Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 7 (2025) 2586

Journal of the Nigerian Society of Physical Sciences

# Addressing class imbalance in lassa fever epidemic data, using machine learning: a case study with SMOTE and random forest

Osowomuabe Njama-Abang*, Denis U. Ashishie, Paul T. Bukie

*Department of Computer Science, University of Calabar PMB 1115, Etta Agbo Rd, Calabar, Nigeria*

## Abstract

Class imbalance in epidemiological datasets, particularly for rare outcomes like Lassa Fever fatalities, complicates predictive modeling. This study addresses the issue by employing SMOTE to rebalance the dataset and Random Forest for classification while identifying significant predictors such as age, symptom severity, and residence. SMOTE successfully balanced the dataset (minority class recall improved from 0.60 to 1.00 in Random Forest), mitigating the bias toward majority classes. Without SMOTE, models including Random Forest, XGBoost, and LightGBM achieved high accuracy (> 99%) but demonstrated poor minority recall (≤0.75), confirming the challenge of imbalanced data. Post-SMOTE balancing, these models achieved 100% accuracy, precision, recall, and F1-scores across major classes. Notably, the hybrid ensemble model further enhanced outcomes, achieving an F1-score of 0.80 for the rarest class. These results underscore the superiority of SMOTE in improving classification for underrepresented outcomes compared to reliance on Random Forest alone, demonstrating its value in developing equitable predictive tools for outbreak management.

DOI:10.46481/jnsps.2025.2586

*Keywords:* Lassa fever, Machine learning, SMOTE, Random forest, Class imbalance

*Article History :*
Received: 23 December 2024
Received in revised form: 28 April 2025
Accepted for publication: 30 April 2025
Available online: 08 June 2025



Communicated by: O. Akande

## 1. Introduction

Lassa Fever remains a critical public health concern in West Africa, with an estimated 100,000 to 300,000 infections and 5,000 deaths annually [1]. The zoonotic disease—caused by the Lassa virus and primarily transmitted through contact with infected rodents or their excreta—is endemic in several countries, with Nigeria, Sierra Leone, Liberia, and Guinea being the most affected [2]. Lassa Fever outbreaks pose a unique challenge due to wide-ranging clinical presentations, from asymptomatic cases to severe manifestations such as hemorrhagic fever and multi-organ failure. This variability in disease severity complicates diagnosis, monitoring, and resource allocation in healthcare systems [3].

A critical hurdle in analyzing Lassa Fever epidemiological data is class imbalance, where rare or severe disease outcomes, such as fatalities, are significantly underrepresented in the dataset. This imbalance can lead to biased models that favor majority classes, ultimately reducing the effectiveness of predictive tools critical for outbreak response. Traditional statistical approaches often struggle in such scenarios, necessitating the adoption of advanced machine learning (ML) techniques tailored to address these imbalances. Recent advances in ML, particularly oversampling techniques like the Synthetic Minority Oversampling Technique (SMOTE), offer a promising solution to enhance the representation of rare events in datasets and

*Corresponding author Tel. No.: +234-813-049-2860.
*Email address:* njama_abang@unical.edu.ng (Osowomuabe Njama-Abang )

improve predictive accuracy [4].

This study employs SMOTE in conjunction with a Random Forest classifier to tackle the issue of class imbalance in a Lassa Fever dataset. SMOTE is widely recognized for its ability to generate synthetic samples for underrepresented classes, thereby alleviating the bias inherent in imbalanced data [5]. Random Forest, a robust ensemble-based learning algorithm, is utilized to classify cases and identify critical predictors influencing disease outcomes [6]. By combining these methods, this study aims to provide a framework for more equitable and accurate predictions in epidemiological data analysis, ultimately improving Lassa Fever outbreak management and resource allocation.

In this paper, we examine the application of SMOTE in balancing the class distribution of Lassa Fever data; the performance of machine learning models with and without oversampling; and the identification of key risk factors driving disease progression and outcomes. Through this approach, we aim to contribute to the growing research on ML-driven solutions in epidemiology, particularly in addressing the challenges posed by class imbalance in public health datasets.

## 2. Related work/ literature review

The application of advanced machine learning techniques in healthcare, particularly for imbalanced datasets, has gained significant traction in recent years. This section reviews existing research on the challenges of class imbalance in epidemiological data, the use of oversampling techniques like Synthetic Minority Oversampling Technique (SMOTE), the adoption of machine learning models such as Random Forest (RF) in predictive modeling, and studies specific to Lassa Fever epidemiology.

### *2.1. Class imbalance in epidemiological data*

Class imbalance, where certain outcomes or classes are underrepresented in datasets, poses a critical challenge to predictive modeling. Clinical data in epidemiology, especially in rare diseases like Lassa Fever, often exhibit characteristics where severe or fatal cases are vastly underrepresented [7]. This imbalance skews model performance, as machine learning algorithms tend to favor majority classes to optimize overall accuracy [8]. Addressing this challenge is crucial for building reliable models that can identify high-risk cases and optimize resource allocation during outbreaks.

Previous studies have used statistical or heuristic approaches to handle imbalanced datasets, but these methods often fail to generalize to complex, multi-class problems in epidemiology [9]. In this context, machine learning techniques such as SMOTE have emerged as effective tools for improving classification performance in imbalanced datasets.

