Harnessing Machine Learning for Predictive Analytics: A Case Study of Lassa Fever Outbreaks in Nigeria

Daniel A. Quezada
Department of Computer Science
California State University Fullerton
Fullerton, USA
0009-0001-4005-8608

Sampson Akwafuo
Department of Computer Science
California State University Fullerton
Fullerton, USA
0000-0001-8255-4127

Samarth Halyal
Department of Computer Science
California State University Fullerton
Fullerton, USA
0000-0001-8923-9031

***Abstract*—In the ongoing battle against global pandemics, understanding the key determinants that fuel outbreaks are of paramount importance. With this focus, our study aims to assess and rank the predictive capabilities of a wide range of socio-economic, eco-climatic, and spatiotemporal variables in predicting Lassa Fever (LF) outbreaks, using data from previous Nigerian outbreaks (2012-2019). Employing machine learning methods, particularly XGBoost and Random Forest, our study aims to offer accurate and robust predictions concerning LF incidence rates. As a crucial add-on, we leverage the innovative SHAP (SHapley Additive exPlanations) technique as a post-processing tool to dissect and better understand the contributions of individual features towards the predictions generated by our machine learning models. This multi-layered approach allowed us to place a pronounced focus on healthcare infrastructure, population demographics, land cover, and other climatic co-variates. Among the models evaluated, XGBoost performed the best; delivering an accuracy of 0.93, and AUC of 0.90, and an F1 score of 0.86 on 2018 data. For 2019 data, it maintained a strong accuracy of 0.90, an AUC of 0.89, and an F1 score of 0.75. Our SHAP analysis further emphasized precipitation seasonality, diagnostic center density, and land cover characteristics as pivotal influencers in predicting LF outbreaks. These findings shed light on the complex interplay between environmental conditions, urbanization, and healthcare infrastructure. Given these promising results, our work sets the stage for the development of an advanced early warning system for Lassa Fever in Nigeria: a system that could efficiently intertwine computational insights with on-ground interventions, ensuring timely and targeted responses to potential outbreaks.**

***Index Terms*—Lassa Fever, Computational Epidemiology, Machine Learning, SHAP, Africa, Public Health, Nigeria**

I. Introduction

Lassa Fever (LF) is a viral hemorrhagic illness, similar to Ebola viral fever [1]. It is prevalent in many regions of West Africa, with Nigeria having the highest number of recurrent outbreaks. It was named after a town in northern Nigeria, where it was first discovered in 1969. Lassa mammarenavirus (LASV) is the etiological agent responsible for LF and poses a significant threat to both human and animal populations. Classified as an old-world arenavirus, LASV is a single-stranded, bipartite RNA virus [2], [3]. Over the years, seven distinct lineages of LASV have been identified across West Africa, three of which have been observed in Nigeria since its initial discovery in the 1960s [4]. The primary reservoir of LASV is Mastomys natalensis, a small rodent species commonly found in Africa [2], [3], [4], [5], [6]. Notably, in a study conducted in Nigeria, found that more than half (53.6%) of captured rodents tested positive for LASV [8]. Interestingly, Mastomys natalensis does not exhibit the typical symptoms of LF experienced by humans but serves as a lifelong carrier, secreting the virus through urine or droppings. Transmission of LASV to humans predominantly occurs through direct or indirect contact with infected rodents.

Once LASV enters the human body, an incubation period of 6 to 21 days elapses before the initial signs of symptoms emerge [4], [7]. The majority of LF infections (80%) present with either mild symptoms or are entirely asymptomatic, manifesting as fever, headaches, and malaise. However, the hospital fatality rate is about 26.5%, and the general rate is believed to be considerably higher, as many cases are usually unreported due to inadequate monitoring and evaluation infrastructure [8], [9]. In approximately 20% of infected individuals, severe symptoms associated with hemorrhagic fevers ensue [7]. These severe manifestations involve internal bleeding in the stomach, small intestines, and brain, as well as inflammation of the liver and kidneys. Additionally, around 29% of patients report temporary or permanent deafness [10]. Recent data indicate an alarming increase in the incidence rate of LF in Nigeria. During 2018, the Nigerian CDC reported a total of 1893 suspected LF cases, of which 423 were laboratory confirmed [11]. The reported case fatality rate stood at an astonishing 25.1%. Previous studies have aimed to quantify and establish relationships between specific ecological and climatic drivers and the rise of LF cases in West Africa [12], [13]. In them, they found that LF transmission is known to be influenced by a multitude of biotic and abiotic factors.

