Journal of Research High School

Vol. 2025 (10) 838 – 849
ISSN 2688-3651 [online]

# The Impact of Carbon Dioxide Emissions on Malaria Incidence in Africa

**Xuyin Dai**[1] *

[1] Yorba Linda High School, Yorba Linda, CA, USA
*Corresponding Author: d3308681@gmail.com

Advisor: Gary Yang, garycloudyang@gmail.com

Received June 8, 2025; Revised July 30, 2025; Accepted August 15, 2025

**Abstract**

Malaria remains a major public health concern in Africa, where environmental and socioeconomic conditions influence its transmission. Continuously rising carbon dioxide ($CO_2$) emissions pose a severe threat to the climate; however, their role in modulating malaria dynamics through climate-related processes remains elusive. This study investigated the relationship between $CO_2$ emissions and malaria cases across African nations by integrating environmental and socioeconomic variables. Using machine learning methods, including random forest and gradient boosting regression, the analysis captured complex, nonlinear interactions that traditional models missed. The findings revealed a positive correlation between $CO_2$ emissions and malaria prevalence, particularly in tropical forest regions where microclimate shifts enhanced mosquito survival. In contrast, this link was weaker in arid areas that were less favorable to mosquito breeding. Socioeconomic factors, such as improved sanitation, consistently mitigate malaria risk across all environments. Correlation analyses highlighted spatial heterogeneity in $CO_2$ effects, while model interpretation tools revealed the relative importance of contributing factors. This study established a framework for understanding how $CO_2$-driven environmental changes and socioeconomic conditions influenced malaria transmission in Africa, offering insights for developing targeted, climate-resilient interventions.

*Keywords: Carbon dioxide emissions, Malaria transmission, Machine learning, Environmental factors, Socioeconomic variables*

## 1. Introduction

Malaria remains one of the most urgent public health challenges worldwide, particularly in sub-Saharan Africa, where the majority of global cases and deaths occur (World Health Organization, n.d.). Over the past few decades, despite significant progress through interventions such as insecticide-treated nets and antimalarial medications, malaria prevention has plateaued in recent years (Pryce et al., 2022). One contributing factor is climate change, largely driven by rising carbon dioxide ($CO_2$) emissions (World Health Organization, 2021). Although precise mechanisms that link climate change to malaria transmission remain unclear (Paaijmans et al., 2009), environmental changes are known to create conditions in which vector-borne diseases can thrive (Beloconi et al., 2023). Vectors, such as mosquitoes, transmit pathogens between humans or from animals to humans and are sensitive to environmental changes (Caminade et al., 2019). While no definitive evidence directly links rising $CO_2$ emissions to increased malaria incidence, numerous studies suggest a positive correlation between global warming and malaria transmission (Nabi & Qader, 2009). This underscores the importance of further research into how climate-related factors may influence malaria transmission. This study aims to investigate how $CO_2$ emissions contribute to the shifting patterns of malaria transmission, thereby informing strategies to adapt disease control measures to a changing climate. Moreover, understanding the role of emissions and climatic conditions such as high temperature and humidity in disease dynamics is particularly important for regions with limited healthcare infrastructure and high malaria burden (Fatima

Vol. 2025 (10) 838 – 849
ISSN 2688-3651 [online]

et al., 2025). This research not only assesses the environmental drivers of malaria but also supports more effective, region-specific public health policies.

A substantial body of past research has examined environmental influences on malaria vector behavior. Ermert et al. (2011) linked greenhouse gas emissions, partially driven by deforestation, to the expansion of malaria zones into higher-altitude regions once considered unsuitable for transmission. Similarly, Himeidan and Kweka (2012) documented how deforestation and agricultural expansion in the East African highlands elevated local temperatures, thereby enhancing the survival and reproductive success of *Anopheles* mosquitoes. Because deforestation both releases stored $CO_2$ and reduces carbon sequestration, it contributes to local warming and global emissions. Furthermore, Le et al. (2019) used the stochastic lattice-based integrated malaria (SLIM) model to show that elevated $CO_2$ concentrations have modified vegetation cover in areas like Kilifi County, Kenya, thereby influencing mosquito breeding and biting patterns. Additional insights come from Costantini et al. (1996), who demonstrated that higher $CO_2$ levels enhance mosquitoes' ability to locate human hosts, and from Takken et al. (2024), who examined how specific environmental stressors, including rising $CO_2$ levels, increased temperatures, and changes in humidity, affect the feeding and mating behaviors of mosquitoes in sub-Saharan Africa. Collectively, these studies highlight how $CO_2$ emissions—whether from fossil fuel combustion or indirectly through deforestation—drive environmental changes that reshape mosquito ecology and influence malaria transmission dynamics by altering key aspects of the vector life cycle, such as the duration of larval development and frequency of blood feeding(Institute of Medicine, 1991).

While climate and environmental variables are essential components of malaria ecology, it is equally important to recognize that socioeconomic factors also impact disease transmission (Castro, 2017). In recent years, studies have emphasized the importance of non-environmental variables such as gross domestic product (GDP), access to clean water, sanitation capacity, urbanization, and healthcare infrastructure on malaria prevalence (Braimah et al., 2024). In particular, regions with limited access to preventive tools, poor health systems, and low income are often more vulnerable to malaria outbreaks (Perera et al., 2022). Rural areas with scarce water sources and limited housing protection are more likely to experience high vector exposure (Sutherst, 2004). Ignoring these factors may lead to an incomplete understanding of the malaria transmission patterns. For example, regions with similar environmental profiles may experience vastly different transmission rates due to dissimilar public health capacity or economic development (Institute of Medicine, 2002). Thus, it is critical to consider confounding variables when studying the relationship between $CO_2$ emissions and malaria, as they can amplify or obscure the impact of climate-related factors (Nissan et al., 2021). To address this, the present study incorporates these socioeconomic factors to produce a more comprehensive model for malaria transmission.

