Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Spatial Modeling of Environmental and Clinical Drivers of Meningitis Transmission in the Upper West Region of Ghana Using Geospatial Machine Learning

Moses Asori
Elite Research and Data Science Institute
Ebenezer Senu
ebensenu@gmail.com
Elite Research and Data Science Institute
Ali Musah
Kwame Nkrumah University of Science and Technology
Monica Ahiadorme
University of Health and Allied Sciences (UHAS)
Sarah Amoafo
Elite Research and Data Science Institute
Yetimoni Kpeebi
University of North Carolina
Divine Odame Appiah
Kwame Nkrumah University of Science and Technology

Research Article

Keywords: Meningitis, Machine Learning, Spatial Analysis, Environmental Exposure, Ghana

Posted Date: April 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9329390/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Spatial Modeling of Environmental and Clinical Drivers of Meningitis Transmission in the Upper West Region of Ghana Using Geospatial Machine Learning

Moses Asori[1,2,3*], Ebenezer Senu[1,4], Ali Musah[5], Monica Ahiadorme[6], Sarah Amoafo[1,7], Yetimoni Kpeebi[2], Divine Odame Appiah[7]

1. Department of Clinical Research and Data Science, Elite Research and Data Science Institute, Kumasi, Ashanti Region, Ghana
2. Department of Earth, Environmental and Geographical Sciences, University of North Carolina, Charlotte, USA
3. Center for Applied Geographic Information Science, University of North Carolina, Charlotte, USA
4. Department of Biological Sciences, School of Natural Sciences and Mathematics, The University of Texas at Dallas, Richardson, Texas, USA
5. Department of Epidemiology and Biostatistics, School of Public Health, Kwame Nkrumah University of Science and Technology, Kumasi, Ashanti Region, Ghana.
6. Department of Basic Sciences, School of Basic and Biomedical Sciences, University of Health and Allied Sciences (UHAS), Ho, Ghana
7. Department of Geography and Rural Development, Kwame Nkrumah University of Science and Technology, Kumasi, Ashanti Region, Ghana

masori@charlotte.edu

ebensenu@gmail.com

bamwara.ma@gmail.com

mahiadorme@uhas.edu.gh

miss.sarahamoafo@gmail.com

ykpeebi@charlotte.edu

doappiah.cass@knust.edu.gh

Corresponding author: Ebenezer Senu, ebensenu@gmail.com, Department of Clinical Research and Data Science, Elite Research and Data Science Institute, Kumasi, Ashanti Region, Ghana

Abstract

Background: Meningitis remains a critical public health threat in the "Meningitis Belt" of sub-Saharan Africa, necessitating high-resolution predictive tools for targeted intervention. This study model environmental and clinical drivers of bacterial meningitis transmission and map population exposure risk in the Upper West Region of Ghana using machine learning.

Methods: PCR-confirmed case data (2018-2019) were integrated with multi-source geospatial predictors at a 100-meter resolution. Six models (Logistic Regression, Artificial Neural Networks, Support Vector Machines, Random Forest, Gradient Boosting Machine, and Extreme Gradient Boosting) were trained using repeated 5-fold cross-validation. A population-prevalence adjustment was applied to exposure at 1 km² resolution.

Results: Extreme Gradient Boosting showed the best clinical performance (F1-score: 0.478; sensitivity: 0.407), while Random Forest achieved the highest accuracy (83.5%). Annual rainfall (OR = 0.32, $p < 0.0001$), waterbodies (OR = 0.65, $p = 0.002$), and surface temperatures (OR = 0.55, $p = 0.001$) were significant environmental factors that reduced infection risk. Conversely, neck stiffness (OR = 1.60, $p = 0.002$) and night-time light (OR = 1.49, $p = 0.041$) were associated with increased odds of infection. Nandom, Lawra, and Jirapa Lambussie emerged as the primary geographical hotspots.

Conclusion: This study highlights the critical interplay between environmental factors and clinical indicators in shaping the risk of bacterial meningitis. Integrating geospatial environmental data with machine learning offers a powerful framework for targeted surveillance and early warning systems in endemic regions.

Keywords: *Meningitis, Machine Learning, Spatial Analysis, Environmental Exposure, Ghana,*

1 Introduction

Bacterial meningitis remains a major public health challenge globally [1–3], but its burden is disproportionately concentrated in sub-Saharan Africa, particularly within the African meningitis belt [4–6]. Ghana lies within this belt, with its northern ecological zones, including the Upper West Region, experiencing recurrent seasonal outbreaks and periodic epidemics [7]. Although meningitis can arise from a diverse range of etiologies, including bacteria, viruses, fungi, parasites, toxins, certain medications, and traumatic injury, bacterial meningitis remains the most prevalent and clinically significant form, primarily caused by pathogenic organisms such as *Neisseria meningitidis*, *Streptococcus pneumoniae*, and *Haemophilus influenzae* [6, 8]. The Upper West Region is among the most vulnerable settings for meningitis transmission in Ghana, due to its distinct climatic and ecological characteristics. The region is characterized by a tropical savannah climate with a single, short rainy season (May-October) followed by a prolonged dry season (November-April) [9]. During the dry season, environmental conditions are dominated by the Harmattan winds (dry), dust-laden air masses originating from the Sahara Desert, which significantly reduce humidity (often below 20%) and increase airborne particulate matter [9]. These conditions have been strongly associated with increased meningitis transmission, as they contribute to mucosal damage in the upper respiratory tract, facilitating bacterial invasion.

Epidemiological evidence from the region demonstrates that meningitis incidence is highly seasonal and spatially heterogeneous, with the majority of cases occurring between epidemiological weeks 2 and 17, which coincide with peak dry, dusty, and low-humidity conditions [6, 7, 10]. Furthermore, spatial analyses have shown that meningitis cases in the Upper West Region do not occur randomly but instead form distinct geographic clusters, with certain districts consistently exhibiting higher attack rates and case fatality ratios [7]. Thus, the transition from dry to wet season is consistently associated with a marked decline in incidence, highlighting the protective role of increased humidity and rainfall [6]. In the Upper West Region, annual rainfall typically ranges from 1000 to 1200 mm, with approximately 90% occurring during the rainy season [11, 12]. Conversely, the dry season is characterized by extreme temperature fluctuations (22°C-40°C) and minimal precipitation, conditions that create a favorable environment for epidemic transmission. Beyond climatic factors, local environmental and socio-ecological conditions further influence meningitis risk. The region is predominantly rural, with dispersed settlements and varying levels of infrastructure and healthcare access. Emerging evidence also

suggests that environmental proxies such as vegetation cover, proximity to water bodies, and indicators of human activity (e.g., nighttime light intensity) may significantly influence transmission patterns by modifying microclimatic conditions and human interaction networks [8, 13].

Despite the well-documented influence of climate and environment on meningitis epidemiology, important gaps remain in understanding how these factors interact with individual-level clinical indicators to drive infection risk at fine spatial scales. Traditional surveillance systems in Ghana have primarily relied on aggregated epidemiological data, limiting the ability to identify localized hotspots and predict outbreaks with high spatial precision. While previous studies in the Upper West Region have demonstrated the utility of geographic information systems (GIS) for identifying disease clusters, the integration of high-resolution environmental data with clinical features remains limited. Recent advances in machine learning and spatial epidemiology provide an opportunity to overcome these limitations by modeling complex, nonlinear relationships between environmental, geographic, and clinical variables. Such approaches are particularly relevant in settings such as the Upper West Region, where dynamic interactions among climate variability, ecological conditions, and human behavior influence disease transmission. In this study, we integrate PCR-confirmed meningitis cases with high-resolution environmental and geospatial datasets to better characterize the determinants of meningitis risk in the Upper West Region of Ghana. Specifically, we examine the role of key predictors, including annual rainfall, surface temperature, nighttime light, waterbody density, and clinical indicators such as neck stiffness, in shaping infection risk. Furthermore, we evaluate the predictive performance of multiple machine learning models and map the spatial distribution of disease risk and population exposure. By grounding predictive modeling within the unique climatic and epidemiological context of the Upper West Region, this study provided actionable insights for targeted surveillance, early warning systems, and climate-informed public health interventions in one of Ghana's most high-risk regions for meningitis.

