AI- Driven Prediction of Lassa Fever Using Evolutional and Random Forest: A Machine Learning Approach for Enhanced Surveillance in West Africa

Osaseri R.O [a,1,*], Usiobaifo A.R [a,2], Ighodaro U.E [b,3]

[a] Department of Computer Science, University of Benin, P.M.B 1154, Benin City Nigeria
[b] Department of Production engineering, University of Benin, P.M.B 1154, Benin City Nigeria
[1] roseline.osaseri@uniben.edu; [2] rosemary.usiobaifo@uniben.edu; [3] uyibis@outlook.com
* corresponding author

ARTICLE INFO

Article history
Received October 1, 2024
Revised March 30, 2025
Accepted June 25, 2025

Keywords
Lassa fever
Artificial intelligence
Machine learning
Evolutionary algorithms
Random Forest
Feature selection
Classification algorithms
Predictive modeling

ABSTRACT

Lassa fever, a viral hemorrhagic fever endemic to West Africa, poses significant public health challenges with annual case estimates ranging from 100,000 to 300,000 infections and mortality rates reaching 15-20% in hospitalized patients. Current surveillance systems rely predominantly on passive case detection and laboratory confirmation, often resulting in delayed outbreak identification and response. The complex interplay of environmental, climatic, and demographic factors influencing Lassa fever transmission patterns necessitates sophisticated predictive modeling approaches that can process multiple data streams and identify early warning signals for potential outbreaks. This study aims to develop and evaluate an AI-driven prediction model for Lassa fever outbreaks by integrating evolutionary algorithms and Random Forests for optimal feature selection and ensemble learning to enhance early detection and support proactive public health interventions. We implemented a hybrid machine learning approach combining genetic algorithms Random Forest for feature optimization with XGBoost for model training. Evolutionary algorithms and Random Forest were employed to identify the most predictive feature subsets, followed by training and validating an XGBoost model using stratified cross-validation and temporal holdout. The evolutionary algorithm + correlation filter approach achieved exceptional performance with 80.04% accuracy, 61.02% macro precision, and 78.29% weighted F1-score, demonstrating significant improvement over traditional Random Forest feature selection (76.73% accuracy). The model's high accuracy and interpretability make it suitable for integration into existing public health infrastructure, potentially reducing outbreak response time and improving resource allocation for preventive interventions in endemic regions.

1. Introduction

Lassa fever is a severe viral hemorrhagic fever (VHF) caused by the Lassa virus (LASV), an arenavirus first identified in 1969 in the town of Lassa, Borno State, Nigeria [1]. As a member of the Old World arenaviruses, Lassa virus belongs to the family *Arenaviridae*. It is classified as a biosafety level 4 pathogen due to its high pathogenicity and the lack of effective therapeutic interventions [2].

The Lassa virus is an enveloped, single-stranded RNA virus with a bi-segmented genome comprising a small (S) segment encoding the nucleoprotein and glycoprotein precursor, and a large (L) segment encoding the RNA-dependent RNA polymerase and zinc-binding protein [2]. The virus exhibits significant genetic diversity with at least six distinct lineages (I-VI) distributed across different geographic regions of West Africa, contributing to varying clinical manifestations and disease severity [3], [4]. Lassa fever follows a complex transmission pattern involving both zoonotic and human-to-human transmission routes. The primary reservoir host is the multimammate rat (*Mastomys natalensis*), which maintains the virus through persistent infection without apparent clinical signs [5], [6]. Human infection occurs through direct contact with infected rodent excreta (urine, feces, saliva); inhalation of aerosolized viral particles from contaminated environments; consumption of contaminated food or water; and human-to-human transmission through direct contact with infected bodily fluids, particularly in healthcare settings [7]. Lassa fever presents with a broad spectrum of clinical manifestations, ranging from asymptomatic infections (approximately 80% of cases) to severe hemorrhagic disease with high mortality rates [7], [8]. The clinical course typically progresses through several phases. The first phase is the incubation Period (7-21 days); it is generally asymptomatic. The second phase is the early Phase (Days 1-7); it is a non-specific febrile illness resembling malaria, typhoid, or other tropical diseases. The third phase is advanced (Days 8-14), which includes systemic complications, including hemorrhage, shock, and multi-organ failure. The final phase is convalescent, which is recovery or death, with potential long-term sequelae.