One study revealed that the peak number of LF infections in Nigeria occurred between December and March from 2016 to 2018, possibly due to rodents' close proximity to humans during the dry season when food is scarce [12]. Another investigation demonstrated a correlation between certain abiotic factors, such as rainfall, temperature, geographic features, and the increased incidence of LF cases in Nigeria [13]. Although studies such as [14], [15], [16] have employed statistical methods to examine the correlation between biological and environmental processes and the rise of LF cases, few, if any, have utilized machine learning models to rank the various abiotic drivers of LF cases specifically in Nigeria.

In this study, we aim to assess the predictive capabilities of socio-economic, eco-climatic, and spatiotemporal variables in predicting LF outbreaks across Nigeria during a two-year period from 2018 to 2019. To achieve this, we employ ensemble machine learning methods, namely XGBoost and Random Forest, to make accurate predictions regarding LF incidence. Subsequently, we employ SHAP as a post-processing technique to determine the most influential drivers for predicting an outbreak. Through our investigation, we endeavor to contribute to the understanding of LF dynamics and prioritize the features that significantly contribute to a positive incidence rate.

II. Methodology

A. Data Collection and Feature Characterization

Our investigation initiated with an in-depth exploration of datasets related to LF in Nigeria. We utilized Google Datasets and employed key search terms such as “Lassa Fever”, “Nigeria”, and “Lassa Infections”. This approach led us to a comprehensive dataset [17], previously curated by another team of researchers, providing epidemiological insights into LF infections spanning from 2012 to 2019. This dataset aggregates diverse data sources, including the Nigerian CDC, the CHELSA climate repository, and various governmental records. The dataset offers a detailed pictures of LF spread. Specifically, the data encompasses weekly epidemiological reports from all 36 states of Nigeria, including the Federal Capital, all recorded at the Local Government Area (LGA) level.

To better understand and address LF outbreaks, the dataset integrates multiple features classes, including geographical, spatiotemporal, health, socioeconomic, and ecoclimatic variables. Each feature class is characterized by distinct data levels, ranging from categorical to numerical measurements, ensuring a multifaceted analysis of LF outbreak dynamics. Table 1 provides a detailed overview of the features, their categories/data levels, feature names, and descriptions.

B. Data Structure and Preprocessing

Once the dataset was chosen, the essential task of preprocessing began. The dataset was initially loaded into a Python DataFrame using Pandas, a data manipulation library. A primary step in our preprocessing involved the conversion of the Date column to a datetime object, enabling us to extract the Year as a separate feature. To facilitate machine learning operations, categorical attributes such as State and LGA names were encoded into numerical formats using a Label Encoder. The Cases feature was transformed into a binary classification target, where any non-zero case count was encoded as 1, representing an outbreak occurrence, and zero cases as 0.

This binary transformation aimed to streamline the model’s focus on predicting the presence of LF outbreaks. Lastly, we addressed potential issues of missing data in our training set by employing a Simple Imputer with a median strategy. This approach involves replacing missing values in each column with the median value of that column, thus preserving the overall distribution of the data.

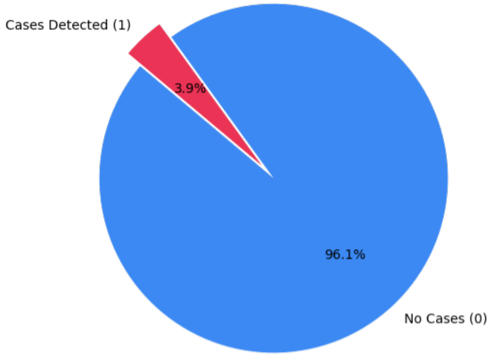

Fig. 1. Distribution of Lassa Fever Case Counts in the Dataset

a) Addressing Class Imbalance in Dataset: One inherent challenge in our dataset is the significant class imbalance, a common issue in epidemiological datasets, particularly those dealing with disease outbreaks. The majority of the records ($\sim$96%) in our dataset reported zero cases of LF, which could lead to a model bias towards predicting the absence of an outbreak. This imbalance needed to be addressed to ensure that our model could effectively identify the less frequent, yet crucial instances of LF outbreaks.

