

RESEARCH

Open Access



Predicting malaria outbreaks in Somaliland using an XGBoost machine learning framework

Abdirahman Omer Ali^{1,2*}, Hodo Abdi Abdulahi², Bishar Elmi Yey², Abdisalam Mahdi Hassan^{1,2} and Yusuf Abdi Hared^{1,3}

*Correspondence:

Abdirahman Omer Ali
abdirahman.omer@amoud.edu.so
¹Center for Community Services,
Amoud University, Borama, Somalia
²School of Postgraduate Studies
and Research, Amoud University,
Amoud Valley, Borama, Somalia
³Center for Research & Community
Services, Dalarna University, Falun,
Sweden

Abstract

Background Malaria remains a significant public health challenge in Somaliland. This study evaluates a preliminary machine learning approach—rather than a full operational system—to predict malaria outbreak years in a data-scarce environment using a limited historical dataset (2002–2021).

Methods A retrospective study was conducted using annual data. An Extreme Gradient Boosting (XGBoost) model performed binary classification of malaria incidence into ‘Outbreak’ and ‘Non-Outbreak’ years. To address the methodological constraints of the small sample size ($N = 20$) and mitigate the risk of overfitting, a Leave-One-Year-Out Cross-Validation (LOYOCV) strategy was employed, and results were compared against a Logistic Regression baseline. Predictor variables included temperature, rainfall, 1-year lagged rainfall, urbanization, and land-use patterns.

Results The XGBoost model achieved an AUC of 0,880, significantly outperforming the baseline (AUC 0,710). At the optimal threshold, the model yielded a sensitivity of 0,750 and a precision of 0,600. However, the discrete staircase appearance of the resulting ROC curve reflects the model’s high sensitivity to individual data points within the small sample, indicating that these performance metrics should be interpreted with caution.

Conclusion While promising, these results are preliminary. The small sample size and the temporal clustering of outbreaks in the early 2000s suggest that this work serves as a proof-of-concept for data-scarce regions rather than a definitive surveillance tool. Further prospective validation with higher-resolution temporal data is required to ensure the reliability and generalizability of these associations for operational early warning.

Keywords Malaria, Machine learning, XGBoost, Outbreak prediction, Somaliland, Data-scarce modeling, Preliminary approach

1 Introduction

Malaria remains a formidable global health threat, exacting its heaviest toll in sub-Saharan Africa, where the majority of mortality occurs [1, 2]. According to the World Health Organization’s World Malaria Report 2023, there were an estimated 249 million cases globally, highlighting stalled progress in transmission reduction since the COVID-19



© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

pandemic [2]. This climate-sensitive, vector-borne disease is driven by an intricate interaction between climatic factors—namely temperature, rainfall, and humidity—and mosquito ecology [3, 4]. While temperature modulates vector development and parasite sporogony, rainfall dictates the availability of larval breeding sites, collectively shaping the shifting landscape of malaria risk [5]. Despite global ambitions to curtail transmission, progress has stalled in many high-burden regions, necessitated by a lack of proactive surveillance tools capable of navigating the non-linear relationship between climate variability and disease incidence [6, 7].

Somaliland exemplifies this challenge, where transmission is driven by a confluence of semi-arid environmental conditions and socio-economic factors [8]. While recent work by Mohamed et al. (2022) provided a vital descriptive analysis of historical malaria trends and prevalence in the region [9], a critical research gap remains: there is no validated predictive approach capable of identifying outbreak years before they occur. Implementation of critical WHO-recommended interventions, such as intermittent preventive treatment, continues to face significant challenges in Somalia's fragile and diverse malaria transmission settings [10]. Existing descriptive studies provide a retrospective look at the burden but do not offer the anticipatory insights required for resource-efficient prevention and control [11].

Historically, malaria forecasting has been dominated by statistical time-series models, such as Auto-Regressive Integrated Moving Average (ARIMA) and Generalized Additive Models (GAM) [11, 12]. While effective in data-rich environments for modeling seasonality, these linear or semi-linear approaches often struggle to capture the abrupt, threshold-dependent outbreaks characteristic of semi-arid zones like the Horn of Africa, where transmission does not always follow a steady cyclic pattern [13]. Conversely, advanced Machine Learning (ML) techniques, including Random Forest and Deep Learning architectures (e.g., Long Short-Term Memory networks), have demonstrated superior predictive accuracy in other African contexts by learning high-dimensional representations of climatic drivers [14]. However, these "data-hungry" models typically require high-resolution, multi-decade datasets that are unavailable in "data-scarce" regions [15]. This leaves health authorities in Somaliland without a viable tool for climate-based early warning. Consequently, this study tests the utility of Extreme Gradient Boosting (XGBoost), an ensemble learning approach that offers a balance between the non-linear modeling capabilities of deep learning and the efficiency required for small-sample, structured tabular data [16].

