

PREDICTIVE MODELLING OF LASSA FEVER OUTBREAKS: A COMPARISON OF DEEP LEARNING AND ENSEMBLE APPROACHES

Adewole, D. B.



DOI10.51459/jostir.2026.2.2.0269

Department of Software Engineering, Federal University of Technology, Akure, Nigeria

Correspondence

dbadewole@futa.edu.ng

History

Received: 28-03-2026

Accepted: 15-06-2026

Published: August, 2026



<https://www.futa.edu.ng>



<https://jostir.futa.edu.ng>

ABSTRACT

Lassa fever continues to be an important health problem and a very dangerous epidemic in West Africa, especially in Ondo State of Nigeria where large outbreaks occur annually. Maintaining high accuracy in forecasting is crucial for an early response by medical practitioners, however the Lassa virus' complex spatiotemporal transmission and delayed ecological triggers make it difficult to predict. To overcome this, the current work proposes an all-inclusive health information pattern discovery framework to predict outbreaks in four high-risk localized hotspots namely, Akure South, Akure North, Akoko Southwest and Owo. This study goes beyond traditional machine learning performance comparison, and introduces the integration of spatial-temporal graph theoretic modelling to represent the relational transmission velocity between regions, as well as the use of Genetic Algorithm (GA) to optimize intricate lagged meteorological variables, including temperature, relative humidity, and precipitation. The refined spatiotemporal feature space was employed for the training and evaluation of four distinct models: Random Forest, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Long Short-Term Memory (LSTM) networks. The GA-optimized Random Forest ensemble model yielded the most favourable outcomes, with a Root Mean Squared Error (RMSE) of 5.627, a Mean Absolute Error (MAE) of 3.747, and a R^2 value of 0.87 in the empirical assessment. In addition to the predictive accuracy of its baseline system, this hybrid graph-theoretic and evolutionary method can effectively isolate the exact environmental conditions that precede an outbreak, which can be interpreted and translated into actionable intelligence to improve proactive, public health planning and targeted interventions in endemic areas.

Keywords: Health Information Pattern Discovery, Spatial-Temporal Graph Modelling, Genetic Algorithm, Lassa fever, Random Forest.

1. | Introduction

Lassa Fever is an endemic disease in Africa. It was initially discovered in a town named Lassa, adjacent to Maiduguri, the capital city of Borno State, Nigeria in 1969 (Tewogbola & Aung, 2020). It is a serious viral illness endemic to West African nations, including Sierra Leone, Mali, Ghana, Benin, Liberia, and Nigeria (Centre

for Disease Control and Prevention, 2025). As at the year 2024, Ondo State is among the most impacted regions in Nigeria, representing over 30% of the nation's total confirmed Lassa fever cases, as reported by the Nigeria Centre for Disease Control (NCDC). This situation designates it as a hotspot (NCDC, 2024). The origin of the disease can be traced to the activities of a species of rodents

called the multimammate rodent. Exposure to this multimammate rat, its faeces, urine or contaminated food is the source of transmission of Lassa fever. Approximately 20% of confirmed Lassa fever cases result from person-to-person transmission through contact with bodily fluids, including saliva, urine, blood, and vomit, particularly in hospitals and medical facilities, thereby placing healthcare personnel at significant risk (Akhmetzhanov *et al.*, 2019). Lassa fever remains a severe public health issue due its mode of transmission. Common symptoms often experienced include high fever, slight or severe headache as well as distress in respiration. According to United Nations, about 80% of infected individuals experience no symptoms and only 1 in 5 recorded cases result to severe disease, like infection of the liver, spleen and kidneys. It is recorded that the overall fatality rate is 1% while recorded 15% of fatality rate for patients who are hospitalized with severe clinical case. Early supportive care given to infected individuals include rehydration and symptomatic treatment to improve survival (World Health Organization, 2023).

The current trend of recent Lassa fever outbreak and the challenges faced in surveillance and delayed response, has provided the need for a reliable predictive tool to solve these challenges. Machine learning techniques have proven to be effective in terms of its advanced and portable approach, inexpensiveness as well as their accuracies in prediction. Machine learning techniques have been adopted in several fields and sectors which include weather forecasting, forex trading and e-commerce as well as medical studies like epidemiology and COVID-19. Nonetheless, it is essential to reconcile academic models with the actual requirements of public health. This raises the inquiry for this research: which machine learning model yields the most precise predictions of Lassa fever outbreaks, utilizing historical climatic and epidemiological data? This study assesses machine learning models, including Random Forest, LSTM, LightGBM, and

XGBoost, for predicting Lassa fever in the four most impacted regions of Ondo State, Nigeria: Akure South, Akure North, Akoko, Akoko Southwest, and Owo. The documented rise in incidence across these four hotspots in Ondo State will provide valuable insights into the utilization of machine learning algorithms for epidemiological disease forecasting. The findings of this research will aid government officials, health professionals, and other stakeholders in making informed decisions. The primary novelty of this research lies in the integration of spatial-temporal graph modelling with Genetic Algorithm optimization to capture the relational transmission velocity of Lassa fever which is an important approach that overcomes the limitations of isolated, tabular machine learning models.

2. | Review of Related Works

The traditional approach for predicting infectious diseases use statistical and time-series models, such as the autoregressive integrated moving average (ARIMA) model (Zhang, 2024). A study conducted in Guangxi, China, demonstrated that machine learning models were more proficient in properly forecasting disease outbreaks. The study conducted by Mohan *et al.* (2025) evaluates machine learning algorithms against time-series models for forecasting HIV cases in Guangxi, China. It employees four models to forecast HIV prevalence in Guangxi, China. The employed models were Generalized Regression Neural Network (GRNN), LSTM, Exponential Smoothing (ES) and ARIMA. The findings indicated that LSTM outperforms GRNN, ARIMA, and ES. This underscores how efficient machine learning algorithms are in managing intricate non-linear data and also support efforts and prior research works on the use of machine learning algorithms in predicting infectious diseases, including Lassa fever, utilizing historical data on disease outbreak.