To counteract this imbalance, we employed the Synthetic Minority Oversampling Technique (SMOTE) [18]. SMOTE is an over-sampling method that generates synthetic samples for the minority class, which in our case was a positive LF incidence rate reported by an LGA. By creating synthetic, yet plausible, instances of outbreak occurrences, SMOTE helps in balancing the dataset. This provides a way for the model to learn from both classes – outbreaks and non-outbreaks.

By augmenting our training data with the SMOTE strategy, we ensured that the model was exposed to a sufficient number of outbreak cases, aiding it its ability to generalize and predict future outbreaks more accurately. This step was particularly crucial given the rarity of LF outbreaks, as it prevented the model from being overwhelmingly influenced by the more common non-outbreak instances.

b) Choice of XGBoost and Random Forest: We anchored our decision on XGBoost [19] and Random Forest [20] due to their ensemble nature. These techniques are renowned for harnessing the strengths of multiple individual models, combining their capabilities to deliver strong overall performance. Specifically, both XGBoost and Random Forest excel in uncovering non-linear relationships between input features and the target variables. Their versatility in their ability to process categorical and numerical data makes them well suited for multifaceted datasets like the one we’re employing.

TABLE I
SUMMARY OF FEATURES USED IN ANALYSIS

Feature Category	Name	Description
		Geographical
Categorical	State	Name of State
Categorical	LGA	Name of Local Government Area
Numerical	AreaKM2	Land Area of LGA in square kilometers
		Health & Medical
Numerical	Cases	Weekly Reported LF cases per LGA
Numerical	NumDiagCentres	Total LF diagnostic centers in LGA
Numerical (km)	LabDist	Average Distance to LF diagnostics in LGA
Numerical	LabTravelTime	Average travel time to nearest LF diagnostic lab
Numerical (km)	HospitalDist	Average Distance to Hospitals in LGA
Numerical (km)	HealthFacilityDist	Average Distance to Health Facilities in LGA
Numerical	HospitalTravTime	Average Travel Time to Nearest Hospital in log minutes
		Demographic & Socioeconomic
Numerical	TotalPop	Total Population of LGA
Numerical	AgriProp	Percentage of LGA land classified as agricultural
Numerical	UrbanProp	Percentage of LGA land classified as urban
Numerical	ForestProp	Percentage of LGA land classified as forest
Numerical	ImprovHousing	Prevalence of Improved Housing in LGA
Numerical	PovertyPropMean	Average Poverty Rate in LGA
Numerical	PovertyPropWeighted	Population-weighted Poverty Rate in LGA
		Ecoclimatic
Numerical (°C)	TempMean	Average Monthly Temperature in LGA
Numerical	TempSeasonality	Standard Deviation of Monthly Mean Temperature in LGA
Numerical (mm)	PrecipTotal	Average Monthly Precipitation in LGA
Numerical	PrecipSeasonality	Standard Deviation of Monthly Mean Precipitation in LGA

C. SHAP: Our Interpretative Tool

In our study, we employ SHAP to interpret the contributions of individual features to the model's predictions [21]. Rooted in cooperative game theory, SHAP is able to quantify the impact of each feature on a prediction.

A key strength of SHAP is its ability to offer local and global explanations. On a local level, it explains the reasoning behind a model's specific prediction for a single instance. Globally, it aggregates all the instances, revealing the overall significance of each feature across the dataset.

In our context, SHAP enables us to identifying the factors that influence the likelihood of LF outbreaks. By understanding which features - such as climate, population density, or availability of health facilities - play a crucial role in the model's predictions, we can provide stakeholders with valuable insights for targeted interventions.

D. Training the Model

The training of the models was a multi-faceted process, incorporating various steps to ensure the effectiveness and accuracy of the predictions. After preprocessing the dataset, we proceeded with the training phase, utilizing data from 2012 up to 2017 to train our models, while the data for years 2018 and 2019 were reserved as separate test sets. This division allowed us to evaluate the performance of our models on unseen data, reflective of more recent conditions.

a) Hyperparameter Tuning for XGBoost: In tuning the XGBoost model, we employed RandomizedSearchCV to explore a diverse set of hyperparameters efficiently. This method was chosen due to its ability to cover a broad parameter space with fewer iterations compared to GridSearchCV, significantly reducing computation. The hyperparameters tuned and rationale for their ranges are as follows:

- **n_estimators**: We varied the number of trees in the ensemble from 50 to 500. This range allows us to explore the trade-off between underfitting and overfitting. Fewer trees can lead to underfitting, whereas more trees increase the model's complexity and potential for overfitting but can capture more detailed patterns in the data.
- **max_depth** (3 to 10): Controlling the depth of each tree is crucial for balancing the model's ability to model complex relationships without overfitting. A maximum depth of 10 ensures the model is deep enough to learn significant interactions but not so deep that it fits overly specific patterns.
- **learning_rate** (0.01 to 0.3): This parameter moderates the impact of each individual tree on the final outcome, helping to prevent overfitting by making the model more conservative. A lower learning rate requires more trees to achieve model convergence, promoting a more robust ensemble by integrating more nuanced patterns.
- **subsample** and **colsample_bytree** (0.6 to 1.0): These parameters determine the fraction of samples and features used for building each tree. By using a subset of the data, the model is less likely to learn noise and more likely to generalize well to new data.
- **min_child_weight**, **gamma**, and regularization terms **reg_alpha** and **reg_lambda** (ranging from 0 to 1): Fine-tuning these parameters adds layers of complexity control. Higher values help in regularizing the model further, preventing overfitting by smoothing the learned patterns.

A total of 25 different parameter combinations were tested, with the primary objective of maxmizing the ROC AUC score, ensuring that the model not only fits well but generalizes well on unseen data.

b) Hyperparameter Tuning for Random Forest: For the Random Forest model, we utilized GridSearchCV, which evaluates all possible combinations of provided hyperparameters. This exhaustive search method was selected because it ensures that the best possible combination is identified, crucial for achieving optimal performance in our predictive modeling. The hyperparameters tuned, along with the reasons for their ranges, include:

- **n_estimators**: The number of trees, tested at 100, 200, and 300, was chosen to determine the optimal count that balances computational efficiency with predictive accuracy. A higher number of trees generally provides better performance but at the cost of increased computational load and potential diminishing returns.
- **max_depth** (10, 20, 50): These values were selected to test various levels of complexity. A depth of 10 may prevent overfitting in scenarios with less complex data, whereas 50 allows for a deeper tree that can capture more complex patterns at the risk of overfitting.
- **min_samples_split** and **min_samples_leaf** (values 2, 5, and 10 for split; 1, 2, and 4 for leaf): These parameters control the minimum number of samples required at a leaf node and a split point. Setting these values ensures that the trees do not grow too deep or too specific, which helps in preventing overfitting and maintaining the generalizability of the model.
- **bootstrap**: The parameter was evaluated both as True and False to assess whether bootstrap sampling (sampling with replacement) enhances the model's accuracy and stability. Bootstrap sampling typically helps in building more diverse trees, reducing the variance component of the model error.

The performance of each configuration was assessed using a 3-fold cross-validation focusing on the ROC AUC score. This validation method was particularly chosen to ensure that our model evaluations are robust and reliable.

III. Results

A. Optimal Hyperparameter Configurations

The optimal values obtained from hyperparameter tuning are presented in Table 2. Note: "-" indicates that the parameter is not applicable for the respective model.

TABLE II
OPTIMAL HYPERPARAMETERS FOR XGBOOST AND RANDOM FOREST

Parameter	**XGBoost**	**Random Forest**
Colsample_bytree	0.6734	-
Gamma	0.1521	-
Learning_rate	0.1674	-
Max_depth	6	10
Min_child_weight	1	-
N_estimators	98	300
Reg_alpha	0.5248	-
Subsample	0.6187	-
Bootstrap	-	True
Class_weight	-	Balanced
Min_samples_leaf	-	4
Min_samples_split	-	2

B. Evaluation of Model Performance

To decide the more suitable model between XGBoost and Random Forest algorithms, we relied upon two primary metrics: F1 score and Area Under the Curve (AUC) score. The AUC score measures the ability of a model to distinguish between classes, specifically in terms of the area under the Receiver Operating Characteristic (ROC) curve. In addition, the F1 score provides a comprehensive measure of a model's performance by factoring in both precision and recall. A high F1 score indicates proficient identification of positive examples and minimizing the false labeling of negatives and positives. The XGBoost model showed impressive results in the year 2018, achieving an accuracy of 93.10%, an AUC score of 0.90 and an F1 score of 0.8611. In 2019, the model maintained a robust performance, though with a decrease in all metrics, recording an accuracy of 90.14%, an AUC of 0.89, and an F1 score of 0.7583. Turning to the Random Forest algorithm, the model achieved an accuracy of 90.86%, an AUC of 0.88, and an F1 score of 0.7139 on 2018 data. For the following year, the model recorded an accuracy of 88.44%, an AUC of 0.89 and an F1 score of 0.7448.

