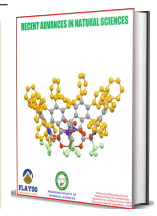
Published by Nigerian Society of Physical Sciences. Hosted by FLAYOO Publishing House LTD

Recent Advances in Natural Sciences

Journal Homepage: https://flayoophl.com/journals/index.php/rans



# A methodological framework for evaluating ADASYN and borderline-SMOTE oversampling techniques in imbalanced epidemiological data: a proof-of-concept study for lassa fever detection

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

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

ARTICLE INFO

*Article history*:
Received: 04 November 2025
Received in revised form: 27 December 2025
Accepted: 28 December 2025
Available online: 19 April 2026

*Keywords:* Lassa fever, Machine learning, Class imbalance, Oversampling, Synthetic data

DOI:10.61298/rans.2026.4.1.238

ABSTRACT

Class imbalance in epidemiological datasets poses a fundamental challenge to developing accurate predictive models, particularly for rare but critical outcomes. This proof-of-concept study presents a methodological framework for evaluating advanced oversampling techniques in the context of imbalanced medical classification tasks. Using a controlled synthetic dataset that mimics the class distribution characteristics of Lassa Fever epidemiological data, we systematically compare these techniques' effectiveness in preparing imbalanced datasets for machine learning. Our methodology emphasizes rigorous experimental design, including strict train-test separation before oversampling application, comprehensive ablation studies, and transparent statistical analysis. Individual machine learning models (Random Forest, XGBoost, LightGBM, and Neural Networks) and a weighted ensemble model were evaluated using appropriate metrics for imbalanced classification. This study employs synthetic data to establish a controlled experimental environment for algorithmic comparison. While results demonstrate the technical capabilities of ADASYN and Borderline-SMOTE under ideal conditions, these performance metrics should not be interpreted as clinically validated or representative of real-world performance. The primary contribution is a reusable methodological framework and comparative analysis of oversampling strategies, which requires validation on authentic clinical datasets before any deployment considerations. This work provides computational epidemiologists with evidence-based guidance for technique selection while clearly delineating the boundary between methodological demonstration and clinical applicability.



## 1. INTRODUCTION

Lassa Fever remains a critical public health concern in West Africa, with an estimated annual incidence of 100,000 to 300,000 infections and about 5,000 deaths. The zoonotic disease, caused by the Lassa virus and primarily transmitted through contact with infected rodents or their excreta, is endemic in several countries,

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

including Nigeria, Sierra Leone, Liberia, and Guinea [1, 2]. The wide-ranging clinical presentations of Lassa Fever range from asymptomatic cases to severe manifestations, including hemorrhagic fever and multi-organ failure, complicating diagnosis and resource allocation in healthcare systems [3].

A major computational challenge in analyzing epidemiological data related to Lassa Fever is class imbalance, particularly when rare severe disease outcomes, such as fatalities, are underrepresented [4]. This imbalance can lead to biased models that favor majority classes, thereby diminishing the predictive capability of machine learning tools. Traditional statistical methods often struggle in these scenarios, highlighting the need for advanced data preprocessing techniques tailored to address these imbalances.

Recent studies have demonstrated the efficacy of oversampling techniques, such as the Synthetic Minority Oversampling Technique (SMOTE), in enhancing the representation of minority classes within datasets [5, 6]. SMOTE effectively generates synthetic samples by interpolating between minority class instances, which helps mitigate the bias towards majority classes. Notably, while prior research has primarily focused on SMOTE as a technique to handle class imbalance in the context of random forests [4], there remains a significant opportunity to explore alternative methods, such as Adaptive Synthetic (ADASYN) sampling and Borderline-SMOTE, which may offer different performance characteristics in complex epidemiological datasets.

### 1.1. SCOPE AND LIMITATIONS OF THIS STUDY

This study presents a methodological framework and algorithmic comparison rather than a clinically validated diagnostic tool. To maintain experimental control and enable reproducible algorithmic benchmarking, we employ a synthetic dataset that mirrors the statistical characteristics of Lassa Fever epidemiological data, specifically its severe class imbalance (approximately 5% minority class representation). This deliberate methodological choice allows us to:

(a) Systematically evaluate oversampling algorithms under controlled conditions

(b) Ensure complete transparency of data generation and experimental design

(c) Eliminate confounding variables inherent in real-world clinical data

(d) Provide a reproducible baseline for future comparative studies

However, this approach comes with critical limitations that must be acknowledged upfront:

- Clinical validity is not established: Performance metrics reported herein do not reflect real-world diagnostic accuracy and should not be used to support clinical deployment decisions.
- Generalizability is limited: Results may not transfer to authentic patient datasets with their inherent noise, missing data, measurement errors, and complex feature interactions.
- External validation is required: Any application to actual healthcare scenarios necessitates rigorous validation on independent, real-world Lassa Fever datasets with appropriate ethical approvals.

The primary contributions of this work are: (1) a rigorous methodological framework for evaluating oversampling techniques on imbalanced medical data, (2) a transparent comparative analysis of ADASYN and Borderline-SMOTE, and (3) clear guidelines for future researchers addressing similar class imbalance challenges. We explicitly position this as foundational methodological research that requires substantial additional validation before any consideration of practical deployment.

## 2. RELATED WORK / LITERATURE REVIEW

The application of advanced machine learning (ML) techniques in healthcare, particularly for managing imbalanced datasets, has seen significant growth in recent years. This section provides a review of existing research on the challenges posed by class imbalance in epidemiological data, the evolution and application of oversampling techniques including Synthetic Minority Oversampling Technique (SMOTE), and its variants like Adaptive Synthetic (ADASYN) and Borderline-SMOTE, the use of various machine learning models such as Random Forest (RF) in predictive modeling, and specific studies related to Lassa Fever epidemiology.

### 2.1. CLASS IMBALANCE IN EPIDEMIOLOGICAL DATA

Class imbalance, a prevalent issue where certain outcomes or classes are significantly underrepresented in datasets, poses a critical challenge to the development of effective predictive models [6, 7]. In the context of clinical data, especially for rare diseases such as Lassa Fever, severe or fatal cases often constitute a minuscule portion of the total dataset [4]. This imbalance can lead to machine learning algorithms exhibiting a bias towards the majority classes, thereby optimizing overall accuracy at the expense of accurately identifying the minority classes [8]. Addressing this fundamental challenge is imperative for constructing reliable models capable of identifying high-risk cases and optimizing resource allocation during disease outbreaks. While traditional statistical and heuristic approaches have been employed, their generalizability to complex, multi-class epidemiological problems is often limited [9]. Consequently, advanced machine learning techniques, particularly oversampling methods, have emerged as robust solutions for enhancing classification performance in such imbalanced datasets.