Given the complex interactions between environmental and socioeconomic factors, traditional statistical methods may not be able to model malaria transmission dynamics (MalariaWorld, 2024). In response to these limitations, machine learning (ML) has emerged as a powerful tool for modeling complicated systems and extracting patterns from high-dimensional data (Surur et al., 2025). Unlike traditional methods, ML models can handle poorly defined data distributions, making them well-suited for public health studies (Panch et al., 2018). ML has already shown promise in forecasting disease outbreaks, optimizing intervention strategies, and identifying risk factors in diseases such as dengue, Zika, and COVID-19 (Al-Hajjar, 2024). By integrating multiple datasets from multiple sources, ML approaches can combine environmental factors such as temperature, rainfall, and other socioeconomic factors to better understand how these elements interact and influence malaria transmission. Building on the proven utility of ML in modeling health systems, this research applied ML techniques to investigate the relationship between carbon dioxide emissions and malaria transmission in Africa.

## 2. Materials and Methods

### 2.1 Data Acquisition and Preprocessing

Datasets used were downloaded from Kaggle (Lydia70. n.d.). Priorities were given to datasets that met the following criteria: (1) institutional provenance (e.g., sourced from the World Health Organization or World Bank), (2) a sufficient number of countries to ensure statistical power. The chosen dataset includes major African countries and

Vol. 2025 (10) 838 – 849
ISSN 2688-3651 [online]

contains information on malaria incidence, total reported malaria cases, carbon dioxide emissions, sanitation, economic indicators, and demographic growth rates from 2007 to 2017.

This study did not select more recent datasets for several reasons. In this context, "recency" indicates that the dataset has an extensive temporal range and includes numerous significant variables, rather than being limited to the most recent year. Moreover, none of the models used the time (2007–2017) as a variable; thus, not including more recent data did not explicitly influence predictions or conclusions. More recent or comprehensive data from official sources such as the World Health Organization or environmental databases could not be obtained, as the available resources were generally restricted to a single year (e.g., 2024), limited by paywalls, or required formal requests for data access. Conversely, the Kaggle dataset was distinctive in its accessibility and comprehensive inclusion of all essential variables required for this analysis. It was considered the most optimal and representative source.

Variates with incomplete country-level data were removed; similarly, countries with missing values for variates were removed. In total, eighteen out of fifty countries were removed. However, because the remaining countries have a wide range of malaria incidence and total reported cases across the studied period, the exclusions of countries likely did not impact the study's conclusion. The filtered dataset contains complete coverage for 32 African countries, including Namibia, Sao Tome and Principe, Sudan, Comoros, Mauritania, Madagascar, Senegal, Kenya, Ethiopia, Zimbabwe, Guinea-Bissau, Tanzania, Zambia, Burundi, Angola, Rwanda, Gabon, Cameroon, Malawi, Ghana, Uganda, Equatorial Guinea, Togo, Nigeria, Liberia, Mozambique, Guinea, Niger, Cote d'Ivoire, Benin, Sierra Leone, and Burkina Faso, across 14 variables encompassing Population, GDP Per Capita (USD), GDP Per Capita PPP (USD), Transportation (Mt), Total CO2 Emission including LUCF (Mt), Rural population growth (annual %), Urban population (% of total population), Urban population growth (annual %), People using at least basic drinking water services (% of population), People using at least basic drinking water services, rural (% of rural population), People using at least basic drinking water services, urban (% of urban population), People using at least basic sanitation services (% of population), People using at least basic sanitation services, rural (% of rural population), and People using at least basic sanitation services, urban (% of urban population).

### 2.2 Data Analysis Environment

All analyses were conducted using Python 3.10. While statistical analyses utilized the Statsmodel library, data processing and numerical computations were performed with NumPy (Harris et al., 2020), Pandas (The pandas development team, 2020), and SciPy (Virtanen et al., 2020). Machine learning models were implemented using Scikit-Learn, and visualizations were generated with Matplotlib (Hunter, 2007) and Seaborn (Waskom, 2021). For interpreting the tree-based regressors, SHAP (SHapley Additive explanations) was employed to provide insight into feature importance and model behavior (Lundberg & Lee, 2017). SHAP values quantify how much each feature positively or negatively contributes to a model's prediction, helping identify which variables are most influential and how they affect the outcome (DataCamp, 2022).

### 2.3 Data Exploratory Analysis

The analysis began by examining malaria incidence and case numbers from 2007 to 2017 across 32 African countries, using a dataset obtained from Kaggle (Lydia70. n.d.). This dataset revealed wide variation across nations and annual fluctuations. It then compared regional trends to explore how environmental factors may differently impact transmission across geographic areas.

### 2.4 Correlation Analysis

To assess whether $CO_2$, the central variable in this study, could explain year-to-year variations in malaria incidence, Pearson correlation coefficients were computed and visualized for each country. Additionally, correlations were also calculated for environmental and socioeconomic factors.

Vol. 2025 (10) 838 – 849
ISSN 2688-3651 [online]

### 2.5 Train-Test Split

To fairly evaluate the performance of the machine learning models, the dataset was randomly divided into training and testing subsets, with 50% of the data allocated for model training and the remaining 50% reserved for unbiased evaluation. This ratio of train-test split was motivated by the fact that the dataset had only 352 data points, and enough samples were needed to do a full and accurate evaluation of how well the classifier worked. This method helped keep performance estimates from being too optimistic and helped get a better idea of how well the model worked on unseen data.

### 2.6 Feature Selection

The feature selection process began with the 14 variables identified in Section 2.1, which provided a comprehensive overview of demographic, environmental, and economic factors across 32 African countries. This prompted forward-backward feature selection, also known as stepwise selection, to improve the model. To keep only highly predictive variables, during each round of selection, this strategy added the most statistically significant variables and removed all insignificant variables. The process checked for multicollinearity and removed variables that are highly correlated to avoid unnecessary repetition. Thus, the routine developed a simplified malaria incidence model with non-redundant, statistically significant variables.