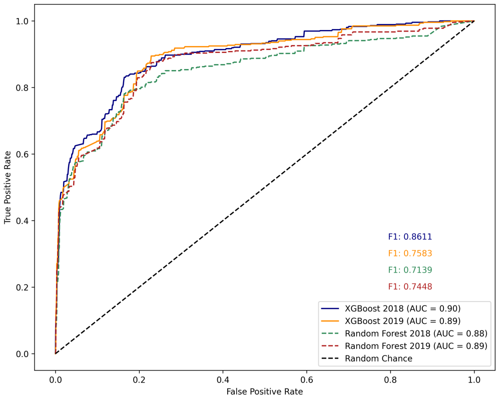

Fig. 2. ROCAUC Plot for XGBoost and Random Forest Models (2018-2019)

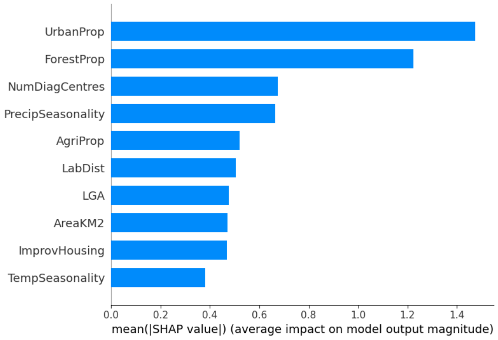

Fig. 3. SHAP Global Bar Plot for XGBoost Model on 2018-2019 Data

C. *Feature Importance Analysis*

Now that we've established XGBoost as the superior model, our next step was to identify the significant factors influencing Lassa Fever infections in Nigeria. For this step, we utilized the SHAP technique for post-processing the XGBoost model to gain a deeper insight into influential features. As depicted in Fig. 1, we can see that SHAP Global Plot for 2018-2019 dataset pinpointed the proportion of land designated as urban area as the most influential feature. This was followed by proportion of land designated as forest area, the number of diagnosis centers, precipitation seasonality, and the proportion of land designated as agriculture.

IV. Discussion

A. *Relevance of SHAP in Understanding Influential Features*

Understanding what drives a machine learning model's prediction is essential, especially when decisions derived from these predictions have substantial real-world implications. SHAP serves as a clarity and helps quantify each feature's contribution to the prediction, enabling us to dissect which factors (like environmental conditions or healthcare accessibility) significantly sway the likelihood of an outbreak. By employing SHAP analysis, policymakers and public health officials can devise targeted interventions, ensuring resource allocation is both efficient and impactful.

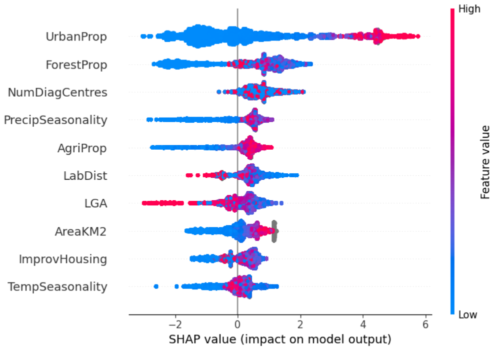

Fig. 4. SHAP Beeswarm Plot for XGBoost Model on 2018-2019 Data

B. *Interpreting SHAP Plots within Lassa Fever's Seasonal Epidemiology*

a) Understanding the SHAP Beeswarm Plot: Just like the global bar plot, the order of features, from top to bottom, indicate the importance in a beeswarm plot while unraveling the direction of each feature's influence. Each dot represents an individual prediction, or sample. It's position along the x-axis represents the SHAP value for that specific sample and the spread of these dots captures the distribution spectrum of SHAP values per feature. Dots leaning towards the right (positive SHAP values) denote a feature amplifying the model's prediction for that sample. In contrast, those on the left suggest a suppressive effect. This positional data, combined with the color coding (blue indicating lower values and red signifying higher values), enables a greater understanding of feature influence.

b) Ecological Influence on Outbreak Predictions: Urban areas, identified as the most significant feature, are characterized by dense populations and increased human-to-human contact, therefore increasing the chance for LF transmission. The model's sensitivity to urbanization, as shown by the far right concentration of red dots, suggests that outbreaks may be more likely or easily detected in these areas.