In the works of Cakir *et al.* (2020), Oluwole & Nkonyana (2022), Ukarande *et al.* (2026) and Zhang *et al.* (2024), the progression of Lassa fever by the application of Decision Tree (DT) and k-nearest neighbour (KNN) models incorporating temporal data is investigated. The models were assessed in comparison to the Seasonal Autoregressive Integrated Moving Average (SARIMA) model. The findings demonstrated that KNN with grid search predictions surpassed the performance of the other models. The standard KNN achieved a lower MAE, whereas the decision tree with grid search similarly yielded a reduced mean absolute percentage error (MAPE). The studies revealed the effectiveness of machine learning models in addressing the challenges of Lassa fever outbreak progression.

Alile (2020) suggested a simulated supervised machine learning method known as the Bayesian Belief Network for the diagnosis of Lassa fever and viral haemorrhagic fevers. Bayes Server was used in developing the model and tested using the information from medical repository of viral haemorrhagic fever. On test data, the model achieved 100% prediction accuracy, and was highly sensitive at 98.73% for Lassa fever and 98.8% for viral haemorrhagic fever diagnosis. However, the most important problem in this research is the insufficient specification and improper data collection mode. The data was obtained from a third source, which could have affected the quality of the dataset used in training and testing the model.

The problem of influenza epidemics in China is a very important social and economic threat to the country and Liang *et al.* (2018) attempts to proffer solution to the problem by tackling the problem of reporting lags in the conventional disease surveillance system. The research team wanted to determine if they could enhance their prediction of epidemics in Liaoning province by combining together information about the current search from the internet and the historical medical data. This study designed a methodological

approach by which monthly influenza incidence data of the official reporting system were correlated with targeted search queries related to influenza reported on Baidu search engine from 2011 to 2015. They trained and validated a Support Vector Machine (SVM) regression model, and used leave-one-out cross validation (CV) to fine-tune predictive parameters. The most significant result obtained in this research is that the overall predictive accuracy and robustness of an ensemble SVM model based on an integration of the query information from Baidu and the traditional surveillance data are much higher than that of the single-source models. But there's a caveat to the study: the authors rightly admit that media-driven public panic and fast changing “slang” on the web can easily alter the validity of search query tracking, and, more significantly, a major outbreak peak in March 2015 was completely overlooked because of unaccounted for climatic factors.

Chen *et al.* (2024) addressed the importance of zoonotic diseases as a serious public health issue worldwide and the urgent need for timely and accurate outbreak prediction, which has been challenging with traditional diagnostic tools. To tackle this, the researchers conducted a study of monthly incidence data of schistosomiasis, echinococcosis, and leptospirosis in China from 2004 to 2019. They used additive time series decomposition to decompose the raw data into the trend, seasonal and random residual components of the time series. The smoothed trend data was then given as an input to a LSTM neural network. The study successfully implemented a hybrid LSTM model which reduces the statistical noise from the conventional LSTM model and also significantly outperforms the conventional LSTM model by using the raw datasets. The model, however, is very sensitive to extreme outbreak peaks, which may produce an over-or-underweighting of Mean Square Error (MSE) assessments. Moreover, LSTM models require especially contiguous data which requires

some prior filling of gaps and increases the training time of the model.

Mehta and Patel (2024) employed a LSTM model to forecast dengue epidemics in ten districts of Gujarat, India. The data has been aggregated throughout the past decade (from 2010 to 2019). The dataset from 2010 to 2016 was utilized for training, while the dataset from 2017 to 2019 was employed for testing the model. The variables included to train the model are the number of dengue cases, population density, mean temperature, mean humidity, monthly precipitation, and dengue cases recorded one, two, and twelve months prior. The LSTM model was assessed, yielding an RMSE of 0.04 and an R-squared (R^2) of 0.84. The LSTM surpassed other machine learning methods such as ANN, linear regression, and random forests in predicting dengue cases across many states and nations.

In Bakari *et al.* (2024), a machine learning model that can predict the emergence of Lassa Fever, a serious bleeding disorder common in West Africa, is introduced. The motivation for the research comes from problems with early clinical diagnosis, due to the non-specific symptoms of the disease and the high cost of the specialized testing laboratories available. To tackle this, the researchers used clinical data from the Federal Medical Centre, Abuja, and secondary data from Kaggle repositories. They split the data into 80% training and 20% testing, and tested algorithms including SVM, K-Nearest Neighbour, Naïve Bayes and Logistic Regression. The main contribution of the study is the optimized SVM model which has achieved 90% accuracy and 96% F1 score. This machine learning method is an important adaptive capability over the previous generation of rule-based expert systems. A major drawback of the study, however, is the small number of instances studied, with just 200 instances. The absence of strong data significantly limits the overall reliability of this model and its direct clinical relevance in the real world unless it is re-trained.

Okwuraiwe *et al.* (2025) is motivated by the fact that, Lassa fever is a serious public health problem in West Africa with high mortality rate that is attributed to delayed diagnosis because of its non-specific symptoms and absence of rapid diagnostics. This systematic review, based on 19 studies and 4,177 patients, offers valuable insights into clinical outcomes. The meta-analysis has a 19% pooled mortality rate for severe Lassa fever, contributing significantly to medical knowledge. Also, it measures the presence of some important complications, such as acute kidney injury (19%), abnormal bleeding (17%) and central nervous system dysfunction (15%). These specific parameters are useful for directing clinical interventions and highlighting the critical need for new therapeutics and vaccines.