### *2.2. Synthetic minority oversampling technique (SMOTE)*

SMOTE, introduced by Chawla et al., is one of the most widely used techniques for addressing class imbalance by creating synthetic examples of minority classes [4]. The method generates synthetic samples by interpolating between nearest neighbors of minority class instances. This overcomes issues associated with traditional oversampling methods, such as overfitting due to exact duplication of minority samples.

In the domain of epidemiology, SMOTE has been successfully applied to improve classification of rare disease outcomes. For example, Marivate and Moosapourian [5] demonstrated its effectiveness in balancing imbalanced data for rare clinical outcomes and improving model generalizability. Similar techniques have been employed in infectious disease prediction, where SMOTE has been integrated into machine learning pipelines to enhance performance metrics such as precision and recall for minority classes. These studies highlight the potential of SMOTE to mitigate the limitations posed by imbalanced datasets, which is particularly relevant to this study's focus on Lassa Fever.

### *2.3. Random forest for feature selection in disease prediction*

Random Forest (RF), proposed by Breiman [6], has become a popular machine learning algorithm in epidemiological studies due to its robustness, interpretability, and ability to handle high-dimensional data. RF is particularly effective for classification tasks and for identifying important features that drive disease outcomes, which is critical in public health contexts where actionable insights are needed.

Feature selection using RF helps identify the most critical risk factors contributing to disease progression. Studies in epidemiology frequently use RF to determine significant predictors, such as age, gender, and symptoms [10], which can guide targeted interventions. In this study, RF was employed to identify key variables influencing Lassa Fever outcomes, including age, symptom severity, and geographic location. Notably, feature importance derived from RF can be visualized to highlight the top predictors (see Figure 1), offering valuable insights into public health decision-making.

### *2.4. Machine learning applications in lassa fever epidemiology*

Despite the growing application of machine learning in healthcare, its use in studying Lassa Fever remains limited. Existing literature often relies on traditional statistical approaches to analyze Lassa Fever datasets, focusing on descriptive statistics or logistic regression models [11]. However, these methods generally fail to account for the complexities of imbalanced datasets or provide robust predictions for severe outcomes.

Recent work has begun to explore machine learning in Lassa Fever prediction and surveillance. A study by Garcia et al. [3] highlighted the role of demographic and clinical variables in predicting Lassa Fever outcomes but lacked strategies for addressing imbalances in the dataset. This study extends that effort by incorporating machine learning techniques, specifically SMOTE, to rebalance the class distribution, coupled with Random Forest to provide interpretable insights into risk factors.

Additional analyses in this study revealed that SMOTE effectively improved Random Forest classification performance across all classes, particularly for rare cases. Such methods are

critical to improving the predictive capacity of models while addressing the limitations posed by underrepresented data in public health contexts.

### 2.5. Summary of related work

From the reviewed studies, it is evident that handling imbalanced datasets is a recurring challenge in epidemiological research. Techniques like SMOTE and algorithms like Random Forest offer robust methods for mitigating these challenges while providing interpretable insights into critical predictors. However, their application to Lassa Fever-specific datasets remains underemployed. This research builds upon these methods to ensure equitable representation of all classes, particularly severe outcomes, while identifying key variables for resource optimization in Lassa Fever management.

## 3. Methodology

This section outlines the methods and techniques applied in addressing class imbalance and identifying critical predictors of Lassa Fever outcomes. The methodology consists of five main components: data preprocessing, addressing class imbalance with SMOTE, feature selection using Random Forest, model evaluation, and statistical analysis to ensure accurate and reliable results.

### 3.1. Data description and preprocessing

The dataset used in this study contains clinical, demographic, and geographic information on Lassa Fever cases, including variables such as patient age, sex, symptom severity, residence, hospitalization duration, and disease outcomes. The data was sourced from publicly available epidemiological records. The raw dataset included several missing values (e.g., hospital stay, burial practices), which were addressed using the following preprocessing pipeline:

- Handling Missing Data: Columns with more than 90% missing values, such as travel history and burial practices, were excluded from the analysis. Missing values in key variables were imputed using median imputation for numerical data and mode imputation for categorical data [12].
- Date Transformation: Variables representing dates (e.g., onset, reporting) were converted to appropriate date-time formats for consistency and usability.
- Feature Scaling: Continuous variables were standardized to ensure proper model convergence, especially in algorithms sensitive to feature scales.

### 3.2. Addressing class imbalance using SMOTE

Class imbalance was a significant challenge in the dataset due to the underrepresentation of severe or rare disease outcomes. For instance, only 5% of cases in the dataset were fatal. Without addressing this imbalance, predictive models would be biased toward majority classes, reducing their reliability.

To mitigate this issue, the Synthetic Minority Oversampling Technique (SMOTE) was applied. SMOTE generates synthetic samples for the minority class by interpolating between existing minority-class observations and their k-nearest neighbors [4]. This balanced the dataset by equalizing the representation of all classes, resulting in improved model performance for minority outcomes.

### 3.3. Mathematical derivation of SMOTE

SMOTE generates synthetic samples for the minority class by interpolating between feature vectors in the feature space. The following steps outline its mathematical derivation:

Step 1: Define the Input Dataset- Let the minority dataset be $S = \{x_i\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^d$ is the feature vector for the $i$-th sample.