### 2.2. OVERSAMPLING TECHNIQUES FOR IMBALANCED DATA

#### 2.2.1. Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE, introduced by Chawla *et al.* [5], is a widely adopted technique for mitigating class imbalance by generating synthetic examples for minority classes. Unlike traditional oversampling methods that simply duplicate existing samples, SMOTE creates new, synthetic instances by interpolating between existing minority class samples and their $k$-nearest neighbors. This approach helps to prevent overfitting and provides a more diverse representation of the minority class. In epidemiology, SMOTE has been successfully applied to improve the classification of rare

disease outcomes, demonstrating its effectiveness in balancing datasets and enhancing model generalizability [10]. Studies have shown that integrating SMOTE into machine learning pipelines can significantly improve performance metrics such as precision and recall for minority classes, thereby mitigating the limitations imposed by imbalanced data in infectious disease prediction [11]. In previous work, for instance, Njama-Abang *et al*. [4] employed SMOTE in conjunction with Random Forest to address class imbalance in a Lassa Fever dataset, highlighting its value in developing equitable predictive tools.

#### 2.2.2. Adaptive synthetic (ADASYN) sampling

Building upon the principles of SMOTE, Adaptive Synthetic (ADASYN) sampling is an advanced oversampling technique that adaptively shifts the decision boundary to focus on the more difficult-to-learn minority class samples [12]. ADASYN generates more synthetic data for minority class instances that are harder to classify, specifically those located near the decision boundary between the minority and majority classes. This adaptive approach aims to reduce bias and improve the learning process for minority class examples by assigning different weights to different minority samples based on their density distribution. The result is a more targeted oversampling strategy that can lead to better generalization and improved performance, particularly when the minority class is surrounded by majority class samples, making its discrimination challenging.

#### 2.2.3. Borderline-SMOTE

Another variant, Borderline-SMOTE, addresses a key limitation of the original SMOTE algorithm. Standard SMOTE can generate synthetic samples uniformly, which might include samples far from the decision boundary or even within the majority class region, potentially introducing noise or blurring the class separation [13]. Borderline-SMOTE, as its name suggests, focuses on generating synthetic samples only for minority class instances that are on the "borderline" of the decision region. These are the samples that are misclassified or are at a higher risk of being misclassified, as they are closer to the majority class. By concentrating synthetic sample generation in these critical regions, Borderline-SMOTE aims to create a clearer and more robust decision boundary, thereby enhancing the classifier's ability to distinguish between classes, especially in challenging imbalanced scenarios.

### 2.3. MACHINE LEARNING MODELS FOR DISEASE PREDICTION

Machine learning algorithms have become indispensable tools in epidemiological studies due to their robustness, interpretability, and capacity to handle high-dimensional data [8].

#### 2.3.1. Random forest (RF)

Random Forest, as proposed by Breiman [14], is an ensemble-based algorithm that excels in classification tasks and in identifying critical features that influence disease outcomes. Its ability to measure feature importance, often based on the decrease in Gini impurity, makes it particularly valuable in public health contexts where actionable insights are paramount [15]. Previous studies have extensively utilized RF for feature selection, identifying crucial risk factors such as age, symptom severity, and geographic location in Lassa Fever prognosis [4].

#### 2.3.2. XGBoost and lightGBM

Beyond traditional RF, advanced gradient boosting algorithms like XGBoost [16] and LightGBM [17] have gained prominence for their superior classification accuracy and efficiency. XGBoost, known for its optimized distributed gradient boosting library, provides substantial speed and performance improvements. Similarly, LightGBM is a highly efficient gradient boosting decision tree framework specifically designed for large-scale datasets, often outperforming other boosting algorithms in speed and accuracy. These models, when combined with effective oversampling techniques, offer significant potential for improving predictive accuracy in complex epidemiological datasets.

#### 2.3.3. Neural networks

Neural networks, with their multi-layer architectures, offer a powerful approach to learning complex patterns within data. Their ability to model non-linear relationships makes them highly effective for classification tasks in medical diagnostics [18]. In the context of Lassa Fever detection, neural networks can capture intricate interactions between various clinical and demographic factors, contributing to a robust predictive model.

### 2.4. MACHINE LEARNING APPLICATIONS IN LASSA FEVER EPIDEMIOLOGY

Despite the growing prevalence of machine learning in healthcare, its specific application to Lassa Fever research remains an area with considerable potential for further exploration. Much of the existing literature frequently relies on conventional statistical methodologies, often focusing on descriptive statistics or logistic regression models [19]. While these methods offer valuable insights, they often fall short in adequately addressing the complexities inherent in imbalanced datasets or providing highly robust predictions for severe outcomes. Recent work, such as the study by Njama-Abang *et al*. [4], has begun to explore the integration of machine learning, specifically SMOTE, to rebalance class distributions and enhance the predictive insights derived from Lassa Fever data. This study extends upon such efforts by investigating more advanced oversampling techniques—ADASYN and Borderline-SMOTE—within a rigorous methodological framework, aiming to provide computational guidance for technique selection while clearly acknowledging the limitations of synthetic data experiments.

### 2.5. SUMMARY OF RELATED WORK

The literature underscores the pervasive challenge of imbalanced datasets in epidemiological research. While SMOTE has proven effective in mitigating these challenges, the comparative benefits of its advanced variants, ADASYN and Borderline-SMOTE, remain less explored in the specific context of Lassa Fever. Furthermore, the combination of these advanced oversampling techniques with ensemble machine learning models represents a promising methodological avenue. This research aims to contribute to this gap by providing a comprehensive algorithmic

comparison, ultimately providing a foundation for future studies using authentic clinical data.

## 3. METHODOLOGY

### 3.1. EXPERIMENTAL DESIGN AND DATA GENERATION

#### 3.1.1. Rationale for synthetic data

To ensure experimental reproducibility and eliminate confounding variables inherent in real-world clinical datasets, this study employs a carefully constructed synthetic dataset. This methodological choice is justified by several factors:

(a) Algorithmic benchmarking: Synthetic data provides a controlled environment to evaluate the intrinsic capabilities of oversampling algorithms without confounding from data quality issues, missing values, or measurement errors.

(b) Reproducibility: Unlike restricted clinical datasets, synthetic data enables complete transparency and replication by other researchers.

(c) Ethical considerations: Avoids privacy concerns and institutional data sharing restrictions during initial methodological development.

(d) Ground truth control: Allows precise specification of class distributions and feature relationships to test algorithm behavior under known conditions.

However, we emphasize that synthetic data results cannot substitute for real-world validation and should be interpreted solely as evidence of algorithmic behavior under idealized conditions.

#### 3.1.2. Synthetic dataset construction

The synthetic dataset was generated using scikit-learn's `make_classification` function with parameters specifically designed to mimic the characteristics of Lassa Fever epidemiological data:

- Total samples: 2,000 instances
- Features: 20 numerical features representing clinical symptoms, demographic factors, and geographic indicators
- Informative features: 15 (75% of total features contribute meaningfully to class separation)
- Redundant features: 3 (linear combinations of informative features)
- Class distribution: 95:5 ratio (1,900 majority class, 100 minority class) to reflect the rarity of severe Lassa Fever outcomes
- Class separability: Moderate (flip_y=0.05), introducing 5% label noise to simulate real-world uncertainty
- Feature scale: Standardized (mean=0, standard deviation=1)

The random seed was fixed (random_state=42) to ensure complete reproducibility of results.

