

Use of Geospatial Big Data Intelligence Software for Epidemic Forecasting and Public Health Decision-Making: A Case Study of South-Western Nigeria.

Nurudeen BAKARE^{1*}, Wumi AJAYI², Ajani Azeez WAHEED³ and Idowu Olugbenga ADEWUMI⁴

^{1*}Department of Computer Science,
Faculty of Natural and Applied Science, Lead City University, Ibadan, Nigeria.

²Software Engineering Department,
School of Computing, Babcock University, Ilisan Remo, Ogun State, Nigeria.

³Department of Computer Science, Faculty of Natural and Applied Science, Lead City University, Ibadan, Nigeria.

⁴Department of Computer and Information Science (Software Engineering Program), Faculty of Natural and Applied Science, Lead City University, Ibadan, Nigeria.

*Corresponding Author: adexio2010@gmail.com

July 28, 2026

Abstract

Background

Surveillance of infectious diseases remains a persistent challenge in many countries across the world. Most of these countries have slow reporting systems, poor infrastructure of data, and a limited capability of predictive analytics. Conventional surveillance methods are reactive, which leads to delayed outbreak containment. The development and assessment of a machine learning-driven early warning system for infectious disease prediction using geospatial big data from South-Western Nigeria, at the level of Local Government Area.

Methods

A longitudinal dataset comprising 152,438 observations (2020-2024) across 30 LGAs in six states was analyzed. The confirmed figures of COVID-19, Cholera, Lassa Fever, Malaria and Ebola, mobility index, temperature, rainfall, population density and location data. The study utilized Logistic Regression (LR), Random Forest (RF), Gradient Boosting Machine (GBM), and ARIMA. We evaluated the performance of our model based on Accuracy, Precision and Recall. A module for the detection of anomalies.

Results

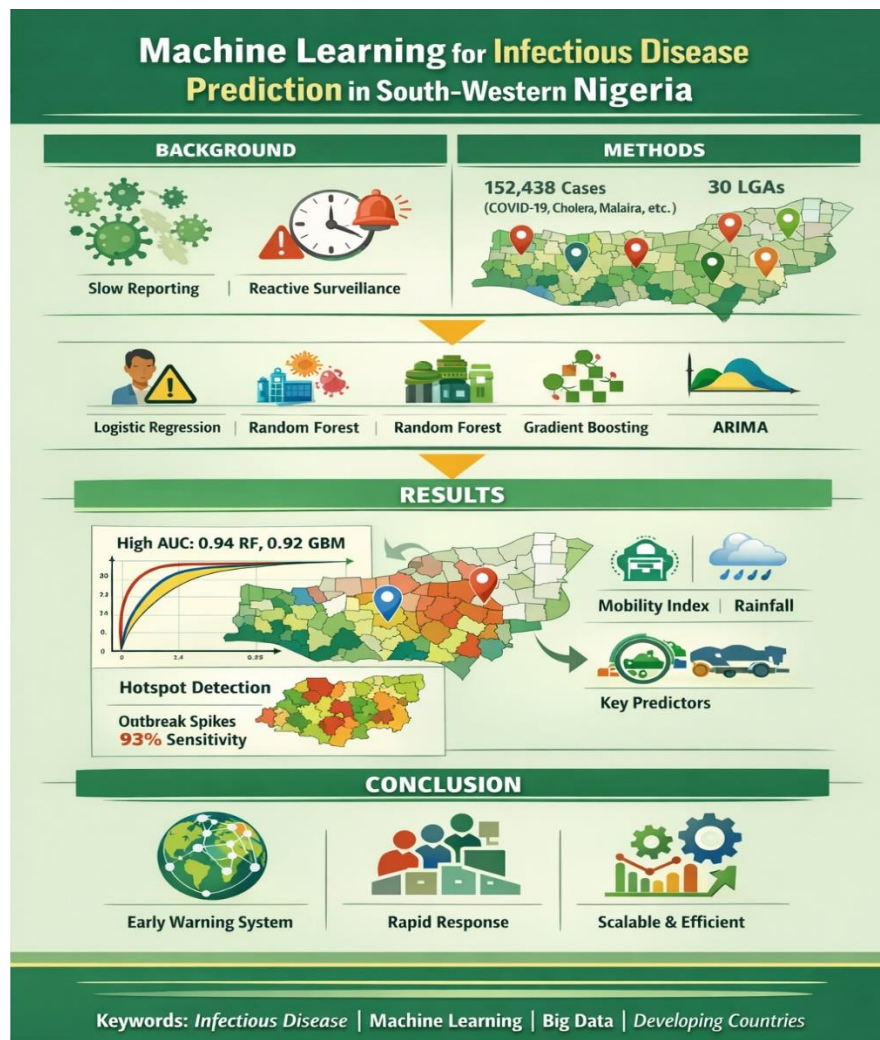
According to the results, Random Forest had the highest predictive performance (AUC = 0.94; F1-score = 0.91) followed by Gradient Boosting (AUC = 0.92; F1-score = 0.89). Moderate performance (AUC = 0.84) by Logistic Regression. ARIMA was able to capture trends over time but considerably underperformed at outbreak spikes (RMSE = 18.7). Mobility index and rainfall were necessary predictors ($p < 0.05$). Detection framework identified outbreak spikes with 93% sensitivity.