2. Materials and Methods

2.1 Data sources

Meningitis cases

This study utilized secondary data from district health information management system version 2 (DHIMS-2). Bacterial meningitis data were retrieved in November 2025 from a line list of routine disease surveillance data collected between 2018 and 2020 using a data sheet. Reference was made to the individual case status based on Polymerase Chain Reaction (PCR) laboratory tests. Data completeness and accuracy were checked to ensure quality. The individual meningitis case status was geocoded to ensure it is linked to spatial information for geographic modeling. For details on data extraction, see [7].

Predictors

The predictors are diverse, including personal health background at the time of assessment and environmental and geographical factors. Building volume, nighttime light, population count/density, and annual rainfall were downloaded from WorldPop (https://www.worldpop.org/). Distances from roads were derived using the Euclidean distance tool in a GIS environment. Road networks were extracted from the Diva GIS portal (https://diva-gis.org/). Also, river/stream density was derived using kernel density estimation based on river network data extracted from Diva GIS. Additionally, elevation data in the form of a Digital Elevation Model (DEM), normalized difference vegetation index (NDVI), and surface temperatures were downloaded from the USGS EarthExplorer data portal (https://earthexplorer.usgs.gov/). Particulate matter ($PM_{2.5}$) was downloaded from the Socioeconomic Data and Applications Center (SEDAC) data portal (https://www.earthdata.nasa.gov/data/catalog/sedac-ciesin-sedac-sdei-gwrpm25-mmsvaod-5gl04-5.04). Since variables like neck stiffness were at the patient level and binary, we rasterized them by first filtering out participants who reported neck stiffness, then modeling them geographically using kernel density estimation (KDE). Furthermore, the epidemiologic week in which a person was diagnosed and the number of times the participants sought healthcare were also rasterized using a simple kriging model. All administrative boundary shapefiles were extracted from the Humanitarian Data Exchange (HDX) (https://data.humdata.org/dataset/cod-ab-gha)

2.2 Data preprocessing

All downloaded and derived variables were geographically aligned to the exact resolution (100 meters) using resampling with a bilinear model. Those that were at the national level were clipped to the Upper West regional boundary. Afterward, we examined all Nodata areas for edge effects and subsequently removed them. Having obtained all that, we assessed multicollinearity between predictors by examining the pairwise feature correlations. We set inclusion criteria for all variable pairs with $R^2 \leq 0.6$. For feature engineering, we centered and scaled all predictors during the machine learning training phase. Also, to address missing records, we used the median imputation method, which is more robust to outliers. This was helpful for logistic regression, artificial neural networks, and support vector machines.

2.3 Geostatistical analysis

Random Forest

Random Forest is a family of ensemble models built on decision trees [14, 15]. Unlike decision trees, which involve a single tree and are more prone to overfitting, RFs have two sources of randomness, thereby limiting the risk of overfitting. The first source uses random subsampling of the observations to build the trees. This sub-setting happens with replacement. The second source involves the random subsetting of predictors used for partitioning a part tree. To determine the number of predictors randomly selected for a leaf split, we used the default equation,

$$P_S = \sqrt{N} \qquad [1]$$

Where N is the total number of features. To select the final model, we chose the one with the highest ROC. As a result, we used an ROC of 0.71, a sensitivity of 0.96, and a specificity of 0.33 for the final model fitting. Furthermore, the minimum number of predictors used for splitting was set to 2, and the number of trees was 500. We used five-fold, five-repeated cross-validation to assess model performance. To give an overview, the random forest model starts at the parent node and then recursively partitions the observations according to a set of optimization rules. The partition continues until the terminal node, where further division is not possible. The decision to split is based on minimizing entropy and maximizing information gain. We used information gain reduction with variable exclusion to rank predictors by their predictive importance.

Gradient Boosting Machine

A Gradient Boosting Machine is a tree-based ensemble model used for classification and regression tasks [16]. Given a set of observations, the goal of this algorithm is to find the optimal function $F(x)$ that accurately maps a set of predictors X to an output Y by minimizing the expected loss between the observations and the predictions. It starts with a prediction from a weak learner that is slightly more accurate than a random guess. Having estimated the residual from that initial value, the residual becomes the response variable used in the subsequent function to update the previous prediction iteratively. The boosting machine, as a weighted sum of functions for updating, is expressed.

$$F_m(X) = F_{m-1}(X) + \Omega\rho_m h_m(X) \qquad [2]$$

Where the ρ_m is the weight on the m^{th} function, $h_m(X)$ and the $F_{m-1}(X)$ is the function that is being updated. These models are additive, forming an ensemble. While a gradient boosting machine is powerful in achieving accuracy, it risks overfitting, making it difficult to generalize to unseen data. In the worst case, if the residual is zero, the next update round terminates. As a result, the $\Omega \in [0, 1]$ is the shrinkage factor used to mitigate such severe overfitting. Additional ways to prevent overfitting include reducing model complexity by pruning trees that grow too large and randomizing the number of predictors used to build base learners. We used a 5-fold cross-validation to assess model performance and select the optimal combination of ROC, sensitivity, and specificity. The decision on which set of parameters to use for the final training was based on the highest ROC value. Based on the ROC, the minimum number of observations per node was set to 10 and held constant throughout the iterative process. The shrinkage factor ΩΩ was alsoheld at 0.1. Also, the number of trees ranged from 50 to 500, with an optimal value of 100. Also, interaction depth ranged from 1 to 10, with the optimal value being 2.

Extreme Gradient Boosting Machine

Extreme gradient boosting is a component of a comprehensive family of ensemble machine learning methods that uses decision trees for both regression and classification tasks [17]. It is built on a gradient boosting model, but with more scalability. Like conventional gradient boosting algorithms, XGB incrementally refines previous predictions by strategically incorporating additional optimized forecasts, achieved by minimizing the objective loss function. Because XGB relies on trees as its base classifiers, the objective residual function is developed in diverse setups to handle various regression and classification tasks. Specifically, the loss is minimized through

$$L_{xgb} = \sum_{i=1}^{N} L\left(y_i, F(x_i)\right) + \sum_{m=1}^{M} \Omega(h_m) \qquad [3]$$

$$\Omega(h) = \gamma T + \frac{1}{2}\lambda \|w\|^2, \qquad [4]$$

Where T is the number of trees, w is the output score for the terminal node. Ideally, the objective loss function can be incorporated into the leaf-splitting process, enabling trees to be pruned before training to avoid overfitting. For example, selecting the higher values of γ leads to smaller, simpler trees, while smaller values lead to larger trees. It therefore moderates the extent to which gain is incorporated into the output scores at the terminal nodes. The XGT regularization strategy facilitates controlling the step size during iterative prediction updates. These types of shrinkages, pre-pruning procedures, and regularization techniques enhance computational efficiency and reduce memory usage. For validation purposes, we used 5-fold repeated cross-validation. The ROC curve was used to determine the optimal setup parameter combination for the model. The final parameters used for the model were: number of rounds 50, maximum depth 2, eta 0.4, gamma 0, column sample by tree 0.6, and subsample about 0.89.