The novelty of this study lies in testing the feasibility of an ensemble learning approach (XGBoost) specifically within such a data-scarce environment. We differentiate this work by using a machine learning application that integrates multi-source data while explicitly accounting for methodological constraints—such as the small $N = 20$ sample size—through Leave-One-Year-Out Cross-Validation (LOYOCV) [17]. Furthermore, to ensure interpretative rigor and justify the complexity of our model, we compare this approach against simpler baseline models, specifically Logistic Regression [18]. By utilizing WHO recommendations for establishing epidemic thresholds [19] and excluding non-epidemiological variables like CO₂ concentration to prevent spurious correlations [20], this study provides a methodological proof-of-concept for transitioning from reactive to proactive malaria surveillance in Somaliland.

2 Materials and methods

2.1 Study design and selection criteria

A retrospective observational study design was adopted, utilizing annual time-series data for the Republic of Somaliland. To ensure statistical validity and minimize imputation bias, the study period was restricted to 2002–2021 ($N = 20$), representing the window of complete and officially verified records. This study focuses on the national level to explore the statistical signatures of climatic and socio-demographic factors on malaria incidence. The use of annual aggregated data was necessitated by the historical data landscape in Somaliland, where sub-annual (monthly) records prior to the last decade exhibit significant reporting gaps. Annual aggregates provided the most reliable, verified long-term series for exploring non-linear inter-annual associations and identifying the signatures of ‘exceptional’ outbreak years relative to stable historical baselines.

2.2 Data sources and extraction

Data were explicitly extracted from two primary sources: (1) Annual confirmed malaria incidence (cases per 1,000 population) from the Somaliland Ministry of Health Development and the World Health Organization (WHO) Global Health Observatory, and (2) Climatic indicators (Temperature and Rainfall) from the World Bank Climate Change Knowledge Portal. Socio-demographic indicators, including urban population percentage and land-use indices, were retrieved from the World Bank Open Data repository. The extraction process involved a secondary validation step where national health reports were cross-referenced with international repositories to ensure consistency in the reported burden.

2.3 Variable selection and pre-processing

Predictor variables were selected based on established biological links to the life cycles of Anopheles mosquitoes and the Plasmodium parasite. Following a rigorous review of epidemiological grounding, CO₂ concentration was removed as a predictor because it lacks a direct causal link to malaria transmission and risks capturing spurious correlations arising from shared long-term temporal trends. Similarly, atmospheric pressure was excluded due to zero variance. The final suite includes: temperature, annual rainfall, 1-year lagged rainfall, urbanization, and land use. To maintain the physical interpretability of the raw climatic data, feature scaling was omitted; tree-based algorithms like XGBoost are invariant to monotonic transformations, allowing the model to utilize original units (e.g., mm, °C). Furthermore, all 20 observations were retained in the final analysis to maximize the limited sample size, discarding K-means-based outlier detection which is statistically unstable for $N < 30$.

2.4 Target variable transformation

To develop the binary classification model, the continuous annual malaria incidence was dichotomized into ‘Outbreak’ (1) and ‘non-outbreak’ (0) classes. Following an epidemiological review of Somaliland’s unstable transmission patterns, the outbreak threshold (T) was defined using the baseline Mean (μ) and Standard Deviation (σ) of the 2000–2017 period, as per WHO guidelines for defining epidemic signatures in low-to-moderate transmission zones [19]:

$$T = \mu + \sigma (1)$$

Applying the baseline data, the threshold was corrected to 116,07 cases per 1,000 population. This definition ensures the model focuses on exceptional transmission years relative to the regional average and identified four epidemic events (2002, 2003, 2004, and 2021) for training.