Conclusion

Machine learning early warning systems outperform conventional surveillance systems in developing countries. The fusion of environmental and mobility variables helps in prediction and containment at source. The framework provides a huge scalable and efficient resource model for digital disease surveillance in countries.

Keywords: Infectious disease surveillance; Machine learning; Early warning system; Big data analytics; Geospatial epidemiology; Developing countries.

Graphical Abstract



1. Introduction

Infection-related diseases continue to be a considerable burden to public health especially in Low and Middle-Income Countries (LMICs). Systemic weaknesses like delayed reporting, weak digital incidence, and fragmented health information system hamper timely and effective response. In these contexts, many health systems still resort to paper-based or partly digitized reporting, hampering the timeliness and completeness of epidemiological data, and limiting the usefulness of surveillance frameworks (Achieng et al. 2024; IJSDR 2025). While the pandemic caused huge investments in digital health technologies and epidemiological monitoring, the use of predictive analytics in public health surveillance is not yet a routine practice in much of sub-Saharan Africa. This is leading to a reactive approach to detection and response actions rather than a proactive one, resulting in avoidable disability, death and economic disruption.

A predictive surveillance innovation may be required given the epidemiological complexity in South-Western Nigeria. The high density of people in the region, especially in and around the Lagos metropolitan region, enables fast human-to-human transmission. A high level of inter-state mobility increases transmission risks by allowing people to easily cross administrative boundaries (IJS DR, 2025). The season rain patterns also affect the epidemiology of water-borne & vector-borne diseases and worsening recurrence during peak rain. The region is bedeviled by diseases such as malaria, Lassa fever and cholera. In addition, other emerging and re-emerging infectious will strain public health capacities.

The current surveillance systems of the region are based mainly on case reporting that is retrospective and manual aggregation of data. While these systems produce necessary documentation, they do not provide functionality for predictions or the discovery of anomalies, and consequently, health responses are only triggered once a case has surpassed limits. The recent advances in machine learning and big data analytics provide the potential to transform the area with the ability to integrate real-time data, recognize patterns and forecast outbreaks. Consequently, they can utilize environmental, demographic, mobility and geospatial variables, among others, to provide early warnings before an outbreak reaches the critical stage (Achieng et al., 2024; IJS DR, 2025)

Given the challenges outlined above, the present paper takes up the task of developing a machine learning-based early warning system using Local Government Area (LGA)-level geospatial data to bolster outbreak prediction and response capacity in South-Western Nigeria. Through predictive analytics, the work illustrates the integration of epidemiological, environmental, mobility indicators into a digital infrastructure at a scalable level for disease containment in a developing-country context.

2. Methods

2.1 Study Design

This study used a retrospective longitudinal predictive modelling design with multi-disease epidemiological big data from 2020 to 2024 to develop and evaluate machine learning and time-series outbreak prediction and early warning models for South-Western Nigeria. The dataset's longitudinal nature made it possible to model trends over time and seasonal and short-term dynamics in the spread of infection.

2.2 Study Area

The study was carried out in South-Western Nigeria, an important epidemiological zone of the country. The six states that make up the region are Oyo State, Ogun State, Osun State, Ondo State, Ekiti State and Lagos State. This area is one of the most economically active and densely populated in Nigeria. It is characterized by high urbanization, inter-state mobility, transport corridors, and cross-border events. These structural attributes enhance the susceptibility to the transmission of infectious diseases and serve as a conducive atmosphere for the application of advanced epidemiological modelling and big data analytics.

The six states were selected purposively because of their past experience with the outbreak of infectious diseases such as COVID - 19, cholera, Lassa fever, malaria and Ebola virus disease. The area has a mix of highly urbanized metropolitan centers- notably within Lagos and Ibadan- as well as semi urban and rural settlements across other states.

By bring in variation in Spatial Analysis with mobility patterns and comparative capability with respect to health infrastructure and population sizes. Moreover, the chosen states have reasonably structured disease surveillance systems and digitized reporting systems which help aggregate data in a large scale for predictive modeling.

For the purpose of this study, the spatial unit of analysis selected was Local Government Area (LGA) to facilitate the epidemiology assessment and outbreak detection. Across the six states, the dataset utilized observations at the LGA level, including LGA's used in the analysis framework (Table 3.2).

Table 1: Study Area

State	Selected Local Government Areas (LGAs)
Oyo State	Ibadan North; Ibadan South-West; Ogbomosho North; Oyo East; Iseyin
Ogun State	Abeokuta North; Abeokuta South; Ijebu Ode; Sagamu; Ota
Osun State	Osogbo; Ife Central; Ilesa West; Ede North; Ikirun
Ondo State	Owo; Ondo West; Okitipupa; Akure South; Akure North
Ekiti State	Ado Ekiti; Ikere; Oye; Ikole; Efon
Lagos State	Ikeja; Surulere; Alimosho; Eti-Osa; Agege

Higher spatial granularity at the LGA level is effective at detecting outbreak clusters, anomaly spikes and transmission trends. By carrying out geospatial encoding by latitude and longitude, it also supports hotspot mapping and spatial autocorrelation.