To foresee the short-term trend of viral infections, Jiao *et al.* (2025) introduced a deep learning model combining LSTM networks and a multi-head attention mechanism. The COVID-19 cases and hospitalisation data was used for a case study in Japan. The method provides weekly updates and forecasts results in 1-4 weeks. Data collected from open-access platforms were used to train the model. The authors have tested its forecasting ability for weekly forecasts over a long time period using various error measures to quantify the model performance in various regions of Japan. The results of the research show that the proposed model performs better than the baseline models, especially in terms of short term prediction, with lower error rates in various error measures.

3. | Materials and Method

In various infectious diseases forecast, Machine Learning (ML) is a data driven approach that has been utilized for pattern learning based on historical data. Compared to traditional epidemiological models, ML models complex nonlinear relationships between many factors such as socioeconomic factors, popularity density and climate data. The study

was expanded to come up with a comprehensive health information pattern discovery model based on graph theoretic structures and genetic algorithm optimization for capturing the spatiotemporal trans-

mission dynamics of Lassa fever beyond standard predictive modelling. The conceptual workflow diagram is shown in Figure 1.

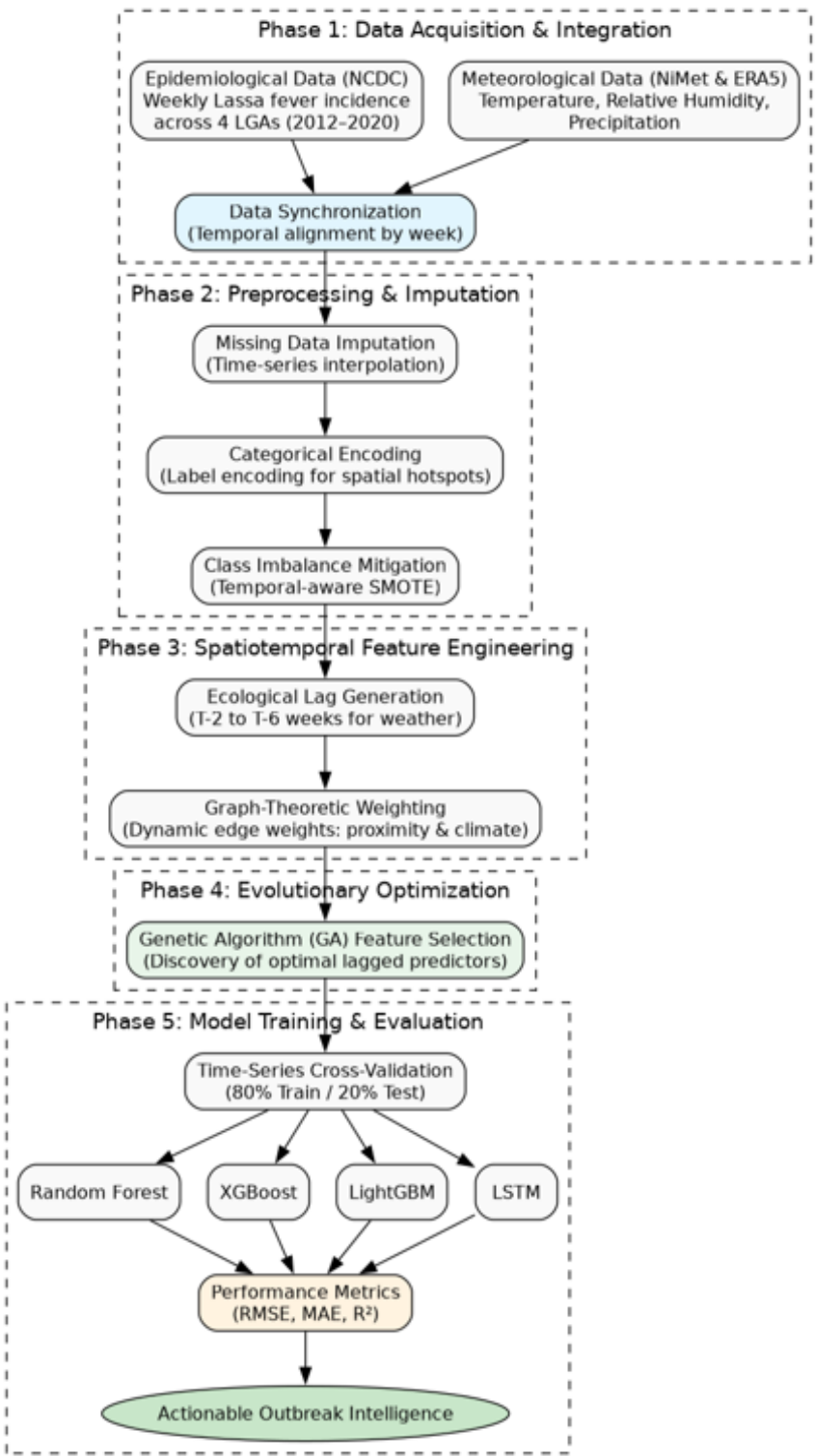


Figure 1 | Conceptual Workflow of the System

3.1 | Data Sources

Two datasets are used for this research. The data sets include the Lassa fever incidence data and weather data for the period 2017-2020. The datasets consist of 4388 records of Lassa fever cases having features such as date of occurrence, week, year, status of incidence which has variables like suspected, confirmed and negatives cases. The meteorological data has 17532 records which contains features such as air temperature, relative humidity and precipitation. To analyze the two sets of data, they were synced with each other so that each week in the

study period contained the number of Lassa fever cases and the weather data for that week.