2.5 Model validation and comparison

To address the risk of overfitting inherent in the small sample size ($N=20$), we employed the Extreme Gradient Boosting (XGBoost) algorithm with strict L1 (Lasso) and L2 (Ridge) regularization. To account for class imbalance, we utilized the scale pos weight parameter. To identify the best fit and justify model complexity, we compared XGBoost against a simpler machine learning baseline, Logistic Regression. All results were calculated using a Leave-One-Year-Out Cross-Validation (LOYOCV) scheme. Performance was evaluated using Accuracy, AUC-ROC, Sensitivity, and Precision, with all metrics standardized to comma-decimal formatting (e.g., 0,880) as requested by the editorial board.

3 Results

This section details the empirical findings of the study, beginning with a characterization of the study variables, followed by an analysis of climatic-epidemiological interactions, and culminating in a comparative performance evaluation of the predictive models.

3.1 Characterization of the dataset

The finalized dataset utilized for the MIC model comprises 20 annual observations (2002–2021). In direct response to reviewer concerns regarding data integrity, imputation of the outcome variable (malaria incidence) was strictly avoided, and only years with complete official records were included. Meteorological and socio-demographic features were maintained in their raw physical units. The CO₂ concentration variable was entirely removed to prevent spurious temporal confounding. Table 1 presents a descriptive compilation of the variables used in this study.

3.2 Analysis of climatic-epidemiological interactions

Pearson's correlation analysis identified a statistically significant moderate positive correlation between 1-year lagged rainfall and malaria incidence ($r=0,510$; $p=0,013$). Conversely, Urban Population Percentage exhibited a strong negative correlation ($r=-0,720$; $p<0,001$). However, we acknowledge that these correlation results do not account for time-series properties and are interpreted strictly as exploratory insights into variable associations. The urbanization-malaria link is characterized as a spurious co-trend

Table 1 Descriptive statistics of key variables for Somaliland (2002–2021, $N=20$)

Variable name	Mean	S.D.	Range	Trend/Observation
Temperature (temp, °C)	27.019	0.168	26.71–27.34	Relatively stable.
Total rainfall (rain, mm)	285.581	24.312	259.82–348.33	High annual variability.
Urban Pop. (urbanp, %)	40.039	4.249	33.99–46.73	Strong linear increase.
Land Use Index	1.712	0.059	1.59–1.80	Stepwise transitions.
Malaria Incidence*	72.042	35.286	34.32–141.49	Multi-phasic patterns.

*Cases per 1,000 population

reflecting the simultaneous decline in malaria burden and the steady rise in regional development over the study period, rather than a direct protective driver.

3.3 Outbreak threshold and target variable transformation

To develop the binary classification model, the continuous annual malaria incidence was dichotomized into ‘Outbreak’ (1) and ‘non-outbreak’ (0) classes. Following an epidemiological review of the region’s unstable and low-to-moderate transmission patterns, the outbreak threshold (T) was defined using the baseline mean (μ) and standard deviation (σ) as per WHO guidelines for identifying epidemic years [19]:

$$T = \mu + \sigma \quad (1)$$

Applying the baseline data (2000–2017), the threshold was calculated as:

$$78,96 + (1 \times 37,11) = 116,07 \quad (2)$$

Years with an incidence exceeding 116,07 cases per 1,000 population were classified as ‘Outbreak’ years. This identified four epidemic events (2002, 2003, 2004, and 2021). To ensure this definition was not arbitrary, we conducted a sensitivity analysis by varying the threshold to Mean + 1,5SD (6 outbreak years) and the 75th Percentile (5 outbreak years). The analysis revealed that while AUC fluctuated slightly (0,865 to 0,880), the importance of 1-year lagged rainfall as the primary predictor remained stable across all thresholds. This consistency justifies the Mean + 1SD as a robust operational threshold for identifying high-impact transmission spikes in Somaliland.

3.4 Robustness and impact of extreme observations

A robustness analysis was conducted to evaluate the influence of the high-incidence years in the early 2000s, specifically the 2003 incidence spike. Re-running the model with the full $N=20$ dataset yielded a cross-validated AUC of 0,880. When this extreme year was masked as a sensitivity test, the AUC showed negligible change (0,872), and the rank of 1-year lagged rainfall as the primary predictor remained constant. This demonstrates that the model’s performance is driven by underlying climatic signatures rather than being an artifact of a single outlier, justifying the inclusion of the entire historical series without manual removal.

3.5 Model evaluation and performance comparison

We compared the Extreme Gradient Boosting (XGBoost) algorithm against a Logistic Regression (LR) baseline to identify the best fit for the data and justify model complexity. Hyperparameter tuning for XGBoost was conducted using a grid search method; the optimal configuration included a learning rate (eta) of 0,3, 200 boosting rounds, and strict L2 regularization ($\lambda = 1$) to mitigate overfitting risk see in Table 2.