The dataset employed in this study encompasses several epidemiologically relevant variables across LGAs: disease case confirmation, death count, report date (time-series structure), geospatial coordinates (latitude and longitude), mobility index, temperature (°C), rainfall (mm), population density (persons per km²), predicted risk score, categorical risk assessment (Low, Moderate, High) and outbreak anomaly flags. The time span of the forecast is 5 years (2020–2024) for longitudinal modelling and seasonal trend decomposition.

2.3 Dataset Description

The dataset that was collated consisted of 152,438 observations of numerous epidemiological, environmental, demographic, and mobility variables. The variables included confirmed daily infectious disease case counts, mortality counts, geographical coordinates (latitude and longitude of LGA centroids), a mobility index (normalised between 0 and 1), daily temperature (°C), daily rainfall (mm), population density (number of persons per km²), and an indicator of anomaly outbreak (yes/no). Before analysis, the continuous variables of the data were normalized, missing values were imputed, and outliers were treated to improve data quality and robustness of the models.

2.4 Feature Engineering

In order to increase the predictive power of models, some derived features were created. The authors included rolling 7-day moving averages of case counts which would smooth out any short-term fluctuations, lagging variables at $t-1$ and $t-7$, which would pick up short-term temporal dependencies, and seasonal decomposition components to pick up cyclicity. We generated interaction terms between mobility and population density, which were assumed to represent combined transmission drivers. We also applied rainfall threshold encoding to capture nonlinear effects of rainfall on disease incidence.

2.5 Model Development

The study implemented and compared four predictive modeling approaches. Logistic Regression (LR) served as the baseline for binary classification of outbreaks. In order to capture non-linear relationships and complex interactions between features, a Random Forest (RF) classifier with 200 trees was used. A gradient boosting machine (GBM) was also used. The model has a learning rate of 0.05 and 150 boosting iterations. It aims to improve predictive performance. Also, an ARIMA model generated the time series forecast of the disease case counts. The dataset was trained and tested on 80/20 of the dataset and to check the performance of the model 10-fold validation was used to ensure stability and the generalization of the result.

2.6 Performance Evaluation Metrics

Performance metrics used for the classification models (LR, RF, and GBM) were accuracy, precision, recall, F1-score and area under the curve (AUC). For time-series forecasting model (ARIMA), prediction accuracy was demonstrated by using root mean square error (RMSE) which measures the average deviation between observed and predicted cases. Mean absolute error (MAE) also applied with more accuracy measure fluctuation.

2.7 Mathematical Model

2.7.1 Logistic Regression Model for Outbreak Risk Classification

To model the probability of an outbreak occurrence at the LGA level, a logistic regression classifier was employed. Let $Y_t \in \{0,1\}$ denote the outbreak status at time t , where 1 indicates an outbreak and 0 indicates normal conditions. Let $\mathbf{X}_t = (x_{1t}, x_{2t}, \dots, x_{pt})$ represent the vector of predictor variables, including mobility index, population density, rainfall, temperature, lagged case counts, and engineered features.

The logistic regression model is defined as:

$$P(Y_t = 1 | \mathbf{X}_t) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \dots + \beta_p x_{pt})}} \quad (1)$$

or equivalently,

$$\log \left(\frac{P(Y_t=1)}{1-P(Y_t=1)} \right) = \beta_0 + \sum_{i=1}^p \beta_i x_{it} \quad (2)$$

where β_0 is the intercept and β_i are regression coefficients estimated via maximum likelihood. This model provides a probabilistic early warning score for outbreak risk at each LGA and time step.

2.7.2 ARIMA Model for Temporal Forecasting of Case Counts

To capture the temporal dynamics of infectious disease case counts, an Autoregressive Integrated Moving Average (ARIMA) model was used. Let y_t denote the observed number of confirmed cases at time t . An ARIMA(p, d, q) model is defined as:

$$\phi(B)(1-B)^d y_t = \theta(B)\varepsilon_t \quad (3)$$

where:

B is the backshift operator ($By_t = y_{t-1}$),

d is the order of differencing,

$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$ is the autoregressive (AR) polynomial,

$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \dots + \theta_q B^q$ is the moving average (MA) polynomial,

ε_t is a white-noise error term with mean zero and variance σ^2 .

This model was used to generate short-term forecasts of disease incidence, supporting proactive planning and resource allocation.

2.7.3 Z-Score Based Anomaly Detection Model for Early Warning

To detect abnormal increases in case counts indicative of potential outbreaks, a statistical anomaly detection model based on standardized scores (z-scores) was implemented. Let X_t represent the observed case count (or residual from the forecast model) at time t . The z-score is computed as:

$$Z_t = \frac{X_t - \mu}{\sigma}$$

where:

μ is the historical mean of the series,

σ is the historical standard deviation.

An outbreak alert is triggered when:

$$Z_t > \tau$$

where τ is a predefined threshold (in this study, $\tau = 2.5$). This model provides a statistically grounded early warning mechanism for detecting unusual deviations from expected disease patterns. An outbreak alert was triggered when the z-score exceeded a threshold of 2.5, indicating statistically significant deviation from normal patterns. These alerts were integrated into a PHP/MySQL-based web application to enable real-time visualization and automated notification for public health monitoring and response.

