from complex datasets for future event prediction. Supervised or unsupervised learning trains models for malaria prediction, employing various techniques such as support vector machines, decision trees, random forests, Extreme Gradient Boosting, logistic regression, K-Nearest Neighbors, and Naïve Bayes [17]

This study specifically focuses on Linear regression, K-Nearest Neighbours, Random Forest, and Neural networks.

2. Research Methodology

The adoption of a quantitative research methodology was justified. A non-experimental associational design, in the form of comparative research was used where we compared conditions without manipulating variables. Our focus was

on Malaria incident rates, Mean temperature, Precipitation, Mosquito net access, Mosquito net use rate, Antimalarial treatment, Indoor residual spraying among Uganda districts from 2000 - 2020.

We employed spatial autocorrelation to reveal any underlying spatial patterns and assessed the potential relationships between these variables. This statistical approach allowed us to explore the interactions between our variables and their impact on malaria prevalence.

Generalized Linear Regression Analysis was employed to comprehensively understand the influence of geospatial factors on our target variables. We conducted a generalized linear regression analysis to reveal any underlying spatial patterns and assessed the potential relationships between these variables.

Regression Analysis: We examined the regression coefficients to assess the impact of individual preventive measures, including Antimalarial Treatment, Indoor Spraying, Net Access, and Net Use Rate, on the malaria incidence rate. Employing multiple linear regression with the malaria incidence rate as the dependent variable, the coefficients in the regression output provided insights into both the strength and direction of the associations between each preventive measure and the malaria incidence rate.

In our research, we employed Gregor's theory Type IV: Explanation and Prediction (EP) to comprehensively understand and address the persistent challenge of malaria transmission. The research applied Theory Type IV effectively, merging explanatory and predictive elements to gain a comprehensive understanding of malaria transmission dynamics and inform practical interventions. The primary goal was to understand the factors influencing malaria incidence rates and predict future occurrences to develop effective interventions [8].

The study aimed to provide detailed explanations for the complexities of how, why, when, and where certain phenomena occurred, navigating the intricate dynamics of malaria transmission. Incorporating a predictive element, the research assessed the relationship between climatic factors, preventive measures, and malaria incidents. This predictive aspect was crucial for devising interventions to control and mitigate the impact of malaria.

Adhering to Theory Type IV, the research emphasized having both testable propositions and causal explanations, ensuring the empirical verification of predictions and enhancing the robustness of findings. The aim was to offer valuable insights for targeted interventions. The combination of explanatory and predictive elements aligned with the practical orientation of Theory Type IV, providing a solid foundation for designing effective strategies against malaria transmission.

2.1. Data Sources

We sourced the dataset from two reputable repositories: the Malaria Atlas Database [18] and the World Bank Climate Change Knowledge [19]. The Malaria Atlas Database provided a comprehensive set of malaria-related information,

encompassing incident rates, mosquito net access, mosquito net use rate, antimalarial treatment, and indoor residual spraying. This data was initially presented in a raster format. On the other hand, the World Bank Climate Change Knowledge Portal housed climate data, specifically mean temperature and precipitation, which was conveniently structured in a tabular format.

2.2. Data Preparation

Figure 1 illustrates the zonal statistics summarizing malaria incident rates across different districts in Uganda. Zonal statistics involve calculating statistics based on the cell values of a raster dataset within defined zones set by another dataset. To perform this analysis, we utilized Zonal Statistics as Table tool, which computes one or multiple statistics using predetermined subsets or all statistics and generates a table as output. Since our data was in raster format, conducting zonal statistics was essential to comprehensively understand the variables and facilitate further analysis.

Using ArcGIS Pro, as seen in Figure 2, we converted Malaria incidents and preventive measures from raster images into a structured dataset for Ugandan districts. Zonal Statistics played a key role in this process. For climate data (Mean temperature and Precipitation), Python was employed to ensure data cleanliness. Transitioning to Python for machine learning, we focused on discerning patterns within specific variables across Ugandan districts from 2000 to 2020, including incident rates, Mean temperature, Precipitation, Mosquito net access, Mosquito net use rate, Antimalarial treatment, and Indoor residual spraying.

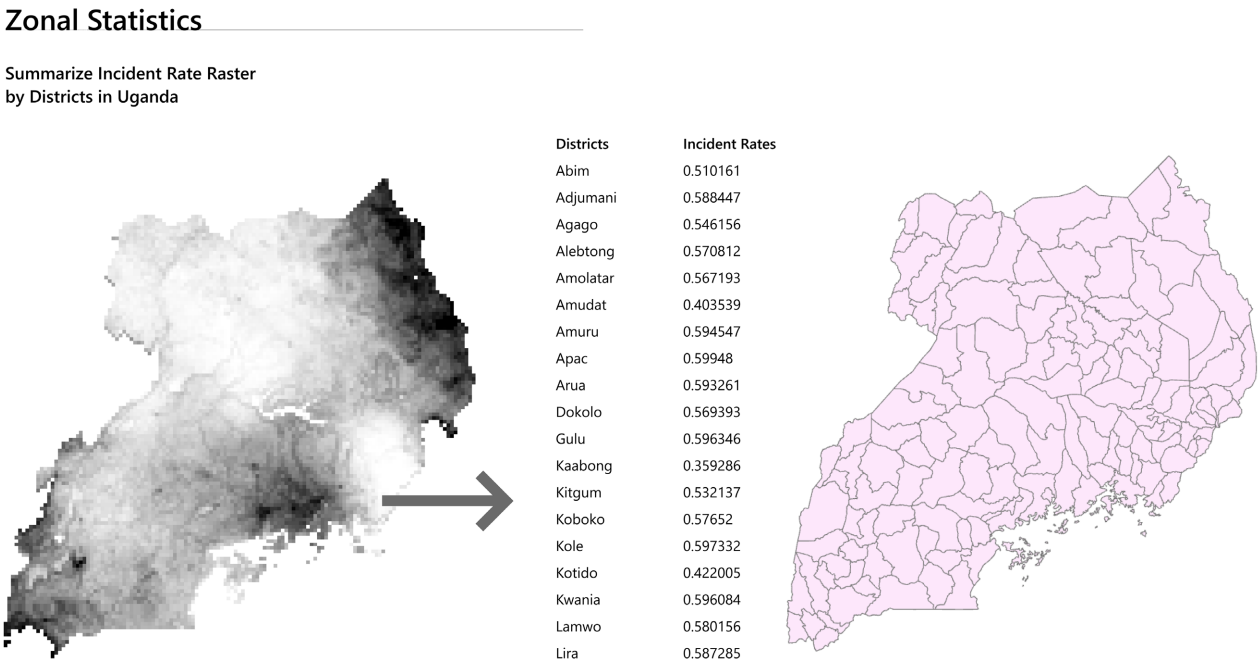


Figure 1. Deriving tabular data from raster images.

