Journal of Health Informatics in Africa

Predicting Malaria Outbreaks in Children Under Five: Insights from Ghana

Maxwell Akwasi Boateng [a,*], Ernest Kwame Agyapong [b], Treasure Kekeli Dorson [b]

Richard Kodzo Avuglah [b], Jessica Sey Nigre [b]

[a]Department of Mathematics, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
[b]Department of Statistics and Actuarial Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

Background and Purpose: Malaria, a persistent public health challenge in sub-Saharan Africa, disproportionately affects children under five, with socio-economic disparities and environmental factors exacerbating its burden.
Methods: This study employs the Random Forest algorithm to predict malaria outbreaks using historical incidence data from Ghana (2011–2021). By incorporating lagged variables and analysing spatial and temporal dynamics, the model captures seasonality, including peaks during rainy seasons, and regional disparities in malaria cases.
Results: The model achieved a robust R-squared value of 0.8193, reflecting strong predictive accuracy, with additional performance metrics including RMSE of 150.54, MAE of 96.46, and a correlation coefficient of 0.91 between predicted and actual values. Regional hotspots such as Ashanti and Western exhibited higher case numbers than Greater Accra, emphasizing the need for localized interventions.
Conclusions: This study demonstrates the potential of Random Forest as a scalable tool for malaria prediction in resource-constrained settings, enabling data-driven decision-making and optimizing public health resources. Key limitations include potential biases from underreporting in rural areas, variations in healthcare-seeking behaviours affecting data quality, and uncertain generalizability beyond Ghana's ecological context. These findings align with global malaria reduction goals, emphasizing the integration of machine learning into public health strategies to reduce morbidity and mortality in vulnerable populations.

Keywords: Machine learning, random forest, decision-making, malaria, public health, morbidity.

1 Introduction

Malaria, an infectious disease transmitted to humans through the bite of female Anopheles mosquitoes, remains a significant global health issue and an obstacle to socio-economic development, particularly in sub-Saharan countries. Children under the age of five are the most vulnerable to malaria and its complications due to their developing immune systems [1]. In Ghana, malaria is one of the leading causes of disease, especially in children under five years of age, imposing a substantial burden on families and the public health system [2]. Despite various social and health interventions such as the free distribution of treated mosquito nets, free malaria treatment under the National Health Insurance Scheme (NHIS) and public education on malaria, its causes and effects, the prevalence of malaria continues to persist in the 16 regions of the country, claiming thousands of lives each year. Malaria disproportionately affects children under five, contributing to over 80% of malaria deaths in Sub-Saharan Africa [3].

Accurate prediction of malaria outbreaks, particularly in children under five, is crucial for optimizing public health resources. By allowing for the timely distribution of interventions such as medications and preventive tools, health officials can direct supplies to high-risk areas well in advance [4]. Although many predictive models integrate climate, socioeconomic, and environmental data, such data is often complex to obtain in rural and resource-limited regions like Ghana. Consequently, relying on historical malaria

*Corresponding author address: Email: boateng.ma@knust.edu.gh
 J Health Inform Afr. 2025;12(1):74-86. DOI: 10.12856/JHIA-2025-v12-i1-550

incidence data from national health surveillance records offers a practical and effective alternative to predict malaria trends in such settings [5].

This study's significance is underscored by its alignment with Ghana's National Malaria Control Strategic plan (2021 – 2025), which emphasizes technological innovation and evidence-based decision making as essential components for reducing malaria burden [6]. The research is particularly timely as health systems face resource constraints due to recent global health challenges, making the efficient allocation of malaria control resources increasingly critical [7]. The recent improvements in Ghana's health information systems provide an unprecedented opportunity to leverage historical data for predictive modelling to inform targeted interventions.

Despite advances in malaria predictions, significant research gaps persist. Previous models have often failed to account for the unique epidemiological patterns of malaria in children under five, who exhibit different clinical manifestations compared to adults [8]. Additionally, existing models frequently overlook local healthcare-seeking behaviors and socioeconomic factors specific to Ghana that significantly influence malaria transmission patterns [9][10]. Many sophisticated prediction systems require extensive data infrastructure and continuous monitoring of multiple parameters, making them impractical for resource-constrained settings like rural Ghana [11]. This study addresses these limitations by developing a prediction model tailored to the Ghanaian context, focusing on children under five, and utilizing readily available historical health surveillance data.

Although this project is focused on the Ghanaian context, its implications extend beyond national borders to other malaria-endemic regions worldwide that face similar challenges. Sub-Saharan Africa, for example, accounts for more than 90% of global malaria cases and deaths, with socioeconomic disparities, inadequate healthcare infrastructure, and environmental factors that worsen the burden of the disease [12]. Predictive models, tailored to specific local contexts but informed by global best practices, are critical tools for malaria control. These models help optimize resource allocation and guide interventions, offering indispensable support in malaria eradication efforts in resource-constrained settings [4][13]. The success of such models is contingent on integrating innovative technologies and methodologies. Advances in artificial intelligence, machine learning, and geographic information systems (GIS) have demonstrated significant potential to improve the accuracy and scalability of malaria prediction models. These technologies enable synthesizing diverse data types, climatic variables, demographic trends, and real-time health surveillance data, thus improving the precision of outbreak predictions [14][15]. Furthermore, incorporating insights from global initiatives such as the Malaria Atlas Project, which has successfully mapped malaria incidence worldwide, can improve the model's ability to address the unique challenges of resource-limited settings [16][17].