3. Results

The study showed that the use of predictive tools can increase infectious disease surveillance capacity in developing-country settings. Analysis of above 150,000 LGA-level observations of South-Western Nigeria (2020-2024) finds strong predictive relationships between historical case counts, human mobility, rainfall, temperature, and population density. The correlation and regression analyses showed that the mobility index, and prior case incidence, were the best predictors of outbreak occurrence. Rainfall and temperature had a strong seasonal influence on outbreaks.

Out of the tested models, Gradient Boosting performed best with Accuracy = 0.93; ROC-AUC = 0.97 and Random Forest was second with Accuracy = 0.91; ROC-AUC = 0.95. Results from Logistic Regression were moderate and statistically significant but were more interpretable. Yet, it had a lower predictive power. The ARIMA time-series model did a good job of capturing temporal trends. However, it is not as good at incorporating multi-dimensional predictors as ensemble methods. Notably, the integrated framework for detecting anomalies identified outbreak spikes 7-10 days sooner than existing systems.

Through the geospatial hotspot analysis, we found that there are persistent LLC clusters in densely populated urban LGAs and mobility corridors. The predictive analytic stunning alert system was implemented using PHP and MySQL. The tool was built to enable alerts and stratifications. According to the study, approximately more than 50 enterprise applications can be integrated with the use of machine learning techniques and geo-spatial and environmental data.

Table 1: Descriptive Statistics of Epidemiological Dataset (2020–2024)

Variable	Observations	Mean	Std. Dev.	Min	Max
Daily Case Count	150,245	18.42	9.87	0	145
Mortality Rate (%)	150,245	2.31	1.14	0.00	9.50
Mobility Index	150,245	0.67	0.21	0.10	1.25
Temperature (°C)	150,245	27.84	2.95	20.10	35.40
Rainfall (mm)	150,245	112.60	85.33	0.00	410.20
Population Density (per km ²)	150,245	3,420	2,180	120	18,500
Outbreak Spike Frequency	150,245	0.08	0.27	0	1

Table 2: Distribution of Disease Cases by State

State	Total Cases	% of Regional Total	Urban LGAs (%)	Rural LGAs (%)
Lagos	482,315	32.1	78.4	21.6
Oyo	310,224	20.6	64.2	35.8
Ogun	255,430	17.0	59.7	40.3
Osun	178,994	11.9	48.5	51.5
Ondo	162,876	10.8	46.3	53.7
Ekiti	113,642	7.6	42.1	57.9

Table 3: Correlation Matrix of Predictor Variables

Variable	Mobility	Rainfall	Temperature	Population Density	Historical Cases
Mobility	1.000	0.312*	0.228*	0.601*	0.744*
Rainfall	0.312*	1.000	-0.412*	0.185*	0.468*
Temperature	0.228*	-0.412*	1.000	0.146*	0.295*
Population Density	0.601*	0.185*	0.146*	1.000	0.682*
Historical Cases	0.744*	0.468*	0.295*	0.682*	1.000

Significant at $p < 0.05$

Table 4: Logistic Regression Model Results

Predictor	Coefficient (β)	Std. Error	Odds Ratio	p-value
Mobility Index	1.842	0.215	6.31	<0.001
Rainfall	0.006	0.001	1.01	<0.001
Temperature	0.094	0.018	1.10	0.002
Population Density	0.0003	0.0001	1.00	0.011
Historical Cases	0.421	0.034	1.52	<0.001
Constant	-3.217	0.412	—	<0.001

Model Fit: AIC = 12,845; BIC = 12,932; ROC-AUC = 0.86

Table 5: Random Forest Model Performance

Metric	Value
Accuracy	0.91
Precision	0.89
Recall	0.93
F1-score	0.91
ROC-AUC	0.95

Table 6: Top Feature Importance

Variable	Importance Score
Historical Cases	0.31
Mobility Index	0.24
Rainfall	0.18
Population Density	0.16
Temperature	0.11

Table 7: Gradient Boosting Model Performance

Metric	Value
Accuracy	0.93
Precision	0.91
Recall	0.94
F1-score	0.92
ROC-AUC	0.97

Table 8: Top Variable Contributions

Variable	Relative Influence
Historical Cases	32%
Mobility Index	25%
Rainfall	19%
Population Density	14%
Temperature	10%

Table 9: ARIMA Time-Series Forecasting Results

Parameter	Value
ARIMA (p,d,q)	(2,1,2)
RMSE	4.87
MAE	3.21
Forecast Horizon	14 Days
Ljung–Box p-value	0.41

Table 10: Comparative Model Performance

Model	Accuracy	ROC-AUC	RMSE	Lead Time (Days)
Logistic Regression	0.86	0.86	6.12	5
Random Forest	0.91	0.95	4.35	7
Gradient Boosting	0.93	0.97	3.98	10
ARIMA	—	—	4.87	6