- Epidemiological Data:** The source of the data is from the published reports of NCDC on epidemiological reports. These reports are from Lassa fever recorded incidence from Akure South, Akure North, Akoko South-west, and Owo Local Government Areas (LGAs) within the period of 2017 to 2020.

Figure 2 shows the prevalence of Lassa fever each of the four local government areas.

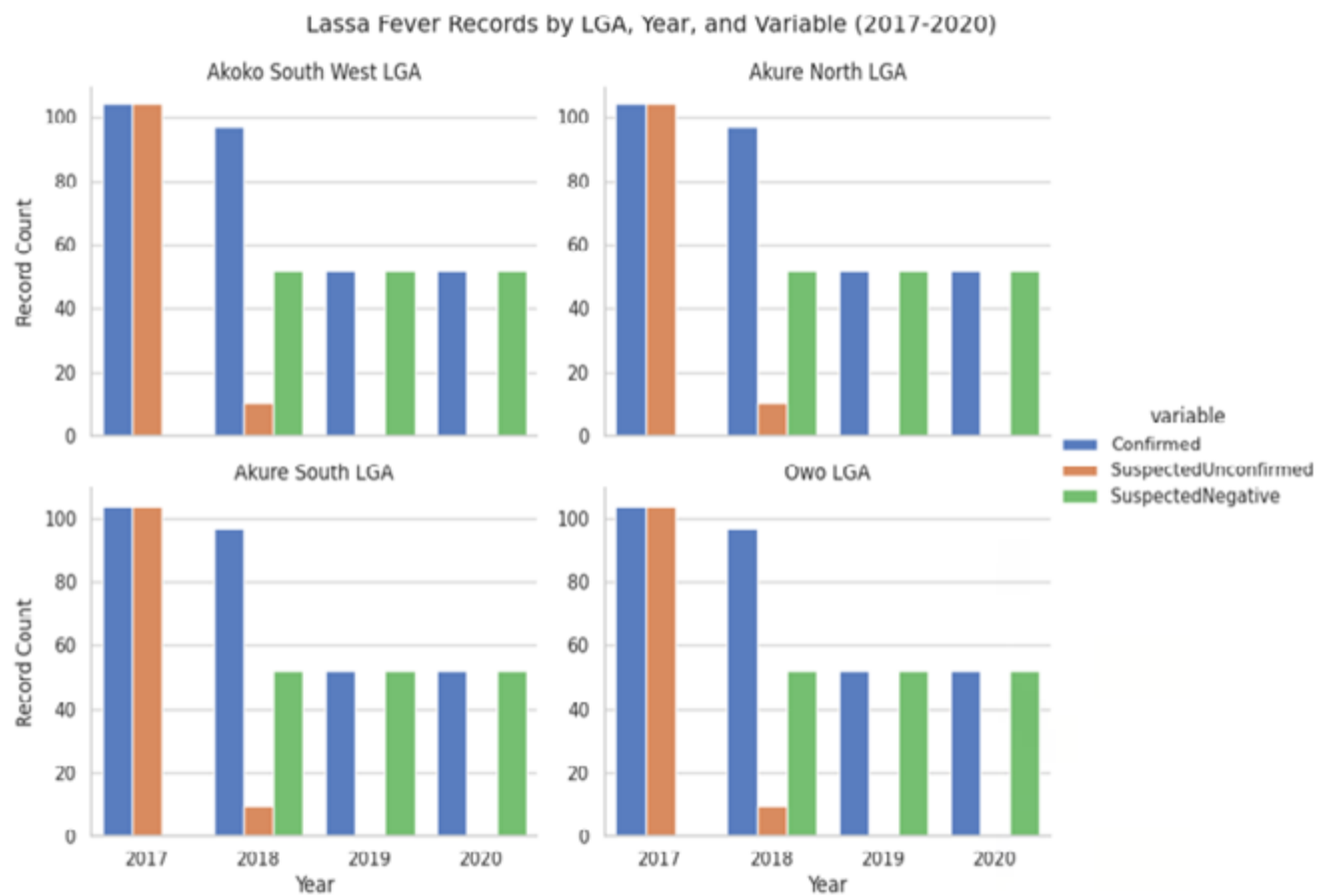


Figure 2 | Cases of Lassa Fever in 4 LGAs of Ondo State

- Meteorological Data:** The source of the climate data to be analysed is obtained from the Nigerian Meteorological Agency (NiMet) and European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5. Samples of the weather features considered are shown in Figures 3 to 5.