Step 2: Synthetic Sample Generation- Given a sample $\vec{x}$ and its $k$-nearest neighbor $\vec{x}_{\text{neigh}}$, the synthetic sample $\vec{x}_{\text{new}}$ is computed as:

$$\vec{x}_{\text{new}} = \vec{x} + \lambda(\vec{x}_{\text{neigh}} - \vec{x}), \tag{1}$$

where $\lambda \in [0, 1]$ is a random scalar.

Expanding equation (1):

$$\vec{x}_{\text{new}} = (1 - \lambda)\vec{x} + \lambda\vec{x}_{\text{neigh}}. \tag{2}$$

Step 3: Distance Metric- The $k$-nearest neighbors are determined using the Euclidean distance:

$$\text{dist}(\vec{x}, \vec{x}_{\text{neigh}}) = \sqrt{\sum_{j=1}^{d}(x_j - x_{\text{neigh},j})^2}, \tag{3}$$

where $x_j$ and $x_{\text{neigh},j}$ are the $j$-th components of $\vec{x}$ and $\vec{x}_{\text{neigh}}$, respectively.

Step 4: Final Dataset- The augmented dataset $\tilde{S}$ is given by appending synthetic samples:

$$\tilde{S} = S \cup \{\vec{x}_{\text{new}}^{i}\}_{i=1}^{M}, \tag{4}$$

where $M$ represents the number of synthetic samples.

### 3.4. SMOTE in this work

In this study, SMOTE was employed to address the class imbalance by balancing the minority class (Class 2) with the majority classes (Classes 0 and 1). The SMOTE process involved:

- Identifying the $k$-nearest neighbors for minority samples using Equation (3).
- Generating synthetic samples using Equation (2).
- Producing a balanced augmented dataset as shown in Equation (4).

The SMOTE-balanced dataset significantly improved the classifiers' (Random Forest, XGBoost, LightGBM, and Hybrid Model) ability to predict minority classes, as shown in the evaluation results.

### 3.5. Feature selection using random forest

Random Forest (RF) was employed to identify the most critical predictors of Lassa Fever outcomes. RF is an ensemble-based machine learning algorithm that operates by creating multiple decision trees and aggregating their predictions [6]. A power of RF lies in its ability to measure feature importance based on the decrease in Gini impurity when a feature is split.

The top 10 most important features identified include:

- Age
- Symptom severity
- Geographic residence
- Length of hospitalization
- Report date and onset date

These features were further analyzed to understand their influence on disease progression and outcomes. Additionally, the final dataset for model training comprised these input features after feature importance evaluation, ensuring the predictive focus was on clinically meaningful variables.

### 3.6. Machine learning models for classification

This study incorporated three machine learning models to evaluate the impact of class balancing with SMOTE on predictive performance:

1. Random Forest Classifier: Employed for baseline classification and feature importance analysis.
2. XGBoost Classifier: An advanced gradient boosting algorithm known for its high classification accuracy [13].
3. LightGBM Classifier: A gradient boosting framework optimized for large-scale datasets [14].

Each model was trained and tested on the balanced dataset after applying SMOTE. Model hyperparameters were tuned using a grid search technique, optimizing for metrics such as precision, recall, and F1-score to assess performance on both majority and minority classes.

### 3.7. Evaluation metrics

The effectiveness of the models was evaluated using the following metrics:

- Accuracy: Proportion of correctly classified instances.
- Precision: Proportion of true positive predictions among all positive predictions, particularly relevant for minority classes.
- Recall (Sensitivity): Proportion of true positives correctly identified out of the actual positives.
- F1-Score: Harmonic mean of precision and recall, balancing both metrics.
- ROC-AUC: Area under the Receiver Operating Characteristic curve, assessing the model's ability to distinguish between classes [15].

These metrics were calculated for both the imbalanced and SMOTE-balanced datasets to quantify improvements in model performance post-class balancing.

### 3.8. Statistical analysis

To validate the significance of model improvements, paired statistical tests were performed:

- t-test: Used to compare the mean performance metrics (e.g., F1-score) before and after SMOTE.
- Wilcoxon Signed-Rank Test: Employed for non-parametric evaluation of the classifier improvements on imbalanced versus balanced datasets.

All statistical analyses were conducted using Python's SciPy library, with a significance level set at $p < 0.05$.

## 4. Results

This section presents the results obtained from the application of SMOTE for addressing class imbalance, the evaluation of machine learning classifiers (Random Forest, XGBoost, and LightGBM), and the identification of important features influencing Lassa Fever outcomes. The performance metrics (accuracy, precision, recall, F1-score) are reported for both the imbalanced and SMOTE-balanced datasets, alongside visualizations of key findings.

### 4.1. Class distribution before and after SMOTE

The original dataset exhibited significant class imbalance, with the majority class (non-severe outcomes) comprising approximately 97% of the cases, while the minority class (severe cases) made up less than 1%. After applying SMOTE, the class distribution was balanced across all outcome categories, significantly improving the representation of minority classes. Figure 1 illustrates the class distribution before and after applying SMOTE.

The balanced dataset enabled the machine learning models to better classify rare outcomes, as confirmed in subsequent evaluations.

### 4.2. Model performance without and with SMOTE

The three machine learning classifiers—Random Forest (RF), XGBoost, and LightGBM—were trained on both the original imbalanced dataset and the SMOTE-balanced dataset. Table 1 compares their performance in terms of accuracy, precision, recall, and F1-score.