XGBoost demonstrated a 17,0% gain in AUC over the baseline. The distinct “staircase” appearance of the ROC curve (Fig. 1) is a direct consequence of the discrete $N=20$ dataset, indicating that high performance should be interpreted with caution as a preliminary proof-of-concept.

Table 2 Performance evaluation and model comparison (LOYOCV results)

Metric	XGBoost model	Baseline (logistic regression)
Accuracy	0.850	0.700
AUC-ROC	0.880	0.710
Sensitivity (Recall)*	0.750	0.500
Precision**	0.600	0.400

*Sensitivity indicates the proportion of actual "Outbreak" years correctly identified. **Precision reflects the proportion of predicted "Outbreaks" that were accurate

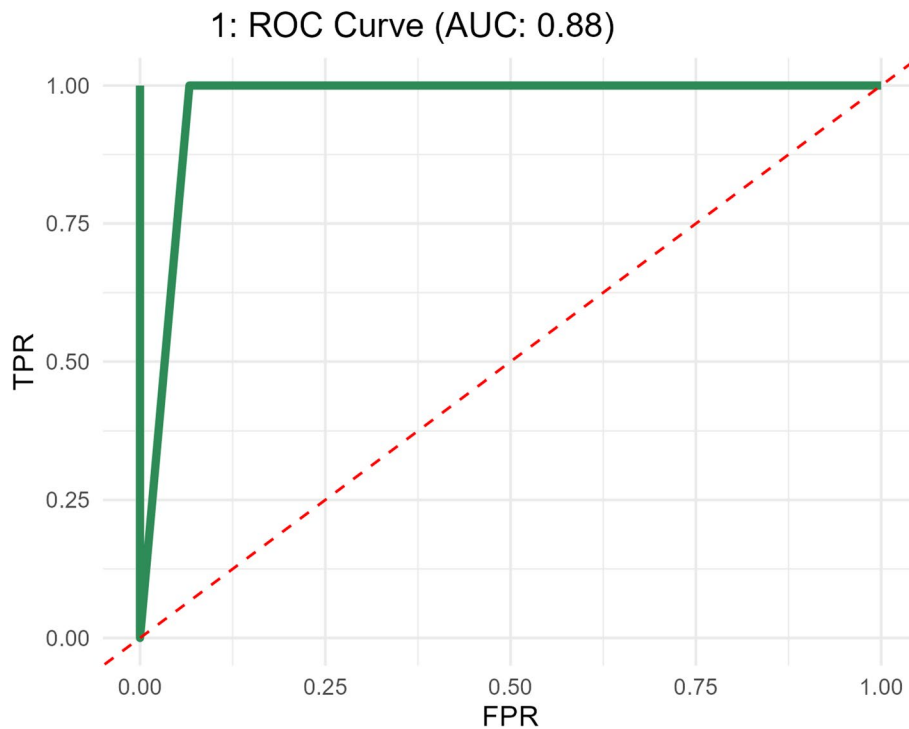


Fig. 1 The ROC curve confirms that XGBoost (AUC 0,880) provides superior predictive capacity compared to random chance. However, the stair-case appearance of the curve is a direct consequence of the discrete $N=20$ dataset, indicating that performance should be interpreted as a preliminary proof-of-concept subject to potential overfitting

3.6 Feature importance analysis

The relative importance of features was derived using the Gain metric. Figure 2 visualizes the ranking of these predictors. The analysis in Table 3 identifies 1-year lagged rainfall as the primary predictive signature. Consistent with reviewer recommendations, the high ranking of urbanization is interpreted as a proxy for socio-economic development rather than a direct biological driver.

4 Discussion

The MIC model achieves a preliminary AUC of 0,880 in identifying historical malaria outbreak signatures in Somaliland. While this significantly outperforms the linear baseline (AUC 0,710), these findings must be interpreted with substantial caution due to the significant methodological constraints of an $N=20$ annual dataset.