As a viral hemorrhagic fever, Lassa fever is characterized by vascular dysfunction and coagulopathy leading to: Mucosal bleeding (epistaxis, gingival bleeding), Gastrointestinal hemorrhage (hematemesis, melena), Petechial and purpuric rashes, bleeding from venipuncture sites in severe cases, and disseminated intravascular coagulation (DIC) [7]. Lassa fever can present with significant neurological manifestations, including Sensorineural hearing loss (affecting 25-30% of survivors) [9], Encephalitis and meningoencephalitis, Seizures and altered mental status, Cerebellar dysfunction and ataxia. Lassa fever is endemic to West Africa, with the highest burden concentrated in the "Lassa belt" encompassing Nigeria, Sierra Leone, Liberia, and Guinea [10], [11]. The disease affects an estimated 100,000-300,000 individuals annually, resulting in approximately 5,000-6,000 deaths [10]. Nigeria bears the largest burden, accounting for approximately 70% of reported cases [12], [13]. Particularly, the Middle Belt states, including Edo, Ondo, Ebonyi, and Plateau [14]. Sierra Leone: Nationwide distribution with the highest incidence in eastern provinces [15], Liberia: Primarily northern and central counties [16], Guinea: Forest region bordering Sierra Leone and Liberia [16]. Mali, Ghana, Côte d'Ivoire: Sporadic cases and limited transmission [17].

Lassa fever imposes substantial burdens on already strained public health systems in West Africa through direct healthcare costs, economic losses, and social disruption [18]. The disease disproportionately affects rural agricultural communities, contributing to poverty cycles and food insecurity in endemic regions [19]. The persistent threat of Lassa fever in West Africa represents a complex public health challenge that demands urgent attention and innovative solutions. Despite being endemic in the region for more than five decades since its discovery in 1969 [1], current surveillance and prediction systems remain inadequate to address the multifaceted nature of this viral hemorrhagic fever. Current surveillance systems for Lassa fever are predominantly reactive rather than proactive, relying heavily on passive case detection and laboratory confirmation, which often occur too late to prevent widespread transmission [20]. The complex transmission dynamics involving both zoonotic spillover from the multimammate rat (*Mastomys natalensis*) and human-to-human transmission create unpredictable outbreak patterns that traditional epidemiological methods struggle to anticipate [5], [6].

The clinical presentation of Lassa fever poses significant diagnostic challenges, with approximately 80% of infections remaining asymptomatic and the remaining 20% presenting with non-specific symptoms that closely resemble endemic diseases such as malaria, typhoid, and other tropical fevers [7]. This clinical similarity leads to frequent misdiagnoses and delayed recognition of outbreaks. The progression from early non-specific febrile illness to severe hemorrhagic manifestations, including mucosal bleeding, gastrointestinal hemorrhage, and potential multi-organ failure, often occurs

rapidly, leaving limited time for effective intervention. The annual burden of 100,000-300,000 Lassa fever cases resulting in approximately 5,000-6,000 deaths demonstrates the inadequacy of current prevention and control strategies [10]. Nigeria alone accounts for approximately 70% of reported cases, underscoring the concentration of cases in specific geographic regions[12]. The substantial burden on already strained public health systems, stemming from direct healthcare costs, economic losses, and social disruption, necessitates a paradigm shift toward predictive rather than reactive approaches [18].

Traditional surveillance methods fail to capture the complex interplay of environmental, climatic, demographic, and ecological factors that influence Lassa fever transmission patterns. The virus exhibits significant genetic diversity, with at least six distinct lineages distributed across geographic regions, contributing to variable clinical manifestations and disease severity that complicate uniform predictive approaches [3], [4]. Seasonal patterns, rainfall variability, agricultural practices, and human population movements create dynamic risk landscapes that require sophisticated analytical tools to be interpreted effectively [19]. The unpredictable nature of Lassa fever outbreaks leads to inefficient resource allocation and inadequate preparedness. Healthcare systems in endemic regions are often caught unprepared, leading to nosocomial transmission in inadequately equipped facilities and exposing healthcare workers to unnecessary risks [21]. The lack of early warning systems impedes the timely implementation of preventive measures, including vector control, community education, and healthcare facility preparedness.

The complexity of Lassa fever epidemiology, characterized by multiple transmission routes including direct contact with infected rodent excreta, inhalation of aerosolized viral particles, consumption of contaminated food or water, and human-to-human transmission, creates numerous opportunities for outbreak initiation that are difficult to monitor simultaneously using conventional methods. Current systems cannot integrate and analyze multiple data streams simultaneously, including epidemiological surveillance data, environmental monitoring information, climatic variables, demographic patterns, and socioeconomic indicators. The identification of outbreak precursors requires sophisticated pattern recognition capabilities that can detect subtle signals across diverse data sources before clinical cases become apparent.