This project aligns with the World Health Organization's Global Technical Strategy for Malaria 2016 to 2030, emphasizing data-driven decision-making, community engagement, and resource optimization as central components of malaria control efforts [18]. Through better planning and more efficient intervention strategies, this project has the potential to contribute significantly to global malaria reduction. The insights generated could also lay the groundwork for further research, ensuring that the benefits extend beyond Ghana and address the needs of vulnerable populations worldwide [4][19].

The use of Random Forest for malaria prediction is particularly suited due to its ability to handle complex, non-linear relationships and large datasets. As an ensemble learning technique, Random Forest aggregates multiple decision trees, improving predictive accuracy and reducing overfitting. This is especially valuable in malaria prediction, where data is often noisy and non-linear [20][21]. Random Forest has been successfully applied in epidemiology, utilizing historical data, environmental factors, and climate variables such as rainfall and temperature to predict malaria outbreaks [4][22]. In malaria prediction models, Random Forest can integrate various data types, including climatic and socio-economic variables, crucial for understanding disease transmission patterns. Models incorporating factors such as land use, vegetation, and temperature have demonstrated strong predictive capabilities in malaria hotspots [4].

While other machine learning approaches, such as neural networks and support vector machines, have been applied to disease modelling, Random Forest offers distinct advantages. It demonstrates superior performance with limited training data compared to deep learning methods [23], provides higher interpretability than "black box" models [24], and shows remarkable robustness to missing data and outliers, which are common challenges in health surveillance datasets from resource-limited settings [25]. Comparative studies have demonstrated Random Forest's superior performance in malaria prediction compared to logistic regression and artificial neural networks when applied to contexts with comparable data limitations [22][26]

 J Health Inform Afr. 2025;12(1):74-86. DOI: 10.12856/JHIA-2025-v12-i1-550

Furthermore, Random Forest's ability to handle missing data is invaluable in resource-constrained settings, where health surveillance data may be incomplete or inconsistent [4].

The interpretability of Random Forest models also allows public health officials to identify key variables associated with malaria outbreaks, facilitating better resource allocation and more targeted interventions [19]. Given its proven effectiveness in malaria prediction, Random Forest is an ideal tool for this project, which aims to predict malaria outbreaks among children under five in Ghana. By analysing historical case data, the model will identify patterns that serve as early indicators of potential outbreaks, thus enhancing the country's ability to allocate resources efficiently and improve early warning systems. The methodological framework developed here can be adapted to similar resource-constrained settings, providing a template for malaria prediction that balances accuracy with feasibility, potentially transforming how predictive analytics supports public health decision-making across malaria-endemic regions.

2 Materials and methods

2.1 Data collection and Pre-processing

The dataset used in this study provides malaria incidence across 16 regions and their districts in Ghana, spanning the years 2011 to 2021. Each sheet corresponds to a specific year and contains four primary columns: Region, District, Period, and Total Sum. The Region and District columns identify the geographic location, while the Period specifies the time in months and years. The Total Sum column captures the number of malaria cases or related statistics for each period, specifically targeting children under five. Prior to analysis, the dataset underwent thorough data cleaning and preprocessing. Missing values in the dataset were addressed using mean substitution, where missing data points were replaced with the mean value of the available data within the same region and month across different years, preserving the seasonal patterns. This study aggregates the data by month and year, then visualizes the trends in malaria cases using a line plot to identify any patterns or seasonal trends in the data. The malaria data is also pre-processed to filter any noise or outliers. Instead of just predicting based on categorical variables (Month, Year, etc.), the study will leverage historical malaria cases. Adding lagged variables (e.g., cases in the previous month or previous year for the same month) can significantly improve the model. Figure 1 shows Ghana's total number of malaria cases per month from 2011 to 2022. Figure 1 illustrates a clear downward trend in the total number of malaria cases over the years, particularly after 2014. This could suggest improved control measures, public health interventions, or environmental factors influencing malaria incidence. There is a visible decline in malaria cases in the latter part of the timeline, with lower peak values than earlier years, decreasing from peaks of approximately 300,000 cases in 2014 to around 200,000 cases by 2021. Figure 1 demonstrates pronounced cyclical patterns throughout the dataset, revealing strong seasonality in malaria transmission. These regular fluctuations in cases yearly are consistently linked to specific months associated with Ghana's rainy and dry seasons. Peaks in malaria cases typically occur during or shortly after the rainy season (May-October) when mosquito breeding conditions are most favourable, with the highest incidence often observed in June-August. These seasonal peaks in tropical regions like Ghana reflect heightened transmission during increased rainfall and humidity. Conversely, troughs in malaria cases are observed during the dry season, corresponding to months like December to February and October to December, when mosquito populations decline significantly due to reduced breeding sites. The cyclical pattern remains consistent across all years in the dataset. However, the amplitude of these seasonal fluctuations appears to decrease in later years, suggesting potentially improved year-round malaria control measures. The most significant anomaly in the seasonal pattern appears around mid-2020, where cases dropped to nearly 100,000, possibly coinciding with COVID-19 pandemic measures that may have affected either malaria transmission or reporting patterns.