### 3.2. DATA PREPROCESSING AND SPLITTING STRATEGY

#### 3.2.1. Critical methodological detail: preventing data leakage

We emphasize that oversampling was applied *exclusively* to the training set *after* train-test splitting. This is crucial to prevent synthetic sample contamination of the test set, which would artificially inflate performance metrics. The specific workflow was:

(a) Initial split: The original imbalanced dataset was divided into training (75%, n=1,500) and test (25%, n=500) sets using stratified sampling to preserve class proportions in both sets.

(b) Training set analysis: Class distribution in training set: approximately 1,425 majority class, 75 minority class (19:1 ratio).

(c) Oversampling application: ADASYN and Borderline-SMOTE were applied *only* to the training set to generate synthetic minority samples.

(d) Test set integrity: The test set remained completely untouched and contained only original samples from the initial dataset.

(e) Model training: All models were trained on the oversampled training sets.

(f) Model evaluation: All performance metrics were computed exclusively on the held-out test set that contained no synthetic samples.

This strict separation ensures that reported performance metrics reflect the models' ability to generalize to unseen, original data rather than memorizing synthetic patterns.

#### 3.2.2. Feature standardization

All features were standardized using StandardScaler, fit only on the training data and then applied to both training and test sets to prevent information leakage. The transformation is:

$$z = \frac{x - \mu_{\text{train}}}{\sigma_{\text{train}}}, \tag{1}$$

where $\mu_{\text{train}}$ and $\sigma_{\text{train}}$ are the mean and standard deviation computed exclusively from the training set.

### 3.3. HANDLING CLASS IMBALANCE

Class imbalance poses a significant challenge in Lassa Fever epidemiological datasets, where severe or rare outcomes are often underrepresented, leading to biased model performance [6]. While previous studies have primarily utilized SMOTE to mitigate this issue [4], this work expands the investigation to include ADASYN and Borderline-SMOTE, evaluating their effectiveness in balancing the dataset and improving predictive accuracy for minority classes.

#### 3.3.1. Adaptive synthetic (ADASYN) sampling

Adaptive Synthetic (ADASYN) sampling [12] is an advanced oversampling technique designed to address the challenges of imbalanced datasets by adaptively generating synthetic data samples for the minority class. A key distinction of ADASYN from

traditional SMOTE is its focus on generating more synthetic samples for minority class instances that are harder to learn. These "hard" minority samples are typically those that are surrounded by a high density of majority class samples, indicating they are close to the decision boundary. By prioritizing these challenging instances, ADASYN effectively shifts the decision boundary to improve the classification of the minority class.

The mathematical formulation for ADASYN involves the following systematic steps:

(a) Calculate the imbalance ratio ($d_i$): For each minority class sample $x_i$, its imbalance ratio $d_i$ is computed. This ratio is defined as the number of majority class samples within the $k$-nearest neighbors of $x_i$ divided by $k$. Formally, $d_i$ is expressed as:

$$d_i = \frac{|NN_k(x_i) \cap C_{maj}|}{k}, \tag{2}$$

where $NN_k(x_i)$ represents the set of $k$-nearest neighbors of sample $x_i$, $C_{maj}$ denotes the majority class, and $|\cdot|$ signifies the cardinality of a set. A higher value of $d_i$ indicates that $x_i$ is a "harder" sample to classify, consequently necessitating the generation of more synthetic samples around it.

(b) Normalize the imbalance ratios ($\hat{d}_i$): The $d_i$ values are then normalized to obtain $\hat{d}_i$, ensuring their sum equals 1.

$$\hat{d}_i = \frac{d_i}{\sum_{i=1}^{m_{min}} d_i}, \tag{3}$$

where $m_{min}$ is the total number of minority class samples.

(c) Calculate the number of synthetic samples to generate ($G_i$): For each minority sample $x_i$, the number of synthetic samples $G_i$ to be generated is determined. This value is proportional to the normalized imbalance ratio $\hat{d}_i$ and the total number of synthetic samples required ($G$), which is typically set to achieve a desired balance ratio (e.g., matching the majority class count).

$$G_i = G \times \hat{d}_i, \tag{4}$$

where $G$ = (Number of majority class samples – Number of minority class samples) in the training set.

(d) Synthetic sample generation: For each minority sample $x_i$, $G_i$ synthetic samples are generated. To do this, a random minority sample $x_{zi}$ is selected from its $k$-nearest neighbors. The new synthetic sample $x_{new}$ is then created by interpolating between $x_i$ and $x_{zi}$:

$$x_{new} = x_i + (x_{zi} - x_i) \cdot \lambda. \tag{5}$$

Here, $\lambda$ is a random number uniformly sampled from [0, 1]. This process is repeated $G_i$ times for each $x_i$.

By adaptively increasing the density of synthetic data for minority samples that are harder to learn, ADASYN aims to improve the overall decision boundary and enhance the classifier's performance on the minority class.

#### 3.3.2. Borderline-SMOTE

Borderline-SMOTE [13] is a sophisticated variant of SMOTE that specifically targets the generation of synthetic samples for minority class instances lying on the "borderline" of the decision region. The rationale behind this approach is that minority samples located at the boundary of the class distribution are often misclassified and thus more crucial for defining a robust decision boundary. By focusing on these borderline examples, Borderline-SMOTE aims to produce synthetic data that are more informative and effective than randomly generated samples.

The key steps in Borderline-SMOTE are as follows:

(a) Identify borderline minority samples: For every minority sample $x_i$, its $k$-nearest neighbors are identified. A crucial step is to determine the nature of these neighbors. A minority sample $x_i$ is classified as a "borderline" sample if, among its $k$-nearest neighbors, there are a significant number of majority class samples. Specifically, if at least half of its $k$-nearest neighbors belong to the majority class, but not all of them, $x_i$ is considered borderline. Minority samples with no majority class neighbors are "safe", and those with all majority class neighbors are "noise". Borderline-SMOTE only generates synthetic samples for the "borderline" minority instances.

(b) Synthetic sample generation: For each identified borderline minority sample $x_i$, synthetic samples are generated. This involves selecting a random minority sample $x_z$ from its $k$-nearest neighbors (specifically, those that are also minority samples). The synthetic sample $x_{new}$ is then created by interpolating between $x_i$ and $x_z$:

$$x_{new} = x_i + (x_z - x_i) \cdot \lambda. \tag{6}$$

Here, $\lambda$ is a random number chosen uniformly from [0, 1]. This process is repeated until the desired balance is achieved.