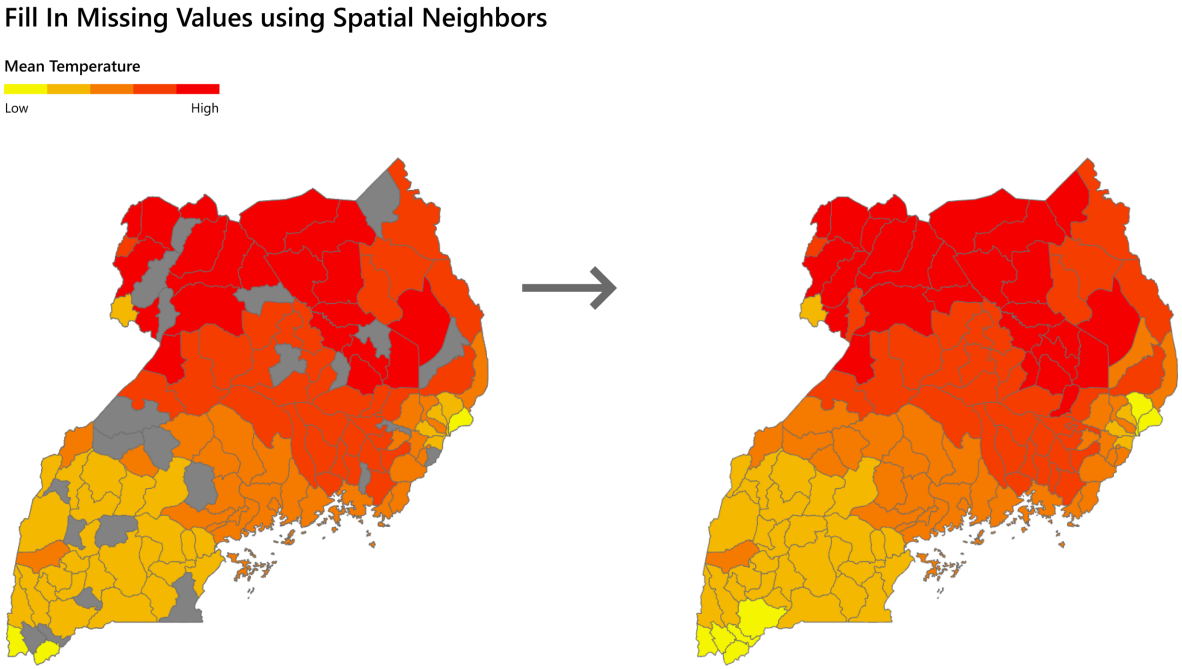


Figure 2. Filling in missing values using spatial neighbors.

An issue arose with the climate data for 23 districts in Uganda, where missing values spanned all 20 years. Typically, missing values could be filled using the mean average of the district; however, this approach was impractical due to the complete absence of data for these districts. Using the national average was not viable due to significant temperature and precipitation variations between the southern and northern regions of Uganda. To address this, ArcGIS Pro offered a tool that utilized spatial neighbors to identify a local average, proving more representative than the national average. Subsequently, the data underwent scaling to accommodate substantial differences in variable values.

The final data was a combined dataset of Precipitation, Meantemperature, Antimalarial treatment, Indoor spraying, Net access, Net use rate, and Incident Rateover 20 years for each Districtin Uganda.

2.3. Part of Training Data

Table 1 displays a segment of the dataset utilized for model training. This dataset comprises the following variables: District name, year, precipitation, Mean temperature, Anti-malarial treatment, Indoor spraying, net access, Net use rate, and incident rate from 2002 to 2020.

3. Results

(Q1) Which ML and AI model provides the most reliable predictions for malaria incident rates in Uganda?

Predictive Modeling Results

Table 2 displays the results of predictive modeling. Among the models tested,

Random Forest, K-Nearest Neighbor, Linear Regression, and Neural Network performed well in that order, as indicated by their respective R-squared scores. Therefore, Random Forest emerges as the most reliable model for predicting malaria incident rates in Uganda, followed by K-Nearest Neighbor, Linear Regression, and Neural Network.

Ordinary Least Squares (OLS) regression analysis Results

Table 3 shows the Ordinary Least Squares (OLS) regression analysis results indicate that the model provides a moderate level of explanation for the variability in the dependent variable, Incident_Rate, with an R-squared value of 0.606. This suggests that approximately 60.6% of the variance in the incident rate can be explained by the independent variables (precipitation, Mean temperature, Anti-malarial treatment, Indoor spraying, net access, Net use rate) included in the model. The F-statistic of 621.5, along with a significant probability value of 0.00, indicates that the overall regression model is statistically significant, implying that at least one of the independent variables has a non-zero effect on the incident rate. The adjusted R-squared value of 0.605 suggests that the model's explanatory power remains consistent when adjusting for the number of predictors. Additionally, the negative AIC (−5168) and BIC (−5120) values suggest that the model is performing well relative to alternative models, indicating a good fit. The positive Log-Likelihood value further supports the adequacy of the model in describing the relationship between the dependent and independent variables.

Table 1. The combined dataset used for training predictive models.

Name	Year	Precipitation	Mean Temperature	Antimalarial Treatment	Indoor Spraying	Net Access	Net Use Rate	Incident Rate
Abim	2000	908.2	24.24	0.35606	0	0.04087	0.718159	0.510161
Abim	2001	882.06	24.44	0.439205	0	0.03045	0.718913	0.523735
Abim	2002	839.06	24.81	0.446969	0	0.023057	0.696928	0.5085
Abim	2003	1018.8	24.64	0.477688	0	0.023119	0.702978	0.497679
Abim	2004	1009.72	24.67	0.527388	0	0.032175	0.746245	0.476498
Abim	2005	862.78	24.68	0.537249	0	0.041348	0.797703	0.464191
Abim	2006	1006.57	24.64	0.516739	0	0.059345	0.849268	0.468501
Abim	2007	1311.59	24.74	0.542825	0	0.145175	0.892662	0.448244
Abim	2008	1090.55	24.7	0.568606	0	0.30419	0.921069	0.417777

Table 2. Predictive modeling results.