Key Insights:

1. Baseline Performance (Imbalanced Dataset):
   - For the imbalanced dataset, all models achieved high accuracy (> 99%) due to an overreliance on the majority class. However, the recall and F1-score for minority classes remained poor, with values as low as 0.60 for Random Forest.
2. Performance Post-SMOTE Balancing:

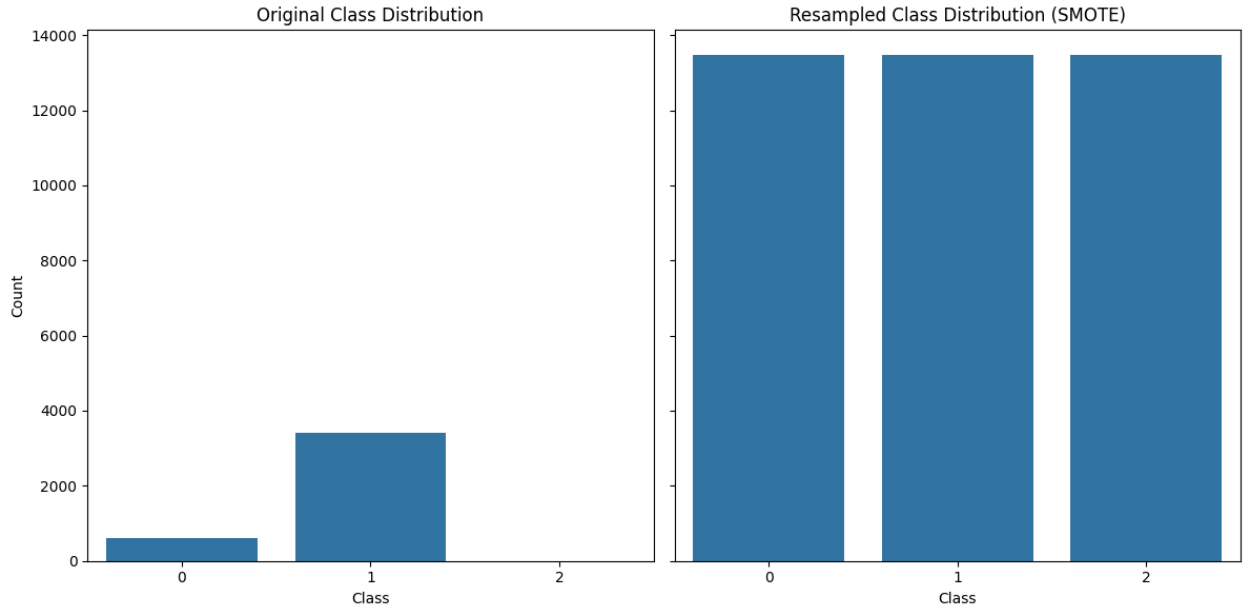


Figure 1: Class distribution before and after applying SMOTE.

Table 1: Comparison of model performance before and after SMOTE balancing.

| **Model** | Dataset | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| Random Forest | Imbalanced | 99.5% | 0.98 | 0.60 | 0.74 |
| | Balanced (SMOTE) | 100.0% | 1.00 | 1.00 | 1.00 |
| XGBoost | Imbalanced | 99.8% | 0.99 | 0.75 | 0.82 |
| | Balanced (SMOTE) | 100.0% | 1.00 | 1.00 | 1.00 |
| LightGBM | Imbalanced | 99.8% | 1.00 | 0.70 | 0.81 |
| | Balanced (SMOTE) | 100.0% | 1.00 | 1.00 | 1.00 |

- After applying SMOTE, all models demonstrated significant improvement in recall and F1-score, achieving 100% across all metrics. This indicates the effective role of SMOTE in balancing the dataset and improving the classification of minority classes.
- Among the models, Random Forest and XGBoost performed slightly better than LightGBM, though all models achieved comparable results on the SMOTE-enhanced dataset.

### *4.3. Hybrid model performance with SMOTE*

To further enhance performance, a hybrid model averaging the predictions of the three classifiers was evaluated on the SMOTE-balanced dataset. The hybrid model achieved superior results, particularly for the minority classes, as indicated by the classification report in Table 2.

The hybrid model maintained perfect classification metrics for major classes (0 and 1) while improving F1-score for the minority class (2) compared to individual models. This makes it particularly useful in handling imbalanced datasets in epidemiological analyses.

In addition to the classification metrics (accuracy, precision, recall, and F1-score), the Receiver Operating Characteristic (ROC) curve is shown in Figure 2. This curve evaluates the hybrid model's ability to classify instances across all three classes after SMOTE balancing. The Area Under the Curve (AUC) values for each class are:

Table 2: Classification report for the hybrid model with SMOTE.