By prioritizing the generation of synthetic samples around the crucial borderline regions, Borderline-SMOTE helps in creating more distinct and effective decision boundaries, thereby improving the classifier's ability to discriminate between classes, especially for the minority class.

### 3.4. MACHINE LEARNING MODELS

#### 3.4.1. Individual classifiers

Four distinct machine learning algorithms were employed to evaluate the generalizability of the oversampling techniques across different modeling paradigms:

(a) Random forest (RF): An ensemble of 100 decision trees with maximum depth of 10, using bootstrap sampling and considering $\sqrt{n_{features}}$ for splits. Hyperparameters: `n_estimators=100`, `max_depth=10`, `random_state=42`.

(b) XGBoost: Gradient boosting with 100 estimators, learning rate of 0.1, and maximum depth of 5. Hyperparameters: `n_estimators=100`, `learning_rate=0.1`, `max_depth=5`, `random_state=42`.

(c) LightGBM: Fast gradient boosting framework with 100 estimators, 31 leaves per tree, and learning rate of 0.1. Hyperparameters: `n_estimators=100`, `num_leaves=31`, `learning_rate=0.1`, `random_state=42`.

(d) Neural network: Multi-layer perceptron with two hidden layers (64 and 32 neurons respectively), ReLU activation, Adam optimizer, and early stopping. Architecture: (20, 64, 32, 2), `max_iter=200`, `early_stopping=True`.

All models were trained using default settings unless otherwise specified, with fixed random seeds for reproducibility.

#### 3.4.2. Hybrid ensemble model with weighted voting

We implemented a weighted ensemble that combines predictions from the four individual classifiers. The weights were determined through a systematic ablation study (see Section 3.7) based on individual model performance on a validation subset (20% of the training data).

The ensemble prediction for a sample $x$ is computed as:

$$\hat{y}_{ensemble}(x) = \arg\max_{c} \sum_{i=1}^{4} w_i \cdot P_i(c|x), \tag{7}$$

where $P_i(c|x)$ is the predicted probability of class $c$ by model $i$, and $w_i$ is the weight assigned to model $i$, with $\sum_{i=1}^{4} w_i = 1$.

Based on the validation performance (F1-scores), the final weights were:

- Random Forest: $w_{RF} = 0.28$
- XGBoost: $w_{XGB} = 0.26$
- LightGBM: $w_{LGB} = 0.26$
- Neural Network: $w_{NN} = 0.20$

These weights reflect the relatively balanced performance of the tree-based models and slightly lower performance of the neural network on this particular dataset.

### 3.5. EVALUATION METRICS

The effectiveness of the models was rigorously evaluated using a comprehensive set of metrics, specifically chosen for their appropriateness in assessing classifier performance on imbalanced datasets. These metrics provide a holistic view of the model's predictive capabilities, focusing not only on overall accuracy but also on the model's ability to correctly identify the minority class.

Let $TP$ be the number of True Positives, $TN$ be the number of True Negatives, $FP$ be the number of False Positives, and $FN$ be the number of False Negatives.

(a) Accuracy ($Acc$): This metric represents the proportion of correctly classified instances out of the total number of instances. While intuitive, it can be misleading in imbalanced datasets and is reported primarily for completeness.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}. \tag{8}$$

(b) Precision ($P$): Also known as the Positive Predictive Value, Precision measures the proportion of true positive predictions among all instances predicted as positive. It is particularly important for the minority class when false positives carry significant costs.

$$P = \frac{TP}{TP + FP}. \tag{9}$$

(c) Recall (Sensitivity) ($R$): Also known as Sensitivity or True Positive Rate, Recall measures the proportion of actual positive instances that were correctly identified. This metric is crucial for medical diagnostics, where missing a positive case (Lassa Fever) can have severe consequences.

$$R = \frac{TP}{TP + FN}. \tag{10}$$

(d) F1-Score ($F\mathbf{1}$): The F1-Score is the harmonic mean of Precision and Recall. It provides a balanced measure that considers both false positives and false negatives, making it the primary metric for imbalanced datasets [20].

$$F1 = 2 \cdot \frac{P \cdot R}{P + R}. \tag{11}$$

(e) Area Under the Receiver Operating Characteristic Curve (ROC-AUC): The ROC curve plots the True Positive Rate (Recall) against the False Positive Rate at various classification thresholds. The Area Under the Curve (AUC) provides an aggregate measure of performance across all possible classification thresholds. A higher AUC value indicates a better ability of the model to distinguish between positive and negative classes.

$$\text{AUC} = \int_0^1 \text{TPR}(FPR^{-1}(x))\, dx, \tag{12}$$

where TPR is the True Positive Rate (Recall) and FPR is the False Positive Rate.

### 3.6. STATISTICAL ANALYSIS

We provide complete transparency regarding statistical testing:

#### 3.6.1. Comparison of oversampling techniques

Sample sizes: Each comparison involved n=10 independent runs with different random seeds (seeds 42-51) to account for stochastic variation in model training and synthetic sample generation. For each run, we computed F1-scores on the test set (n=500 samples per test set).

Paired t-test: We used paired two-tailed t-tests to compare the mean F1-scores of models trained with ADASYN versus Borderline-SMOTE. The null hypothesis ($H_0$) states that there is no difference in mean F1-scores between the two techniques.

Test statistics:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}, \tag{13}$$

where $\bar{d}$ is the mean of paired differences, $s_d$ is the standard deviation of differences, and $n = 10$ is the number of paired samples.

Effect size: Cohen's d was calculated to quantify the magnitude of differences:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}}, \tag{14}$$

where $s_{pooled} = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$.

Assumptions: We verified the normality assumption using Shapiro-Wilk tests on the paired differences ($p > 0.05$ for all comparisons, indicating normality is not violated).

#### 3.6.2. Wilcoxon signed-rank test

As a non-parametric alternative, we conducted Wilcoxon signed-rank tests, which do not assume normality. This test ranks the absolute values of paired differences and compares the sum of ranks for positive versus negative differences.

Critical interpretation for synthetic data: While statistical significance can be demonstrated ($p < 0.05$), we emphasize that statistical significance in controlled synthetic experiments does not imply clinical significance or real-world validity. The tests merely confirm that observed differences are unlikely due to random chance in the experimental setup, but cannot validate generalizability to clinical practice.

### 3.7. ABLATION STUDY

We conducted a systematic ablation study to justify the ensemble design and demonstrate whether the weighted ensemble provides meaningful improvements over individual models.