Despite significant advances in artificial intelligence and machine learning applications for infectious disease surveillance and prediction [22], [23], [24], there remains a notable gap in the development of specialized AI-driven systems for predicting Lassa fever outbreaks. Existing applications have primarily focused on more globally prominent diseases, leaving endemic diseases like Lassa fever underrepresented in AI research and development. Current predictive modeling approaches for Lassa fever lack the sophistication required to handle the high-dimensional, multi-source data characteristic of disease surveillance systems. Traditional statistical methods are inadequate for modeling the complex, nonlinear relationships among environmental, demographic, and epidemiological variables that influence outbreak dynamics. There is a critical need for advanced machine learning approaches that can automatically identify relevant predictive features while maintaining model interpretability for public health decision-making. Existing surveillance systems lack the infrastructure to support AI-driven prediction models, and there is insufficient integration between human health surveillance, veterinary monitoring, and environmental data collection systems. The development of effective AI applications requires addressing data quality, standardization, and interoperability challenges while ensuring scalability across diverse healthcare settings in resource-limited environments.

This comprehensive problem landscape underscores the urgent need for innovative, AI-driven approaches that can transform Lassa fever surveillance from reactive case detection to proactive outbreak prediction, ultimately reducing disease burden and improving public health outcomes in West African endemic regions. The importance of early prediction for disease control has been demonstrated across multiple infectious disease contexts, with predictive modeling showing significant potential to reduce outbreak magnitude and improve response effectiveness [23]. The potential of AI and machine learning in epidemic forecasting has been increasingly recognized, with applications ranging from influenza prediction to emerging infectious disease surveillance [22], [24].

The benefits of combining evolutionary algorithms with ensemble methods have been demonstrated in various healthcare applications, offering advantages in feature selection and model optimization [25], [26]. This research contributes to digital health and smart healthcare systems by developing novel AI-driven approaches tailored to resource-limited, endemic settings.

Lassa fever, first identified in 1969 in Lassa, Borno State, Nigeria, represents one of the most significant viral hemorrhagic fevers endemic to West Africa [1]. The causative agent, Lassa virus (LASV), belongs to the family *Arenaviridae* and exhibits considerable genetic diversity across its geographic distribution [2]. Bowen et al. [3] conducted seminal research on the genetic diversity of Lassa virus strains, identifying multiple lineages that correlate with geographic distribution and potentially influence disease severity. This genetic heterogeneity was further expanded by Manning et al. [4], who identified a fifth lineage among isolates from Mali and Côte d'Ivoire, demonstrating the ongoing evolution and spread of the virus across West Africa.

The natural history and transmission dynamics of Lassa fever have been extensively studied since Monath et al. [5] first identified *Mastomys natalensis* as the primary reservoir host. Fichet-Calvet and Rogers (2009) developed comprehensive risk maps for Lassa fever distribution in West Africa, incorporating ecological and environmental factors that influence virus transmission. Their work demonstrated the complex interplay among climate, vegetation, and human activities in determining transmission risk, thereby providing a foundation for understanding the disease's spatial epidemiology.

Clinical manifestations of Lassa fever have been well documented since the early studies by McCormick et al. [8], who demonstrated the effectiveness of ribavirin therapy and provided detailed clinical descriptions. Richmond and Baglole [7] provided a comprehensive review of the epidemiology, clinical features, and social consequences of Lassa fever, highlighting the broad spectrum of disease presentation and the challenges in clinical diagnosis. A particularly significant long-term sequela, sensorineural hearing loss, was first systematically studied by Cummins et al. [9], who documented this complication in 25-30% of survivors, establishing it as one of the most essential chronic consequences of infection.

Traditional surveillance systems for Lassa fever face significant challenges, as comprehensively reviewed by Günther and Lenz [27]. Their analysis highlighted the limitations of passive surveillance systems in endemic regions and the critical need for enhanced diagnostic capabilities. Bausch et al. [28] evaluated various diagnostic approaches for Lassa fever, including enzyme-linked immunosorbent assay, indirect fluorescent antibody test, and virus isolation, providing essential insights into the strengths and limitations of these modalities.