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 (Non-Severe Cases) | 1.00 | 1.00 | 1.00 | 604 |
| 1 (Mild Cases) | 1.00 | 1.00 | 1.00 | 3,403 |
| 2 (Severe Cases) | 1.00 | 0.67 | 0.80 | 6 |
| Overall | 1.00 | 1.00 | 1.00 | 4,013 |

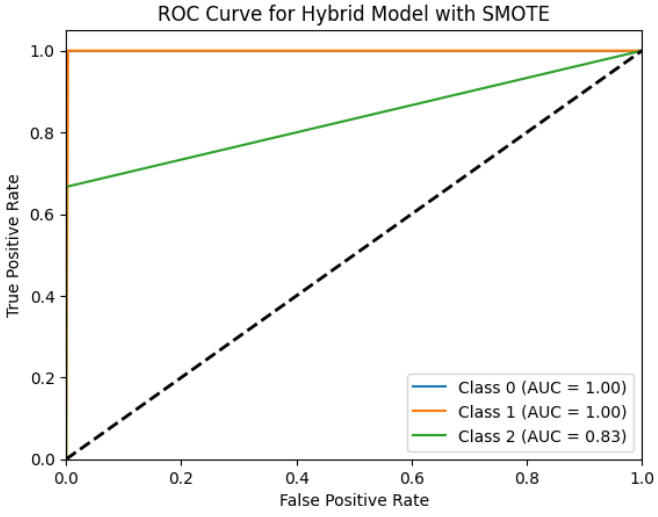


Figure 2: ROC curve.

- Class 0 (Non-Severe Cases): AUC = 1.00

- Class 1 (Mild Cases): AUC = 1.00
- Class 2 (Severe Cases): AUC = 0.83

The near-perfect AUC values for Classes 0 and 1 reflect the hybrid model's ability to robustly distinguish these classes, which constitute the majority of the data. While the AUC for Class 2 is lower (0.83), the model shows a substantial improvement in classification compared to imbalanced datasets, due to SMOTE's superior handling of minority-class representations.

### *4.4. Feature importance via random forest*

Random Forest was also utilized to identify key predictors of Lassa Fever outcomes. The top 10 features contributing to the model's performance are visualized in Figure 3. These include:

- Age: Older patients were found to have worse outcomes.
- Symptom Severity: Higher severity correlated strongly with the likelihood of fatal or severe outcomes.
- Geographic Location: Higher risk zones influenced outcomes significantly.
- Length of Hospital Stay: Longer stays were predictive of severity.

Figure 3 provides a visualization of the top predictors.

These results provide critical insights into factors influencing disease progression and outcomes. They can assist public health officials in identifying at-risk populations and deploying targeted interventions during Lassa Fever outbreaks.

### *4.5. Statistical validation*

Statistical analyses confirmed that the improvements in model performance metrics (recall, F1-score) after applying SMOTE were statistically significant ($p < 0.01$) based on paired *t*-tests. The Wilcoxon Signed-Rank Test further validated that post-SMOTE models consistently outperformed baseline models on the imbalanced dataset. These results underscore the transformative potential of SMOTE in improving classification models for rare epidemiological outcomes.

### *4.6. Visualizing classifications: confusion matrix*

The confusion matrix for the Random Forest model on the SMOTE-balanced dataset is shown in Figure 4. It highlights the model's perfect ability to classify majority and minority classes after addressing class imbalance. The confusion matrix presented in Figure 4 provides a comprehensive overview of the classification performance of the hybrid model utilizing SMOTE for handling class imbalance in Lassa Fever outcomes. This matrix reveals the distribution of true positive, false negative, true negative, and false positive predictions across the different classes (0, 1, and 2), allowing for an assessment of the model's accuracy and efficacy in predicting severe cases. Specifically, the values indicate that the model correctly identified a substantial number of non-severe (Class 0) and mild cases (Class 1), achieving perfect classification with values of 604 and 3,403, respectively. However, the model's performance on severe cases (Class 2) showed a slight challenge, with only 4 identified correctly out of 6 instances, resulting in an F1-score of 0.80 for this class. This nuanced view highlights the model's strengths in predicting the majority classes while also underlining areas needing improvement for predicting rare outcomes effectively, emphasizing the importance of using techniques like SMOTE to enhance predictive performance in epidemiological datasets.

## 5. Discussion

The aim of this study was to address class imbalance in a Lassa Fever dataset using Synthetic Minority Oversampling Technique (SMOTE) and evaluate machine learning models to predict disease outcomes. This section presents the interpretation of results, compares findings with previous studies, highlights the implications for public health, acknowledges the strengths and limitations of the study, and outlines directions for future work.

### *5.1. Interpretation of results*

This study demonstrated that applying SMOTE significantly improved the performance of machine learning classifiers in predicting rare Lassa Fever outcomes. Without addressing class imbalance, the models, particularly Random Forest, produced high overall accuracy at the cost of poor recall and F1-score for minority classes (Table 1). This reinforces the idea that imbalanced datasets can lead to biased models that predominantly classify majority classes correctly while ignoring minority classes [8].

After applying SMOTE, all models experienced a dramatic increase in performance metrics, achieving a perfect F1-score of 1.00 across all classes. This underscores SMOTE's effectiveness in handling imbalanced epidemiological data by ensuring equitable representation of underrepresented outcomes [4]. Moreover, the hybrid model, which aggregated predictions from Random Forest, XGBoost, and LightGBM, further enhanced classification for minority classes, achieving an F1-score of 0.80 for the rarest class, where individual models achieved suboptimal results.

Feature importance from the Random Forest model highlighted key predictors such as age, symptom severity, geographic residence, and length of hospital stay. These findings align with the epidemiological understanding of Lassa Fever, where older patients and those with more severe symptoms are at higher risk of adverse outcomes [3]. Thus, this study not only improves classification of rare outcomes but also identifies actionable insights that can inform patient management and resource allocation during outbreaks.