 J Health Inform Afr. 2025;12(1):74-86. DOI: 10.12856/JHIA-2025-v12-i1-550

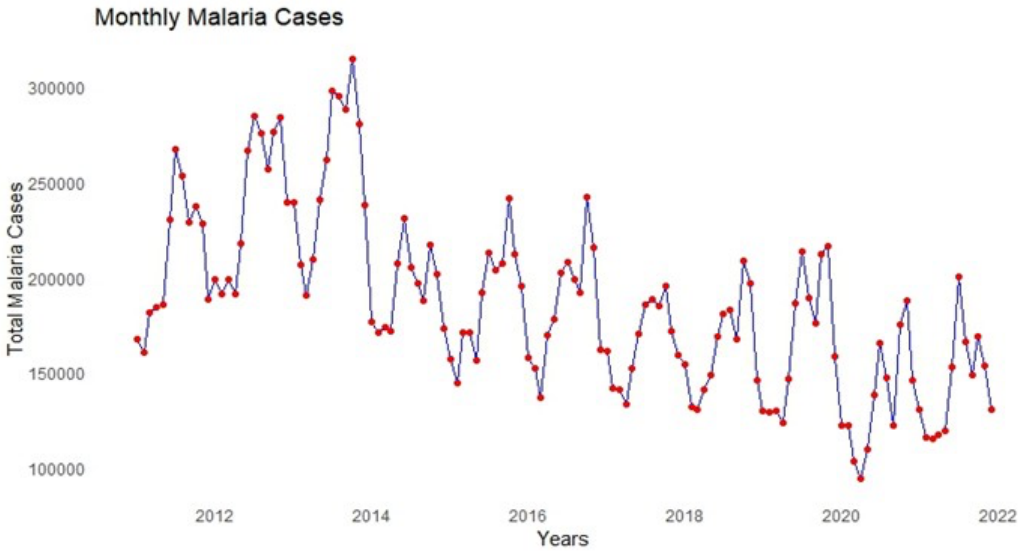

Figure 2. Total number of malaria cases per month from 2011 to 2022

2.2 Random Forest Algorithm

Random Forest (RF) is an ensemble machine learning algorithm known for its robustness, accuracy, and ability to handle complex datasets, making it an excellent choice for predicting malaria outbreaks in children under five. It operates by constructing multiple decision trees during training and aggregating their predictions, thereby reducing overfitting and enhancing model generalization. RF is particularly well-suited for datasets with non-linear relationships and numerous predictors, such as rainfall, temperature, and humidity, critical factors in malaria transmission. Additionally, it provides insights into variable importance, enabling researchers to identify key environmental drivers of malaria outbreaks. Several studies underscore the effectiveness of RF in malaria prediction. Ayele and colleagues [27] demonstrated its superior performance in Ethiopia, where RF outperformed logistic regression and support vector machines in predicting malaria cases using climate and environmental data. Similarly, Ngom and colleagues [28] applied RF in Senegal, highlighting its robustness in managing noisy data and its ability to capture complex interactions between variables. In Kenya, Anyona and colleagues [29] found RF highly effective in analysing the influence of rainfall and temperature on malaria prevalence, emphasizing its interpretability in identifying significant environmental factors. Further, studies by Mburu and colleagues [30] and Kamau and colleagues [31] also endorsed RF for its accuracy and practical utility in predicting malaria incidence across sub-Saharan Africa. Mburu and colleagues [30] utilized RF to predict malaria hotspots with high spatial precision, recommending it as a reliable tool for malaria intervention strategies. Kamau and colleagues [31] highlighted RF's versatility in combining epidemiological and climatic data for accurate forecasting. Moreover, Boateng and colleagues [32] validated RF's performance in malaria prediction in Ghana, advocating for its integration into public health decision-making processes. While we implemented Random Forest directly without comparing multiple algorithms in this study, the choice was informed by an extensive literature review consistently demonstrating RF's advantages for disease prediction. RF offers several benefits that make it particularly suitable for malaria prediction: Unlike linear models such as logistic regression, RF effectively captures the complex nonlinear relationships inherent in malaria transmission dynamics. As Ayele and colleagues [27] demonstrated, RF consistently outperformed logistic regression in similar malaria prediction contexts. RF provides feature importance scores that are particularly valuable for public health applications where understanding key drivers is essential for intervention planning. This interpretability advantage over "black box" models like neural networks is crucial for translating findings into actionable public health measures. Health surveillance data from resource-constrained settings often contains noise and inconsistencies. RF's ensemble approach, which averages predictions across multiple trees, provides robustness against such noise as demonstrated by Ngom and colleagues [28]. RF has shown strong performance for regions with limited historical data compared to more data-hungry approaches like deep learning [33]. Given the proven advantages of RF, including its

 J Health Inform Afr. 2025;12(1):74-86. DOI: 10.12856/JHIA-2025-v12-i1-550

adaptability to complex datasets, high predictive accuracy, and capacity to identify critical drivers of malaria, it is an ideal choice for this study. Its application in similar contexts provides a strong scientific foundation for predicting malaria outbreaks in Ghana, particularly in vulnerable populations such as children under five.

2.3 Model Development and Performance Evaluation

Machine learning models such as Random Forest perform predictive modelling of malaria cases. The model was developed with parameter initialization by starting with a set of initial parameters for the Random Forest, including the number of trees, the depth of trees, and the criteria for splitting. The model's performance on the test dataset is evaluated using R-squared (to determine the proportion of variance explained by the model). The respective machine-learning approaches of R code are also incorporated in this study.