### 2.7 Linear Regression Modeling

Statsmodels were used to do multivariate ordinary least squares (OLS) regression, which yields statistical significance and independent variable coefficients (Seabold & Perktold, 2010). Many input variables were statistically insignificant in the initial results. Therefore, the model was re-fitted using the reduced set of features identified through the forward-backward selection process described in Section 2.6. The simplified model kept only four variables out of fourteen but had comparable predictive accuracy, demonstrating that these features captured the most significant information. Despite this improvement, the OLS model had a low test-set correlation, prompting more flexible machine learning methods.

### 2.8 Tree-based Regression Modeling

To better capture the nonlinear relationships between $CO_2$ emissions, socioeconomic factors, and malaria incidence, this study applied random forest regression (Ho, 1995) and gradient boosting regression (Friedman, 2001). Both models were run with the Scikit-Learn default parameters, except for the max depth and random state. Specifically, both models were configured with the number of trees equal to 100, maximum depth equal to 3, and random state equal to 0. The shallow tree depth was meant to avoid overfitting, and setting the random state was to ensure reproducibility. The input features were the same as the ones chosen via feature selection in section 2.6. Gradient boosting achieved a Pearson correlation coefficient of 0.8 on the test set, indicating strong predictive accuracy. To gain deeper insight into how each feature influenced the models' predictions, SHAP analysis was subsequently performed and visualized. SHAP values helped explain how much each feature contributes to a model's prediction, allowing for a clearer interpretation of variable importance and direction of influence (DataCamp, 2022). Ultimately, this analysis confirmed the initial hypothesis by revealing that $CO_2$ emissions tend to increase malaria transmission, whereas improved sanitation consistently reduces it.

## 3. Results

Malaria incidence trends remained relatively constant or gradually declined across most African countries between 2007 and 2017, as shown in Figure 1A. However, countries like Rwanda experienced a dramatic surge in incidence, from approximately 250 to nearly 600 cases per 1,000 population between 2014 and 2016, followed by a slight decrease. In terms of incidence magnitude, Burkina Faso, Sierra Leone, Niger, and Côte d'Ivoire consistently

Vol. 2025 (10) 838 – 849
ISSN 2688-3651 [online]

reported some of the highest incidence rates, ranging between 400 and 550 cases per 1,000 population throughout the recorded period. Nevertheless, modest declines could be seen toward the end of 2017. Meanwhile, countries like Namibia, São Tomé, Príncipe, and Comoros maintained low incidence rates (<50 cases per 1,000 population).

Figure 1C illustrates within-continent variation in malaria incidence, revealing a pronounced geographic contrast between regions such as West, Central, and North Africa. The incidence rates ranged from virtually zero cases per 1,000 population in Morocco, Algeria, and Cabo Verde to over 500 cases in Burkina Faso. A clear regional difference emerged between West and Central African nations and North African countries, with the former generally reporting substantially higher malaria burden than the latter. Mid-range incidence rates (approximately 100-300 cases per 1,000 population) were observed throughout East African countries like Kenya, Ethiopia, Tanzania, and Zimbabwe. This pattern linked malaria incidence to regional annual temperature profiles, which were partly driven by carbon emissions and other environmental factors.

As depicted in Figure 1B, total reported malaria cases increased substantially from 2007 to 2017, even in countries where incidence rates remained stable or declined. A striking increase began around 2012-2013, with Nigeria, Mozambique, and Burkina Faso experiencing a particularly steep upward trend. Notably, Nigeria reported over 12 million malaria cases in 2016, representing a more than tenfold increase from 2007. This substantial growth in reported cases, despite a relatively flat incidence trend, was likely due to improvements in case detection and reporting systems rather than a true increase in malaria burden. Even though most countries experienced an increase in total reported cases, Namibia, São Tomé and Príncipe, and Comoros maintained low absolute case numbers throughout the study period.

This study used incidence rate to measure malaria burden because it normalized for population size and reflected exposure to malaria vectors. This ensured trends represent transmission dynamics rather than demographics. Incidence also offered a more accurate view of transmission intensity and is the primary metric recommended by the WHO in its Global Technical Strategy 2016–2030 (World Health Organization, 2015). Accordingly, the modeling focused on malaria incidence and its relationship with $CO_2$ emissions and other variables.