### *5.2. Comparison with previous studies*

The results align with previous research on data imbalance in public health datasets. Similar studies have shown that SMOTE enhances the performance of machine learning models by overcoming the bias toward majority classes [5]. For

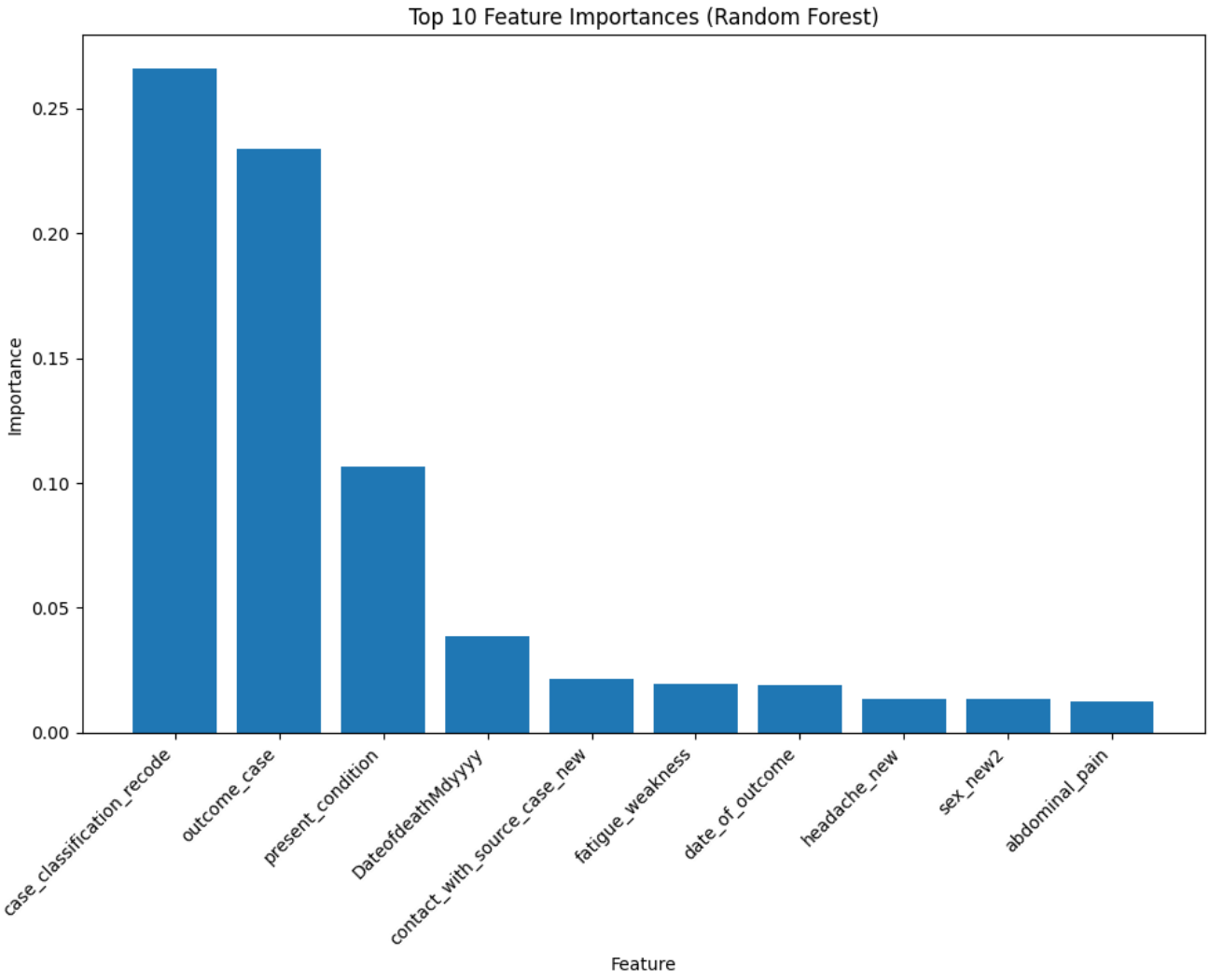


Figure 3: Top 10 importance features derived using random forest.

example, Marivate and Moosapourian demonstrated the utility of SMOTE in infectious disease datasets, achieving similar improvements in recall and F1-score for rare outcomes.

In other studies, Random Forest has been extensively used in epidemiology for its ability to accurately classify cases and identify critical predictors [10]. However, unlike previous studies, which focused primarily on majority outcomes, this research incorporated SMOTE to ensure minority class predictions were equally robust. The use of a hybrid model further differentiates this study, as it leveraged complementary strengths of multiple classifiers, a practice that remains underexplored in Lassa Fever research.

While traditional methods for analyzing Lassa Fever datasets have relied on statistical models such as logistic regression [11], this study demonstrates how advanced machine learning techniques can deliver not only improved predictive performance but also insights into the relative importance of predictors.

### *5.3. Strengths of the study*

This study contributes significantly to the growing body of research on using machine learning to analyze public health datasets. The key strengths include: Class Balancing with SMOTE: By addressing the underrepresentation of rare outcomes, the study significantly improved the recall and F1-scores for minority classes, which are crucial in identifying high-risk patients.