The 'ForestProp' feature, representing the proportion of forested land within an LGA, displayed a complex pattern on the beeswarm plot. A mixture of red and blue dots near center right and a string of blue dots extending to the left suggests that impact of forestation on LF predictions may not be unidirectional. High forest cover (red dots) sometimes correlates to increased predictions of outbreaks as dense vegetation may provide a habitat for the rodent hosts. On the other hand, blue dots on the left signifies instances where greater forestation may be associated with reduced prediction of outbreaks, perhaps because of less human interactions in these regions.

Similarly, 'PrecipSeasonlity' and 'TempSeasonlity' offer insights into the environmental context of LF outbreaks. The variability in precipitation and temperature could affect the survival and breeding patterns of rodent populations, potentially influencing incidence of LF. A wide dispersion of SHAP values for these features indicates that both higher and lower seasonality in precipitation and temperature play a role and impact on the disease dynamic.

c) Healthcare Accessibility and Living Standards as Predictors: The SHAP analysis reveals surprising correlations between healthcare infrastructure and predicted outbreak frequencies. The NumDiagCentres variable, representing the number of diagnostic centers within an LGA, emerged as a significant predictor. As seen in the SHAP beeswarm plot (Fig. 3), it shows a predominantly positive SHAP value, indicating that a higher number of centers tends to correlate with an increase prediction of outbreaks. At a glance, this might appear counterintuitive, as one could assume that regions with more centers would be better prepared to diagnose and manage infections. However, upon considering the potential for reporting bias, it becomes clear that areas with fewer diagnostic centers may be underreporting due to their restricted capabilities. Consequently, our model is inclined to highlight regions reporting a surge in cases, not necessarily because they inherently experience more cases, but rather due to their enhanced detection capabilities.

For LabDist, the average distance to the nearest diagnostic lab, we observe a concentration of higher SHAP values for smaller distances. This pattern aligns with expectations that increased distances to healthcare facilities could be a barrier to timely diagnosis and treatment, potentially allowing for greater spread of the disease before containment measures can be implemented.

This nuanced understanding of the relationship between Lassa Fever incidence and variables related to healthcare access and housing standards is crucial. It suggests that enhancing healthcare infrastructure and improving housing conditions, while inherently beneficial, must be part of a broader, more coordinated public health strategy that considers the local context and an array of intersecting factors.

C. Conclusion

This study has demonstrated the power of machine learning in enhancing our understanding and prediction of Lassa Fever outbreaks in Nigeria. By employing XGBoost and Random Forest models, enhanced with SHAP for interpretability, our research offers a novel, robust approach to predicting LF incidence rates based on a wide range of socio-economic, eco-climatic, and spatiotemporal variables.

Our findings underscore the significance of certain predictors, such as urban land cover, diagnostic center density, and precipitation seasonality, in influencing LF outbreaks. The high performance of the XGBoost model, achieving an accuracy of up to 93.10%, an AUC of 0.90, and an F1 score of 0.8611, confirms the efficacy of our modeling approach. These results not only enhance our understanding of the disease's dynamics but also pave the way for the development of an advanced early warning system for LF in Nigeria.

Looking forward, we recommend the integration of real-time data and the consideration of changing climate patterns in our models to improve responsiveness and adaptability to outbreaks. Furthermore, the development of localized models that account for regional characteristics could provide more precise predictions and support targeted public health interventions.

The implications of our study extend beyond the academic realm into practical applications. By incorporating our findings into public health strategies, policymakers and healthcare providers can optimize resource allocation and intervention strategies, ensuring timely and targeted responses to outbreaks. Moreover, our approach can serve as a blueprint for other regions grappling with similar infectious diseases.

In conclusion, while our study has made significant strides in using machine learning for disease prediction, ongoing research is essential to refine these models and adapt them to new challenges posed by global health dynamics. Enhancing data collection practices and embracing interdisciplinary collaboration will be critical in advancing our capability to predict and manage infectious disease outbreaks effectively.

V. Acknowledgment

The authors would like to express their gratitude for the support received from the National Science Foundation (NSF) under Grant No. HRD-182640, which contributed to the research in this paper.

References

[1] S. Akwafuo, T. Abah, and J. Oppong, "Evaluation of the Burden and Intervention Strategies of TB-HIV Co-Infection in West Africa," Journal of Infectious Diseases and Epidemiology, vol. 6, Oct. 2020, doi: 10.23937/2474-3658/1510143.