Model	Mean Squared Error (MSE):	R-squared (R^2) Score:
Linear Regression	0.0065	0.73
Random Forest	0.0534	0.88
K-Nearest Neighbor	0.4261	0.7956
Neural network	**0.0816**	**0.723**

Table 3. Ordinary least squares results.

Dependent Variable:	Incident_Rate
Method:	Least Squares
R-squared:	0.606
Adj. R-squared:	0.605
F-statistic:	621.5
Prob (F-statistic):	0.00
Log-Likelihood:	2591.8
AIC:	−5168
BIC:	−5120

Table 4 presents the Ordinary Least Squares (OLS) regression results for the independent variables. Each coefficient represents the estimated effect of the corresponding independent variable on the dependent variable, Incident_Rate. The constant term, with a coefficient of 22.5588 and a t-value of 12.735, signifies the expected value of the dependent variable when all other independent variables are held constant at zero. The negative coefficient for the "Year" variable (−0.0115) suggests a decreasing trend in the incident rate over time, supported by its significant t-value (−12.894). The coefficients for "Precipitation", "Mean_Temp", "Indoor_Spraying", "Net_Access", and "Net_Use_Rate" are positive, indicating that an increase in these variables is associated with higher incident rates, while their respective significant t-values and p-values (<0.05) confirm their statistical significance. However, the coefficient for "Antimalarial_Treatment" is not statistically significant (p-value > 0.05), suggesting that this variable does not have a significant impact on the incident rate when controlling for other variables. Overall, these results provide insights into the relationships between the independent variables and the incident rate of malaria in Uganda.

3.1. Geospatial Analysis Results

Generalized Linear Regression: Model Type: Continuous (Gaussian/OLS)

Table 5 summarizes the results of the Generalized Linear Regression model, which is a continuous model type (Gaussian/OLS). Each coefficient in the table represents the estimated effect of the corresponding independent variable on the dependent variable. The "Precipitation" and "Mean_Temp" variables show positive coefficients of 0.0001 and 0.044 respectively, indicating that an increase in precipitation and mean temperature is associated with higher incident rates of malaria. The "Antimalarial_Treatment" variable has a negative coefficient of −0.4978, suggesting that areas with higher levels of antimalarial treatment have lower incident rates of malaria. Similarly, "Indoor_Spraying," "Net_Access," and "Net_Use_Rate" also exhibit negative coefficients, indicating that increased indoor spraying, net access, and net use rates are associated with lower incident rates of malaria. The intercept term, with a coefficient of −0.2783, represents the

Table 4. OLS Results for independent variables.

	Coefficients	Std Error	t-value	p-Value
Constant	22.5588	1.771	12.735	0.000
Year	−0.0115	0.001	−12.894	0.000
Precipitation	7.548e−05	7.69e−06	9.809	0.000
Mean_Temp	0.0420	0.001	37.246	0.000
Antimalarial_Treament	0.0209	0.061	0.340	0.734
Indoor_Spraying	−0.0519	0.016	−3.278	0.001
Net_Access	−0.0704	0.013	−5.471	0.000
Net_Use_Rate	−0.0914	0.019	−4.850	0.000

Table 5. Summary of generalized linear regression.

Variable	Coefficients	Std Error	t-value	Probability
Precipitation	0.0001	0	7.1761	0
Mean_Temp	0.044	0.0011	38.3707	0
Antimalarial_Treatment	−0.4978	0.0478	−10.4238	0
Indoor_Spraying	−0.0632	0.0163	−3.8854	0.0001
Net_Access	−0.1783	0.0101	−17.732	0
Net_Use_Rate	−0.1105	0.0193	−5.7155	0
Intercept	−0.2783	0.0333	−8.3567	0

expected incident rate when all independent variables are zero. Each coefficient's standard error, t-value, and probability are provided to assess the coefficient's significance. All variables, except for precipitation, have significant t-values ($p < 0.05$), indicating their statistical significance in predicting the incident rate of malaria in Uganda.

3.2. GLR Diagnostics

Table 6 presents the results of diagnostics for the Generalized Linear Regression (GLR) model. The multiple R-squared value of 0.583 indicates that approximately 58.3% of the variance in the dependent variable (incident rate of malaria) can be explained by the independent variables (precipitation, Mean temperature, Anti-malarial treatment, Indoor spraying, net access, Net use rate) included in the model. The adjusted R-squared value, which considers the number of predictors in the model, is slightly lower at 0.5821 but still reflects a good fit of the model to the data. Akaike's Information Criterion (AIC) is a measure of the relative quality of a statistical model, with lower values indicating a better fit. The negative AIC value of −5005.5618 suggests that the GLR model is performing well relative to alternative models, further indicating its adequacy in describing the relationship between the independent variables and the incident rate of malaria in Uganda. Overall, these diagnostics support the validity and reliability of the GLR model in predicting malaria incidence rates.

Table 6. Results for GLR diagnostics.

Property	Value
Multiple R-Squared	0.583
Adjusted R-Squared	0.5821
Akaike's Information Criterion (AIC)	−5005.5618

3.3. Spatial Autocorrelation

The results of the Global Moran's I test, as summarized in Table 7, indicate significant spatial autocorrelation in the dataset. The Moran's Index value of 0.319029 suggests a positive spatial autocorrelation, indicating that neighboring areas tend to have similar values for the incident rate of malaria. The expected index value of −0.000353 indicates what the Moran's Index would be if the spatial distribution were completely random. The high z-score of 53.954655 suggests that the observed Moran's Index is significantly higher than what would be expected under spatial randomness, indicating a highly clustered pattern in the data. The p-value of 0.000000 further confirms the significance of the spatial autocorrelation, indicating that there is less than a 1% likelihood that this clustered pattern could be the result of random chance. In summary, these results suggest that there is a significant spatial clustering of malaria incident rates in the dataset, with neighboring districts exhibiting similar levels of malaria incidence.