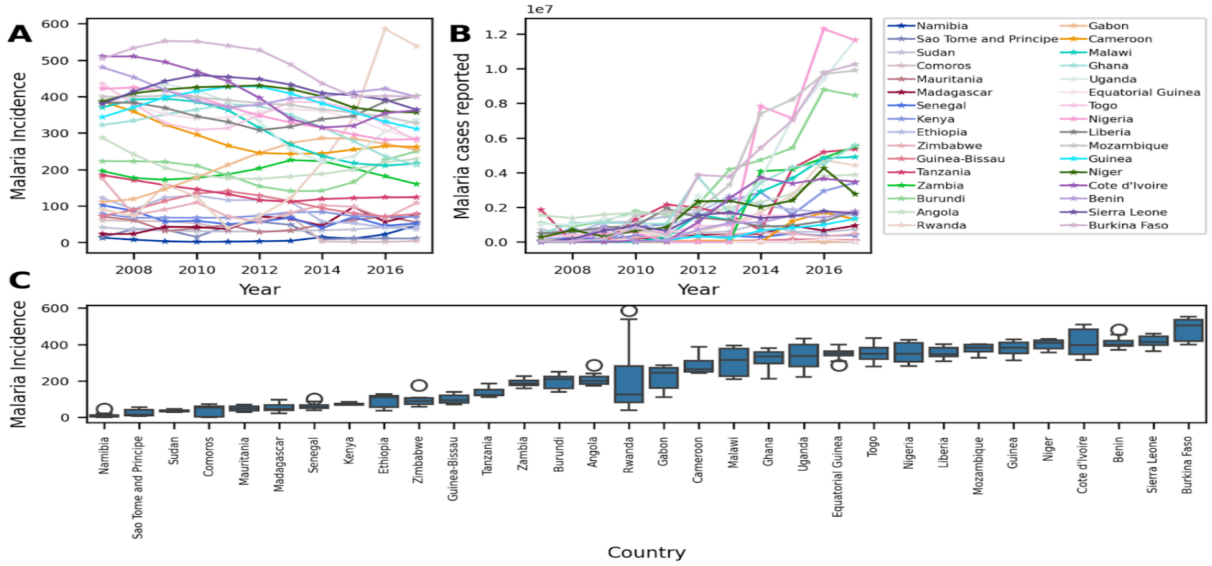


Figure 1. Malaria incidence and total Cases across selected African countries in 2007-2017. Figure 1A shows temporal trends in malaria incidence rates (per 1,000 population at risk), and Figure 1B illustrates total reported malaria cases across African countries from 2007 to 2017. Figure 1C demonstrates the intra-country malaria incidence variation.

As shown in Figure 2A, correlation coefficients were computed between carbon dioxide emissions and malaria incidence across African countries, revealing strong variation in both direction and strength of association. This underscored the complex and diverse nature of malaria transmission, which likely depends on additional environmental and socioeconomic variables. These correlations were then binned by magnitude: ones with an absolute magnitude greater than 0.5 were classified as strong, those between 0.3 and 0.5 were moderate, and the rest were

Vol. 2025 (10) 838 – 849
ISSN 2688-3651 [online]

considered weak. Several countries, including Angola and Malawi, demonstrated a strong negative correlation, whereas countries like Cameroon and Gabon exhibited a strong positive correlation. Conversely, Kenya and Rwanda displayed negligible correlation. It should be noted that the weak correlation cannot simply be attributed to the countries having a small range of incidence rates. A parabolic-shaped relationship could be seen for Liberia and Rwanda, where both countries had a large incidence range. Because carbon emissions cannot foster and inhibit malaria transmission simultaneously, one could only conclude that carbon emissions alone cannot fully capture variations of malaria incidence. Malaria transmission dynamics likely resulted from a complex interplay between carbon emissions and many other factors. This is further supported by the observation that geographically proximal countries, such as Cameroon and Nigeria, showed opposing correlations between carbon emissions and malaria incidence.