[2] D. A. Asogun, S. Günther, G. O. Akpede, C. Ihekweazu, and A. Zumla, "Lassa Fever: Epidemiology, Clinical Features, Diagnosis, Management and Prevention," Infect Dis Clin North Am, vol. 33, no. 4, pp. 933–951, 2019, doi: https://doi.org/10.1016/j.idc.2019.08.002.

[3] O. Ogbu, E. Ajuluchukwu, and C. J. Uneke, "Lassa fever in West African sub-region: an overview," J Vector Borne Dis, vol. 44, no. 1, pp. 1–11, Mar. 2007.

[4] P. Tewogbola and N. Aung, "Lassa fever: history, causes, effects, and reduction strategies," Virus, vol. 2, p. 16, 2020.

[5] A. N. Happi, C. T. Happi, and R. J. Schoepp, "Lassa fever diagnostics: past, present, and future," Curr Opin Virol, vol. 37, pp. 132–138, 2019, doi: https://doi.org/10.1016/j.coviro.2019.08.002.

[6] S. E. Akwafuo, A. Hussain, and C. Ihinegbu, "Recurrent Lassa Fever Outbreaks: Spatiotemporal Analysis and Modelling of Environmental Intervention Strategies," in Proceedings of the 2023 9th International Conference on Control, Decision and Information Technologies (CoDIT), Rome: IEEE Xplore, Jul. 2023.

[7] C. Houlihan and R. Behrens, "Lassa fever," BMJ, vol. 358, 2017, doi: 10.1136/bmj.j2986.

[8] C. A. Yaro, E. Kogi, K. N. Opara, and G. E. S. Batiha, "Infection pattern, case fatality rate and spread of Lassa virus in Nigeria," BMC Infect Dis, vol. 21, no. 1, pp. 1–9, 2021, doi: 10.1186/s12879-021-05837-x.

[9] S. Akwafuo, X. Guo, and A. Mikler, "Epidemiological modelling of vaccination and reduced funeral rites interventions on the reproduction number , R 0 of Ebola virus disease in West Africa .," International Journal of Infectious and Tropical Diseases, vol. 2, pp. 7–11, Oct. 2018.

[10] "Lassa Fever: Signs and Symptoms," Center for Disease Control and Prevention.

[11] E. A. Ilori et al., "Epidemiologic and Clinical Features of Lassa Fever Outbreak in Nigeria, January 1-May 6, 2018," Emerg Infect Dis, vol. 25, no. 6, pp. 1066–1074, Jun. 2019.

[12] A. R. Akhmetzhanov, Y. Asai, and H. Nishiura, "Quantifying the seasonal drivers of transmission for Lassa fever in Nigeria," Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 374, no. 1775, p. 20180268, 2019, doi: 10.1098/rstb.2018.0268.

[13] D. W. Redding et al., "Geographical drivers and climate-linked dynamics of Lassa fever in Nigeria," Nat Commun, vol. 12, no. 1, p. 5759, 2021, doi: 10.1038/s41467-021-25910-y.

[14] A. S. Oluwole and T. Nkonyana, "Forecasting Lassa Fever Outbreak Progression with Machine Learning," in 2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Nov. 2022, pp. 1–5. doi: 10.1109/ICECCME55909.2022.9987787.

[15] S. O. Alile, "A supervised machine learning approach for diagnosing Lassa fever and viral Hemorrhagic fever types reliant on observed signs," Life, vol. 3, p. 4.

[16] M. M. Ojo, B. Gbadamosi, T. O. Benson, O. Adebimpe, and A. L. Georgina, "Modeling the dynamics of Lassa fever in Nigeria," Journal of the Egyptian Mathematical Society, vol. 29, no. 1, p. 16, 2021, doi: 10.1186/s42787-021-00124-9.

[17] D. Redding, I. Abubakar, K. Jones, R. Gibb, C. Ihekweazu, and C. Dan-Nwafor, "Spatiotemporal analysis of systematic surveillance data enables climate-based forecasting of Lassa fever." figshare, 2021. doi: 10.6084/M9.FIGSHARE.9777656.V1.

[18] Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. Journal of Machine Learning Research 2017;18(17):1–5.

[19] Chen T, Guestrin C. Xgboost: A Scalable Tree Boosting System. CoRR 2016;abs/1603.02754.

[20] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 2011;12:2825–2830.

[21] Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems 30. Curran Associates, Inc., 2017; 4765–4774.