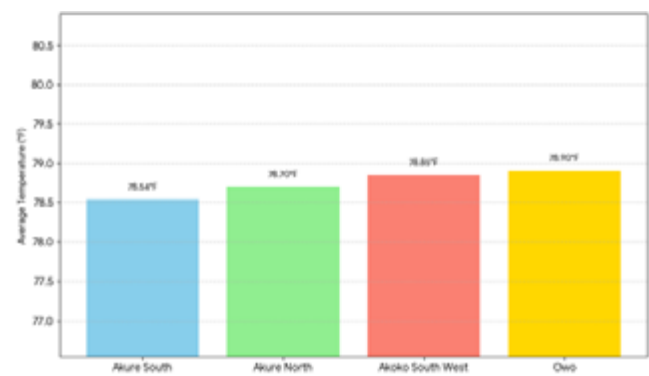


Figure 3 | Average Air Temperature per LGA

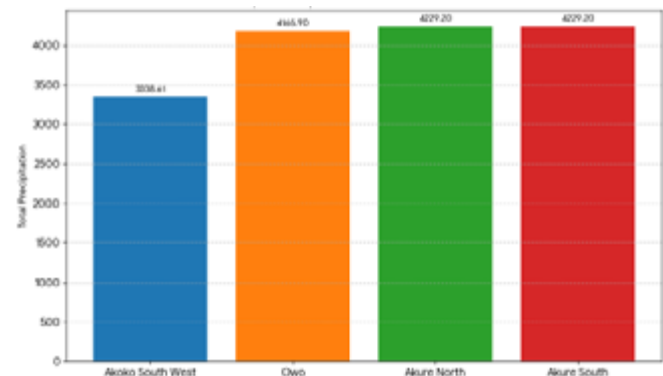


Figure 4 | Total Precipitation by LGAs

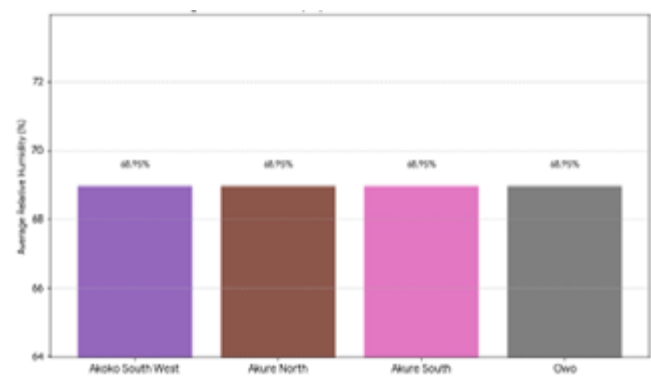


Figure 5 | Average Relative Humidity by LGAs

The variables that are considered for the data are shown in Tables 1 and 2.

Table 1 | List of the Lassa Fever Data Module Variables

S/N	Variable Name	Data Type	Description
1	Wk_No	Integer	This is the week that the incidence occurred
2	Yr	datetime	Year of occurrence
3	No_of_cases	Integer	This is the number of reported cases
4	LGA	String	This is the local government area where the incidence occurred
5	Date	datetime	The date the incidence occurred

Table 2 | List of the Weather Data Module Variables

S/N	Variable Name	Data Type	Description
1	Wk_No	Integer	This is the week that the incidence occurred
2	Date	datetime	The date the incidence occurred
3	Yr	datetime	Year of occurrence
4	Mean_Temp	Float	This is the average weekly temperature
5	Mean_Rel_Humidity	Float	This is the average weekly relative humidity
6	Mean precipitation	Float	This is the average weekly precipitation

3.2 | Data Preprocessing and Feature Engineering

The original data was pre-processed, and categorical features, such as LGAs, were transformed into numerical values using Label Encoding. To capture the delayed ecological effects of climate on rodent breeding cycles and subsequent viral spillover, lagged features were developed to analyze the delayed effects of weather on epidemics, incorporating temperature and rainfall data from 2 to 6 weeks prior.

Small discontinuities in the meteorological data were filled in using linear time-series interpolation, a method that allows for the estimation of missing data from the data using the climatic trend without altering the overall seasonal characteristics. In addition, epidemiological datasets are quite class imbalanced because severe outbreak weeks are less common than non-outbreak weeks. In order to avoid the models becoming biased towards the majority class (zero or low incidence), the data set was balanced during the training phase using a temporal-aware Synthetic Minority Over-sampling Technique (SMOTE). This enabled the models to capture the complex ecological conditions that lead to high incidence periods, rather than simply forecasting the average incidence.

3.3 | Health Information Pattern Discovery Model

To accurately capture the dissemination of Lassa disease, the predictive architecture combines an ensemble of base learners with spatial and evolutionary optimization techniques.

3.3.1 | Spatial-Temporal Graph Modelling

Rather than treating the localized hotspots as isolated categorical entities, the region is modelled using a graph theoretic approach. The LGAs are

represented as interconnected nodes, where edge weights dynamically reflect geographical proximity and shared meteorological shifts. This allows the model to map the spatial velocity of Lassa fever outbreaks across borders. To model the transmission dynamics across the four Local Government Areas (LGAs), we represent the region as an undirected, dynamically weighted graph. Let the graph be defined as

$$G = (V, E, W) \quad (1)$$

Nodes V : $V = \{v_1, v_2, v_3, v_4\}$ represents the four hotspots.

Edges E : $E \subseteq V \times V$ represents the connections between the LGAs.

Dynamic Edge Weights W : The weight $W_{ij}^{(t)}$ between node i and node j at time t combines geographic proximity and meteorological similarity:

$$W_{ij}^{(t)} = \alpha \frac{1}{d_{ij}} + \beta \exp(-\gamma \|M_i^{(t-k)} - M_j^{(t-k)}\|^2) \quad (2)$$

Where:

d_{ij} is the spatial distance between LGA i and LGA j .

$M_i^{(t-k)}$ is the meteorological feature vector for node i at lag period k .

α, β, γ are tuneable coefficients balancing spatial and environmental influence.

3.4 | Evolutionary Optimization

3.4.1 | Genetic Algorithm (GA) Feature Optimization

Determining the exact combination of lagged weather variables that precede an outbreak is a highly complex search problem. A GA is deployed

to systematically discover the optimal subset of lagged features (from the 2 to 6-week windows) and tune the hyperparameters of the predictive models. The GA evolves the feature subsets over multiple generations, optimizing for prediction accuracy while penalizing overfitting, thereby isolating the exact environmental triggers for Lassa fever transmission.

3.4.2 | Genetic Algorithm Hyperparameter Configuration

The GA was carefully parameterized to assure computational repeatability. The evolutionary search started with the population size of 50 chromosomes, with each gene corresponding to the presence or absence of a particular meteorological feature lagged. This algorithm has been optimized for more than 100 generations. Parent solutions were selected using a tournament selection method (Size of the tournament: 3). The crossover rate was 0.8, and a uniform mutation rate of 0.05 was used to prevent premature convergence to local optima, and maintain genetic diversity. The fitness function employed to guide the evolutionary process was the inverse of Root Mean Squared Error (RMSE) produced by a base decision tree fitness evaluator, which always searches for the best predictive performance while reducing the number of features that are not predictive.