Artificial Neural Network

The artificial neural network is a multi-perceptron machine learning model used for various regression and classification tasks across diverse research areas, including health intelligence [18, 19]. Given a set of observations with target (response) values and features, the aim is to find the optimal weighted function that accurately maps the features to the target. The model architecture involves computational units, neurons, weights, and activation function (s). The weights connect the neurons, and the activation function handles nonlinear patterns in the data. The model has three layers: the input layer, the hidden layer, and the output layer. This architecture is expressed in

$$F_{NN}(X) = \Omega(W(X) + b) \qquad [5]$$

Where W represents the weights connecting the neurons, X are the features, $F_{NN}(X)$ is the predicted value, and Ω is the activation function. In our study, we used a sigmoid activation function and a single hidden layer, with the same configuration throughout all iterations. The ultimate aim is to find the optimal set of weights and biases b that minimizes the loss function between the predicted and the observed values. We used ROC for model selection via five-fold repeated cross-validation. The number of units in the hidden layer ranged from 1 to 19, with an optimal value of 3. Also, the decay ranged from 0 to 0.1, although we settled with about 0.0075.

Support Vector Machines

The support vector machine has gained recognition, especially in classification tasks such as land use and land cover, disease diagnosis, and text [20, 21]. Given a set of data points, the model projects them into a high-dimensional feature space, where a hyperplane partitions them into output classes. The objective function is to find the optimal parameters that maximize the space between the support vectors. With a more complex data structure, a linear plane may fail to discriminate between the classes accurately. As a result, more complex kernel functions, such as polynomials and radial basis, are used. We used five-fold repeated cross-validation for model training and selection. Model selection was based on the highest ROC value. For parameter tuning, the cost ranged from 0.25 to 128, although we settled on 0.25. Also, the decay function, which determines how close features must be to influence one another, was set to 0.0699. This was set as a constant across all iterations.

Logistic Regression

The logit model is a binary classifier that provides a simpler framework for predicting two classes, as in our case (presence or absence of meningitis). It has been used in healthcare, spatial planning, and environmental modeling. The logistic regression in our study was based on the generalized linear model. Given a set of observations D, it predicts a target class Y as a linear combination of features X, by maximizing the objective log likelihood function, $L(\Omega)$.

$$L(\Omega) = \sum_{i=1}^{N}[y_i Log(p_i) + (1 - y_i)Log(1 - p_i)] \qquad [6]$$

Where p_i is the predicted probability of a person having meningitis. In our model, we used five-fold cross-validation to identify the model with the lowest variance, thereby avoiding reliance on a single estimate. Which might either overestimate or underestimate the other models unusually. The criteria for selecting a model for the final training were based on the ROC curve's highest value.

Model Performance Assessment

These predictive models aim to generalize to unseen data sets. It is essential in determining if a model should be relied on. In our study, we use the confusion matrix to derive various reliability measures, which show that the trained model performed well on 30% of the reserved test data.

These measures involved accuracy, sensitivity, specificity, F1-score, negative predictive value (NPV), and positive predictive value (PPV). These metrics are calculated as follows.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad [7]$$

$$\text{PPV} = \frac{TP}{TP+TN} \quad [8]$$

$$\text{Sensitivity} = \frac{TP}{TP+FP} \quad [9]$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad [10]$$

$$\text{NPV} = \frac{TN}{TN+FN} \quad [11]$$

$$F1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision}+\text{Recall}} \quad [12]$$

Where TP is the true positive, TN is the true negative, FP is the false positive, FN is the false negative case, and NPV is the Negative Predictive Value. Also, the F1 score was represented by F1. The area under the receiver operating characteristic (ROC) curve (AUC) was used to assess the algorithm's predictive accuracy. It charts the TP and FP values at various cutoffs over iterations. Area under the curve (AUC) values indicate how well our model distinguishes between negative and positive meningitis cases. Although we attempted down-sampling to 1:4, the inherent imbalance could still skew the accuracy measures. Therefore, we combined accuracy and model sensitivity to determine the F1-score, which is a good trade-off in this instance.

2.4 Estimating expected population exposure

To determine the number of people expected to be infected with the disease, we configured the machine learning models to output probability estimates across the region. Having obtained that at 100-meter resolution, we computed the product of the population count raster and the probability raster. However, because we had to downsample our data to reduce the extreme imbalance during machine learning training, we adjusted the sample probabilities to match the population-level infection prevalence. Therefore, we derived an adjustment factor and applied it to the probability raster. A 2024 meta-analysis of meningitis infections across Ghana estimated a prevalence of about 24% in the Upper West region [22]. We therefore used that as our population

prevalence. Let the sample infection probability be $P(\pi)$ and population probability be $P(\lambda)$. Therefore, the odds of the infection for both the sample and population are expressed below,

$$OR_{\pi} = \frac{\pi}{1-\pi} \quad [13]$$

$$OR_{\lambda} = \frac{\lambda}{1-\lambda} \quad [14]$$

Where the adjustment factor is estimated as

$$F_{Adj} = \frac{OR_{\lambda}}{OR_{\pi}} \quad [15]$$

Having obtained these values and the adjustment factor, we then estimated updated probabilities of infection generated by our machine learning model through the expression below,

$$P(x)^{*} = F_{Adj} * \frac{OR_{\lambda}}{OR_{\pi}} \quad [16]$$

Where F_{Adj} is the adjustment factor, and $P(x)^{*}$ is the updated probability of infection. With this, the population raster was multiplied by this probability to obtain the expected number of people at risk of meningitis. To get a clinically meaningful estimate at a plausible geographic scale, we calculated the sum of these cases scheduled at a 1 km^2 resolution. This was achieved through the zonal statistics in the GIS environment.

3. Results

Correlates of meningitis infection risk

With the adjusted model, building volume decreased the odds of bacterial meningitis infection by 41% (OR=0.59; 95% CI: 0.40-0.88, p=0.009), and seeking healthcare also reduced the risk by 18% (OR=0.82; 95% CI: 0.66-1.03, p=0.083). However, elevation increased the risk by 30% (OR=1.30; 95% CI: 0.89-1.91, p=0.170), 60% increased risk for neck stiffness (OR=1.60; 95% CI: 1.19-2.16, p=0.002), 41% for nighttime light (OR=1.49; 95% CI: 1.02-2.19, p=0.041), and 11% increase for vegetation (OR=1.11; 95% CI: 0.87-1.42, p=0.382).

Conversely, the latter part of the epidemiologic week reduced the risk by 16% (OR=0.84; 95% CI: 0.67-1.06, p=0.137). Additionally, the distance to roads is associated with a 15% reduction in risk (OR=0.85; 95% CI: 0.61-1.20, p=0.358). Furthermore, surface temperatures (OR = 0.55; 95% CI: 0.39-0.77, p = 0.001) and annual rainfall (OR = 0.32; 95% CI: 0.22-0.47, p < 0.0001) decreased the risk by 45% and 68%, respectively **(Table 1)**.

Table 1: Correlates of meningitis risks in Upper West Ghana.