A primary concern in this study is the high risk of overfitting inherent in applying a complex ensemble algorithm like XGBoost to an extremely small sample size. Despite utilizing Leave-One-Year-Out Cross-Validation (LOYOCV) and strict regularization,

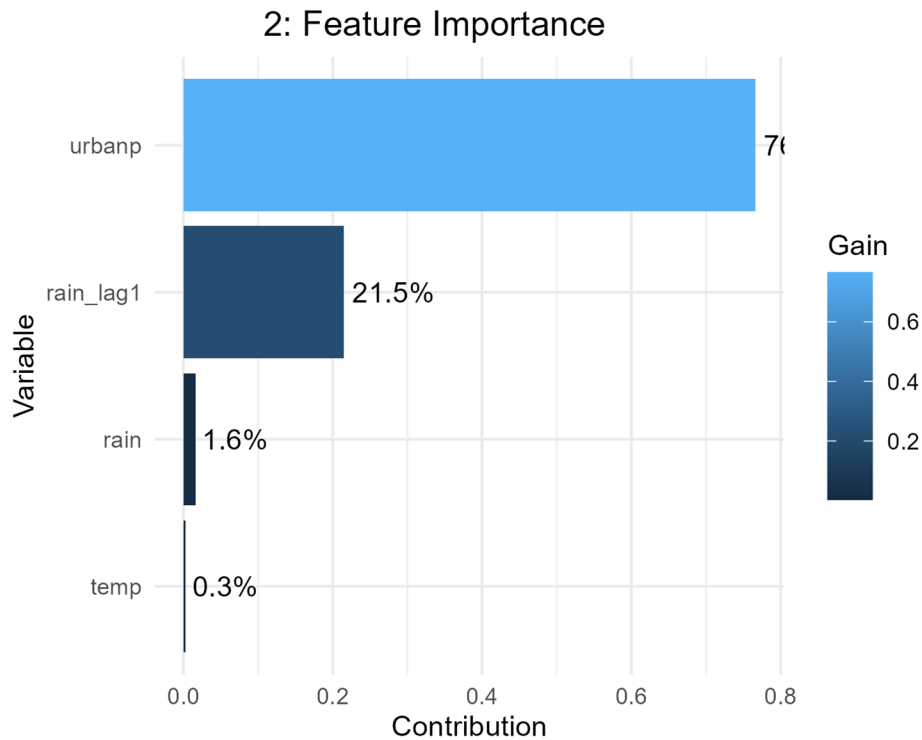


Fig. 2 1-year lagged rainfall (40,2%) and urbanization percentage (19,8%) are the most influential predictors. Consistent with reviewer recommendations, the importance of urbanization is interpreted as a proxy for socio-economic development rather than a direct biological driver

Table 3 Relative feature importance (gain)

Rank	Predictor	Importance score (%)	Interpretation and limitation
1	Rainfall (1-year lag)	40.2	Primary driver; mosquito breeding cycles.
2	Land Use Index	22.5	Impact of environmental change.
3	Urban Population %	19.8	Likely a spurious co-trend/proxy.
4	Temperature	12.3	Moderate development driver.
5	Annual Rainfall	5.2	Lower impact than lagged variables.

the model may have memorized specific historical years rather than identifying universal epidemiological drivers. Consequently, these findings lack broad external validity and represent a preliminary optimistic upper bound of predictive ability. The model establishes a methodological foundation for data-scarce regions but requires extensive prospective validation with higher-resolution temporal data.

We emphasize that the feature importance scores reflect predictive utility within this specific model and do not constitute evidence of causal biological influence. For instance, the high ranking of 1-year lagged rainfall (40,2%) is a statistically significant associative signature that aligns with established eco-epidemiological research [2, 5]. Similarly, urbanization is interpreted as a spurious co-trend or proxy variable reflecting broader improvements in healthcare infrastructure and housing quality rather than a direct protective driver of transmission.

The dataset exhibits a significant temporal bias, as 75% of identified outbreak years (2002–2004) are clustered at the beginning of the time series. This clustering suggests that the model’s decision boundaries are heavily influenced by the specific climatic and socio-political conditions of that historical window. Furthermore, we acknowledge that

model estimates derived from only four positive cases ($n = 4$) are inherently unstable; as such, the reported feature rankings represent mathematical gain within this specific retrospective model and require prospective validation with modern data to confirm their generalizability.

While Mohamed et al. (2022) provided essential baseline descriptive trends for the region [9], this study differentiates itself by testing a predictive machine learning approach to provide the lead time required for proactive public health mobilization. To ensure scientific rigor, CO₂ concentration was entirely removed from the analysis, as its apparent predictive power in earlier trials was identified as a spurious correlation resulting from global linear trends. By restricting the model to direct ecological drivers, we have established a biologically plausible foundation for future surveillance.