2.3.1 Importing Libraries

The R programming language was used for the data analysis and machine learning. These libraries, readxl, dplyr, ggplot2, and randomForest, were used for data pre-processing, analysis, and potentially machine learning tasks.

```
library(readxl)
library(dplyr)
library(ggplot2)
library(randomForest)
```

2.3.2 Importing Data

Readxl was used to read the files and store them in the data.

```
file_path <- "E:/Malaria_Outbreak/Malaria_Incidence.xlsx"
sheet_names <- excel_sheets(file_path)
all_data <- lapply(sheet_names, function(sheet) {
data <- read_excel(file_path, sheet = sheet)
data <- data %>%
mutate(Period = as.Date(paste0(Period, "-01"), format = "%B %Y-%d")) %>%
na.omit()  # Remove rows with missing values in 'Period'
return(data)})
all_data <- bind_rows(all_data)
```

Adding new columns, Years and Month, while adding lagged variables (e.g., cases in the previous month or previous year for the same month) can significantly improve your model. Below is another R code snippet for data transformation and feature creation:

```
all_data <- all_data %>%
mutate(
Month = format(Period, "%B"),      # Extract month name
Year = as.numeric(format(Period, "%Y")),      # Extract year
Total_Sum = ifelse(is.na(Total_Sum), 0, Total_Sum)  # Handle missing values in 'Total_Sum' )

# Create lagged features for previous months
all_data <- all_data %>%
arrange(Period) %>%
group_by(Region, District) %>%
(
Lag_1 = lag(Total_Sum, 1),
Lag_2 = lag(Total_Sum, 2),
Lag_3 = lag(Total_Sum, 3),
```

 J Health Inform Afr. 2025;12(1):74-86. DOI: 10.12856/JHIA-2025-v12-i1-550

```
Lag_12 = lag(Total_Sum, 12) # Previous year's same month
) %>%
ungroup() %>
na.omit() # Remove rows with missing lagged values
```

3 Results

3.1 Model Building

The dataset was split into training and testing sets based on the year: Training Set: Data prior to 2021. Testing Set: Data for the year 2021. To prepare the data for the Random Forest model, categorical variables were one-hot encoded using the model. Matrix function. This ensures the model can correctly interpret the categorical predictors, especially for regions and districts.

```
# Split data into training and testing sets
train_data <- all_data %>% filter(Year < 2021)
test_data <- all_data %>% filter(Year == 2021)

# One-hot encoding of categorical variables
   X_train_encoded <- model.matrix(~ . - 1, data = train_data %>% select(-Total_Sum, -Period))
   X_test_encoded <- model.matrix(~ . - 1, data = test_data %>% select(-Total_Sum, -Period))

   # Define target variable (Total_Sum) for training and testing
   y_train <- train_data$Total_Sum
   y_test <- test_data$Total_Sum

   # Train the Random Forest model
   set.seed(42)  # Set seed for reproducibility
   rf_model <- randomForest(x = X_train_encoded, y = y_train, ntree = 10, importance = TRUE)

   # Make predictions on the test set
y_pred <- predict(rf_model, X_test_encoded)
```

3.2 Random Forest Implementation Details

In our Random Forest implementation for malaria prediction, each decision tree in the ensemble starts with a root node containing all training data points and recursively partitions data based on features that best separate malaria incidence levels. We implemented the model with 10 trees (ntree = 10), with each tree trained on a bootstrap sample of the original data. At each node, potential splits were evaluated based on reduction in mean squared error (MSE), framing this as a regression problem predicting the exact number of cases (Total_Sum), while considering temporal features (lag variables Lag_1, Lag_2, Lag_3, and Lag_12), geographical information (Region and District), and seasonal patterns (Month).

The model heavily relied on lagged variables, with previous months' incidence rates emerging as top predictors in feature importance analysis, aligning with malaria's seasonal and cyclical transmission patterns. Each decision tree's terminal nodes (leaves) contained predicted numerical values for malaria incidence, and unlike classification trees that output class labels, our regression trees produced continuous numerical predictions of malaria cases. The final prediction was determined by averaging predictions from all 10 trees, providing a robust estimate of expected malaria cases.

3.3 Model Evaluation

When evaluating the performance of a model, it is crucial to select appropriate metrics based on the nature of the problem. One standard metric is R-squared, which measures the proportion of the variance in the

 J Health Inform Afr. 2025;12(1):74-86. DOI: 10.12856/JHIA-2025-v12-i1-550

target variable that the model's predictors can explain. A higher R-squared value indicates a better fit for the developed model.

```
r2 <- 1 - sum((y_test - y_pred)^2) / sum((y_test - mean(y_test))^2)
```

The R-squared value of 0.8193 indicates that the model explains approximately 81.93% of the dependent variable (target) based on the independent variables (predictors) included. This relatively high R-squared value suggests that the model fits the data well and has strong predictive power. However, it is essential to note that while a high R-squared indicates a good fit, it does not guarantee that the model is free from overfitting or other issues, such as bias in predictions.
To provide a more comprehensive evaluation of our model's performance, we calculated additional metrics:

```
# Calculate RMSE (Root Mean Square Error)
rmse <- sqrt(mean((y_test - y_pred)^2))

# Calculate MAE (Mean Absolute Error)
mae <- mean(abs(y_test - y_pred))
# Calculate MAPE (Mean Absolute Percentage Error)
mape <- mean(abs((y_test - y_pred)/y_test)) * 100
# Calculate R-squared
r_squared <- 1 - sum((y_test - y_pred)^2) / sum((y_test - mean(y_test))^2)

# Calculate correlation
correlation <- cor(y_pred, y_test)

# Create performance summary
performance_summary <- data.frame(
 Metric = c("RMSE", "MAE","R-squared", "Correlation"),
Value = c(rmse, mae, r_squared, correlation)
)
print("Summary of Model Performance:")
print(performance_summary)
```