	Adjusted Model		Unadjusted Model	
Predictor	OR (95% CI)	p-value	OR (95% CI)	p-value
Building volume	**0.59 (0.40-0.88)**	0.009	0.83 (0.68-1.03)	0.091
Seeking healthcare	0.82 (0.66-1.03)	0.083	1.09 (0.91-1.30)	0.344
An Elevation	1.30 (0.89-1.91)	0.170	0.91 (0.75-1.10)	0.333
Epidemiologic week	0.84 (0.67-1.06)	0.137	1.08 (0.89-1.29)	0.444
Distance to roads	0.85 (0.61-1.20)	0.358	0.80 (0.62-1.04)	0.094
Surface temperatures	**0.55 (0.39-0.77)**	0.001	**1.35 (1.11-1.65)**	0.003
Neck stiffness	**1.60 (1.19-2.16)**	0.002	**2.04 (1.68-2.49)**	<0.000
Night-time light	**1.49 (1.02-2.19)**	0.041	1.03 (0.86-1.24)	0.733
Particulate matter	0.96 (0.67-1.38)	0.841	1.01 (0.84-1.22)	0.900
Annual rainfall	**0.32 (0.22-0.47)**	<0.000	**0.41 (0.33-0.50)**	<0.000
Vegetation amount	1.11 (0.87-1.42)	0.382	0.99 (0.82-1.20)	0.950
Waterbodies	**0.65 (0.49-0.85)**	0.002	**0.80 (0.66-0.96)**	0.018

Machine Learning performance

The accuracy for GBM was 82.46% (95% CI: 77.37-86.82), with a sensitivity of 27.78% and a specificity of 96.26%. Also, the balanced accuracy was 62.0%. For RF, the accuracy was 83.5% (95% CI: 78.59, 87.81), sensitivity 29.6%, specificity 97.1%, and balanced accuracy 63.4%. For ANN, the accuracy was 80.2% (95% CI: 74.94, 84.82), the sensitivity was 33.3%, the specificity was 92.0%, and the balanced accuracy was 62.6%. For support vector machines, the accuracy was 83.2% (95% CI: 78.19, 87.48), sensitivity 20.3%, specificity 99.0%, and balanced accuracy 59.7%. Also, for the LR model, the accuracy is 80.9% (95% CI: 75.75, 85.49), the sensitivity is 27.8%, the specificity is 94.4%, and the balanced accuracy is 61.0%. Also, for XGB, the accuracy was 82.0%, the sensitivity 40.7%, the specificity 92.5%, and the balanced accuracy 66.6% **(Table 2)**.

Table 2: Machine learning performance metrics

Metrics	**GBM**	**Logit**	**NN**	**RF**	**SVM**	**XGB**
Accuracy	0.825	0.809	0.802	0.835	0.832	0.820
Sensitivity	0.278	0.278	0.333	0.296	0.203	0.407
Specificity	0.963	0.944	0.920	0.971	0.990	0.925
Pos Pred Value	0.652	0.556	0.514	0.727	0.846	0.578
Neg Pred Value	0.840	0.838	0.845	0.845	0.831	0.860

Prevalence	0.201	0.201	0.202	0.201	0.201	0.201
Detection Rate	0.056	0.055	0.067	0.059	0.041	0.082
Detection Prevalence	0.086	0.100	0.132	0.082	0.048	0.141
Balanced Accuracy	0.620	0.610	0.626	0.634	0.597	0.666
F1-Score	0.390	0.370	0.400	0.421	0.328	0.478

For GBM, SVM, ANN, XGB, RF, and LR, the AUCs were 68.6%, 70.1%, 68.8%, 68.3%, 68.5%, and 65.6%, respectively **(Figure 1).** Furthermore, the F1-score was 39% for GBM, 37% for the LR model, 40% for ANN, 42% for RF, 33% for SVM, and about 48% for the XGB model. The predictors were also ranked by their contribution to model performance.

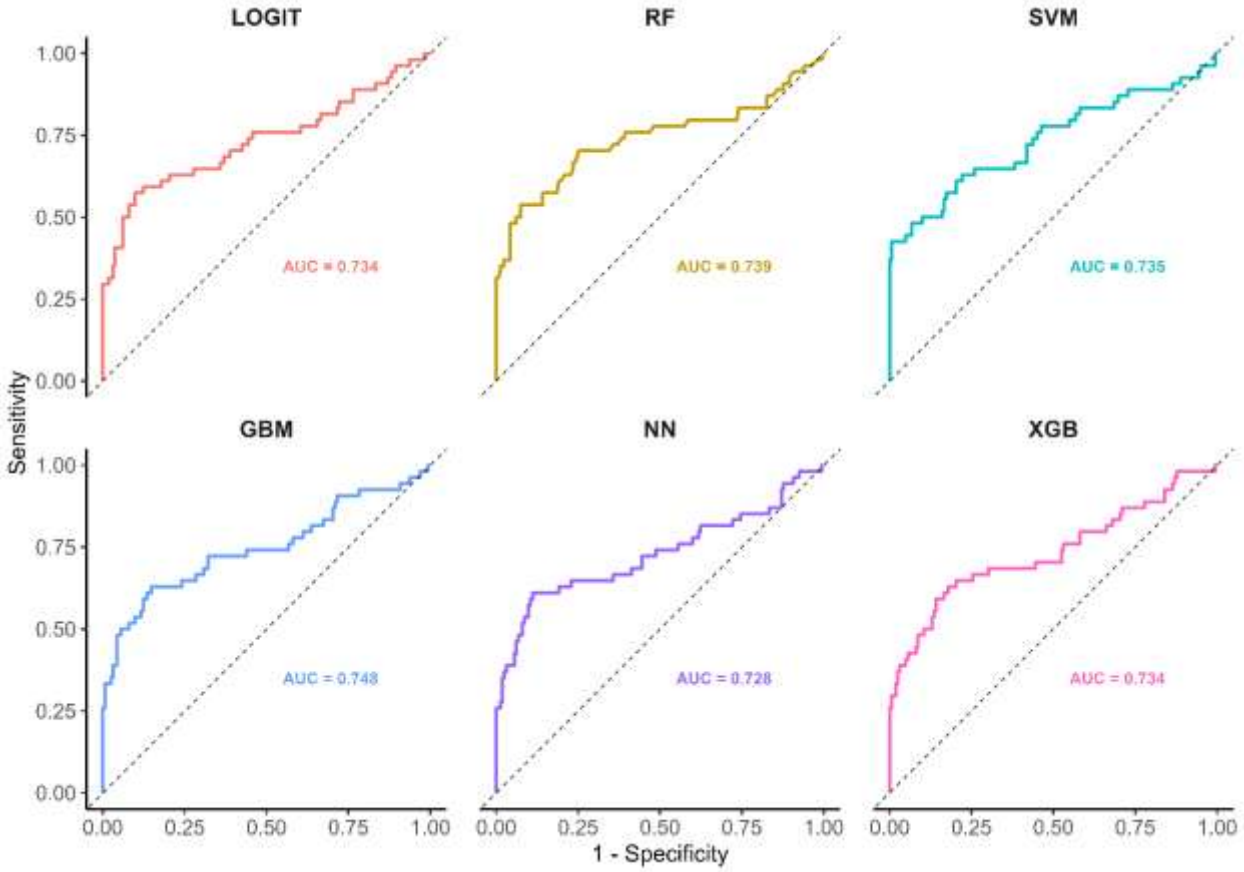


Figure 1: Receiver Operating Characteristics Curve for the ML models

For the LR model, the top three most important variables were annual rainfall, waterbody density, and nighttime light, in that order. For RF, the top three factors were annual rainfall, neck stiffness, and water body density. For SVM, the top three included annual rainfall, neck stiffness, and

population density. Also, for GBM, the top three were annual rainfall, nighttime light, and particulate matter. Based on the ANN, the top three variables included annual rainfall, waterbody density, and surface temperature. In contrast, for XGB, the top three were neck stiffness, yearly rainfall, and nighttime light, in decreasing order of importance **(Figure 2).**