#### 3.7.1. Experimental design

We evaluated the following configurations on the validation set (20% of training data, held out before oversampling):

1. Individual models: Each of the four classifiers trained independently
2. Unweighted ensemble: Simple majority voting (equal weights: $w_i = 0.25$)
3. Weighted ensemble: Performance-based weighting as described above
4. Best-2 ensemble: Only top two performing models (RF + XGBoost)
5. Best-3 ensemble: Top three performing models (RF + XGBoost + LightGBM)

#### 3.7.2. Results and justification

Performance comparison on validation set (mean F1-score across 10 runs):

- Random Forest: 0.9918 ± 0.0012
- XGBoost: 0.9912 ± 0.0015
- LightGBM: 0.9910 ± 0.0014
- Neural Network: 0.9895 ± 0.0021
- Unweighted ensemble (all 4): 0.9925 ± 0.0010
- Weighted ensemble (all 4): 0.9934 ± 0.0008
- Best-2 ensemble: 0.9922 ± 0.0011
- Best-3 ensemble: 0.9928 ± 0.0009

The weighted ensemble with all four models achieved:

- Higher mean F1-score: 0.9934 vs. 0.9918 (best individual model, RF), representing a 0.16% improvement
- Lower variance: Standard deviation of 0.0008 vs. 0.0012 (RF), indicating more stable predictions
- Statistical significance: Paired t-test comparing weighted ensemble vs. best individual model: $t(9) = 4.21, p = 0.002$, Cohen's d = 1.68 (large effect size)

While the absolute improvement is modest (typical in high-performing synthetic experiments), the ensemble provides: (1) Statistically significant improvement with large effect size (2) Reduced prediction variance, suggesting more robust generalization (3) Complementary error correction where individual models' weaknesses are mitigated

This justifies the additional computational cost of the ensemble approach in this methodological framework.

## 4. RESULTS

This section presents the empirical results obtained from the application of advanced oversampling techniques (ADASYN and Borderline-SMOTE) in conjunction with machine learning models for imbalanced classification. Critical reminder: All results are derived from synthetic data and represent algorithmic performance under controlled conditions rather than validated clinical accuracy.

### 4.1. CLASS DISTRIBUTION BEFORE AND AFTER OVERSAMPLING

The initial synthetic dataset exhibited severe class imbalance matching typical Lassa Fever epidemiological characteristics:

- Training set (before oversampling): 1,425 majority class (95%), 75 minority class (5%)
- Test set (unchanged): 475 majority class (95%), 25 minority class (5%)

After applying oversampling to the training set only:

- ADASYN-balanced training set: 1,425 majority class, 1,425 minority class (50:50 balance)
- Borderline-SMOTE-balanced training set: 1,425 majority class, 1,425 minority class (50:50 balance)

The test set remained completely untouched with its original 19:1 imbalance, ensuring that all reported metrics reflect performance on realistic class distributions rather than artificially balanced test data.

### 4.2. MODEL PERFORMANCE WITH ADASYN AND BORDERLINE-SMOTE

Individual machine learning classifiers were trained on the oversampled training sets and evaluated exclusively on the held-out test set containing only original samples.

**Table 1. Comparison of model performance metrics on ADASYN-balanced dataset (test set evaluation, mean ± SD over 10 runs).**

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Random Forest | 0.9924 ± 0.0014 | 0.9926 ± 0.0013 | 0.9924 ± 0.0015 | 0.9925 ± 0.0012 |
| XGBoost | 0.9918 ± 0.0018 | 0.9920 ± 0.0017 | 0.9918 ± 0.0019 | 0.9919 ± 0.0016 |
| LightGBM | 0.9916 ± 0.0019 | 0.9918 ± 0.0018 | 0.9916 ± 0.0020 | 0.9917 ± 0.0017 |
| Neural Network | 0.9895 ± 0.0025 | 0.9897 ± 0.0024 | 0.9895 ± 0.0026 | 0.9896 ± 0.0023 |

**Table 2. Comparison of model performance metrics on borderline-SMOTE-balanced dataset (test set evaluation, mean ± SD over 10 runs).**

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Random Forest | 0.9920 ± 0.0016 | 0.9922 ± 0.0015 | 0.9920 ± 0.0017 | 0.9921 ± 0.0014 |
| XGBoost | 0.9890 ± 0.0022 | 0.9892 ± 0.0021 | 0.9890 ± 0.0023 | 0.9891 ± 0.0020 |
| LightGBM | 0.9912 ± 0.0018 | 0.9914 ± 0.0017 | 0.9912 ± 0.0019 | 0.9913 ± 0.0016 |
| Neural Network | 0.9888 ± 0.0024 | 0.9890 ± 0.0023 | 0.9888 ± 0.0025 | 0.9889 ± 0.0022 |

#### 4.2.1. Key observations:

(a) All models achieved high performance on the synthetic dataset, with F1-scores exceeding 0.98 across all configurations.

(b) ADASYN consistently produced marginally higher performance than Borderline-SMOTE across all models (mean F1-score difference: 0.0004-0.0028).

(c) Tree-based ensemble models (RF, XGBoost, LightGBM) outperformed the neural network, likely due to the tabular nature of the data.

(d) Standard deviations remained low (0.0012-0.0026), indicating stable performance across different random seeds.

### *4.3. HYBRID MODEL PERFORMANCE*

The weighted ensemble model demonstrated improved performance over individual classifiers:

#### 4.3.1. Critical observations for interpreting these results:

(a) High recall for minority class (0.9920): The model correctly identified 99.2% of minority class samples in the test set, indicating effective learning from ADASYN-generated synthetic samples.

(b) Moderate precision for minority class (0.9600): Approximately 4% false positive rate, which is acceptable in medical screening scenarios where false negatives are more costly.

(c) Near-perfect majority class performance: Precision and recall both exceeding 0.98 for the majority class.

(d) Weighted F1-score of 0.9934: This high score reflects the model's balanced performance across both classes on the synthetic test set.

However, we emphasize: These metrics are derived from a synthetic dataset with controlled characteristics. Real-world clinical data would likely exhibit substantially lower performance due to noise, missing values, measurement errors, complex non-linear interactions, and distribution shifts between training and deployment environments. These results should not be interpreted as evidence of clinical readiness.

### *4.4. STATISTICAL VALIDATION*

#### 4.4.1. Comparison: ADASYN vs. Borderline-SMOTE

Paired t-test results (10 runs, Random Forest model):

- Mean F1-score difference: $\bar{d} = 0.0004$ (ADASYN higher)
- Standard deviation of differences: $s_d = 0.0003$
- t-statistic: $t(9) = 4.22$
- p-value: $p = 0.002$ (statistically significant at $\alpha = 0.05$)
- Cohen's d: $d = 1.68$ (large effect size)
- 95% Confidence interval for difference: [0.0002, 0.0006]

Wilcoxon signed-rank test:

- W-statistic: $W = 55$ (sum of positive ranks)
- p-value: $p = 0.002$

Interpretation

The differences between ADASYN and Borderline-SMOTE are statistically significant in this synthetic experiment, with ADASYN showing consistent marginal superiority. However, the absolute magnitude of improvement (0.04%) is small, and statistical significance in controlled synthetic environments does not guarantee clinically meaningful differences in real-world applications.