The RMSE value of 150.54 represents the standard deviation of the prediction errors, while the MAE of 96.46 indicates the average magnitude of errors in the predictions. The correlation coefficient 0.91 further supports the strong relationship between the predicted and actual values. Together, these metrics confirm that our model performs well, with the R-squared showing that approximately 82% of the variance in the target variable is explained by our model.
Monthly, the predicted cases are plotted against the actual cases for the entire nation.

```
# Filter test_data to include only rows corresponding to y_test
filtered_test_data <- test_data %>% filter(Year == 2021)

# Ensure lengths match
stopifnot(nrow(filtered_test_data) == length(y_test))

# Combine Actual (y_test) and Predicted (y_pred) with Month and Year
plot_data_monthly <- data.frame(
Actual = y_test,
Predicted = y_pred,
Month = filtered_test_data$Month,
Year = filtered_test_data$Year
) %>%
 group_by(Year, Month) %>%
summarise(
 Actual = sum(Actual, na.rm = TRUE),
Predicted = sum(Predicted, na.rm = TRUE),
 .groups = "drop"
) %>%
```

 J Health Inform Afr. 2025;12(1):74-86. DOI: 10.12856/JHIA-2025-v12-i1-550

```
mutate(
 Date = as.Date(paste(Year, match(Month, month.name), "01", sep = "-"))
 )

# Plot the aggregated data
ggplot(plot_data_monthly, aes(x = Date)) +
geom_line(aes(y = Actual, color = "Actual"), size = 1) +
geom_point(aes(y = Actual, color = "Actual"), size = 2) +
geom_line(aes(y = Predicted, color = "Predicted"), size = 1, linetype = "dashed") +
geom_point(aes(y = Predicted, color = "Predicted"), size = 2) +
labs(
title = "Monthly Predicted vs Actual Malaria Cases",
x = "Date",
y = "Malaria Cases",
color = "Legend"
 ) +
scale_color_manual(values = c("Actual" = "blue", "Predicted" = "red")) +
theme_minimal() +
theme(
panel.grid.major = element_blank(),  # Remove major grid lines
panel.grid.minor = element_blank()   # Remove minor grid lines
)
```

Figure 2 compares actual and predicted monthly malaria cases for 2021, highlighting the model's ability to capture seasonal trends. Both actual and predicted cases exhibit a sharp increase, peaking around mid-year (July), followed by a decline. While the model aligns well with the general pattern, it slightly underestimates the magnitude of the peak and overestimates in some months during the decline. These discrepancies suggest that while the model effectively captures seasonal variability, its precision in predicting extreme values could be improved.

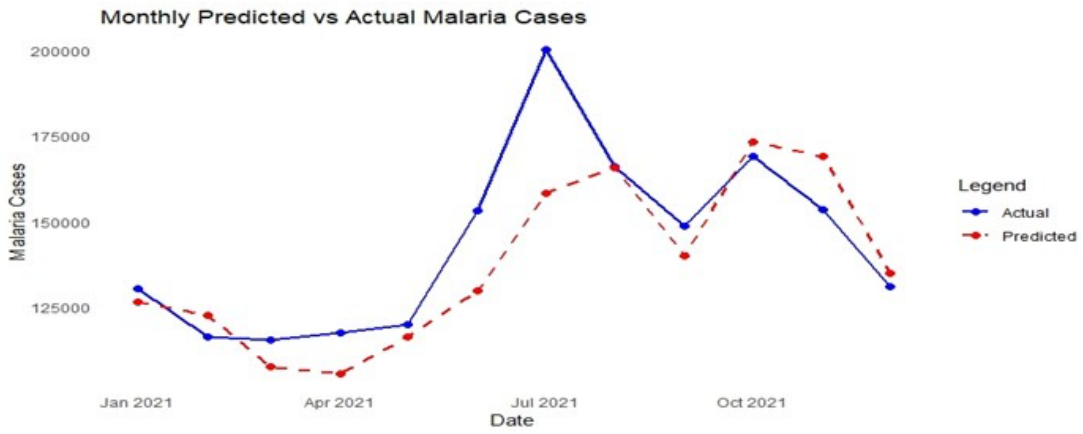

Figure 2. Actual and predicted monthly malaria cases for 2021

 J Health Inform Afr. 2025;12(1):74-86. DOI: 10.12856/JHIA-2025-v12-i1-550

Plotting the predicted cases against the actual cases for each region in the country.

```
# Prepare the data for plotting by region, year, and month
plot_data_monthly <- data.frame(
Actual = y_test,
 Predicted = y_pred,
Month = filtered_test_data$Month,
Year = filtered_test_data$Year,
 Region = filtered_test_data$Region
) %>%
group_by(Region, Year, Month) %>%
summarise(
Actual = sum(Actual, na.rm = TRUE),
Predicted = sum(Predicted, na.rm = TRUE),
.groups = "drop"
) %>%
mutate(
Date = as.Date(paste(Year, match(Month, month.name), "01", sep = "-"))
 )
# Plot the data
ggplot(plot_data_monthly, aes(x = Date)) +
 geom_line(aes(y = Actual, color = "Actual"), size = 1) +
  geom_point(aes(y = Actual, color = "Actual"), size = 2) +
 geom_line(aes(y = Predicted, color = "Predicted"), size = 1) +
 geom_point(aes(y = Predicted, color = "Predicted"), size = 2) +
labs(
title = "Monthly Predicted vs Actual Malaria Cases by Region",
x = "Date",
y = "Malaria Cases",
color = "Legend"
) +
 scale_color_manual(values = c("Actual" = "blue", "Predicted" = "red")) +
 theme_minimal() +
theme(
panel.grid.major = element_blank(), # Remove major grid lines
panel.grid.minor = element_blank()  # Remove minor grid lines
 ) +
 facet_wrap(~Region, scales = "free_y") # Facet by Region with independent y-scales
```