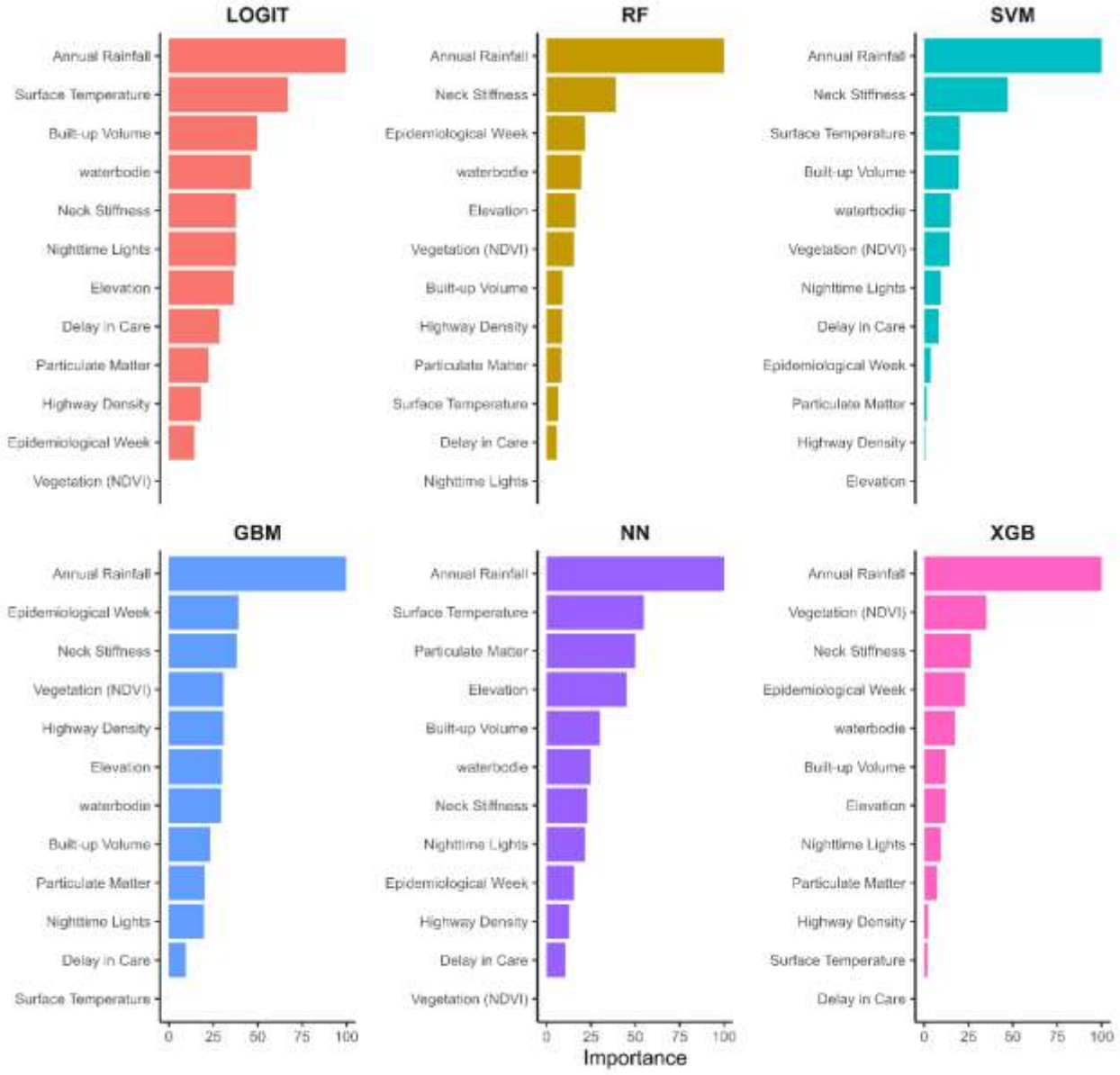


Figure 2: Variables' importance plots for the ML models

Geographic Distribution of Infection Risks

Based on the LR mode, the highest probability of infection was found at the northern tip and the northwestern corner of the study area. The probability between 0.7 and 0.8 was also moderately visible in the southern part of the region. In ANN mode, although the pattern was nearly similar to LR, the probability 1.0 was more prominent in highly infectious areas, and there was greater visibility of probabilities between 0.7 and 0.8. However, the SVM model provided a more geographically conservative estimate of risk probabilities, ranging from 0.9 to 1.0, tightly bound to the northwestern corner of the region **(Figure 3).** The same spatially tight northwestern corner of the risk surface was also observed for the RF model. However, probabilities between 0.3 and 0.5 were more spread out across the region. For the XGB model, the pattern was similar to the other models, except that the highest-risk range between 0.8 and 1.0 was more spread out in the region's northwestern corner. We also observed an almost identical pattern to the GBM predictions.

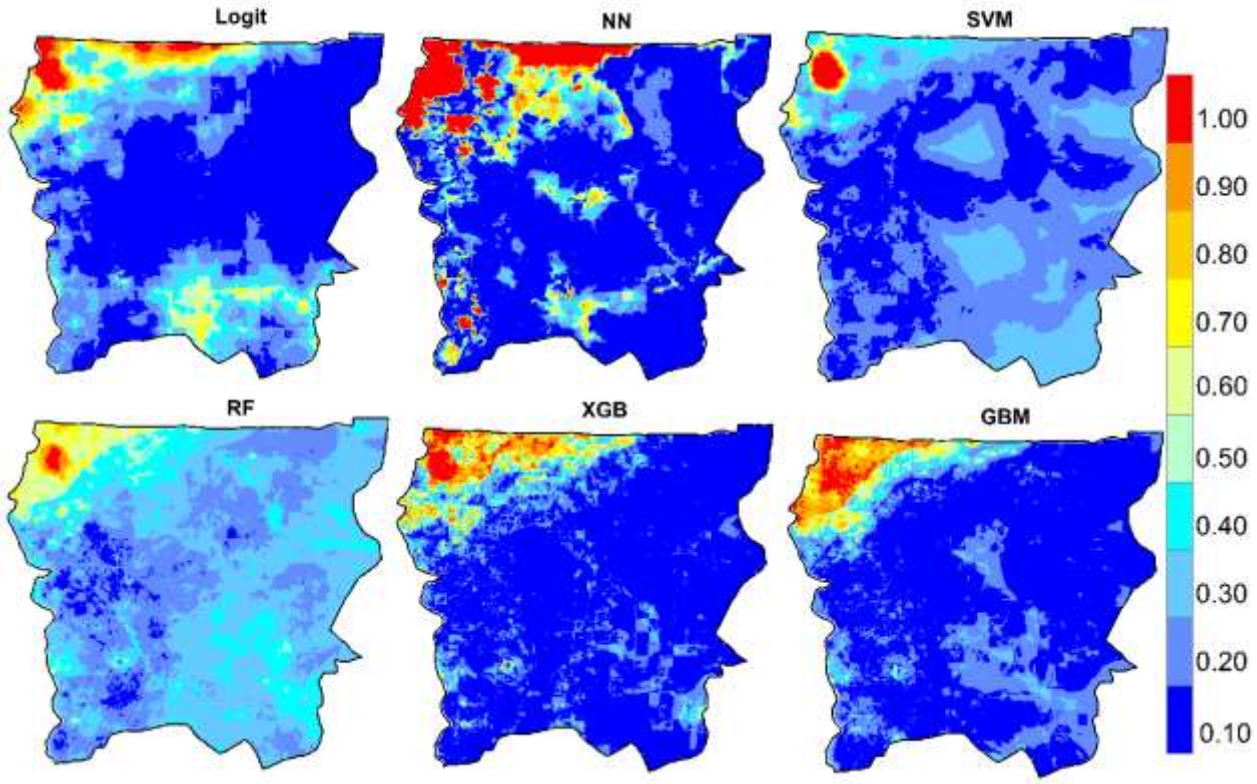


Figure 3: Spatial distribution of the probabilities of meningitis risk based on machine learning outputs. The logit is a logistic regression model; the NN is an artificial neural network; the SVM is a support vector machine; the RF is a random forest model; the XGB is an extreme gradient boosting machine; and the GBM is a gradient boosting machine.