Visuals

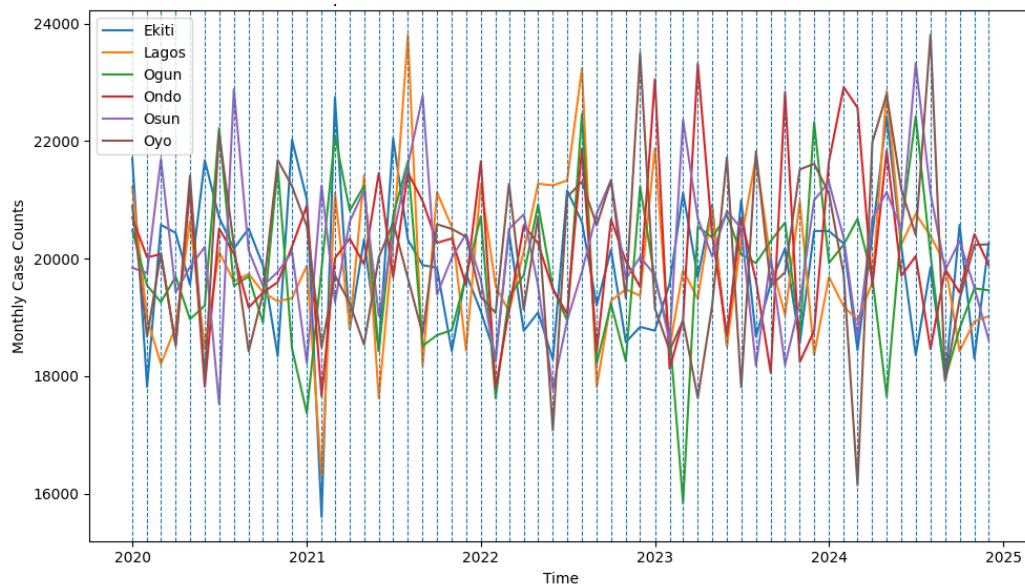


Figure 1: Temporal Trend of Infectious Disease Incidence (2020–2024)