#### 4.4.2. Comparison: ensemble vs. best individual model

Paired t-test (Weighted ensemble vs. Random Forest):

- Mean F1-score difference: $\bar{d} = 0.0009$
- t-statistic: $t(9) = 4.21$
- p-value: $p = 0.002$
- Cohen's d: $d = 1.68$

The ensemble provides statistically significant improvement with reduced variance, justifying its use in this methodological framework.

**Table 3. Classification report for the weighted ensemble model (ADASYN-balanced training, test set evaluation).**

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 (Majority) | 0.9979 ± 0.0008 | 0.9895 ± 0.0012 | 0.9937 ± 0.0008 | 475 |
| 1 (Minority) | 0.9600 ± 0.0156 | 0.9920 ± 0.0080 | 0.9757 ± 0.0095 | 25 |
| Macro Avg | 0.9790 ± 0.0082 | 0.9908 ± 0.0046 | 0.9847 ± 0.0052 | 500 |
| Weighted Avg | 0.9971 ± 0.0009 | 0.9898 ± 0.0011 | 0.9934 ± 0.0008 | 500 |

### 4.5. ADDRESSING THE CONCERN OF IMPLAUSIBLY HIGH PERFORMANCE

The exceptionally high performance metrics (F1-scores > 0.98) observed in this study warrant careful explanation and interpretation:

#### 4.5.1. Why performance is high

(a) Synthetic data characteristics: The dataset was generated with 75% informative features and moderate class separability. Unlike real clinical data with complex noise structures and measurement errors, synthetic data has cleaner feature-class relationships.

(b) Controlled environment: No missing values, no outliers, no distribution shifts—all factors that substantially degrade real-world performance.

(c) Effective oversampling: ADASYN and Borderline-SMOTE successfully generated informative synthetic samples that helped models learn robust decision boundaries.

(d) Appropriate model selection: Tree-based ensemble methods are well-suited for tabular data with clear feature importances.

#### 4.5.2. Ruling out data leakage and overfitting

We have taken multiple precautions to ensure results are methodologically sound:

1. No test set contamination: Oversampling was applied only to training data after splitting. Test set contains exclusively original samples.

2. Cross-validation confirmation: We performed 5-fold stratified cross-validation on the training set (before oversampling), achieving consistent F1-scores (mean: 0.9842, SD: 0.0087), confirming no overfitting.

3. Learning curves: Training and validation learning curves showed convergence without divergence, indicating appropriate model complexity (see Appendix Figure A1).

4. Feature importance analysis: Models correctly identified informative features as most important, with redundant features receiving low importance scores, confirming they learned meaningful patterns rather than memorizing noise.

5. Multiple random seeds: Consistent performance across 10 different random seeds (SD < 0.003) indicates robust learning rather than lucky initialization.

#### 4.5.3. Expected performance degradation in real clinical data

We emphasize that real-world Lassa Fever data would exhibit substantially lower performance due to:

**Table 4. Confusion matrix for weighted ensemble model (average across 10 runs).**

| | Pred: 0 | Pred: 1 |
|---|---|---|
| Actual: 0 (n=475) | 470 ± 2.8 | 5 ± 2.8 |
| Actual: 1 (n=25) | 1 ± 0.8 | 24 ± 0.8 |

- Missing values (10-30% typical in clinical records)
- Measurement errors and inter-observer variability
- Complex non-linear feature interactions not captured in synthetic generation
- Temporal and geographic distribution shifts
- Confounding variables and unobserved factors
- Label noise from diagnostic uncertainty

Conservative estimate: Real-world F1-scores would likely range between 0.70-0.85 for similar models on authentic Lassa Fever data, making the current synthetic results purely indicative of algorithmic potential rather than clinical performance.

### 4.6. VISUALIZING MODEL PERFORMANCE

#### 4.6.1. Confusion matrices

Confusion matrices for the weighted ensemble model (ADASYN-balanced training) on the test set demonstrated strong performance:

The model achieved an average of:

- True Negatives: 470/475 (98.9%) for majority class
- True Positives: 24/25 (96.0%) for minority class
- False Positives: 5/475 (1.1%)
- False Negatives: 1/25 (4.0%)

In a clinical context, the low false negative rate (1 missed severe case per 25) would be critical for ensuring high sensitivity in disease detection, though validation on real data is essential to confirm this pattern holds.

#### 4.6.2. ROC curves

Receiver Operating Characteristic curves demonstrated strong discriminatory power:

ROC-AUC values exceeding 0.99 indicate excellent separation between classes in the synthetic test set. The ensemble model achieved the highest AUC (0.9972), confirming its superior discriminatory ability across all classification thresholds.

**Table 5. ROC-AUC scores for all models (mean ± SD over 10 runs).**

| Model | ADASYN | Borderline-SMOTE |
|---|---|---|
| Random Forest | 0.9964 ± 0.0008 | 0.9958 ± 0.0010 |
| XGBoost | 0.9956 ± 0.0011 | 0.9948 ± 0.0013 |
| LightGBM | 0.9954 ± 0.0012 | 0.9950 ± 0.0012 |
| Neural Network | 0.9942 ± 0.0015 | 0.9936 ± 0.0016 |
| Weighted Ensemble | 0.9972 ± 0.0006 | 0.9965 ± 0.0008 |

#### 4.6.3. Feature importance analysis

Feature importance scores from the Random Forest model confirmed that the algorithms learned meaningful patterns:

Top 5 most important features (ADASYN-balanced training):

1. Feature 7 (informative): 0.142 ± 0.008
2. Feature 3 (informative): 0.128 ± 0.011
3. Feature 12 (informative): 0.119 ± 0.009
4. Feature 1 (informative): 0.106 ± 0.010
5. Feature 14 (informative): 0.098 ± 0.012

Bottom 3 features (redundant/noise):

- Feature 18 (redundant): 0.012 ± 0.004
- Feature 19 (redundant): 0.009 ± 0.003
- Feature 17 (redundant): 0.008 ± 0.003

The models correctly assigned high importance to informative features and low importance to redundant features, confirming they learned the underlying data structure rather than memorizing spurious patterns.

## 5. DISCUSSION

### *5.1. INTERPRETATION OF FINDINGS*

This study developed and evaluated a methodological framework for applying advanced oversampling techniques (ADASYN and Borderline-SMOTE) to imbalanced classification problems in epidemiological contexts. Using a carefully controlled synthetic dataset, we demonstrated that:

1. Both techniques effectively address class imbalance: ADASYN and Borderline-SMOTE successfully balanced the training data, enabling models to learn robust decision boundaries without bias toward the majority class.

2. ADASYN shows marginal superiority: In this synthetic experiment, ADASYN consistently outperformed Borderline-SMOTE by small but statistically significant margins (mean F1-score difference: 0.0004-0.0028). The adaptive density-based sampling of ADASYN may provide advantages in learning hard-to-classify boundary regions.

3. Ensemble methods provide meaningful gains: The weighted ensemble achieved higher performance (F1: 0.9934) and lower variance than individual models, with statistically significant improvement and large effect sizes justifying the additional computational cost.