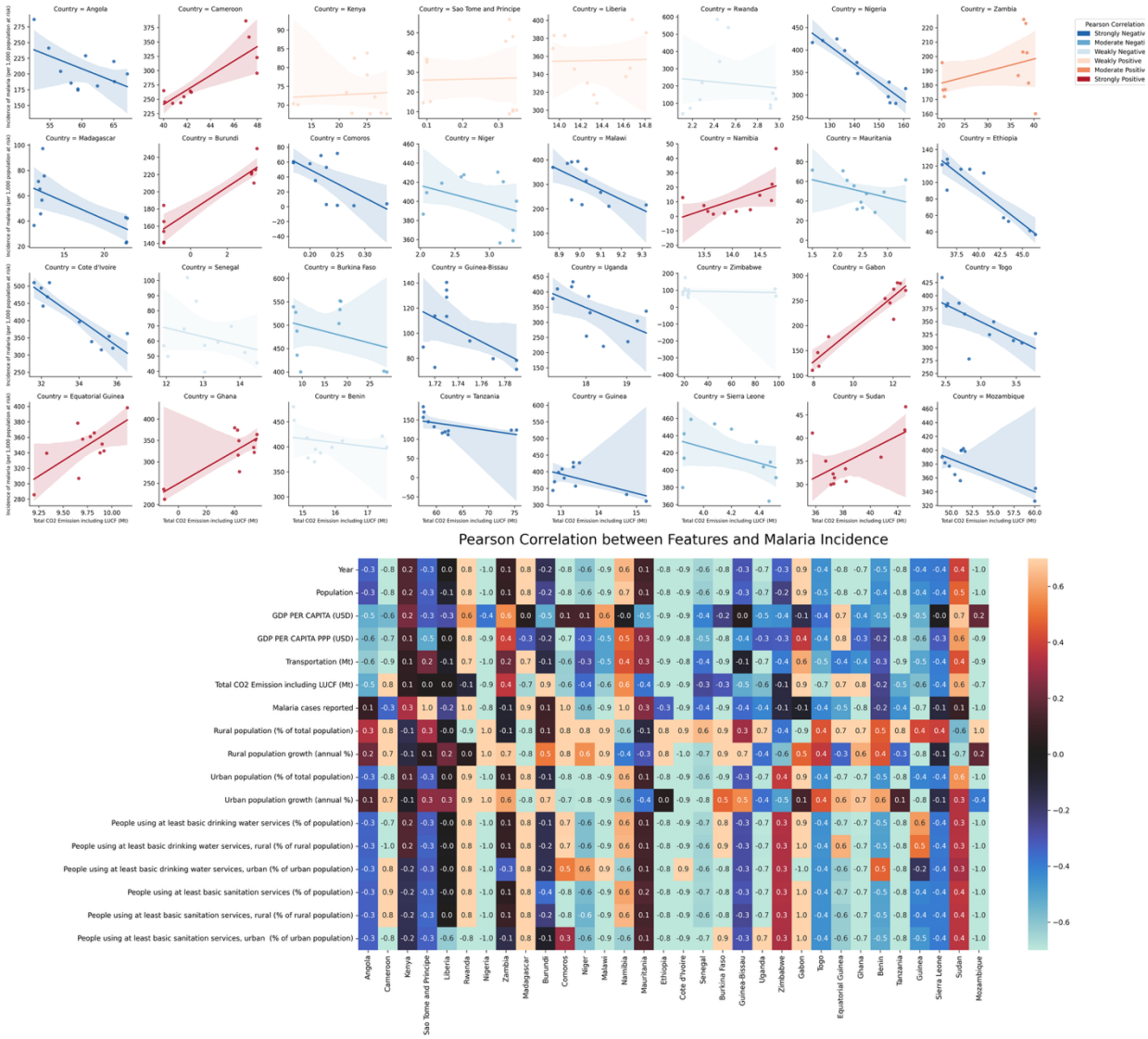


Figure 2. The association between environmental and socioeconomic factors and malaria incidence across African countries. Figure 2A illustrates the country-specific relationships between total $CO_2$ emissions (including LUCF) and malaria incidence across multiple African nations. Figure 2B presents a comprehensive correlation heatmap illustrating the relationships between various socioeconomic and environmental factors with malaria incidence across numerous African countries.

A similar analysis was then conducted on other variables to investigate whether other variables can explain the observed variations in malaria incidence. Figure 2B displays the Pearson correlation between malaria incidence and other variables, such as GDP per capita and access to sanitation, highlighting similarly inconsistent patterns across countries. The Pearson correlation coefficient, which ranges from -1 to 1, quantifies the strength and direction of a

Vol. 2025 (10) 838 – 849
ISSN 2688-3651 [online]

linear association between two variables. A value closer to 1 or -1 indicates a stronger positive or negative linear relationship, respectively. These correlations reveal how closely malaria incidence aligns with a given factor across countries. Not a single variable was consistently positively or negatively correlated with malaria incidence across all analyzed African countries. In particular, gross domestic product (GDP) per capita displayed strong positive correlations (R=0.7) in Sudan and Equatorial Guinea and negative correlations (R=-0.9) in Ethiopia and Côte d'Ivoire. Variables like access to basic sanitation services and rural population percentage also displayed similar ranges of Pearson correlations. The lack of consistency in the direction of correlation coefficients across all countries demonstrated that these features also could not explain the variations in malaria incidence.

As shown in Figure 3A, a multivariate linear regression model (OLS) was constructed using all independent variables to explore their combined effect on malaria incidence across African countries. Ordinary Least Squares (OLS) is a linear method that minimizes the sum of squared differences between actual and predicted values. The model had a moderate predictive power, indicated by a test set Pearson correlation of 0.42. Figure 3B demonstrates the estimated regression coefficients and associated p-values from a Student's t-test, indicating the statistical significance of each variable's relationship with malaria incidence. The p-value reports the probability that the observed effect occurred by random chance. A low p-value (typically $p \leq 0.05$) suggests that the relationship is statistically significant and unlikely due to random variation. Computing the p-value between independent and dependent variables helps identify which independent variables truly have a correlational relationship with malaria transmission dynamics. Regression coefficients represent the estimated change in the dependent variable (malaria incidence) for a one-unit change in the independent variable, assuming other variables are constant. While the regression coefficients provide a direct understanding of whether the feature positively or negatively impacts the malaria transmission, the p-values assess statistical significance and certainty of the estimated relationship between the variables and incidence. Because only half of the variables were statistically significant ($p \leq 0.05$) and only 2 out of 14 independent variables had a p-value less than 0.001, interpreting the direction and magnitude of the regression model inferred coefficients was not very meaningful. I hypothesized that this poor significance was caused by the severe feature redundancy, such as GDP per capita and GDP per capita based on purchasing power parity (PPP).

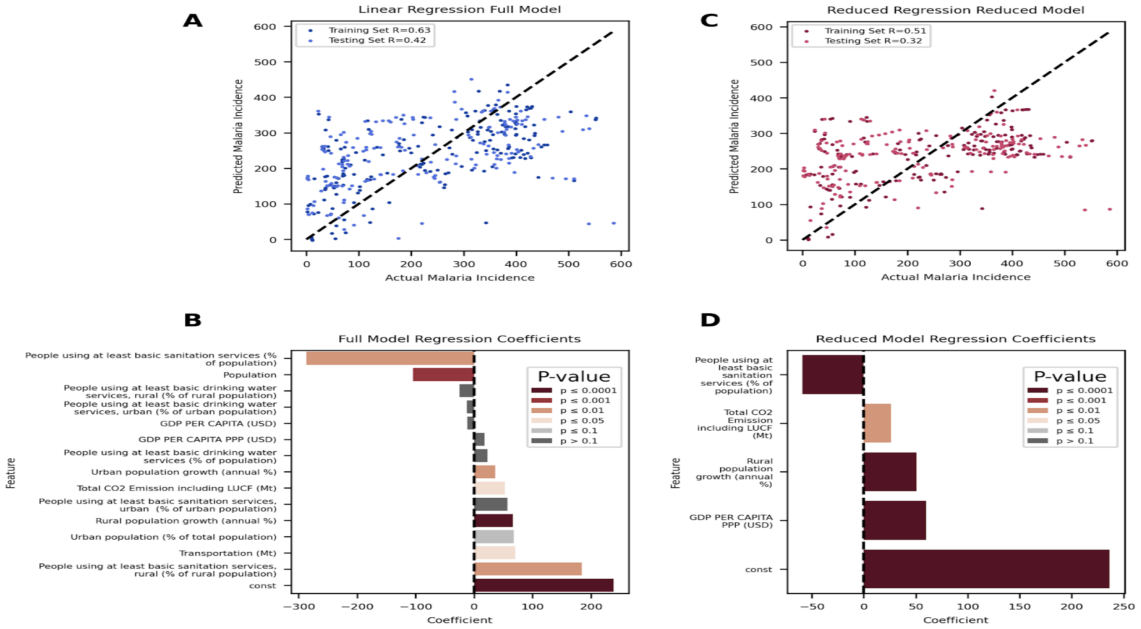


Figure 3. Linear regression fails to capture malaria transmission dynamics. Figure 3A shows the linear regressor only captured a weak association between actual malaria incidence and predicted malaria incidence, with R=0.42 on the testing set. Figure 3B demonstrates a comprehensive and general visualization of regression coefficients colored by statistical significance levels (p-values). Figure 3C presents regression coefficients only for key predictors of malaria incidence, color-coded by their statistical significance (p-values). Figure 3D demonstrates the 4 most important features that were selected through the feature selection routine mentioned in the Materials and Methods section.