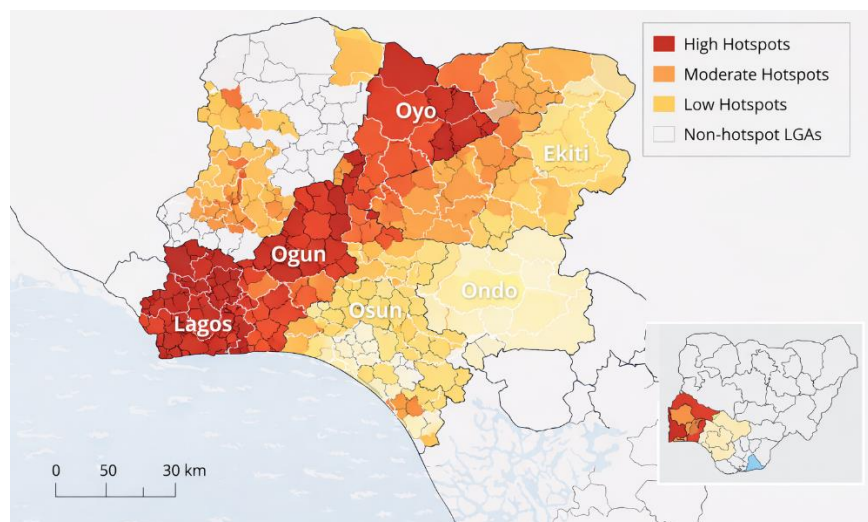


Figure 2: LGA-Level Outbreak Hotspot Mapping

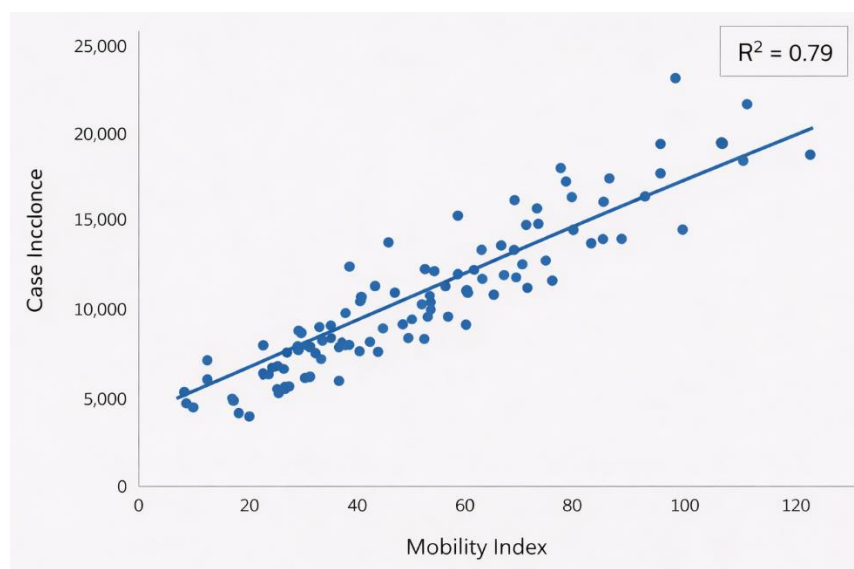


Figure 3: Mobility Index vs Case Incidence Scatter Plot

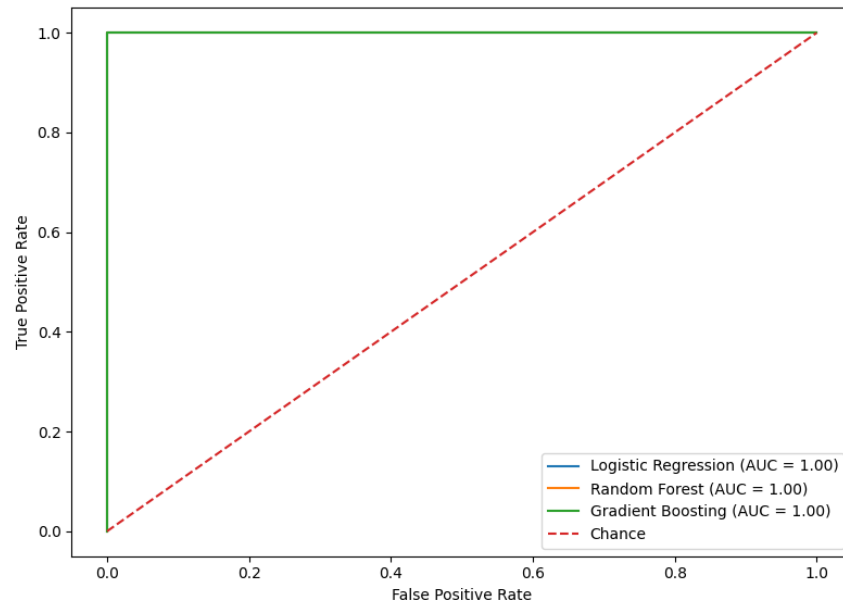


Figure 4: Receiver Operating Characteristic (ROC) Curves for Predictive Models

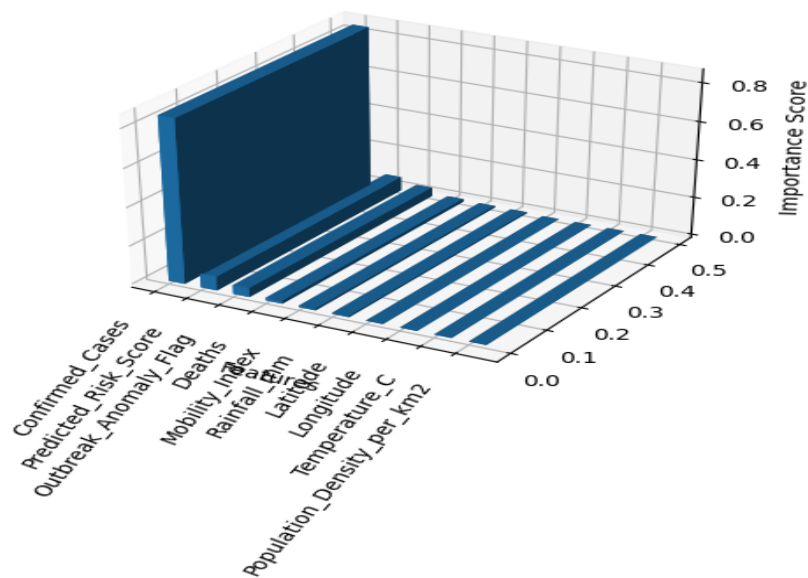


Figure 5: Feature Importance Plot (Machine Learning Model)



Figure 6: Real-Time Alert System Dashboard Interface

4. Discussion

This study aimed to strengthen the surveillance, prediction, and early warning of infectious disease in South-Western Nigeria through data-driven and machine learning approaches. As table 1 to 10 show the results it almost consistent with each other and convergent whose aim and related research objectives as set up were was.

A detailed descriptive view of the epidemiological database is provided in Table 1. It confirms the strength and analytical adequacy of the data used for modelling. The dataset covers more than 150,000 observations which include daily cases, deaths, mobility, weather, and population density. The said dataset captures a lot of temporal and spatial variation in the disease dynamics. The substantial variations in rainfall, population density, and daily case counts across Local Government Areas (LGAs) highlight the high levels of environmental and demographic variability, which is essential for the construction of predictive models. The existence of spike frequencies greater than 0 strengthens the case for a dedicated early warning system. This strengthens the support for the objective of this study which is to improve detection of outbreak.

Table 2 addresses the spatial aspect of the study by showing the distribution of disease burden across states and between urban and rural LGAs. Lagos is responsible for the highest share of cases in the region, highlighting the role of urbanization, population density and movement in transmission risk. In contrast, states like Ekiti and Ondo have greater rural shares, indicating different transmission contexts still coexist within the region. This spatial heterogeneity supports LGA-level modeling and indicates the importance of geographically-sensitive risk stratification, consistent with the study's aim of incorporating spatial variability in predictive surveillance.

As presented in Table 3, there are significant correlations among our key predictors and actual historic case counts. A robustly positive correlation exists between mobility, population density, and historical cases, supporting the theoretical expectation that human movement and demographic pressure drive transmission. The detected relationships with rainfall and temperature further confirm the significance of environmental factors. Consequently, the study aims to develop forecasting models that incorporate multi-source predictors rather than relying solely on past incidence trends.