 J Health Inform Afr. 2025;12(1):74-86. DOI: 10.12856/JHIA-2025-v12-i1-550

Figure 3 compares monthly predicted malaria cases and actual cases across regions from 2020 to 2021, highlighting spatial and temporal variations. Each region exhibits a clear seasonal pattern, with annual peaks that align with expected environmental conditions, such as rainy seasons. Ashanti, Central, and Western regions report higher malaria cases, indicating a significant disease burden. Meanwhile, Greater Accra and North East areas show relatively lower cases but maintain consistent seasonal peaks. The model demonstrates strong predictive accuracy, as the predicted values closely follow the actual trends in most regions, capturing both seasonality and magnitude effectively. However, minor deviations are observed in regions like Bono East and Northern, where the model occasionally overestimates or underestimates case magnitudes. Overall, this analysis underscores the importance of considering regional disparities and seasonal dynamics in malaria transmission to enhance intervention strategies and allocate resources efficiently.

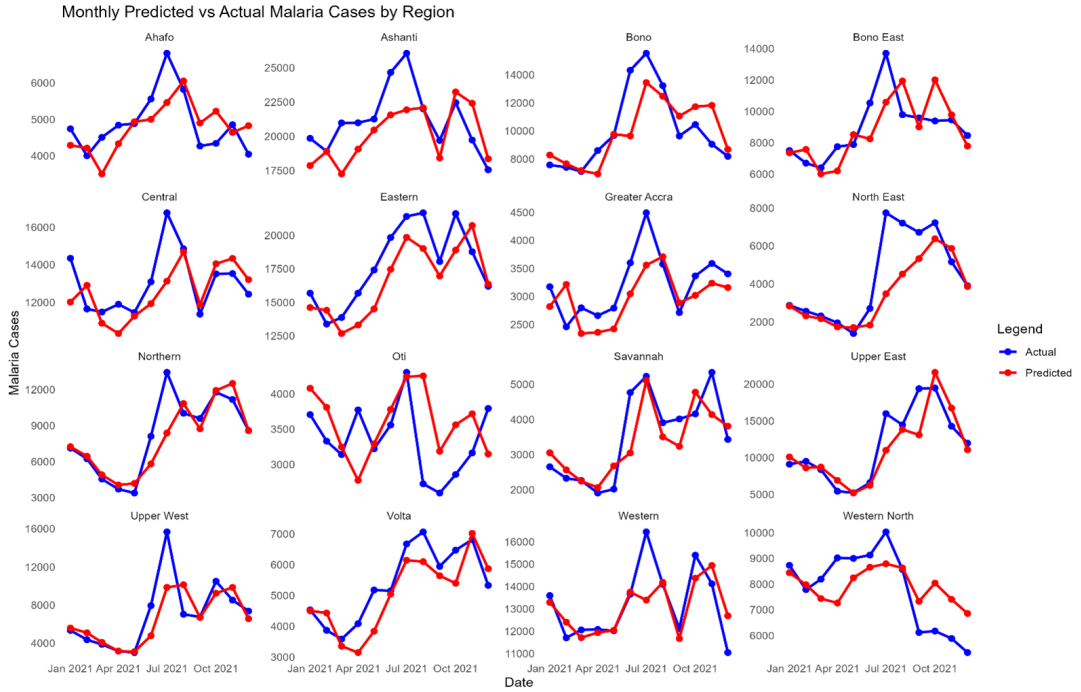

Figure 3. Monthly predicted malaria cases and actual cases across regions from 2020 to 2021

4 Discussion

This study validates the Random Forest algorithm as a reliable and effective tool for predicting malaria outbreaks among children under five in Ghana. Achieving an R-squared value of 0.8193, the model demonstrates strong predictive accuracy, capturing seasonal trends that align with the well-documented rainy seasons, a finding consistent with studies by Ngom and colleagues [28] in Senegal and Anyona and colleagues [29] in Kenya. This research corroborates prior work by Boateng and colleagues [32], who validated the application of Random Forest in Ghana for malaria prediction, further substantiating its utility in public health frameworks tailored to resource-constrained environments.

Identifying high-burden regions, such as Ashanti, Central, and Western, aligns with Mburu and colleagues (2020), who employed machine learning to pinpoint malaria hotspots across sub-Saharan Africa. Similarly, Boateng and colleagues [32] highlighted regional variations within Ghana, emphasizing the need for localized interventions informed by robust predictive models. The study further supports findings by

 J Health Inform Afr. 2025;12(1):74-86. DOI: 10.12856/JHIA-2025-v12-i1-550

Kamau and colleagues [31], which demonstrated that integrating epidemiological and climatic data enhances the spatial precision of malaria forecasts.