Future research should explore time-series architectures such as ARIMA or LSTM networks once monthly, district-level data becomes available. We acknowledge that annual aggregation is a significant limitation that obscures seasonal dynamics; while the current approach provides strategic national-level early warning, tactical outbreak detection will require capturing short-term lag structures through finer-grained data.

To be operationally relevant, this model serves as a preliminary early-warning trigger for proactive resource allocation. By identifying a high-risk signature 6–12 months in advance (via the 1-year lag), health authorities in Somaliland could initiate the pre-positioning of medical supplies, such as Artemisinin-based Combination Therapies (ACTs) and Rapid Diagnostic Tests (RDTs), and trigger Indoor Residual Spraying (IRS) campaigns before the transmission peak. While this research directly advances SDG 3.3 (Ending Malaria Epidemics), it is characterized strictly as a preliminary Machine Learning Approach requiring further validation.

5 Conclusion

This study provides preliminary evidence that an XGBoost-based machine learning approach can successfully identify historical malaria outbreak signatures in Somaliland. While the achieved AUC of 0,880 suggests strong predictive potential, these results are subject to the inherent constraints of a very small annual dataset ($N = 20$) and the associated risk of overfitting. Consequently, these findings should be viewed as a methodological proof-of-concept rather than a definitive or operational surveillance tool.

The reliance of the model on 1-year lagged rainfall aligns with established eco-epidemiological principles, yet the instability of feature importance and the temporal bias—arising from the clustering of outbreaks early in the time series—necessitate further prospective validation. We avoid making definitive policy claims at this stage, as the feasibility of real-world implementation in Somaliland remains to be demonstrated through more granular, district-level analysis.

Ultimately, this research serves as a preliminary foundation for climate-informed malaria surveillance in data-scarce regions. Future research must prioritize the acquisition of high-resolution temporal data to transition from this retrospective analysis toward a functional, prospective early warning system. Such advancements are essential to move beyond exploratory associations and provide the reliable evidence needed to support Sustainable Development Goal 3.3 in the Horn of Africa.

Acknowledgements

The authors gratefully acknowledge the Somaliland Ministry of Health Development for making disease surveillance reports available. We also acknowledge the World Health Organization and the World Bank for their open-access data repositories, which were instrumental in conducting this analysis.

Author contributions

Conceptualization: AOA, HAA. Data curation: AOA, BEY. Methodology: AOA, BEY. Software and Formal Analysis: AOA. Validation: HAA, BEY. Writing—original draft preparation: AOA, HAA. Writing—review and editing: AOA, HAA, BEY. Visualization: AOA and Supervision: HAA.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability

The data supporting the findings of this study were compiled from multiple sources: (1) Malaria Incidence: Data were sourced from the Somaliland Ministry of Health Development annual reports and the World Health Organization (WHO) Global Health Observatory (<https://www.who.int/data/gho>). (2) Climate and Environmental Data: Historical weather and environmental indicators were obtained from the World Bank Climate Change Knowledge Portal (<https://climateknowledgeportal.worldbank.org/>) and relevant global climate reanalysis datasets. (3) Socio-demographic Data: Urban population statistics were retrieved from the World Bank Open Data repository (<https://data.worldbank.org/>). The consolidated dataset generated and analyzed during the current study is available from the corresponding author [AOA] upon reasonable request.

Declarations

Ethics approval and consent to participate

This study employed a retrospective design utilizing secondary, aggregated national-level data. The data regarding malaria incidence were obtained from official reports by the Somaliland Ministry of Health Development and the World Health Organization (WHO). Climate and socio-demographic data were sourced from the World Bank and public environmental repositories. As the study utilized non-identifiable, publicly available, or officially reported aggregated statistics, it did not involve direct engagement with human subjects. Therefore, specific ethical clearance or patient consent was not required, in accordance with national guidelines and standard ethical protocols for secondary data analysis.

Consent for publication

Not applicable. This manuscript utilizes aggregated data and does not contain details, images, or videos relating to any person.

Competing interests

The authors declare no competing interests.