(Q2) What is the impact of different independent variables on malaria incident rates?

Spatial Exploration of Independent Variables for the year 2020

Figure 3 displays the distribution of precipitation and mean temperature across districts in Uganda in the year 2020. Darker areas on the map represent regions with higher levels of precipitation and temperatures, while lighter areas indicate lower levels. This visualization allows for an understanding of the spatial variation in climate conditions within Ugandan districts, highlighting areas with potentially higher rainfall and temperatures compared to others.

Figure 4 illustrates the distribution of malaria preventive measures across Uganda in the year 2020. Darker regions on the map indicate higher utilization rates of preventive measures such as anti-malarial treatment, indoor spraying, net access, and net use. Conversely, lighter, or lower intensity areas represent districts with reduced access or utilization of these malaria preventive measures. This visualization provides insights into the spatial distribution of malaria prevention efforts throughout Ugandan districts, highlighting areas where preventive measures are more extensively adopted and those where there may be room for improvement.

Figure 5 depicts Ugandan districts identified as malaria incidence hotspots with a confidence level exceeding 90%. These districts are marked in darker red, indicating significantly higher rates of malaria incidents compared to those depicted in lighter colors. This map serves as a tool to identify areas requiring immediate attention and intervention efforts to address the heightened malaria burden within those regions.

Table 7. Global Moran's I summary.

Moran's Index	0.319029
Expected Index	−0.000353
Variance	0.000035
z-score	53.954655
p-value	0.000000

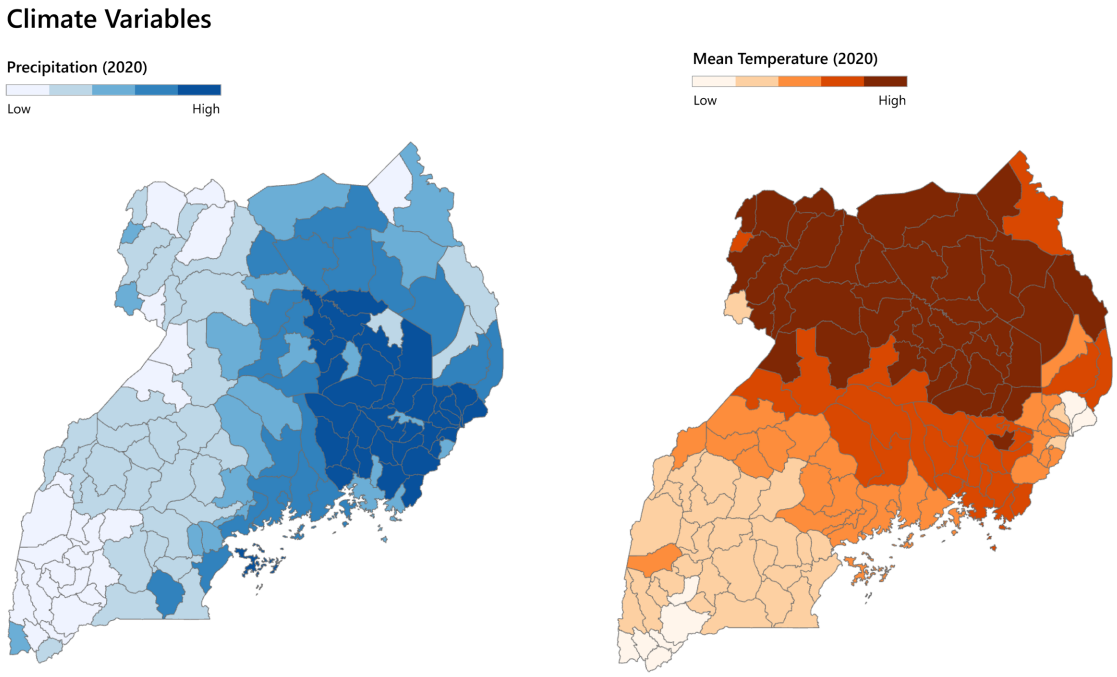
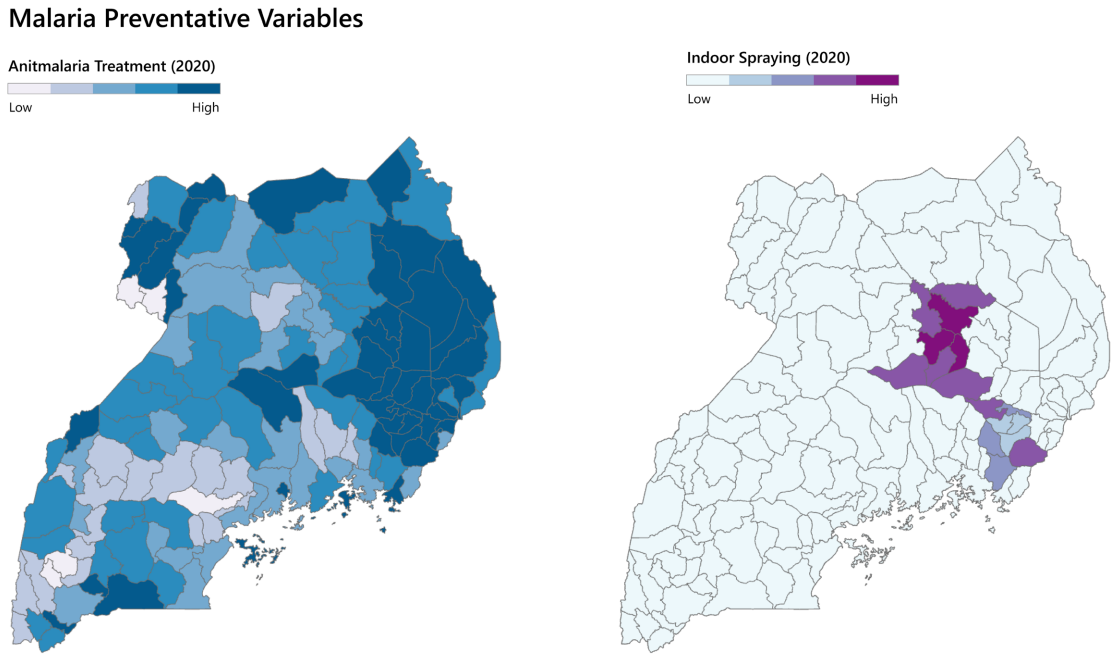


Figure 3. Distribution of precipitation and mean temperature for districts in Uganda 2020.