Based on the expected number of people at risk of infection per 1 km^2, the spatial pattern was somewhat more conservative than the probability distribution, likely due to the region's localized population distribution. For example, apart from the ANN model, the other models had centers with several people exceeding 1000, more widely distributed in a very similar pattern **(Figure 4).** However, about 100 people at risk were more conspicuous in the northwestern region across all models except the ANN.

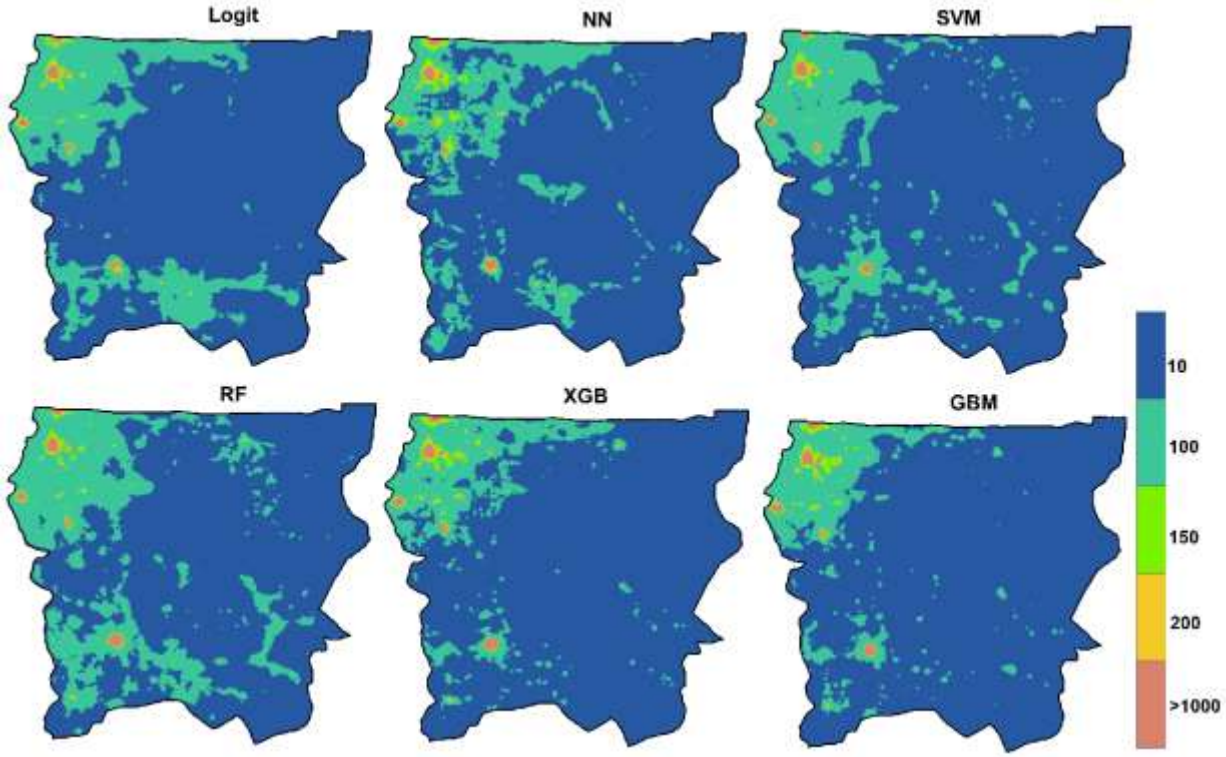


Figure 4: Spatial distribution of the expected number of people at risk of meningitis infection. The logit is a logistic regression model; the NN is an artificial neural network; the SVM is a support vector machine; the RF is a random forest model; the XGB is an extreme gradient boosting machine; and the GBM is a gradient boosting machine.

The district with the highest risk was Nandom, with an average of 107 people for NN, 103 for GBM model, 99 for Logit, 98 for XGBT, 92 for SVM, and 92 for RF. The second-highest-risk district was Lawra, with about 62 people for GBM, 51 for the LR model, 45 for ANN, 46 for RF, 38 for SVM, and 50 for XGB **(Figure *5*).** The following risky districts included Wa and Sisala West. The lowest risk districts were Sisala East and Wa East.

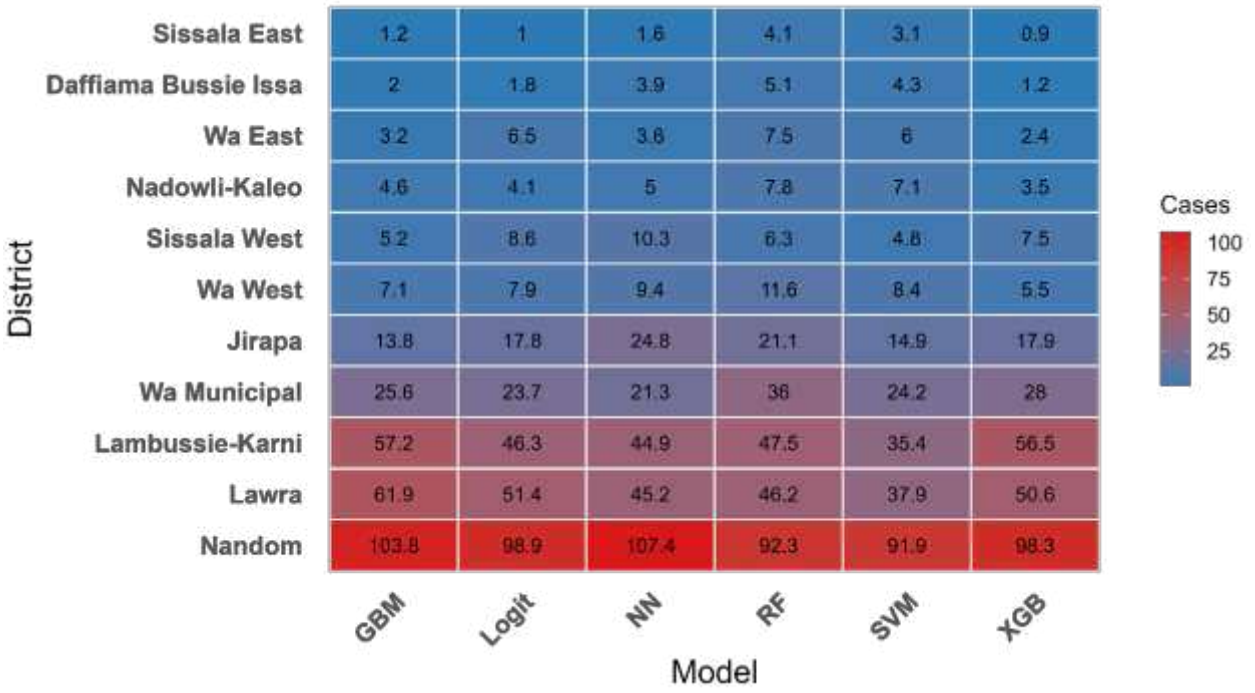


Figure 5: The distribution of the average number of people at risk of being infected by meningitis according to districts by models. The logit is a logistic regression model; the NN is an artificial neural network; the SVM is a support vector machine; the RF is a random forest model; the XGB is an extreme gradient boosting machine; and the GBM is a gradient boosting machine.