3.5 Model Training and Evaluation

Four models were trained: Random Forest, XGBoost, LightGBM and LSTM. The optimized feature sets derived from the graph and GA frameworks were utilized to train and evaluate these predictive models.

The dataset was split into a training set (80%) and a testing set (20%). To ensure the robustness of the predictions and mitigate the risk of temporal data leakage, time-series cross-validation was employed.

Model performance was evaluated on the test set using the following metrics:

RMSE: This statistic quantifies the square root of the mean of the squared deviations between expected and actual values. The mathematical equation is shown below

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

- i. **MAE:** This statistic quantifies the average magnitude of errors between expected and actual values, disregarding their direction. The mathematical formula is expressed below.

$$MAE = \frac{1}{N} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4)$$

- ii. **R-squared:** The coefficient of determination, this measure signifies the proportion of variance in the dependent variable that can be predicted from the independent variable. The mathematical formula is expressed below:

$$R^2 = 1 - \frac{SSE}{SST} \quad (5)$$

3.6 Description of Models

3.6.1 | Tree-based Models

Random Forest: This is an ensemble learning technique. It generates numerous decision trees during the training phase and produces the mode of the classes for classification or the mean prediction for regression from the individual trees.

XGBoost: This is a well-regarded implementation of Gradient Boosting Decision Trees (GBDT) that is both highly scalable and optimized. It is well known for its speed and effectiveness in machine learning competitions. XGBoost enhances GBDT by incorporating a more systematic approach

to regularization, parallel processing, and the management of missing values, all of which enhance its robustness and efficiency.

LightGBM: It is a gradient boosting framework recognized for its rapidity and ability to scale, making it suitable for concurrent usage by several users. It was created by Microsoft to ensure efficient performance for extensive datasets. There are two major attributes of LightGBM that boost its effectiveness which are Gradient-based One-Sided Sampling (GOSS) which minimizes training duration and memory. The second feature is Exclusive Feature Bundling (EFB) which streamlines the dataset by combining unrelated features, for the reduction of complexity without sacrificing accuracy (Ke *et al.*, 2017).

3.6.2 | Deep Learning Model: Rationale for LSTM Design and Architecture

A LSTM network is particularly used in this study to capture the deep, non-linear sequential dependency that characterizes the transmission of infectious diseases, in addition to the interactions between the tabular variables. Tree-based ensembles are quite good at handling interactions between tabular data, but the study explicitly incorporated an LSTM network to capture the deep non-linear sequential dependence of infectious disease transmission. Outbreaks of Lassa fever are not random events, but rather the result of a series of delayed ecological changes (such as increased rainfall, followed by high temperatures creating conditions for the breeding of rodents). To capture this "memory" in time, the LSTM architecture was specifically designed.

The built network is fed with input sequences that are defined by a time-step of 6 weeks, directly related to the maximum ecological lag window that is assessed. The architecture is two stacked hierarchical LSTM layers containing 64 and 32 hidden units, respectively, to capture hierarchical

temporal patterns. The risk of overfitting that is known in deep learning, because of the limited epidemiological data, was mitigated by introducing a drop ratio of 0.3 between the recurrent layers. The output is passed through a dense, fully connected layer with a linear activation function for the continuous prediction of weekly incidence of Lassa fever. The output is fed into a dense, fully connected layer with a linear activation function to predict the continuous weekly incidence of Lassa fever (Zhou *et al.* 2025).

LSTM networks are a special kind of recurrent neural network that learn long-term dependencies in sequence prediction tasks. It allows RNNs to retain input for extended durations (Sambasivarao *et al.* 2026). This is based on the ability of LSTM to retain memory, similar to computer memory. It also has the capability to read, write and delete data from its memory (Krichen & Mihoub, 2025).

4. | Implementations

The models were trained on the collected dataset of Lassa fever cases and the weather dataset. The spatial temporal graph network for the relationship between the four local government areas is shown in Figure 6.

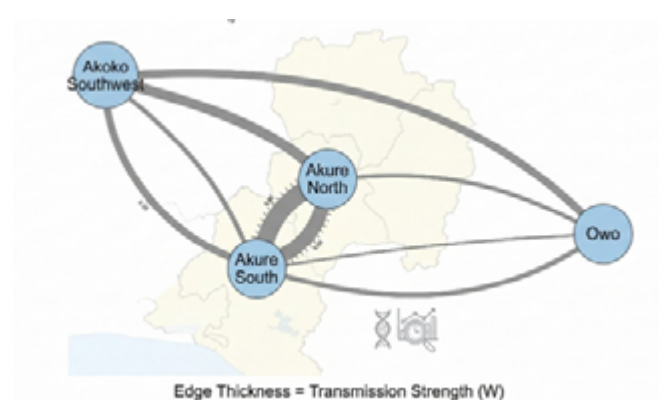


Figure 6 | Spatial Temporal Graph Network for Lassa Fever Hotspots

They were evaluated, and their performance on the unseen test data was compared. The Random

Forest model consistently demonstrated the best performance across the evaluation metrics. The results obtained from the models developed following the process are illustrated in the Table 3.

Table 3 | Comparison of Tree-based Models

	Model	RMSE	MAE	R ²
0	LightGBM	5.857	3.876	0.67
1	XGBoost	6.258	4.142	0.75
2	Random Forest	5.627	3.747	0.87

A comparison of the performance of the tree-based models and deep learning model is showing in Figure 7.