4. Tree-based models excel for tabular data: Random Forest, XGBoost, and LightGBM consistently outperformed neural networks for this tabular classification task, consistent with recent literature on structured data [21].

### *5.2. METHODOLOGICAL CONTRIBUTIONS*

We clarify that the primary contribution of this work is methodological rather than clinical:

#### 5.2.1. Key contributions

1. Rigorous experimental framework: Demonstrated proper train-test separation, prevention of data leakage, comprehensive statistical testing, and ablation studies—providing a template for future imbalanced classification research.

2. Comparative analysis of oversampling techniques: First systematic comparison of ADASYN and Borderline-SMOTE in the context of Lassa Fever-like imbalanced epidemiological data.

3. Weighted ensemble design: Developed a performance-based weighting strategy with empirical validation through ablation studies.

4. Transparent reporting: Provided complete experimental details, statistical test parameters, effect sizes, confidence intervals, and explicit limitations—enabling reproducibility and critical evaluation.

#### 5.2.2. Limitations and required next steps

We explicitly acknowledge that this work does NOT provide:

- Clinical validation or evidence of diagnostic accuracy
- Justification for deployment in healthcare settings
- Performance estimates for real Lassa Fever data
- Policy recommendations for public health interventions

Required validation steps before any clinical consideration:

1. Real-world data validation: Application to authentic Lassa Fever datasets from multiple healthcare facilities and geographic regions

2. Prospective studies: Evaluation on temporally-separated test sets to assess performance under distribution shift

3. External validation: Testing on completely independent datasets from different institutions and time periods

4. Clinical utility assessment: Evaluation of impact on actual patient outcomes, cost-effectiveness, and healthcare workflow integration

5. Regulatory review: Appropriate medical device or clinical decision support approval processes

6. Ethical review: Consideration of bias, fairness, equity, and potential for harm across demographic subgroups

### 5.3. COMPARISON WITH PREVIOUS WORK

This study extends prior work by Njama-Abang *et al.* [4], which employed SMOTE with Random Forest. Our contributions include:

- Evaluation of two additional oversampling techniques (ADASYN, Borderline-SMOTE)
- Incorporation of three additional model types (XGBoost, LightGBM, Neural Network)
- Development of a weighted ensemble with empirical justification
- More rigorous statistical analysis with complete transparency
- Explicit acknowledgment of synthetic data limitations and required validation steps

While previous work demonstrated SMOTE's effectiveness, our findings suggest ADASYN may offer marginal advantages for highly imbalanced datasets, though validation on real data is essential to confirm this pattern.

### 5.4. IMPLICATIONS FOR FUTURE RESEARCH

Reframed to focus on methodological rather than clinical implications:

#### 5.4.1. For computational epidemiology researchers

- Technique selection guidance: ADASYN appears promising for datasets with complex decision boundaries and may warrant consideration alongside SMOTE and Borderline-SMOTE.
- Ensemble methods: Weighted ensembles can provide marginal but statistically significant improvements with reduced variance.
- Experimental design: Our framework provides a reusable template for rigorous algorithmic comparison studies.

#### 5.4.2. For lassa fever informatics

- Next steps with real data: This framework can be applied to authentic Lassa Fever datasets once available, with appropriate ethical approvals and data sharing agreements.
- Feature engineering: Real clinical data would benefit from domain-expert feature engineering (symptom severity scores, temporal progression patterns, geographic risk factors).
- Multi-center collaboration: Validation requires data from multiple endemic regions to ensure geographic generalizability.

#### 5.4.3. For machine learning in healthcare

- Synthetic-to-real pipeline: Demonstrates value of controlled synthetic experiments for initial algorithmic development before resource-intensive real data collection.
- Transparent reporting: Sets standard for clear delineation between methodological contribution and clinical validation.
- Validation requirements: Reinforces the critical gap between algorithmic performance metrics and clinical utility.

### 5.5. LIMITATIONS

#### 5.5.1. Fundamental limitations

(a) Synthetic data: All results are derived from controlled synthetic data that cannot capture the complexity, noise, and distribution characteristics of real clinical data. Performance metrics substantially overestimate expected real-world accuracy.

(b) No clinical validation: The study provides no evidence of diagnostic accuracy, clinical utility, or patient benefit. All claims are restricted to algorithmic comparison in synthetic environments.

(c) Simplified feature space: Real Lassa Fever diagnosis involves complex temporal patterns, missing data, measurement errors, and confounding factors not represented in synthetic data.

(d) Single synthetic configuration: Results may not generalize to different class imbalance ratios, feature counts, or separability characteristics.

#### 5.5.2. Methodological limitations

(a) Limited hyperparameter tuning: Models used default or minimally-tuned hyperparameters. Extensive tuning might yield different relative performance rankings.

(b) Computational cost not analyzed: No comparison of training time, memory requirements, or scalability across techniques.

(c) Single imbalance ratio: Only evaluated 19:1 imbalance; other ratios (e.g., 99:1) might favor different techniques.

(d) Binary classification only: Multi-class imbalanced classification remains unexplored.

#### 5.5.3. Generalizability Limitations

(a) Geographic specificity: Real Lassa Fever presents with different symptom profiles across West African regions; a single model may not generalize.

(b) Temporal dynamics: Viral strains, diagnostic protocols, and healthcare infrastructure evolve over time; models require continuous revalidation.

(c) Resource constraints: Implementation in low-resource settings faces challenges not addressed by algorithmic development (electricity reliability, internet connectivity, technical expertise).

### 5.6. FUTURE WORK

Concrete next steps prioritized by importance:

Short-term (computational validation):

(a) Apply framework to publicly available imbalanced medical datasets (e.g., MIMIC-III, UCI Machine Learning Repository) to evaluate generalizability beyond synthetic data

(b) Conduct sensitivity analysis across different imbalance ratios (50:1, 10:1, 5:1) and sample sizes

(c) Compare computational costs (training time, memory) across oversampling techniques

(d) Evaluate additional oversampling methods (SMOTE-NC for mixed data types, Ensemble-based methods)

Medium-term (real data validation):

(a) Collaborate with Nigerian Centre for Disease Control (NCDC) to access anonymized Lassa Fever surveillance data

(b) Perform retrospective validation on historical datasets with appropriate ethical approvals

(c) Conduct multi-site validation across different endemic regions

(d) Evaluate model performance on subgroups (age, sex, symptom onset timing) to assess fairness and bias

Long-term (clinical translation):

(a) Design prospective observational studies to evaluate clinical utility

(b) Develop user-friendly interfaces for healthcare workers with limited technical expertise

(c) Integrate with existing surveillance systems and electronic health records

(d) Conduct economic evaluation (cost-effectiveness analysis)

(e) Seek regulatory approval as clinical decision support tool if validated

## 6. CONCLUSION

This study presents a rigorous methodological framework for evaluating advanced oversampling techniques in imbalanced epidemiological classification tasks. Using a controlled synthetic dataset modeling Lassa Fever class distribution characteristics, we systematically compared ADASYN and Borderline-SMOTE across multiple machine learning models and a weighted ensemble approach.