Vol. 2025 (10) 838 – 849
ISSN 2688-3651 [online]

Therefore, a forward-backward feature selection routine was implemented. For further analyses, this heuristic recommended only 4 variables, which were people with basic sanitation services, total CO2 emission, rural population growth, and GDP per capita PPP. As shown in Figure 3C, the reduced regression model had a slight decrease in predictive accuracy (test set R dropped from 0.42 to 0.32), while Figure 3D confirms that all four selected variables were statistically significant ($p \leq 0.05$). The model suggested that access to sanitized resources inhibited the spread of malaria, whereas carbon emissions, rural population, and GDP all promoted the spread of this disease. The bias term was also important, with its value around 250, suggesting that many malaria incidence values were around 250 cases out of 1,000 population. The bias term, also known as the intercept, is the baseline value predicted when all input variables are set to zero. In this case, it reflected a baseline malaria incidence of around 250 per 1000 people. This reflected an underlying level of malaria burden that persisted regardless of the values of $CO_2$ emissions, sanitation access, rural population growth, or GDP. The intercept's significant p-value indicated that this baseline was not due to random variation—it had a consistent and meaningful presence in the model. This supported the observation that malaria incidence in many countries clusters around this level and reinforced the idea that even in the absence of strong external drivers, malaria transmission remains an endemic issue in much of Africa. This is further supported by the fact that the predicted malaria incidence largely fell between 200 to 350, even though the actual incidence range was 0 to 500.

Although OLS considered the remaining four independent variables significant, their weak predictive power undermined the predicted relationship between malaria incidence and the features of interest. The previous observation of the linear model's inability to predict a full range of malaria incidence suggested the oversimplicity of the model. Therefore, two tree-based machine learning models were employed. Unlike OLS, tree-based methods are more expressive because they can model nonlinear relationships between input and output variables. As illustrated in Figure 4A, the random forest regression model achieved a test-set correlation of 0.55 between predicted and actual malaria incidence, demonstrating improved performance. Test-set correlation refers to the Pearson correlation between predicted and actual values on unseen data, serving as a measure of the model's predictive accuracy. Despite the improved performance, it was obvious that the model was still limited because many incidence values were predicted to be around 360 cases, yet their actual range was between 150 and 500. Therefore, I further implemented a gradient boosting algorithm. Figure 4B shows that the gradient boosting model further improved predictive performance, reaching a test-set correlation of 0.82, without cases of severe over-fitting.

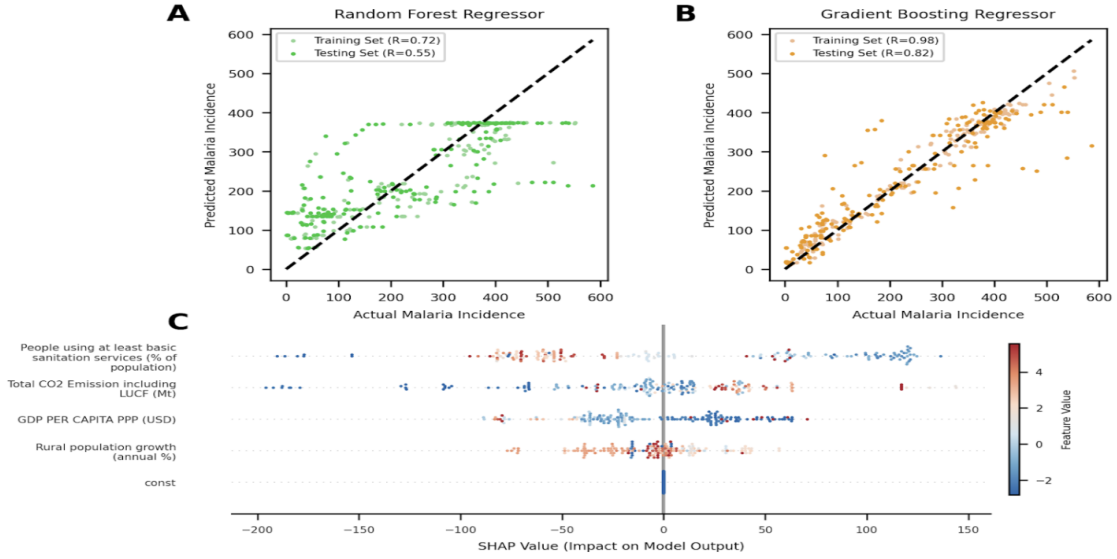


Figure 4. The outcomes of predictive modeling and the key insights derived from employing machine learning on malaria incidence data. Figure 4A presents a comprehensive evaluation of three machine learning approaches for predicting malaria incidence, along with their comparative performance visualization. Figure 4B shows actual malaria incidence vs. predicted malaria incidence value reached in the training set (R=0.72) and the testing set (R=0.55) correlations. Figure 4C shows feature importance from Gradient Boosting regressor for malaria incidence prediction.