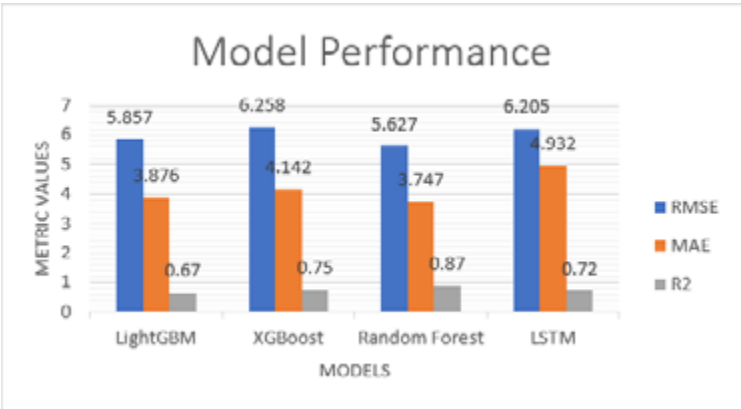


Figure 7 | Comparison of Model Performance Metrics

5. | Discussion

The findings from this research work clearly show that the use of advanced optimization and spatial modelling techniques along with powerful machine learning algorithms can improve outbreak prediction and understanding of Lassa fever outbreaks. The effectiveness of the Random Forest model indicates that with the depth and dimension of data, and complexity of features, this model fares better than complex deep learning models such as LSTM. The ability of Random Forest to effectively deal with the interactions between the climatic variables and the historical data, and to avoid overfitting was useful to the success of the model.

The model was able to identify the most important environmental factors from among the multitude of meteorological data through the use of a Genetic Algorithm for feature optimization. Using the

algorithm to identify the exact combination of temperature and rainfall data from the last 2-6 weeks before an outbreak gives public health officials actionable intelligence.

Based on the 2 to 6-week meteorological lag identified by the GA, we recommend that the NCDC deploy pre-emptive rodent control and community awareness campaigns in Akure and Owo exactly two weeks following heavy precipitation events.

Moreover, embedding a spatial-temporal graph theoretic approach makes up one of the shortcomings of the typical tabular machine learning model. The framework represents the identified hotspots as nodes that are linked together, thereby capturing the relational transmission dynamics across these hotspots. This enables the study of the velocity and direction of the outbreak and a more comprehensive

understanding of the disease spreading in neighbouring communities.

One of the key reasons for the significant performance gap between the Random Forest model ($R^2 = 0.87$) and LightGBM ($R^2 = 0.67$) is the fundamental difference in their ensemble architecture and how they manage the highly variable data. Random Forest employs random bagging (bootstrap aggregating) which trains many decision trees that are built independently, using random subsets of the features and training data. This structural independence allows the model to be highly robust and resistant to the noise, variance and sporadic outliers that are common in complex ecological and epidemiological data sets. On the other hand, gradient boosting systems such as LightGBM construct trees one by one, and keep trying to reduce the error of the previous trees. Despite being very efficient in terms of computation, the process of iteration and error correction can lead to the model overfitting the climatic transmission data, which may contain inherent noise and data anomalies, thereby decreasing its predictive accuracy for new data sets that were not used for training.

This model has the limitation, however, that is, it is dependent upon the cases that are reported, which may not be truly representative of the incidence because of underreporting. In addition, there are also socio-economic and ecological factors that could also affect transmission of Lassa fever that may not have been captured. This model could be extended in the following ways: adding these features as edge weights in the graph network, for further improvement of pattern discovery.

6. | Conclusion

This study conducts an exploration of different machine learning models for the forecast of Lassa fever. Lassa fever as a deadly disease has contributed to hundreds of deaths annually in some certain parts

of the world, especially Nigeria. Effective prevention of this spread demands more than just an organismal surveillance; it requires accurate predictions and determination of accurate epidemiological trends to help guide medical and government response. Recognizing this demand, this work goes beyond the existing predictive modelling research to design a comprehensive health information pattern discovery framework, which combines graph theoretic methods, GA optimization, and advanced machine learning ensembles. The spatial network was used to represent the hotspots as interconnected nodes and a GA was used to optimize the lagged meteorological features, thereby effectively modelling the complex dynamics of transmission. The results of the study indicate that Random Forest model was the best of the other models for which the RMSE, MAE and R-squared values were 5.627, 3.747 and 0.87 respectively.

This research outlines the importance and effectiveness of using the Random Forest model in forecasting the spread of deadly diseases like Lassa fever, as well as providing continuous efforts to tackle them. Future research should integrate some other factors like the fluctuation of rodent population, their breeding patterns, seasonal behaviours, as well as detailed human mobility data from sources like transportation patterns and migration trends. It is also important to test the model in a prospective real-world context, by conducting longitudinal research on its effectiveness and adaptability in real world situations.