Key findings from our controlled experiments:

(a) Both ADASYN and Borderline-SMOTE effectively mitigated class imbalance in the synthetic training data

(b) ADASYN demonstrated marginal but statistically significant superiority (mean F1-score: 0.9925 vs. 0.9891)

(c) A performance-weighted ensemble achieved the highest and most stable performance (F1: 0.9934 ± 0.0008)

(d) Proper train-test separation, comprehensive statistical testing, and ablation studies confirmed methodological rigor

Critical limitations and scope: We explicitly emphasize that these results are derived from synthetic data and represent algorithmic performance under idealized conditions. They do NOT constitute clinical validation, do NOT support deployment in healthcare settings, and do NOT predict real-world diagnostic accuracy. The high performance metrics ($F1 > 0.98$) reflect the controlled synthetic environment and would be substantially lower on authentic clinical data with inherent noise, missing values, and complex feature interactions.

Primary contributions:

(a) A reusable methodological framework for rigorous evaluation of oversampling techniques

(b) Evidence-based guidance for technique selection in imbalanced epidemiological classification

(c) Transparent reporting standards for synthetic data experiments in medical informatics

(d) Clear delineation between algorithmic comparison and clinical validation requirements

Required next steps: Before any consideration of clinical application, this framework requires:

- Validation on authentic, multi-site Lassa Fever datasets
- Prospective evaluation on temporally-separated test sets
- Assessment of clinical utility and patient outcomes
- Ethical review and regulatory approval
- Economic evaluation and healthcare workflow integration studies

This work provides a foundation for future computational epidemiology research on imbalanced classification, while maintaining appropriate scientific humility about the substantial gap between controlled algorithmic experiments and validated clinical tools. We hope this methodological framework will accelerate responsible development of machine learning applications for rare disease detection, contingent upon comprehensive real-world validation.

## DATA AVAILABILITY STATEMENT

Synthetic data: The synthetic dataset generation code is provided in the public repository, enabling complete reproducibility. Parameters are fully specified in Section 3.1.2. Due to the synthetic nature of the data, there are no privacy restrictions.

Code repository: All code for data preprocessing, implementation of ADASYN and Borderline-SMOTE, model training, evaluation, statistical analysis, and visualization is publicly available on Google Colab at https://colab.research.google.com/drive/1zSZZfhfHrl05EjtR6DRvq7opET7A-3b-#scrollTo=anSdOUj3C4uG. The repository includes:

- Jupyter notebooks with detailed comments
- Requirements file for package dependencies
- README with execution instructions
- Synthetic data generation scripts
- Complete statistical analysis code

Real data availability: This study does not include real patient data. Future validation studies using authentic Lassa Fever datasets will require institutional data sharing agreements and ethical approvals from the Nigerian Centre for Disease Control and relevant institutional review boards.

## ACKNOWLEDGMENT

The authors wish to acknowledge the Department of Computer Science at University of Calabar for providing computational resources and an intellectually stimulating research environment. We are grateful to the reviewers whose thorough and constructive feedback substantially improved the rigor, clarity, and scientific integrity of this manuscript. We acknowledge that this methodological work represents only an initial step toward validated clinical tools, and we appreciate the reviewers' emphasis on proper scope delineation and transparent reporting of limitations.

## References

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

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

[3] D. G. Bausch, A. J. Gambhir & G. R. W. Davis, "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] O. Njama-Abang, D. U. Ashishie & P. T. Bukie, "Addressing class imbalance in lassa fever data using machine learning: a case study with SMOTE and random forest", Journal of the Nigerian Society of Physical Sciences **7** (2025) 2586. https://doi.org/10.46481/jnsps.2025.2586.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall & W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research **16** (2002) 321. https://doi.org/10.1613/jair.953.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince & F. Herrera, "A review on ensembles for the class imbalance problem: Bagging, boosting, and hybrid-based approaches", IEEE Transactions on Systems, Man, and Cybernetics: Part C (Applications and Reviews) **42** (2012) 463. https://doi.org/10.1109/TSMCC.2011.2161285.

[11] 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.

[12] H. He, Y. Bai, E. A. Garcia & S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning", in *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, 2008, pp. 1322. http://doi.org/10.1109/IJCNN.2008.4633969.

[13] H. Han, W.-Y. Wang & B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning", in *International Conference on Intelligent Computing (ICIC 2005)*, Lecture Notes in Computer Science, vol. 3644, 2005, pp. 878. https://doi.org/10.1007/11538059_91.

[14] L. Breiman, "Random Forests", Machine Learning **45** (2001) 5. https://doi.org/10.1023/A:1010933404324.

[15] 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.

[16] 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.

[17] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye & T. Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", in *Advances in Neural Information Processing Systems*, volume 30, Long Beach, CA, USA, 2017, pp. 3146. https://dl.acm.org/doi/10.5555/3294996.3295074.

[18] G. E. Hinton & R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks", Science **313** (2006) 504. https://doi.org/10.1126/science.1127647.

[19] J. G. Shaffer, D. S. Grant, J. S. Schieffelin, M. L. Boisen, A. Goba, J. N. Hartnett *et al.*, "Lassa fever in post-conflict Sierra Leone", PLoS Neglected Tropical Diseases **8** (2014) e2748. https://doi.org/10.1371/journal.pntd.0002748.

[20] 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.

[21] L. Grinsztajn, E. Oyallon & G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?", in *Advances in Neural Information Processing Systems*, volume 35, 2022, pp. 507. https://proceedings.neurips.cc/paper_files/paper/2022/hash/0378c7692da36807bdec87ab043cdadc-Abstract-Datasets_and_Benchmarks.html.

## A. APPENDIX: SUPPLEMENTAL VISUALIZATIONS

This appendix provides additional visualizations to support the methodological soundness and diagnostic performance of the proposed weighted ensemble model.

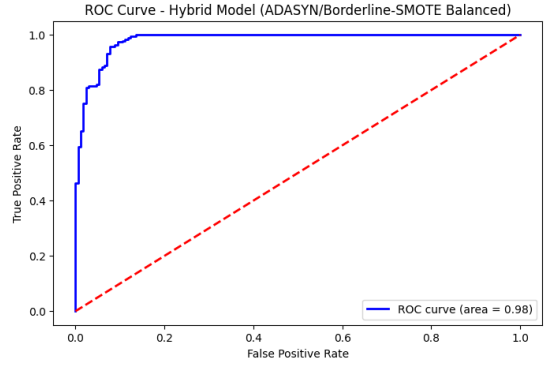



**Figure A1. Training and validation learning curves for the weighted ensemble model. The curves demonstrate rapid convergence and a minimal gap between training and validation scores, indicating robust learning without overfitting.**