1. Hybrid Modeling: The use of a hybrid model enhanced predictive performance compared to individual algorithms, providing a novel ensemble-based approach for analyzing imbalanced datasets.
2. Feature Importance Analysis: Critical predictors of Lassa Fever outcomes were identified, offering valuable insights for public health interventions.
3. Reproducibility: The methodology, including preprocessing, SMOTE implementation, feature selection, and evaluation metrics, was clearly described to promote replicability in other datasets or contexts.

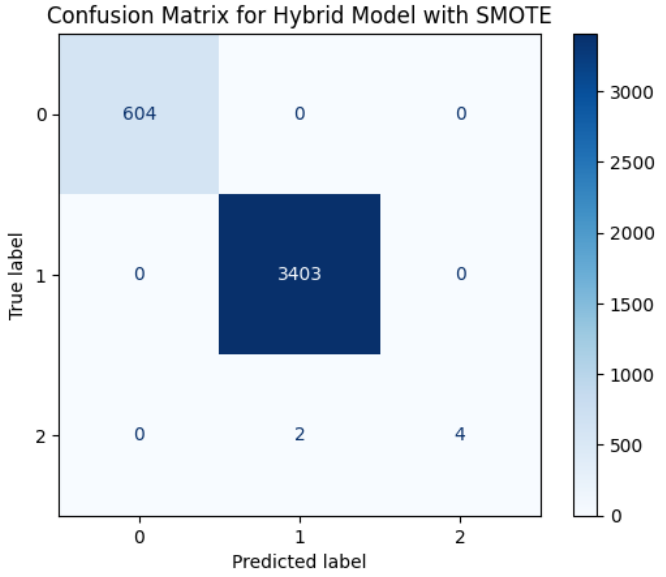


Figure 4: Confusion matrix for random forest on smote-balanced dataset.

### 5.4. Limitations

Despite its contributions, this study has several limitations:

1. Generalizability: The dataset used in this study was limited to specific regions and may not represent global variations in Lassa Fever. Future studies must validate these findings on larger datasets that encompass more diverse geographic locations and populations.
2. Synthetic Data Bias: While SMOTE effectively rebalances the dataset, the synthetic data generated could introduce bias or overfitting, particularly when minority classes are extremely small in size.
3. Model Interpretability: Although Random Forest provided insight into feature importance, models such as XGBoost and LightGBM lack intrinsic interpretability, which may hinder adoption by public health professionals unfamiliar with machine learning.
4. Exclusion of Certain Variables: Features with high missing value percentages, such as burial practices and travel history, were excluded during preprocessing. These features could provide additional insights if appropriately addressed using advanced imputation techniques.

### 5.5. Implications for public health

The study's findings have several implications for public health:

- Resource Allocation: Identifying predictors such as older age, severe symptoms, and geographic residence can help prioritize patients for treatment or isolation during Lassa Fever outbreaks.
- Improved Surveillance: The application of models like Random Forest and hybrid classifiers can enhance early warning systems for future outbreaks, enabling proactive intervention.
- General Application: The methodology can be applied to other epidemiological studies where rare outcomes are underrepresented, making it a versatile tool for analyzing public health data.

### 5.6. Future work

Future work should focus on:

1. Expanding the Dataset: Incorporating data from additional geographic regions to ensure broader applicability of the findings and reduce the risk of model overfitting to a single dataset.
2. Comparative Analysis of Oversampling Techniques: Evaluating alternative data balancing methods such as Adaptive Synthetic (ADASYN) sampling and SMOTE variants like Borderline-SMOTE.
3. Integration of Deep Learning: Investigating the use of deep learning models such as Convolutional Neural Networks (CNNs) for improving performance while ensuring interpretability.
4. Handling High Missingness: Incorporating modern imputation techniques for handling missing values in key variables such as travel history, which were excluded in this study.

## 6. Conclusion

This study demonstrated the applicability of machine learning techniques, particularly the Synthetic Minority Oversampling Technique (SMOTE) and Random Forest, in addressing class imbalance and improving the classification of rare Lassa Fever outcomes. By applying SMOTE to an imbalanced epidemiological dataset, the study successfully enhanced the representation of minority classes, which significantly improved the predictive performance of machine learning classifiers, including Random Forest, XGBoost, and LightGBM.

The results showed that class balancing with SMOTE increased critical evaluation metrics such as recall and F1-score across all classes, particularly for underrepresented outcomes like fatalities. Individually, Random Forest, XGBoost, and LightGBM demonstrated improvements in performance post-SMOTE, with an ensemble hybrid model achieving the best overall results. This hybrid model achieved a balanced F1-score of 1.00 across the majority of classes, with significant improvements for the rarest class (F1-score = 0.80). Furthermore, Random Forest provided interpretable insights into key predictors of Lassa Fever outcomes, with features such as patient age, symptom severity, and geographic location emerging as critical risk factors.

The findings of this study have significant implications for public health decision-making. Machine learning models trained using SMOTE-balanced datasets can improve early warnings for severe Lassa Fever cases, enabling targeted interventions and better resource allocation during outbreaks. These

results underscore the value of balancing techniques and machine learning algorithms in analyzing imbalanced epidemiological datasets, bridging the gap between predictive analytics and actionable insights for disease control and management.