Vol. 2025 (10) 838 – 849
ISSN 2688-3651 [online]

To interpret the model's predictions, SHAP analysis was conducted, as illustrated by Figure 4C. This visualization reveals how each feature contributed to individual malaria predictions, highlighting the impact of sanitation, carbon dioxide emissions, GDP, and rural population growth. SHAP values come from game theory and help interpret machine learning models by showing how much each feature contributes to a specific prediction. In this case, a negative SHAP value means the feature helps lower predicted malaria incidence, while a positive value means it increases it. One can see that basic sanitation services were mostly high with a negative SHAP value, suggesting that higher percentages of people with basic sanitation services tend to inhibit the spread of malaria. Furthermore, the total carbon dioxide emission was mostly high with a strong positive SHAP value, supporting the initial hypothesis that increased emissions will lead to heightened malaria transmission. GDP and rural population growth rate had evenly distributed SHAP values between positive and negative, indicating that the effects of these two variables on malaria incidence were not one-sided and more likely to be country-specific.

## 4. Discussion

The study revealed complex trends in malaria incidence across Africa. While rates stayed stable or declined in most countries, total reported cases rose sharply. Regional patterns varied—West Africa had the highest incidence, North Africa the lowest, and East Africa was intermediate. Conflicting correlations with factors like $CO_2$ or GDP showed that no single variable consistently explained transmission. Tree-based models performed well, with gradient boosting achieving a 0.82 Pearson correlation. SHAP analysis showed sanitation lowers malaria risk, while $CO_2$ increases it. GDP and rural population growth showed mixed effects, highlighting malaria as a product of intertwined environmental, demographic, and infrastructure factors.

This study built on prior research to deepen understanding of malaria transmission in Africa. It extended findings from Ermert et al. (2011) and Le et al. (2019), showing that $CO_2$ emissions lack a consistent linear association with malaria incidence. This ML model also identified sanitation access as a strong predictor, supporting Perera et al.'s (2022) claims about infrastructure's role in disease control. In contrast to Braimah et al. (2024), I found GDP's effect on malaria to be inconsistent. Lastly, by using SHAP to interpret complex interactions, this study affirmed Surur et al.'s (2025) view on ML's value in public health research.

This study has several important limitations. First of all, the analysis relied on secondary datasets from Kaggle, which were collected by other researchers rather than through a controlled, long-term study that is manually collected. Thus, while these datasets provide general, broad coverage, they may lack standardized methodologies and criteria across countries. This would potentially introduce biases, such as inconsistent malaria reporting practices in different countries. Secondly, the original dataset contained a certain amount of missing values across variables, countries, and time; the reduction in sample size and potential selection bias, including omitted countries that may have unique environmental or socioeconomic profiles, could affect the generalizability of my findings. Last but not least, the machine learning models were generated under the assumption that observed relationships between the independent variables (i.e., $CO_2$ emissions, GDP, sanitation, etc) and malaria incidence are stable over time. So the dynamic nature and potential changes of climate and disease transmission across these African countries were not considered significantly in this study.

The results had immediate policy relevance. They supported emissions-aware surveillance to detect emerging microclimates and emphasized prioritizing sanitation investments in high-burden areas. They also highlighted how ML tools (Panch et al., 2018) can optimize malaria intervention targeting. Addressing gaps identified by Pryce et al. (2022), this study developed a framework combining ML and SHAP-based interpretation to assess how environmental and socioeconomic factors—$CO_2$ emissions, sanitation, GDP, and rural population growth—influence malaria incidence across Africa. The framework captured complex interactions and revealed region-specific drivers of transmission. This marked a key advance in malaria control planning using explainable models.

Multiple analytical methods converged to support a positive correlation between $CO_2$ emissions and malaria incidence. SHAP analysis showed positive SHAP values for $CO_2$, and linear regression indicated a statistically significant positive coefficient. In contrast, GDP and rural population growth exhibited context-dependent effects varying by region. Correlation analysis highlighted $CO_2$'s inconsistent regional impact: in tropical forest nations like

Vol. 2025 (10) 838 – 849
ISSN 2688-3651 [online]

Gabon, emissions likely altered microclimates, boosting mosquito survival and showing strong positive correlations; in arid countries like Niger, where environmental conditions already restrict mosquito populations, correlations were weak or negative, regardless of $CO_2$ levels.

Sanitation access consistently emerged as a protective factor across all analytical techniques, being negatively associated with malaria incidence, regardless of the country. Although these data-driven trends are clear, targeted field experiments are needed to uncover the biological mechanisms—particularly how emission-induced vegetation changes (Le et al., 2019) interact with vector life cycles. Future research will examine how $CO_2$ emissions influence malaria via (1) field studies comparing mosquito populations and malaria cases in high vs. low emission areas, and (2) lab experiments testing how elevated $CO_2$ affects mosquito breeding and survival. These experiments will validate model findings and guide integrated control strategies.

## 5. Conclusion

Overall, the findings revealed that malaria transmission in Africa was shaped by complex environmental and socioeconomic factors. $CO_2$ emissions are positively linked to incidence—especially in tropical regions like Cameroon and Gabon—while access to sanitation consistently reduces risk. Gradient Boosting models ($R = 0.82$) outperformed linear models by capturing non-linear patterns. Regional differences, particularly the higher burden in West/Central Africa, underscored the need for country-specific strategies. Effective malaria control under climate change will require integrated approaches combining emissions monitoring with improved sanitation. Further research should explore how $CO_2$ impacts vector ecology.