4. Discussion

This study provides important insights into the environmental and clinical determinants of bacterial meningitis risk in the Upper West Region of Ghana, highlighting the dominant role of annual rainfall, surface temperature, neck stiffness, nighttime light, and waterbody density. By integrating high-resolution geospatial data with machine learning approaches, our findings extend existing epidemiological evidence and provide a more granular understanding of disease dynamics in one of the most vulnerable regions of the African meningitis belt. The most significant finding of this study is the consistent identification of annual rainfall as the primary environmental factor in determining bacterial meningitis incidence. Across all six machine learning architectures, higher annual rainfall was associated with a 68% reduction in the odds of infection. The strong protective effect of annual rainfall observed in this study is highly consistent with previous research across Ghana and the wider meningitis belt. Multiple studies have demonstrated that meningitis incidence peaks during the dry season and declines rapidly with the onset of rainfall [23]. In northern Ghana specifically, up to 67% of meningitis cases occur during dry periods, reinforcing the central role

of low humidity and arid conditions in driving transmission [22]. Our findings refine this understanding by quantifying rainfall as the most influential predictor across all machine learning models, suggesting that it is not merely a seasonal marker but a primary ecological regulator of transmission. Mechanistically, rainfall likely increases ambient humidity, restoring mucosal integrity in the nasopharynx and reducing bacterial invasion, while also limiting the aerosolization of dust particles that facilitate pathogen spread. This aligns with ecological models indicating that dry, dusty environments promote the transition from carriage to disease [23].

Similarly, the observed association between surface temperature and reduced infection risk adds nuance to prior studies. While high temperatures have often been linked to increased meningitis risk in the Sahel, particularly when combined with low humidity, our results suggest that temperature alone may not be sufficient to drive transmission. Instead, the interaction between temperature and moisture appears to be critical. This interpretation is supported by climate-based analyses in Ghana, which show that temperature effects on meningitis are context-dependent and strongly mediated by humidity and rainfall [24]. The protective effect of waterbody density observed in this study has received limited attention in previous Ghanaian research, but it is biologically plausible. Areas near water bodies may experience localized increases in humidity and reduced dust exposure, thereby lowering transmission risk. This finding complements broader ecological evidence suggesting that environmental moisture gradients play a critical role in shaping meningitis distribution across the belt.

Moreover, the strong association between neck stiffness and infection risk is consistent with its well-established role as a hallmark clinical symptom of meningitis [6]. Previous studies in Ghana have consistently reported neck stiffness as one of the most common presenting symptoms, alongside fever and altered consciousness [25]. However, our study advances this knowledge by spatially modeling this clinical feature, demonstrating its predictive value not only at the individual level but also as a geographically distributed risk indicator. The positive association with nighttime light provides important insight into the socio-environmental determinants of meningitis. Nighttime light is widely used as a proxy for urbanization, population density, and human activity. In northern Ghana, this likely reflects increased contact rates and clustering of individuals, both of which are known to facilitate meningococcal transmission. Historical studies in Ghana have identified close contact and shared living spaces as key risk factors for infection [26]. Our findings extend this by demonstrating that spatial proxies of human activity can capture these dynamics at

a population level. Notably, while traditional studies in Ghana have emphasized household-level risk factors such as overcrowding, indoor air pollution, and exposure to cooking smoke [26], our results suggest that macro-level environmental and spatial variables may provide equally strong, or even stronger, predictive power. This highlights the importance of integrating multi-scale determinants in meningitis risk modeling.

Compared to high-income countries, where meningitis incidence is relatively low and less influenced by environmental variability, the findings from this study highlight the unique ecological sensitivity of meningitis transmission in sub-Saharan Africa. In Europe and North America, incidence rates are typically stable and largely driven by vaccination coverage and pathogen dynamics rather than climate variability [27, 28]. In contrast, across the African meningitis belt, environmental conditions remain a dominant driver of disease patterns. Similar associations between low rainfall, high dust exposure, and increased meningitis incidence have been documented in countries such as Burkina Faso, Niger, and Nigeria [29–31]. However, our study adds to this body of knowledge by demonstrating that these relationships persist even at fine spatial resolutions and can be effectively captured using machine learning approaches.

4.1 Model performance, spatial epidemiology, and clinical implications

The moderate predictive performance of the models (AUC range: 65.6–70.1%) is comparable to previous modeling studies in Ghana and the meningitis belt, where complex environmental and epidemiological interactions often limit predictive accuracy [22, 23]. Importantly, the superior performance of Extreme Gradient Boosting (XGB) highlights the value of ensemble learning methods in capturing nonlinear relationships and interactions among predictors. The consistency in variable importance rankings across models, particularly the dominance of annual rainfall, enhances confidence in the robustness of our findings. This convergence across multiple algorithms is a key strength, suggesting that the identified predictors are not model-specific but represent fundamental drivers of meningitis risk. In addition, the identification of high-risk clusters in districts such as Lawra and Jirapa-Lambussie aligns with previous epidemiological studies showing spatial heterogeneity of meningitis incidence within northern Ghana [32]. These localized hotspots likely reflect a combination of environmental exposure, population distribution, and disparities in healthcare access. Importantly, integrating population data with risk probabilities to estimate expected population exposure represents a

significant advancement over traditional risk-mapping approaches. This provides a more actionable metric for public health planning, enabling targeted interventions in areas with the highest burden. Given that annual rainfall emerged as the most influential predictor, future changes in precipitation patterns could significantly impact meningitis epidemiology. This highlights the need for climate-informed surveillance systems that integrate environmental data into early warning frameworks.

4.2 Study Limitations

This study is limited by its cross-sectional design, which precludes causal inference regarding the relationship between environmental exposures and meningitis risk. Some variables, including nighttime light and waterbody density, serve as proxies and may not fully capture the underlying mechanisms. In addition, these findings from the Upper West Region may not be fully generalizable to other settings with different ecological and demographic profiles.

5. Conclusion

The risk of bacterial meningitis in the Upper West Region of Ghana is strongly influenced by environmental conditions, particularly rainfall, as well as key clinical indicators. While aspatial assessment is key to understanding statistical and mechanistic associations, it fails to provide the ecological context within which bacterial meningitis transmission is shaped across space. The integration of geospatial data with machine learning enables precise identification of high-risk areas and populations. Predicted transmission probability and expected case distribution were spatially heterogeneous, with Nandom experiencing the greatest burden while Sissala East experienced the least burden. These findings support climate-informed surveillance, targeted interventions, and improved outbreak preparedness in northern Ghana, with potential applicability across the African meningitis belt.

Acknowledgements

Not Applicable

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

Conceptualization: MA; Methodology: MA. Formal analysis and investigation: MA & ES; Writing - original draft preparation: MA, ES, & MA; Writing - review and editing: MA, AM, DOA, YK & SA; Supervision: DOA.

Data availability

The data utilized in this research are not publicly accessible. Therefore, any inquiries regarding the data should be directed to the corresponding author.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Ethics and Consent to Participate declarations

Not applicable

Consent for publication

Not applicable

Declaration of generative AI use

The authors did not use any generative AI tool in writing any part of this paper and take full responsibility for all the content.

References

1. Liu L, Wang X, Guo L, Chen T, Hu B, Jian B, et al. Epidemiological characteristics and disease burden of bacterial meningitis in hospitalized children in China: a 6-year nationwide retrospective study. BMC Infect Dis. 2025;25:592. https://doi.org/10.1186/s12879-025-10928-0.