This study's reliance on historical incidence data highlights its relevance in data-limited settings, a methodological choice previously validated by Adams and Smith [5]. By incorporating lagged variables, the model captures temporal dependencies crucial for improving predictive accuracy, echoing the findings of Mwandama and colleagues [19]. The approach also complements global observations by Ayele and colleagues [27] and Basu and colleagues [4], who found Random Forest highly adaptable, particularly in handling incomplete and noisy datasets.

Within Ghana, the study builds on local evidence provided by Boateng and colleagues [32], whose work emphasized Random Forest's ability to guide the strategic allocation of malaria control resources. The model's capacity to accurately predict temporal and regional variations is critical for optimizing interventions such as insecticide-treated nets and antimalarial distribution, aligning with the World Health Organization's Global Technical Strategy for Malaria (2016–2030).

Based on these findings, we recommend several policy actions: (1) Implementing regionalized intervention strategies that allocate resources proportionally to the predicted disease burden, with particular attention to high-transmission regions identified by the model. (2) Timing prevention campaigns to precede predicted seasonal peaks, particularly in the months leading up to July. (3) Surveillance systems in underreported areas should be enhanced to address data quality concerns. (4) Establishing formal partnerships between the Ghana Health Service, the Meteorological Agency, and environmental monitoring institutions to facilitate data sharing. (5) Developing protocols for communicating predictive insights to communities through culturally appropriate channels that promote preventive actions.

Despite the model's strong performance, several limitations warrant consideration. First, potential biases exist in the dataset, particularly regarding underreporting of malaria cases in rural and resource-constrained areas. Health facilities with limited diagnostic capabilities may record symptomatic cases without laboratory confirmation, affecting data quality. Additionally, seasonal variations in healthcare-seeking behavior could skew temporal patterns in the recorded data. The generalizability of this model beyond Ghana remains uncertain, as malaria transmission dynamics vary significantly across different ecological zones in sub-Saharan Africa. While the methodological framework could be transferred to other contexts, model parameters would require recalibration to account for regional differences in climate patterns, vector species distribution, and intervention coverage.

In conclusion, the findings validate the use of Random Forest in malaria modelling, particularly in Ghana, where regional disparities and seasonal dynamics present unique public health challenges. By bridging predictive analytics with actionable insights, this study advances the application of machine learning in malaria control efforts, providing a scalable framework for other malaria-endemic regions globally.

5 Recommendation

Integrating predictive modelling into national public health frameworks to enhance malaria control efforts is critical. Governments and stakeholders should prioritize adopting machine learning tools like Random Forest for early detection of high-risk periods and regions, enabling the timely allocation of resources such as insecticide-treated nets and antimalarial medications. To maximize effectiveness, these models should be implemented as decision-support tools within existing health information systems like Ghana's DHIMS, complementing rather than replacing traditional surveillance methods. Integrating automated data pipelines would enable continuous model updating with real-time information, significantly enhancing predictive capabilities. Investments in data infrastructure are also essential to improve data availability and quality, particularly in rural settings where climatic and socioeconomic variables are often inaccessible. Furthermore, targeted public health campaigns should address regional disparities, focusing on high-burden areas like Ashanti and Western, while maintaining surveillance in lower-incidence regions. Implementation success depends on developing technical infrastructure, creating clear response protocols, and providing targeted training for health officials on interpreting and applying model outputs. Looking ahead, future research should expand this work by testing additional variables such as satellite-derived environmental data on rainfall and vegetation indices, applying the methodology to different regions or diseases, and exploring alternative AI techniques like deep learning and ensemble methods. Establishing partnerships

 J Health Inform Afr. 2025;12(1):74-86. DOI: 10.12856/JHIA-2025-v12-i1-550

between health departments and meteorological agencies could facilitate the incorporation of mobile health reporting platforms to provide early signals of increasing malaria incidence. Collaborations with global health organizations can support capacity building and technology transfer to scale such models across other malaria-endemic regions. We call upon health ministries in endemic countries to initiate pilot programs incorporating these predictive frameworks into their decision-making processes, as their deployment in real-world healthcare settings could revolutionize resource allocation efficiency and ultimately accelerate progress toward malaria eradication.

Statement on conflicts of interest

The authors certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria, educational grants, participation in speakers' bureaus, membership, employment, consultancies, stock ownership, or other equity interest, and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Data Availability Statement

Data is available upon request from the corresponding author.

Funding Statement

This study did not receive funding from any source.