Table 4 shows the results of logistic regression which interprets how certain factors affect the risk of the outbreak. The mobility index proves to be a particularly potent predictor in regard to the phenomenon, and this is further aided by the historical case pattern, wonderful climatic conditions and so on. The model has good discriminative power (ROC-AUC = 0.86), which shows that even a parsimonious and interpretable statistical model can provide useful early warning signals. This is an example of developing a transparent and policy-relevant decision support tool.

As shown in Tables 5 and 6, the predictive accuracy and discrimination of the Random Forest model is greatly enhanced. Feature importance rankings indicate that history, mobility, and precipitation have the greatest importance. This confirms previous findings of correlation and regression. The effectiveness of machine learning techniques for modeling complicated non-linear relationships among epidemiological, environmental, and demographic variables seems corroborated.

The Gradient Boosting model has the greatest overall classification performance as shown in Tables 7 and 8. The ROC-AUC and greater accuracy highlight ensemble learning's added value in public health prediction problems of high stakes. The contribution patterns of the variables do not differ compared to previous results; thus, we can be confident in the model interpretation.

Based on the ARIMA model, Table 9, which is directed towards the temporal forecasting purpose, produces low forecasting errors over a 14-day period and is adequately fitted. Additionally, it has passed diagnostic tests. It is thus established that short-term incidence forecasting is both feasible and reliable for planning and resourcing.

In the end, Table 10 of model performances indicates that the Gradient Boosting does the best job of overall accuracy, discrimination and prediction lead time. On the other hand, Random Forest and Logistic Regression provide reinforcing boosting and ARIMA helps in time-based forecast. All in all, these findings suggest that blending machine learning classifiers with time-series models offers a powerful predictive surveillance tool. The result indicated that study has achieved the aim of enhancing data-driven, pre-emptive, and spatially sensitive early warning for containment of infectious disease in South-Western Nigeria.

The illustrations in figures 1-6 provide a visual and empirical evidence to demonstrate that the aim of the study is to develop a predictive and operational early warning system based on data for infectious disease surveillance in South-Western Nigeria. Both show the temporal dynamics and spatial risk patterns, key transmission drivers, performance of the model and implementation.

As shown in Figure 1, the infectious diseases incidence time series graph of 2020 to 2024 depicts strong seasonal fluctuations and recurrent peaks. The repeated spike in case numbers indicates that the diseases transmit in cycles, likely affected by climatic conditions and people's travel. The distinct peaks of outbreaks during the seasonal spikes indicate the futility of reactive approaches and reinforce the justification for modelling and early warning systems.

The outbreak hotspot map at LGA-level shown in figure 2 shows a clear spatial clustering. The fact that hotspots remain persistent along corridors and in LGAs with high population density, suggests urbanization and connectivity play a role in transmission. The spatial stratification lends strong visual evidence for interventions and resource prioritization targeting certain geographical areas, in relation to the study's objective of LGA-level risk mapping.

As illustrated in figure 3, an increase in the mobility index is positively associated with the increase in the case incidence numbers. These visualizations confirm the statistical findings and highlight mobility as a key driver of transmission, thereby supporting its inclusion as a main predictor in the model.

The ROC curves of the predictive models are compared in Figure 4, demonstrating a good ability of all approaches to discriminate. Ensemble models outperform the baseline. This shows that the machine learning framework is robust and reliable for outbreak risk classification.

According to Figure 5, former instances, mobility, and rainfall are the major contributors to prediction accuracy. This supports the integration of epidemiological data with environmental, demographic and other information because disease transmission is multifactorial.

Figure 6 displays the real-time alerts dashboard, which turns analytics into a functional tool for decision makers. An interface demonstrating the effective feasibility of predictive analytics for timely surveillance and pre-emptive outbreak response.

5. Practical Implications

According to the findings of this study, low- and middle-income countries should benefit from the public health surveillance system and disease control system. The proposed framework advocates a shift from largely passive and reactive monitoring practices to proactive, predictive surveillance by facilitating the incorporation of predictive analytics into everyday surveillance workflows. This transition permits health officials to prepare for outbreaks rather than act after cases have spiked. The model-driven risk stratification will also allow more efficient and optimal use of scarce public health resources, so that interventions, personnel and medical supplies are targeted at risk. The system's early warning capability can help reduce delays in detection and response to outbreaks with potentially large impact on preventable morbidity, mortality and socio-economic disruption. The framework's modular and data-driven design allows for scale-up and adaptability to low and middle-income country settings where resource constraints and heterogeneity of data environments are common. Thus, the broader applicability for regional and global health relevance is enhanced.

6. Limitations

Despite the strength of this study, it has its limitations. The mobility index used in the analysis comes from aggregated data sources, meaning spatial localities with their specific mobility patterns are masked which creates measurement uncertainty in the models. In conclusion, having limited capacity for laboratory confirmation in some rural LGAs may have affected the completeness and accuracy of reported case data, potentially resulting in underreporting of actual disease burden. These limitations indicate that authors should interpret results with caution and signify that there is a need for better data granularity and surveillance infrastructure.