## References

Al-Hajjar, S. (2024). Breakthroughs in artificial intelligence for combating infectious diseases. *International Journal of Pediatrics and Adolescent Medicine, 11*(3), 55–57. https://doi.org/10.4103/ijpam.ijpam_101_24

Beloconi, A., et al. (2023). Malaria, climate variability, and interventions: Modelling transmission dynamics. *Scientific Reports, 13*, 7367. https://doi.org/10.1038/s41598-023-33868-8

Braimah, J. O., et al. (2024). The fight against malaria in Edo-North, Edo State, Nigeria: identifying risk factors for effective control. *PeerJ*, *12*, e18301. https://doi.org/10.7717/peerj.18301

Caminade, C., McIntyre, K. M., & Jones, A. E. (2019). Impact of recent and future climate change on vector-borne diseases. *Annals of the New York Academy of Sciences*, *1436*(1), 157–173. https://doi.org/10.1111/nyas.13950

Castro M. C. (2017). Malaria Transmission and Prospects for Malaria Eradication: The Role of the Environment. *Cold Spring Harbor perspectives in medicine*, *7*(10), a025601. https://doi.org/10.1101/cshperspect.a025601

Costantini, C., et al. (1996). Mosquito responses to carbon dioxide in a west African Sudan savanna village. *Medical and veterinary entomology*, *10*(3), 220–227. https://doi.org/10.1111/j.1365-2915.1996.tb00734.x

DataCamp. (2022, May 23). *Introduction to SHAP values for machine learning interpretability*. https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability

Ermert, V., et al. (2012). The impact of regional climate change on malaria risk due to greenhouse forcing and land-use changes in tropical Africa. *Environmental health perspectives*, *120*(1), 77–84. https://doi.org/10.1289/ehp.1103681

Fatima, S. H., et al. (2025). Impact of temperatures on malaria incidence in vulnerable regions of Pakistan: empirical evidence and future projections. *Epidemiology and infection*, *153*, e33. https://doi.org/10.1017/S0950268825000111

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189–1232.

Vol. 2025 (10) 838 – 849
ISSN 2688-3651 [online]

Harris, C. R., et al. (2020). Array programming with NumPy. *Nature, 585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Himeidan, Y. E., & Kweka, E. J. (2012). Malaria in East African highlands during the past 30 years: impact of environmental changes. *Frontiers in physiology*, *3*, 315. https://doi.org/10.3389/fphys.2012.00315

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278–282).

Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment*. *Computing in Science & Engineering, 9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Institute of Medicine (US) Committee for the Study on Malaria Prevention and Control. (1991). *Malaria: Obstacles and opportunities* (S. C. Oaks Jr., V. S. Mitchell, G. W. Pearson, et al., Eds.). National Academies Press. https://www.ncbi.nlm.nih.gov/books/NBK234322/

Institute of Medicine (US) Committee on Guidance for Designing a National Healthcare Disparities Report. (2002). *Guidance for the national healthcare disparities report* (E. K. Swift, Ed.). National Academies Press. https://www.ncbi.nlm.nih.gov/books/NBK221045/

Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1705.07874

Lydia70. (n.d.). *Malaria in Africa* [Data set]. Kaggle. https://www.kaggle.com/datasets/lydia70/malaria-in-africa

MalariaWorld. (2024, August 20). *Shifting landscape: Climate change's impact on malaria*. https://www.malariaworld.org/blogs/shifting-landscape-climate-change-s-impact-on-malaria

Nabi, S., & Qader, S. (2009). Is Global Warming likely to cause an increased incidence of Malaria?. *The Libyan journal of medicine*, *4*(1), 18–22. https://doi.org/10.4176/090105

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

Paaijmans, K. P., Read, A. F., & Thomas, M. B. (2009). Understanding the link between malaria risk and climate. *Proceedings of the National Academy of Sciences of the United States of America, 106*(33), 13844–13849. https://doi.org/10.1073/pnas.0903423106

Panch, T., Szolovits, P., & Atun, R. (2018). Artificial intelligence, machine learning and health systems. *Journal of global health*, *8*(2), 020303. https://doi.org/10.7189/jogh.08.020303

Pryce, J., Medley, N., & Choi, L. (2022). Indoor residual spraying for preventing malaria in communities using insecticide-treated nets. *The Cochrane database of systematic reviews*, *1*(1), CD012688. https://doi.org/10.1002/14651858.CD012688.pub3

Ricci F. (2012). Social implications of malaria and their relationships with poverty. *Mediterranean journal of hematology and infectious diseases*, *4*(1), e2012048. https://doi.org/10.4084/MJHID.2012.048

Surur, F. M., et al. (2025). *Unlocking the power of machine learning in big data: A scoping survey*. *Data Science and Management*. Advance online publication. https://doi.org/10.1016/j.dsm.2025.02.004
Sutherst R. W. (2004). Global change and human vulnerability to vector-borne diseases. *Clinical microbiology reviews*, *17*(1), 136–173. https://doi.org/10.1128/CMR.17.1.136-173.2004

Takken, W., Charlwood, D., & Lindsay, S. W. (2024). The behaviour of adult Anopheles gambiae, sub-Saharan Africa's principal malaria vector, and its relevance to malaria control: a review. *Malaria journal*, *23*(1), 161. https://doi.org/10.1186/s12936-024-04982-3

Vol. 2025 (10) 838 – 849
ISSN 2688-3651 [online]

The pandas development team. (2020). *pandas-dev/pandas: Pandas* (Version latest) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.3509134

Virtanen, P., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods, 17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Waskom, M. L. (2021). *Seaborn: Statistical data visualization*. *Journal of Open Source Software, 6*(60), 3021. https://doi.org/10.21105/joss.03021

World Bank. (2025, January 24). *Incidence of malaria (per 1,000 population at risk)*. *World Development Indicators*. Retrieved July 26, 2025, from https://databank.worldbank.org/reports.aspx?source=2&series=SH.MLR.INCD.P3&country=SSF

World Health Organization. (n.d.). *Malaria*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/malaria

World Health Organization. (2021, July 19). *Global technical strategy for malaria 2016–2030: 2021 update*. World Health Organization. https://www.who.int/publications/i/item/9789240031357

World Health Organization. (2023, October 12). *Climate change*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/climate-change-and-health