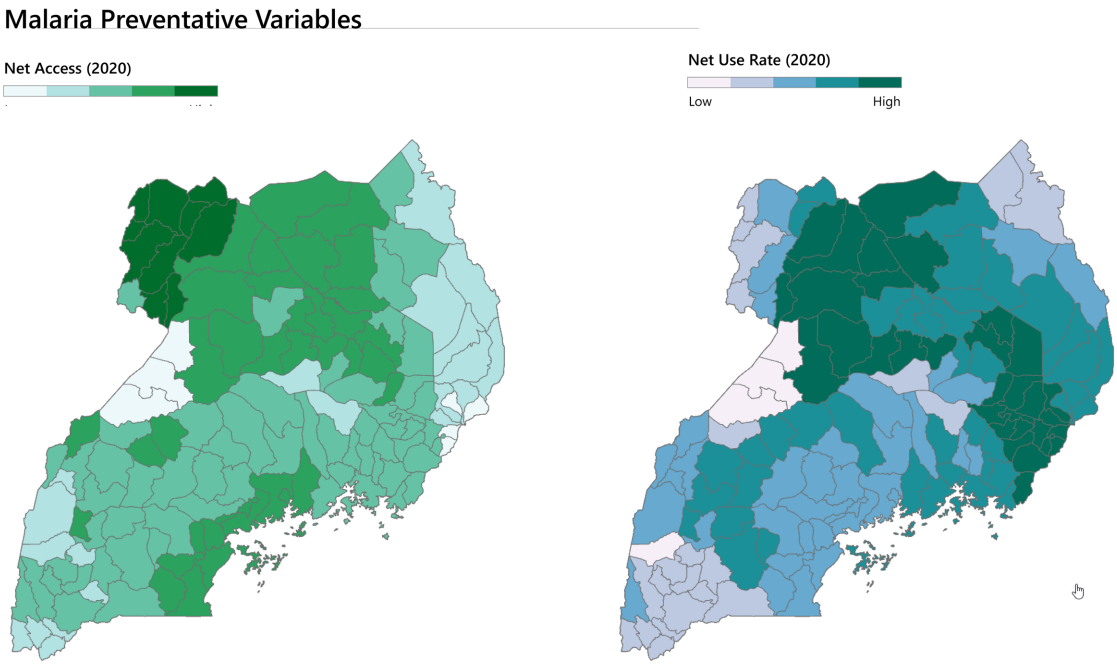


Figure 4. Malaria preventive Measure distribution across Uganda for the year 2020.

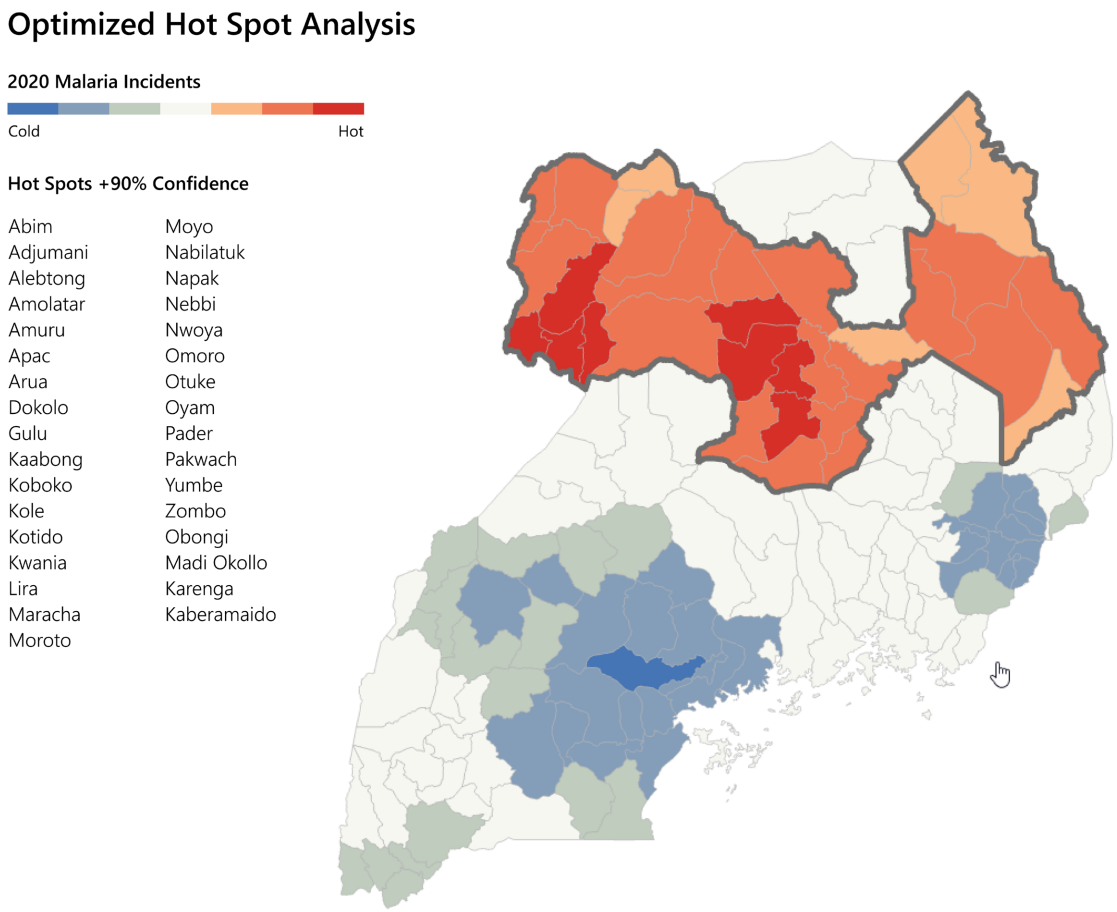


Figure 5. Ugandan districts with malaria incidence hot spots (+90% Confidence).