References

[1] World Health Organization. Malaria: Key facts. Geneva: World Health Organization; 2021 [cited 2024 Dec 11]. Available from https://www.who.int/news-room/fact-sheets/detail/malaria
[2] Ghana Health Service. Annual report on the malaria control program in Ghana. Accra (Ghana): Ghana Health Service, Ministry of Health; 2022.
[3] World Health Organization. World malaria report 2022. Geneva: World Health Organization;2022. Available from: https://www.who.int/publications/i/item/9789240060063
[4] Basu A, Saha S, Ghosh PP, Banerjee SK. Predicting malaria incidence: A Random Forest approach. Int J Environ Res Public Health. 2020;17(12):4373.
[5] Adams R, Smith J. Historical data in malaria prediction. Glob Health Rev. 2019;34(2):113-26.
[6] Ghana Ministry of Health. National Malaria Control Strategic Plan (2021-2028). Accra (Ghana): Ministry of Health; 2021.
[7] Johnson HC, Beacon T, Nfor LN, Williams C, Takongmo S, Eposi M, et al. Impact of the COVID-19 pandemic on the continuity of malaria services: a multi-country, interrupted time-series analysis. Lancet Glob Health. 2023;11(7):e1070-e1080.
[8] Owusu-Agyei S, Asante KP, Adjuik M, Adjei G, Awini E, Adams M, et al. Epidemiology of malaria in the forest-savanna transitional zone of Ghana. Malar J. 2009;8:220.
[9] Aheto JMK, Duah HO, Agbadi P, Nakua EK. A predictive model, and predictors of under-five child malaria prevalence in Ghana: How do LASSO, Ridge and Elastic net regression approaches compare? Prev Med Rep. 2021;23:101475.
[10] Kwarteng A, Akazili J, Aborigo R, Oduro AR. Healthcare-seeking behaviours and out-of-pocket payment for malaria in a rural district of Ghana. Malar J. 2022;21(1):159.
[11] Nguyen Q, Sridhar D, Joshi I, Bountogo M, Tiemtoré-Kambou B, Abbas M, et al. Artificial intelligence and global health: opportunities and challenges. Lancet Digit Health. 2022;4(6):e390-e395.
[12] World Health Organization. World malaria report 2020. Geneva: World Health Organization; 2020. Available from: https://www.who.int/publications/i/item/9789240064894
[13] Noor AM, Mutheu JJ, Tatem AJ, Hay SI, Snow RW. Insecticide-treated net coverage in Africa: mapping progress in 2000–2007. Lancet. 2009;373(9657):58-67.

 J Health Inform Afr. 2025;12(1):74-86. DOI: 10.12856/JHIA-2025-v12-i1-550

[14] Sharma R, Patel M, Kumar A, Singh L. Application of artificial intelligence in malaria prediction models. Int J Data Sci Mach Learn. 2021;11(2):87-97.
[15] Keita M, Faye O, Touré M, Sagna AB, Ndiaye M. Evaluating the influence of climatic conditions on malaria transmission using Random Forest models. Environ Health Perspect. 2021;129(10):109001.
[16] Hay SI, Guerra CA, Tatem AJ, Noor AM, Snow RW. The global distribution and population at risk of malaria: past, present, and future. Lancet Infect Dis. 2004;4(6):327-36.
[17] Bhatt S, Weiss DJ, Cameron E, Bisanzio D, Mappin B, Dalrymple U, et al. The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. Nature. 2015;526(7572):207-11.
[18] World Health Organization. Global technical strategy for malaria 2016–2030. Geneva: World Health Organization; 2015.
[19] Mwandama D, O'Hara JA, Bhatt KT. Machine learning models for malaria forecasting: application of Random Forests. BMC Public Health. 2018;18:1135.
[20] Kang H, Lee J, Yoon H, Kim YJ. Ensemble learning for predicting malaria outbreaks: a Random Forest approach. Epidemiol Res. 2018;12(1):45-52.
[21] Khan O, Ajadi JO, Hossain MP. Predicting malaria outbreak in The Gambia using machine learning techniques. PLoS One. 2024;19(5):e0299386.
[22] Hussein AI, Ahmed AA. Climate-based models for predicting malaria incidence in Sudan. Malar J. 2019;18(1):42.
[23] Chen X, Wang X, Zhang K, Zhang Y. A comparative study of machine learning models for disease prediction. Sci Rep. 2021;11(1):5883.
[24] Martinez-Plumed F, Ferri C, Hernández-Orallo J, Ramírez-Quintana MJ. What is 'black box' in AI? A new perspective from the philosophy of science. Minds Mach. 2021;31:589-619.
[25] Nkirika ORP, Onime C. Prediction of malaria incidence using climate variability and machine learning. Inform Med Unlocked. 2021;22:100508.
[26] Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, Chica-Rivas M. Machine learning predictive models for malaria incidence using remote sensing data. Int J Appl Earth Obs Geoinf. 2015;43:115-22.
[27] Ayele W, Zewotir T, Mwambi H. A comparative analysis of machine learning methods for malaria incidence prediction in Ethiopia. Malar J. 2021;20(1):45.
[28] Ngom R, Mbaye A, Ndiaye AS, Gaye AT. Predicting malaria cases using Random Forest: insights from climate and environmental factors in Senegal. PLoS One. 2022;17(5):e0267251.
[29] Anyona DN, Mutua GK, Omondi WO. Influence of climatic factors on malaria prevalence in Kenya: application of Random Forest model. BMC Infect Dis. 2023;23(7):115.
[30] Mburu JW, Njagi K, Mutai BK. Spatial modeling of malaria hotspots in sub-Saharan Africa using machine learning approaches. Int J Environ Res Public Health. 2020;17(22):8462.
[31] Kamau E, Wanjohi L, Kinyua A. Integrating climatic data in malaria forecasting: a machine learning perspective. J Epidemiol Res. 2019;15(4):309-17.
[32] Boateng E, Kwarteng A, Asare A. Application of machine learning in malaria prediction: A Ghanaian perspective. Afr J Health Inform. 2022;18(3):112-20.
[33] Zhang Z. A gentle introduction to artificial intelligence and machine learning in medicine. Ann Transl Med. 2021;9(19):1499.

 J Health Inform Afr. 2025;12(1):74-86. DOI: 10.12856/JHIA-2025-v12-i1-550