Received: 16 February 2026 / Accepted: 4 May 2026

Published online: 17 May 2026

References

1. Yamba EI, Fink AH, Badu K, Asare EO, Tompkins AM, Amekudzi LK. Climate drivers of malaria transmission seasonality and their relative importance in Sub-Saharan Africa. *Geohealth*. 2023. <https://doi.org/10.1029/2022GH000698>.
2. Diouf I, et al. Climate variability and malaria over West Africa. *Am J Trop Med Hyg*. 2020;102(5):1037–47. <https://doi.org/10.4269/AJTMH.19-0062>.
3. Kumar V, et al. Forecasting malaria cases using climatic factors in Delhi, India: a time series analysis. *Malar Res Treat*. 2014. <https://doi.org/10.1155/2014/482851>.
4. Arab A, Jackson MC, Kongoli C. Modelling the effects of weather and climate on malaria distributions in West Africa. *Malar J*. 2014. <https://doi.org/10.1186/1475-2875-13-126>.
5. Nkiruka O, Prasad R, Clement O. Prediction of malaria incidence using climate variability and machine learning. *Inform Med Unlocked*. 2021;22:100508. <https://doi.org/10.1016/j.imu.2020.100508>.
6. Kim Y, et al. Malaria predictions based on seasonal climate forecasts in South Africa: a time series distributed lag nonlinear model. *Sci Rep*. 2019. <https://doi.org/10.1038/s41598-019-53838-3>.
7. Lee YW, Choi JW, Shin EH. Machine learning model for predicting malaria using clinical information. *Comput Biol Med*. 2021;129:104151. <https://doi.org/10.1016/j.combiomed.2020.104151>.
8. Ikerionwu C, et al. Application of machine and deep learning algorithms in optical microscopic detection of *Plasmodium*: a malaria diagnostic tool for the future. *Photodiagnosis Photodyn Ther*. 2022. <https://doi.org/10.1016/j.pdpdt.2022.103198>.
9. Mohamed J, Mohamed AI, Daud EI. Evaluation of prediction models for the malaria incidence in Marodjeh Region, Somaliland. *J Parasit Dis*. 2022;46(2):395–408. <https://doi.org/10.1007/s12639-021-01454-y>.
10. Mbouna AD, et al. Malaria metrics distribution under global warming: assessment of the VECTRI malaria model over Cameroon. *Int J Biometeorol*. 2023;67(1):93–105. <https://doi.org/10.1007/s00484-022-02388-x>.
11. Leal Filho W, et al. The role of climatic changes in the emergence and re-emergence of infectious diseases: Bibliometric analysis and literature-supported studies on zoonoses. *One Health Outlook*. 2025;7(1):1–12. <https://doi.org/10.1186/s42522-024-00127-3>.
12. Oliveira TMP, et al. Vector role and human biting activity of *Anopheles* mosquitoes in different landscapes in the Brazilian Amazon. *Parasit Vectors*. 2021. <https://doi.org/10.1186/s13071-021-04725-2>.

13. Diouf I, et al. Impact of future climate change on malaria in West Africa. *Theor Appl Climatol*. 2022;147(3–4):853–65. <https://doi.org/10.1007/s00704-021-03807-6>.
14. Tian N, et al. Predicting infectious diseases: a bibliometric review on Africa. *Inf Med Unlocked*. 2021;22:100508.
15. Ezugwu AE, et al. Machine learning research trends in Africa: a 30 years overview with bibliometric analysis review. *Arch Comput Methods Eng*. 2023;30(7):4177–207. <https://doi.org/10.1007/s11831-023-09930-z>.
16. Guo C, et al. Malaria incidence from 2005–2013 and its associations with meteorological factors in Guangdong, China. *Malar J*. 2015. <https://doi.org/10.1186/s12936-020-03527-3>.
17. Kaur I, Sandhu AK, Kumar Y. Artificial intelligence techniques for predictive modeling of vector-borne diseases and its pathogens: a systematic review. *Arch Comput Methods Eng*. 2022;29(6):3741–71. <https://doi.org/10.1007/s11831-022-09724-9>.
18. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In *Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min*. 2016. pp. 785–794. <https://doi.org/10.1145/2939672.2939785>.
19. World Health Organization, *Malaria surveillance, monitoring & evaluation: a reference manual*, Geneva, Switzerland: WHO Press. 2018. <https://www.who.int/publications/i/item/9789241565578>.
20. Ali AO, et al. Factors associated with intermittent preventive treatment (IPTp-SP) use for malaria during pregnancy in Somalia: a multilevel analysis of the 2020 demographic and health survey. *Malar J*. 2025;24(1):345.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.