3.4. Change in Malaria Incidence Rates

Figure 6 illustrates the change rate for districts where malaria incidents have increased from 2016 to 2020. The first map represents incident rates per 1000 population in the year 2020, with darker blue color indicating higher rates and lighter yellow color indicating lower rates. The second map features arrows indicating the change in rates from 2016 to 2020. Arrows pointing upwards signify an increase of 80%, while arrows pointing downwards indicate a decrease of negative 45%. This visualization provides insights into the spatial distribution and magnitude of changes in malaria incident rates over the specified period, aiding in the identification of areas requiring targeted interventions to address the rising burden of malaria.

3.5. Model Validation

In this study, the predictive modeling of malaria incident rates in Uganda was conducted using four machine learning models: Linear Regression, Random Forest, K-Nearest Neighbors (KNN), and Neural Network. Rigorous model validation included data preprocessing steps such as cleaning, feature selection, data splitting into training and testing sets, and normalization/standardization. Each model underwent training with the processed dataset, with Linear Regression serving as a baseline and other models optimized through hyperparameter tuning. Evaluation metrics, included Root Mean Squared Error (RMSE) and R-squared (R^2), were employed to assess model performance. The models were compared, and Random Forest emerged as particularly robust in meeting the study's objectives, despite Neural Network showing promise. This comprehensive validation methodology ensures the reliability and actionability of insights gained from predicting malaria incidence patterns in Uganda.

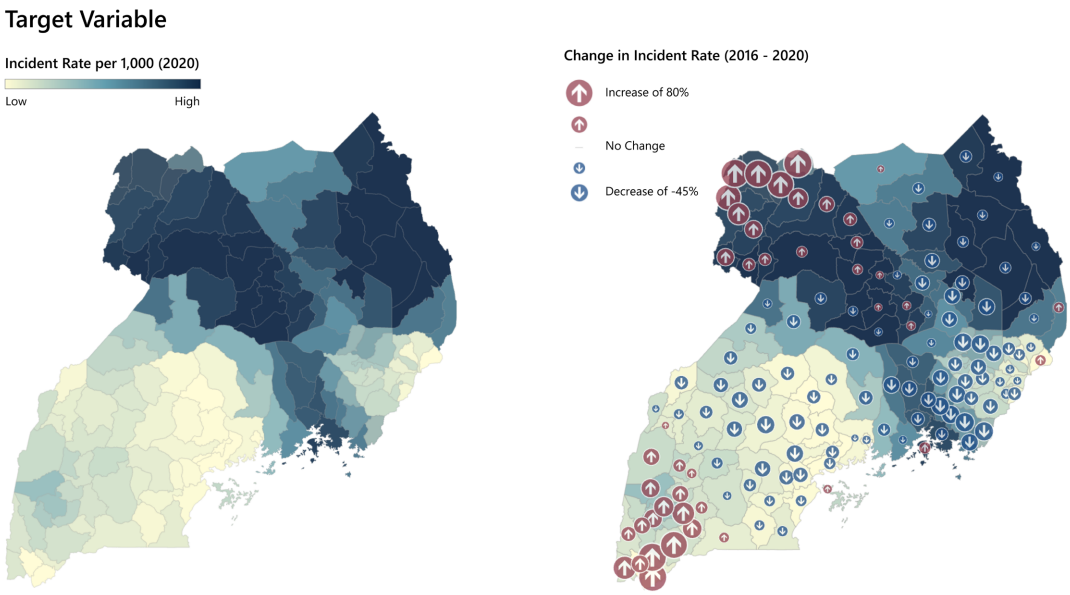


Figure 6. Change rate for districts where malaria incidents have increased from 2016 to 2020.

4. Discussion and Areas for Further Study

The reproductive cycle of Anopheles mosquitoes and the prevalence of Plasmodium parasites are significantly influenced by climate-related environmental conditions, impacting the life cycle of malaria vectors. Temperature and rainfall are pivotal factors in this process, with alterations in climate directly affecting malaria incidence. Additionally, socio-economic factors, encompassing shifts in land use, public health interventions, housing changes, human migration, and community practices, are substantial contributors to malaria spread [20] [21]. Our findings highlight the critical role of mean temperature in increasing incident rates, while rainfall's impact is nuanced. Increased rainfall does not necessarily correlate with higher incident rates, as it may wash away mosquito breeding areas. Conversely, sparse rainfall can elevate incidence rates by creating stagnant water, fostering mosquito breeding. Emphasizing the importance of preventive measures, community awareness, and usage rates is crucial for malaria control. The intricate interplay between environmental and socio-economic factors underscores the complex dynamics of malaria.

4.1. Interpretation of an Ordinary Least Squares (OLS) Regression Analysis

The model focuses on predicting Malaria incident rates, with an R-squared value of 0.606, indicating that approximately 60.6% of the variability in incident rates is explained by the model. The Adjusted R-squared accounts for the number of predictors, with a value of 0.605. The high F-statistic of 621.5, coupled with a low probability (Prob (F-statistic)) close to zero, emphasizes the overall statistical significance of the model. Coefficients for predictors such as "Year," "Precipitation," "Mean_Temp," "Indoor_Spraying," "Net_Access," and "Net_Use_Rate" are statistically significant, as indicated by low p-values. Diagnostic tests for residuals provide insights into normality, autocorrelation, and heteroscedasticity. The large Condition Number of 2.34e+06 suggests potential multicollinearity, which should be considered. In summary, the model is statistically significant, and the coefficients offer valuable information about the relationships between independent variables and incident rates.

4.2. Interpretation of Predictive Models

Utilizing linear models with stepwise regression, it was noted that an increase in daily precipitation and mean temperature significantly augments the probability of exposure to malaria [22]. Our study employed various models, including Linear Regression, Random Forest, K-Nearest Neighbor, and Neural Network. The Linear Regression Model, with an R^2 score of 0.73, demonstrated a commendable ability to explain a substantial portion of the variance in malaria incident rates. Further exploration revealed the Random Forest model outperforming Linear Regression, boasting a higher R^2 score and lower Mean Squared Error (MSE), signifying its efficacy in capturing the complex relationships within the

data. The K-Nearest Neighbor Model exhibited an R^2 of approximately 0.80, indicating a noteworthy level of predictive accuracy. Lastly, the Neural Network, with an R^2 of 0.723, demonstrated a strong predictive power. Conclusively, the Random Forest model, with an impressive R^2 of approximately 0.88, emerged as the optimal fit, elucidating about 88% of the variance in malaria incident rates in our study. Therefore, the recommendation would be to prioritize and utilize the Random Forest model for predicting malaria incident rates in Uganda. Its ability to capture the complex relationships within the data and explain about 88% of the variance in malaria incident rates suggests that it is a robust and reliable model for this specific context.