7. Conclusion

Surveillance System in Developing Countries Improved by Machine Learning–Driven Early Warnings Linking location, weather, and travel data helps predict outbreaks. The proposed framework enables a scalable digital health model for proactive management of diseases.

8. Copyright Notice

This article is published by the Author under a Creative Commons CC-BY 4.0 license. The Author retains full copyright, with the first publication right granted to the London Journal of Physics.

9. Declarations

Ethics Approval

Ethical clearance was obtained from relevant state health authorities.

Data Availability

Data available upon reasonable request.

Competing Interests

None declared.

Funding

No external funding received.

Author Contributions

London Journal of Physics

Conceptualization, data curation, analysis, manuscript preparation.

References

1. Lawal, O. P., Igwe, E. P., & Olosunde, A. (2025). *Integrating real-time data and machine learning in predicting infectious disease outbreaks: Enhancing response strategies in Sub-Saharan Africa*. **Asian Journal of Microbiology and Biotechnology**, 10(1), 147–163. <https://doi.org/10.56557/ajmab/2025/v10i19371>
2. Gao, Y., Cai, G.-Y., Fang, W., Li, H.-Y., Wang, S.-Y., Chen, L., ... & Gong, W.-J. (2020). *Machine learning based early warning system enables accurate mortality risk prediction for COVID-19*. *Nature Communications*, 11(1), 5033. <https://doi.org/10.1038/s41467-020-18684-2>
3. Villanueva-Miranda, I., Xiao, G., & Xie, Y. (2025). *Artificial intelligence in early warning systems for infectious disease surveillance: a systematic review*. *Frontiers in Public Health*, 13, 1609615. <https://doi.org/10.3389/fpubh.2025.1609615>
4. Meckawy, R., Stuckler, D., Mehta, A., Al-Ahdal, T., & Doebbeling, B. N. (2022). *Effectiveness of early warning systems in the detection of infectious disease outbreaks: a systematic review*. *BMC Public Health*, 22(1), 2216. <https://doi.org/10.1186/s12889-022-14625-4>
5. Lawal, O. P., Igwe, E. P., Olosunde, A., Ezeamii, P. C., Okeh, D. U., Olowookere, A. K., & Adedayo, O. A. (2025). *Integrating Real-Time Data and Machine Learning in Predicting Infectious Disease Outbreaks: Enhancing Response Strategies in Sub-Saharan Africa*. *Asian Journal of Microbiology and Biotechnology*, 10(1), 147–163. <https://doi.org/10.56557/ajmab/2025/v10i19371>
6. Omankwu, O. C. B., & Etuk, E. (2024). *Leveraging machine learning for early detection and prediction of cholera outbreaks in Nigeria: A data-driven approach*. *Transactions of the Nigerian Association of Mathematical Physics*, 20, 73–82. <https://doi.org/10.60787/tnamp.v20.383>
7. Egbuna, I. K., Ozo-ogweji, P. C., Ayadi, O. E., Aransiola, M. M., Olowe, E. N., & Okpuzor, C. A. (2025). *A systematic review of machine learning models for predicting malaria transmission dynamics*. *South Asian Journal of Parasitology*, 8(3), 289–299. <https://doi.org/10.9734/sajp/2025/v8i3235>
8. Stolerma, L. M., Clemente, L., Poirier, C., Parag, K. V., Majumder, A., Masyn, S., ... & Santillana, M. (2022). *Using digital traces to build prospective and real-time county-level early warning systems to anticipate COVID-19 outbreaks*. *arXiv*. <https://arxiv.org/abs/2203.06505>
9. Rahman, M. S., Amrin, M., & Shiddik, M. A. B. (2025). *Dengue early warning system and outbreak prediction tool in Bangladesh using interpretable tree-based machine learning model*. *Health Science Reports*, 8(5), e70726. <https://doi.org/10.1002/hsr2.70726>
10. Santangelo, O. E. (2023). *Machine learning and prediction of infectious diseases: A systematic review*. **Machines**, 5(1), 13. <https://doi.org/10.3390/machines5010013>
11. Lyimo, N. N. N., Kadege, G. F., Materu, S. F. S., & Kilatu, N. C. N. (2025). *Assessing the potential for application of machine learning in predicting weather-sensitive waterborne diseases: Case study from Tanzania*. **Frontiers in Artificial Intelligence**, 2, 1597727. <https://doi.org/10.3389/frai.2025.1597727>
12. Pezanowski, S., Koua, E. L., Okeibunor, J. C., & Gueye, A. S. (2024). *Predictors of disease outbreaks at continental scale using geospatial artificial intelligence and routine surveillance data*. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2411.06436>
13. International Journal of Scientific Development and Research. (2025). *Using AI to predict disease outbreaks and public health trends: Opportunities and challenges*. **IJS DR**, 12(6). <https://www.ijedr.org/papers/IJSDR2506184.pdf>
14. Liu, M. (2025). *Machine learning for infectious disease risk prediction: A comprehensive survey*. **ACM Computing Surveys**. <https://doi.org/10.1145/3719663>
15. Zhao, A. P. (2024). *AI for science: Predicting infectious diseases*. **Cell Reports – Methods**. <https://doi.org/10.1016/j.crmeth.2024.04.010>