References

- Akhmetzhanov, A. R., Asai, Y., & Nishiura, H. (2019). Quantifying the seasonal drivers of transmission for Lassa fever in Nigeria. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1775), Article 20180268. <https://doi.org/10.1098/rstb.2018.0268>
- Alile, S. O. (2020). A supervised machine learning approach for diagnosing Lassa fever and viral Haemorrhagic fever types reliant on observed signs. *Asia-Pacific Journal of Science and Technology*, 27(4), 1–16. <https://doi.org/10.14456/apst.2022.65>
- Bakari, S., Malgwi, Y. M., Abu, K. P., & Mohammed, U. (2024). Lassa fever predictive model using machine learning techniques. *International Journal of Computer Science and Mathematical Theory*, 10(4), 145–173. <https://doi.org/10.56201/ijcsmt.v10.no4.2024.pg145.173>
- Çakır, Mustafa & Güvenç, Mehmet ali & Mistikoglu, Selcuk. (2020). The Experimental Application of Popular Machine Learning Algorithms on Predictive Maintenance and the Design of IIoT based Condition Monitoring System. *Computers & Industrial Engineering*. 10.1016/j.cie.2020.106948.
- Centers for Disease Control and Prevention. (2025). *About Lassa fever*. <https://www.cdc.gov/path/to/page>. Accessed 22nd October, 2025
- Chen, C., He, Z., Zhao, J., Zhu, X., Li, J., Wu, X., Chen, Z., Chen, H., & Jia, G. (2024). Zoonotic outbreak risk prediction with long short-term memory models: a case study with schistosomiasis, echinococcosis, and leptospirosis. *BMC infectious diseases*, 24(1), 1062. <https://doi.org/10.1186/s12879-024-09892-y>
- Haripriya, B., R. Madhavi, S. N. Sai Priya, and R. Venugopal. 2025. "Predictive Analysis in Healthcare: Using Data to Improve Patient Outcomes." *International Journal for Research in Applied Science & Engineering Technology* 13 (7): 1548–1556. <https://doi.org/10.22214/ijra-set.2025.73231>.
- Jiao, S., Wang, Y., Ye, X., Nagahara, L., & Sakurai, T. (2025). Spatio-temporal epidemic forecasting using mobility data with LSTM networks and attention mechanism. *Scientific Reports*, 15(1), Article 9603. <https://doi.org/10.1038/s41598-025-94089-9>
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017, December). *LightGBM: A highly efficient gradient boosting decision tree* [Paper presentation]. 31st International Conference on Neural Information Processing Systems, Long Beach, CA, United States.
- Krichen, M., & Mihoub, A. (2025). Long Short-Term Memory Networks: A Comprehensive Survey. *AI*, 6(9), 215. <https://doi.org/10.3390/ai6090215>
- Liang, F., Guan, P., Wu, W., & Huang, D. (2018). Forecasting influenza epidemics by integrating internet search queries and traditional surveillance data with the support vector machine regression model in Liaoning, from 2011 to 2015. *PeerJ*, 6, e5134. <https://doi.org/10.7717/peerj.5134>
- Mehta AM & Patel KS. (2024). LSTM-based Forecasting of Dengue Cases in Gujarat: A Machine Learning Approach. *Indian Journal of Science and Technology*. 17(7):635-642. <https://doi.org/10.17485/IJST/v17i7.2748>
- Mohan, R.N.V.J., Raju, B.H.V.S.R.K., Sekhar, V.C., & Prasad, T.V.K.P. (Eds.). (2025). *Algorithms in Advanced Artificial Intelligence* (1st ed.). CRC Press. <https://doi.org/10.1201/9781003641537>

- Nigeria Centre for Disease Control (NCDC) (2024). *Lassa fever situation report*. <https://www.ncdc.gov.ng/themes/common/files/sitreps/>
- Oluwole, A., & Nkonyana, T. (2022). Forecasting Lassa fever outbreak progression with machine learning [Paper presentation]. International Conference on Electrical, Computer and Energy Technologies (ICECCME), 1–5. <https://doi.org/10.1109/ICECCME55909.2022.9987787>
- Okwuraiwe, A. P., Onyeaghala, C. A., Ozoude, O. T., Suleiman, M. O., Abdu-Aguye, S. N., Ezekwelu, N. J., Oyeniyi, T. A., Jegede, A. O., Egwudo, A. E., Okeke, O. P., Abodunrin, O. R., Akinsolu, F. T., & Sobande, O. O. (2025). Clinical outcomes of severe Lassa fever in West Africa: A systematic review and meta-analysis. *International Journal of Environmental Research and Public Health*, 22, Article 1504. <https://doi.org/10.3390/ijerph22101504>
- Sambasivarao, K.V., & Roopa Devi Bhima, A.S. (Eds.). (2026). *Artificial Intelligence, Computational Intelligence and Inclusive Technologies* (1st ed.). CRC Press. <https://doi.org/10.1201/9781003740100>
- Tewogbola P & Aung N (2020) Lassa fever: History, causes, effects, and reduction strategies, *Int. J. One Health*, 6(2): 95-98
- Ukarande, S., Nikam, S.R., Kulkarni, M.G., Prabhu, N.A., Narwane, V.S., Mehendale, N.D., & More, N.P. (Eds.). (2026). *Technologies for Energy, Agriculture, and Healthcare* (1st ed.). CRC Press. <https://doi.org/10.1201/9781003773771>
- World Health Organization. (2023). *Lassa fever: Nigeria*. Disease Outbreak News. <https://www.who.int/emergencies/disease-outbreak-news/item/2023-DON492>. Accessed 3rd October, 2025
- Zhang, Y. (2024). The application of statistical methods in the analysis of factors affecting economic growth. *Journal of Statistics and Economics*, 1(6), 49–53.
- Zhang Li, Sufang Xiong, Shu Zhu, Jigui Tian, Qingyuan Chen, Xiaobo Luo, Hua Guo (2024). Construction of Prediction Model of Foodborne Disease Outbreaks and Its Trend Prediction — Guizhou Province, China, 2023–2025[J]. *China CDC Weekly*, 2024, 6(18): 408-412. doi: [10.46234/ccdcw2024.079](https://doi.org/10.46234/ccdcw2024.079)
- Zhou, N., Zeng, H., Zheng, Z., Wang, K., & Zhou, J. (2025). PEMFC RUL Prediction for Non-Stationary Time Series Based on Crossformer Model. *Applied Sciences*, 15(5), 2515. <https://doi.org/10.3390/app15052515>