4.3. Spatial Analysis Discussion

Utilizing optimized hot spot analysis, we identified potential "hot spots" for malaria cases. Despite a significant rise in mosquito net utilization, particularly insecticide-treated nets (ITNs) and long-lasting insecticidal nets (LLINs), these measures alone may prove inadequate. The implementation of Indoor Residual Spraying (IRS) is crucial to achieving Uganda's malaria goals [23] [24]. Our results indicate that the most affected areas with hot spots are not employing IRS, which could help reduce incident rates. The outcome variables in our analysis highlighted districts with high malaria rates as of 2020, necessitating considerable attention, as depicted in **Figure 5**. Enhancing these control measures with education on the appropriate and consistent use of ITNs and LLINs, coupled with promoting safe living habits like minimizing outdoor activities during peak mosquito-biting hours, can significantly contribute to reducing the malaria burden in Uganda.

4.4. Generalized Linear Regression (GLR) Results

The Generalized Linear Regression (GLR) results, focusing on a Continuous (Gaussian/OLS) model, provided insights into the impact of various independent variables on the dependent variable (Incident Rate) as seen in **Table 8** above. Coefficients in the GLR analysis represented the change in the dependent variable for a one-unit change in the corresponding independent variable, with the magnitude indicating the strength of this impact. Notably, "Antimalarial_Treatment" emerged as the most influential factor, exhibiting a significant negative impact on incident rates; a one-unit increase in antimalarial treatment correlated with a noteworthy decrease in incident rates. "Net_Access" also showed a substantial negative impact, implying that increased access to mosquito nets was associated with a significant reduction in incident rates. Furthermore, "Mean_Temp" had a considerable impact, suggesting that higher temperatures were linked to an increase in incident rates. These findings underscored the importance of antimalarial treatment and mosquito net access as influential factors in mitigating malaria incidents, while also highlighting the role of temperature in affecting incident rates.

Table 8. GLR Coefficient and Interpretation for variable.

Variable	Coefficient	Interpretation
Precipitation	0.0001	For a one-unit increase in precipitation, the incident rate is expected to increase by 0.0001 units.
Mean_Temperature	0.044	For a one-unit increase in mean temperature, the incident rate is expected to increase by 0.044 units
Antimalarial_Treatment	−0.4978	For a one-unit increase in antimalarial treatment, the incident rate is expected to decrease by 0.4978 units
Indoor_Spraying	−0.0632	For a one-unit increase in indoor spraying, the incident rate is expected to decrease by 0.0632 units.
Net_Access	−0.1783	For a one-unit increase in net access, the incident rate is expected to decrease by 0.1783 units.
Net_Use_Rate	−0.1105	For a one-unit increase in net use rate, incident rate is expected to decrease by 0.1105 units.
Intercept	−0.2783	The intercept represents the expected value of the incident rate when all independent variables are zero.

Given the significant impact of antimalarial treatment and mosquito net access in reducing incident rates, it is recommended that public health interventions prioritize and promote the accessibility and proper utilization of antimalarial treatments and mosquito nets. Additionally, awareness campaigns should emphasize the importance of consistent and appropriate use of these preventive measures. Temperature management strategies, especially during periods of heightened risk, should also be considered to further address and mitigate the impact of malaria incidents.

5. Limitations

The limitation lies in the scarcity of monthly malaria and climate data for Uganda, posing a challenge to the comprehensive establishment of a strong connection between climate variability and malaria transmission. Initial data transformation involved zonal statistics from raster data as seen in Figure 1, converting them into summarized mean data for each variable.

After merging the tables, 23 climate variables had missing data. This was addressed by incorporating spatial neighbors' averages as seen in Figure 2. Despite efforts, the final dataset spans 20 years, with inherent limitations. Transforming raster to structured data incurs data loss, and filling missing climate data with spatial neighbors' averages, while superior to a national average, remains an approximation.

The primary challenge in data validation lies in limited and inaccurate data, attributed to collection methods. Much of the data, often incomplete, was rasterized using methods like interpolation, impacting the overall coverage accuracy. These limitations necessitate careful consideration during the model-building process and subsequent use of these models.

6. Conclusion

Our study utilized various models to predict malaria incident rates in Uganda, with the Random Forest model proving most effective. We recommended prioritizing this model for future predictions and proposed implementing Indoor Residual Spraying (IRS) in areas with malaria hotspots lacking coverage. Additionally, districts with high malaria rates in 2020 required immediate attention. Emphasizing control measures like education on Insecticide-Treated Net (ITN) and Long-Lasting Insecticidal Net (LLIN) use is crucial for reducing malaria in Uganda. Our research significantly advanced understanding of malaria dynamics, identifying hotspots and effective prevention measures. We highlighted the importance of considering environmental variables like temperature and integrating geospatial analysis techniques for effective disease control. Overall, our study aimed to contribute to evidence-based decision-making in malaria control strategies, aiming to reduce transmission rates and save lives in Uganda and beyond.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] World Health Organization (WHO) (2024) World Malaria Report 2022. https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2022

[2] Bagcchi, S. (2023) Locally Acquired Malaria Cases in the USA. *The Lancet Infectious Diseases*, **23**, e401. https://doi.org/10.1016/S1473-3099(23)00581-9

[3] Nabet, C., Chaline, A., Franetich, J.F., Brossas, J.Y., Shahmirian, N., Silvie, O., Tannier, X. and Piarroux, R. (2020) Prediction of Malaria Transmission Drivers in Anopheles Mosquitoes Using Artificial Intelligence Coupled to MALDI-TOF Mass Spectrometry. *Scientific Reports*, **10**, Article No. 11379. https://doi.org/10.1038/s41598-020-68272-z

[4] World Health Organization (WHO) (2023) Fact Sheet about Malaria. https://www.who.int/news-room/fact-sheets/detail/malaria