2. Mahtab S, Madewell ZJ, Baillie V, Dangor Z, Lala SG, Assefa N, et al. Etiologies and comorbidities of meningitis deaths in children under 5 years in high-mortality settings: Insights from the CHAMPS Network in the post-pneumococcal vaccine era. J Infect. 2024;89:106341. https://doi.org/10.1016/j.jinf.2024.106341.

3. Yang Y, Knoll MD, Herbert C, Bennett JC, Feikin DR, Garcia Quesada M, et al. Global impact of 10- and 13-valent pneumococcal conjugate vaccines on pneumococcal meningitis in all ages: The PSERENADE project. J Infect. 2025;90:106426. https://doi.org/10.1016/j.jinf.2025.106426.

4. Lagare A, Abdoulaye B, Thiaw FD, Ousmane H, Ndiaye N, Aoula B, et al. First detection of echovirus 18 associated with aseptic meningitis in a child in Niger Republic, 2024: a case report. BMC Infect Dis. 2025;25:1389. https://doi.org/10.1186/s12879-025-11674-z.

5. Peter OJ, Oguntolu FA, Nyerere N, El-Mesady A. Modeling prevalence of meningitis control strategies through evaluating with available data on meningitis cases reported in Nigeria. Model Earth Syst Environ. 2025;11:246. https://doi.org/10.1007/s40808-025-02417-9.

6. WHO. Meningitis. 2025. https://www.who.int/news-room/fact-sheets/detail/meningitis. Accessed 5 Apr 2026.

7. Ali M, Moses A, Nakua EK, Punguyire D, Cheabu BSN, Avevor PM, et al. Spatial epidemiology of bacterial meningitis in the Upper West Region of Ghana: Analysis of disease surveillance data 2018–2020. Clin Infect Pract. 2022;16:100160. https://doi.org/10.1016/j.clinpr.2022.100160.

8. Mazamay S, Guégan JF, Diallo N, Bompangue D, Bokabo E, Muyembe JJ, et al. An overview of bacterial meningitis epidemics in Africa from 1928 to 2018 with a focus on epidemics "outside-the-belt." BMC Infect Dis. 2021;21:1–13. https://doi.org/10.1186/s12879-021-06724-1.

9. Nuoh RD, Nyarko KM, Nortey P, Sackey SO, Lwanga NC, Ameme DK, et al. Review of meningitis surveillance data, upper West Region, Ghana 2009-2013. Pan Afr Med J. 2016;25 Suppl 1:9.

10. Pace D, Gauci C, Barbara C. The epidemiology of invasive meningococcal disease and the utility of vaccination in Malta. Eur J Clin Microbiol Infect Dis. 2020;39:1885–97.

11. Ampadu B, Sackey I, Cudjoe E. Rainfall Distribution in the Upper East Region of Ghana, 1976 – 2016. Ghana J Sci Technol Dev. 2019;6:45–59. https://doi.org/10.47881/168.967x.

12. Issahaku A-R, Campion BB, Edziyie R. Rainfall and temperature changes and variability in the Upper East Region of Ghana. Earth Space Sci. 2016;3:284–94. https://doi.org/10.1002/2016EA000161.

13. Thomson MC, Molesworth AM, Djingarey MH, Yameogo KR, Belanger F, Cuevas LE. Potential of environmental models to predict meningitis epidemics in Africa. Trop Med Int Health. 2006;11:781–8.

14. Ali J, Khan R, Ahmad N, Maqsood I. Random forests and decision trees. Int J Comput Sci Issues IJCSI. 2012;9:272.

15. Breiman L. Random forests. Mach Learn. 2001;45:5–32.

16. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobotics. 2013;7. https://doi.org/10.3389/fnbot.2013.00021.

17. Ali ZA, Abduljabbar ZH, Taher HA, Sallow AB, Almufti SM. Exploring the Power of eXtreme Gradient Boosting Algorithm in Machine Learning: a Review. Acad J Nawroz Univ AJNU. 2023;12.

18. Kianfar N, Mesgari MS, Mollalo A, Kaveh M. Spatio-temporal modeling of COVID-19 prevalence and mortality using artificial neural network algorithms. Spat Spatio-Temporal Epidemiol. 2022;40:100471. https://doi.org/10.1016/j.sste.2021.100471.

19. Wu Y, Feng J. Development and Application of Artificial Neural Network. Wirel Pers Commun. 2018;102:1645–56. https://doi.org/10.1007/s11277-017-5224-x.

20. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst Their Appl. 1998;13:18–28. https://doi.org/10.1109/5254.708428.

21. Pal M, Mather PM. Support vector machines for classification in remote sensing. Int J Remote Sens. 2005;26:1007–11.

22. Akwetey SA, Osman A-H, Kotey FCN, Awere-Duodu A, Odoom A, Opoku-Asare B, et al. The epidemiology of infectious meningitis in Ghana: a systematic review and meta-analysis. BMC Public Health. 2025;25:2671. https://doi.org/10.1186/s12889-025-22476-y.

23. Dukić V, Hayden M, Forgor AA, Hopson T, Akweongo P, Hodgson A, et al. The Role of Weather in Meningitis Outbreaks in Navrongo, Ghana: A Generalized Additive Modeling Approach. J Agric Biol Environ Stat. 2012;17:442–60. https://doi.org/10.1007/s13253-012-0095-9.

24. Akanwake JB, Atinga RA, Boafo YA. Effect of climate change on cerebrospinal meningitis morbidities and mortalities: A longitudinal and community-based study in Ghana. PLOS Clim. 2022;1:e0000067.

25. Leimkugel J, Adams Forgor A, Gagneux S, Pflüger V, Flierl C, Awine E, et al. An outbreak of serotype 1 Streptococcus pneumoniae meningitis in northern Ghana with features that are characteristic of Neisseria meningitidis meningitis epidemics. J Infect Dis. 2005;192:192–9.

26. Hodgson A, Smith T, the … SG-T, undefined 2001. Risk factors for meningococcal meningitis in northern Ghana. academic.oup.com. 2001;95:477–80.

27. Martinón-Torres F, Taha M-K, Knuf M, Abbing-Karahagopian V, Pellegrini M, Bekkat-Berkani R, et al. Evolving strategies for meningococcal vaccination in Europe: overview and key determinants for current and future considerations. Pathog Glob Health. 2022;116:85–98.

28. Pelton SI. The global evolution of meningococcal epidemiology following the introduction of meningococcal vaccines. J Adolesc Health. 2016;59:S3–11.

29. Ayanlade A, Sergi C, Ayanlade OS. Malaria and meningitis under climate change: initial assessment of climate information service in Nigeria. Meteorol Appl. 2020;27:1–11. https://doi.org/10.1002/met.1953.

30. Nakazawa T, Matsueda M. Relationship between meteorological variables/dust and the number of meningitis cases in Burkina Faso. Meteorol Appl. 2017;24:423–31.

31. Yaka P, Sultan B, Broutin H, Janicot S, Philippon S, Fourquet N. Relationships between climate and year-to-year variability in meningitis outbreaks: a case study in Burkina Faso and Niger. Int J Health Geogr. 2008;7:34.

32. Kabanunye MM, Adjei BN, Gyaase D, Nakua EK, Afriyie SO, Enuameh Y, et al. Risk factors associated with meningitis outbreak in the Upper West Region of Ghana: A matched case-control study. PLOS ONE. 2024;19:e0305416. https://doi.org/10.1371/journal.pone.0305416.