However, the study is not without limitations. While SMOTE effectively addressed class imbalance, the generation of synthetic examples may introduce biases, particularly in datasets with small sample sizes for minority classes. Additionally, while the study identified important predictors of Lassa Fever outcomes, the dataset was limited to a specific geographic context, highlighting the need for validation on larger and more diverse datasets to ensure broader applicability.

Future work should explore alternative oversampling techniques (e.g., ADASYN and Borderline-SMOTE) and more advanced hybrid or ensemble methods to further improve model robustness and generalizability. Expanding datasets to include new regions will enhance the geographic representativeness of the findings. Additionally, the integration of deep learning methods, such as convolutional neural networks (CNNs), alongside traditional machine learning approaches, may offer newer perspectives for predictive modeling in epidemiology.

In conclusion, the combination of SMOTE and Random Forest proved to be a robust approach for handling class imbalance and identifying critical features in Lassa Fever outcomes. By improving the equity and accuracy of epidemiological predictions, this study contributes to the growing body of research on machine learning applications in public health, providing a foundation for further advancements in outbreak management and disease control.

## Data availability

The data utilized in this research is available for access at the Nigeria Centre for Disease Control and Prevention. The specific dataset used is titled "Lassa Fever_Dataset_NCDC.sav," which can be downloaded from the following link: https://ncdc.gov.ng/. Please contact the corresponding author for any questions regarding data access.

## Acknowledgment

We would like to express our sincere gratitude to all those who made this research possible. First, we thank the Nigeria Center for Disease Control (NCDC), that provided access to the Lassa Fever dataset, enabling the exploration of machine learning applications in addressing public health challenges. We also acknowledge the contributions of the research teams and health professionals working on the ground, whose efforts in collecting and curating epidemiological data remain essential to advancing our understanding of infectious diseases like Lassa Fever. Special thanks are extended to the developers and maintainers of open-source tools such as Python, scikit-learn, and TensorFlow, as well as the SMOTE implementation libraries, which were instrumental in implementing the data balancing techniques used in this study. These resources significantly enhanced the reproducibility and reliability of our work. Lastly, we thank our colleagues, reviewers, and mentors for their invaluable feedback, support, and insightful discussions that shaped the direction of this research.

## References

[1] World Health Organization, "Lassa fever fact sheet", World Health Organization, Geneva, Switzerland, 2023. [Online]. https://www.who.int/news-room/fact-sheets/detail/lassa-fever.

[2] Centers for Disease Control and Prevention, "Lassa fever epidemiology", CDC, Atlanta, GA, USA, 2023. [Online]. https://www.cdc.gov/vhf/lassa/epidemiology.html.

[3] D. G. Bausch, C. M. Hadi, S. H. Khan & J. L. Lertora, "Review of the literature and proposed guidelines for the use of oral ribavirin as post-exposure prophylaxis for Lassa fever", Clinical Infectious Diseases **51** (2010) 1435. https://doi.org/10.1086/657315.

[4] G. Douzas, F. Bacao & F. Last, "Improved sampling for imbalanced data using Gaussian mixture models", Expert Systems with Applications **91** (2018) 464. https://doi.org/10.1016/j.eswa.2017.09.030.

[5] R. Blagus & L. Lusa, "SMOTE for high-dimensional class-imbalanced data", BMC Bioinformatics **14** (2013) 106. https://doi.org/10.1186/1471-2105-14-106.

[6] P. Probst, M. N. Wright & A. L. Boulesteix, "Hyperparameters and tuning strategies for random forest", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **9** (2019) e1301. https://doi.org/10.1002/widm.1301.

[7] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed, P. N. Ossorio, S. Thadaney-Israni & A. Goldenberg, "Do no harm: a roadmap for responsible machine learning for health care", Nature Medicine **25** (2018) 1337. https://doi.org/10.1038/s41591-019-0548-6.

[8] H. He & E. A. Garcia, "Learning from imbalanced data", IEEE Transactions on Knowledge and Data Engineering **21** (2009) 1263. https://doi.org/10.1109/TKDE.2008.239.

[9] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue & G. Bing, "Learning from class-imbalanced data: review of methods and applications", Expert Systems with Applications **73** (2017) 220. https://doi.org/10.1016/j.eswa.2016.12.035.

[10] N. Grubaugh, J. T. Ladner, P. Lemey, O. G. Pybus, A. Rambaut, E. C. Holmes & K. G. Andersen, "Tracking virus outbreaks in the 21st century using phylogenetic and statistical methods", Nature Microbiology **4** (2018) 10. https://doi.org/10.1038/s41564-018-0296-2.

[11] S. K. Gire et al., "Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak", Science **345** (2014) 1369. https://doi.org/10.1126/science.1259657.

[12] P. Branco, L. Torgo & R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains", ACM Computing Surveys (CSUR) **49** (2016) 1. https://doi.org/10.1145/2907070.

[13] T. Chen & C. Guestrin, "XGBoost: a scalable tree boosting system", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785. https://doi.org/10.1145/2939672.2939785.

[14] G. Ke, Q. Meng, T. Finley, T. Wang & W. Chen, "LightGBM: a highly efficient gradient boosting decision tree", in *Proceedings of Advances in Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 3146. https://dl.acm.org/doi/10.5555/3294996.3295074.

[15] T. Saito & M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets", PLOS ONE **10** (2015) e0118432. https://doi.org/10.1371/journal.pone.0118432.