[5] Kigozi, R.N., Bwanika, J., Goodwin, E., Thomas, P., Bukoma, P., Nabyonga, P., Isabirye, F., Oboth, P., Kyozira, C., Niang, M., Belay, K., Sebikaari, G., Tibenderana, J.K. and Gudoi, S.S. (2021) Determinants of Malaria Testing at Health Facilities: The Case of Uganda. *Malaria Journal*, **20**, Article No. 456. https://doi.org/10.1186/s12936-021-03992-9

[6] Murad, A. and Khashoggi, B.F. (2020) Using GIS for Disease Mapping and Clustering in Jeddah, Saudi Arabia. *ISPRS International Journal of Geo-Information*, **9**, Article 328. https://doi.org/10.3390/ijgi9050328

[7] Esri (2023) Seeking to Eradicate Malaria, Zambia Turns to Geospatial Data. https://www.esri.com/about/newsroom/arcnews/seeking-to-eradicate-malaria-zambia-turns-to-geospatial-data/

[8] Gregor, S. (2006) The Nature of Theory in Information Systems. *MIS Quarterly*, **30**, 611-642. https://doi.org/10.2307/25148742

[9] Rogers, D.J. and Randolph, S.E. (2003) Studying the Global Distribution of Infectious Diseases Using GIS and RS. *Nature Reviews Microbiology*, **1**, 231-237. https://doi.org/10.1038/nrmicro776

[10] Gebre, S.L., Temam, N. and Regassa, A. (2020) Spatial Analysis and Mapping of Malaria Risk Areas Using Multi-Criteria Decision Making in Didessa District, South West Ethiopia. *Cogent Environmental Science*, **6**, Article ID: 1860451. https://doi.org/10.1080/23311843.2020.1860451

[11] Kalipe, G., Gautham, V. and Behera, R.K. (2018) Predicting Malarial Outbreak Using Machine Learning and Deep Learning Approach: A Review and Analysis. 2018 *International Conference on Information Technology* (*ICIT*), Bhubaneswar, 19-21 December 2018, 33-38. https://doi.org/10.1109/ICIT.2018.00019

[12] Wiru, K., Oppong, F.B., Gyaase, S., Agyei, O., Abubakari, S.W., Amenga-Etego, S., Zandoh, C. and Asante, K.P. (2021) Geospatial Analysis of Malaria Mortality in the Kintampo Health and Demographic Surveillance Area of Central Ghana. *Annals of GIS*, **27**, 139-149. https://doi.org/10.1080/19475683.2020.1853231

[13] Mbunge, E., Milham, R.C., Sibiya, M.N. and Takavarasha, S. (2023) Machine Learning Techniques for Predicting Malaria: Unpacking Emerging Challenges and Opportunities for Tackling Malaria in Sub-saharan Africa. In: Silhavy, R. and Silhavy, P., Eds., *Artificial Intelligence Application in Networks and Systems*, Springer, Cham, 327-344. https://doi.org/10.1007/978-3-031-35314-7_30

[14] Yeka, A., Gasasira, A., Mpimbaza, A., Achan, J., Nankabirwa, J., Nsobya, S., Staedke, S.G., Donnelly, M.J., Wabwire-Mangen, F., Talisuna, A., Dorsey, G., Kamya, M.R. and Rosenthal, P.J. (2012) Malaria in Uganda: Challenges to Control on the Long Road to Elimination. I. Epidemiology and Current Control Efforts. *Acta Tropica*, **121**, 184-195. https://doi.org/10.1016/j.actatropica.2011.03.004

[15] Kazeem, I. and Adebanji, S. (2021) A Model for Predicting Malaria Outbreak Using Machine Learning Technique. *Scientific Annals of Computer Science*, **19**, 9-15.

[16] Kaur, I., Sandhu, A.K. and Kumar, Y. (2022) Artificial Intelligence Techniques for Predictive Modeling of Vector-Borne Diseases and Its Pathogens: A Systematic Review. *Archives of Computational Methods in Engineering*, **29**, 3741-3771. https://doi.org/10.1007/s11831-022-09724-9

[17] Parveen, R., Jalbani, A.H., Shaikh, M., Memon, K.H., Siraj, S., Nabi, M. and Lakho, S. (2017) Prediction of Malaria Using Artificial Neural Network. *International Journal of Computer Science and Network Security*, **17**, 79-86.

[18] Bbosa, F.F., Nabukenya, J., Nabende, P. and Wesonga, R. (2021) On the Goodness of Fit of Parametric and Non-Parametric Data Mining Techniques: The Case of Malaria Incidence Thresholds in Uganda. *Health and Technology*, **11**, 929-940. https://doi.org/10.1007/s12553-021-00551-9

[19] (2023) Malaria Atlas Project. https://malariaatlas.org/

[20] (2023) World Bank Climate Change Knowledge Portal. https://climateknowledgeportal.worldbank.org/

[21] Tiu, L.A., Wahid, W.E., Andriani, W.Y., Mirnawati, and Tosepu, R. (2021) Literature Review: Impact of Temperature and Rainfall on Incident Malaria. *IOP Conference Series*: *Earth and Environmental Science*, **755**, Article ID: 012084. https://doi.org/10.1088/1755-1315/755/1/012084

[22] Wang, C., Thakuri, B., Roy, A.K., Mondal, N., Qi, Y. and Chakraborty, A. (2023) Changes in the Associations between Malaria Incidence and Climatic Factors across Malaria Endemic Countries in Africa and Asia-Pacific Region. *Journal of Environmental Management*, **331**, Article ID: 117264.

https://doi.org/10.1016/j.jenvman.2023.117264

[23] Makinde, O.S. and Abiodun, G.J. (2020) The Impact of Rainfall and Temperature on Malaria Dynamics in the KwaZulu-Natal Province, South Africa. *Communications in Statistics*: *Case Studies*, *Data Analysis and Applications*, **6**, 97-108. https://doi.org/10.1080/23737484.2019.1699000

[24] Roberts, D. and Matthews, G. (2016) Risk Factors of Malaria in Children under the Age of Five Years Old in Uganda. *Malaria Journal*, **15**, Article No. 246. https://doi.org/10.1186/s12936-016-1290-x