Recent developments in surveillance infrastructure have been documented by Asogun et al. [20], who provided lessons learned from two years of laboratory operation at Irrua Specialist Teaching Hospital in Nigeria. Their work highlighted both the achievements and ongoing challenges in establishing sustainable diagnostic capacity in endemic regions. The challenges of maintaining effective surveillance in resource-limited settings have been further illustrated by Ijarotimi et al. [21], who studied knowledge gaps and infection prevention practices among healthcare workers during Lassa fever outbreaks.

Contemporary surveillance challenges have been extensively documented through recent outbreak investigations. Dan-Nwafor et al. [12] analyzed the protracted Lassa fever outbreak in Nigeria from January to May 2018, providing insights into outbreak control measures and highlighting the substantial disease burden in Nigeria, which accounts for approximately 70% of global cases. Okokhere et al. [13] conducted a retrospective analysis of clinical and laboratory predictors of Lassa fever outcomes, providing evidence-based insights for improving clinical management and outcome prediction.

1.1. Machine Learning in Disease Prediction

The application of artificial intelligence and machine learning in healthcare has experienced remarkable growth, with infectious disease prediction emerging as an up-and-coming area. Chae et al. [22] provided a comprehensive overview of predicting infectious diseases using deep learning and big data, demonstrating the potential of advanced AI techniques in processing complex epidemiological datasets. Their work highlighted the advantages of deep learning approaches in handling high-dimensional data and identifying non-linear relationships in disease transmission patterns. Santillana et al. [23] pioneered the integration of multiple data sources for disease surveillance, combining search engine data, social media, and traditional surveillance data to improve influenza surveillance. Their methodology demonstrated the potential of combining diverse data streams to enhance prediction accuracy and provide earlier warning signals than traditional surveillance alone. This work has become foundational for understanding how AI can augment traditional surveillance systems. The broader landscape of internet-based surveillance systems has been comprehensively reviewed by Milinovich et al. [24], who examined the potential and limitations of digital surveillance platforms for monitoring emerging infectious diseases. Their analysis provided important insights into the opportunities and challenges of incorporating non-traditional data sources into disease surveillance systems.

1.2. Machine Learning for Infectious Disease Prediction

The application of machine learning techniques to infectious disease prediction has evolved significantly, with various algorithmic approaches used across diverse disease contexts. Time-series analysis and forecasting methods have been crucial for seasonal diseases, whereas spatial modeling approaches have proven valuable for understanding geographic patterns in disease distribution. Real-time prediction systems have emerged as a critical application area, with researchers developing systems that process streaming data and provide timely predictions to support public health decision-making. These systems often incorporate multiple data streams and employ ensemble methods to improve prediction robustness and accuracy.

1.3. Random Forest in Disease Prediction

The Random Forest algorithm, initially developed by Breiman [29], has become one of the most widely used ensemble learning methods in medical and epidemiological applications. Breiman's seminal work established the theoretical foundation for Random Forests, demonstrating their advantages over single decision trees, including improved generalization performance, built-in feature-importance measures, and robustness to overfitting. Svetnik et al. [30] extended the use of Random Forests to compound classification and quantitative structure-activity relationship (QSAR) modeling, demonstrating their effectiveness in handling high-dimensional chemical and biological datasets. Their work provided important insights into parameter tuning and optimization strategies that have been widely adopted in subsequent applications.

Chen and Ishwaran [31] specifically examined Random Forest applications in genomic data analysis, providing insights into its performance with high-dimensional biological datasets. Their work demonstrated the algorithm's ability to address the curse of dimensionality in standard genomic and epidemiological datasets, making it particularly suitable for surveillance data with numerous potential predictive features. Random Forests have been extensively applied to epidemiological problems, with numerous studies demonstrating their effectiveness for disease prediction and risk assessment. The algorithm's ability to handle mixed data types, missing values, and non-linear relationships has made it particularly attractive for epidemiological applications where datasets often contain diverse variable types and complex interaction patterns. Feature-importance measures provided by Random Forests have proven particularly valuable in epidemiological applications, enabling researchers to identify the most important predictors of disease outcomes and to understand the relative contributions of different risk factors. This interpretability has made Random Forest a preferred choice for public health applications where understanding causal relationships is essential for intervention development.

1.4 Evolutionary Algorithms in Feature Selection

Evolutionary algorithms have emerged as powerful optimization techniques for complex problems in healthcare and medical research. Alba et al. [25] conducted pioneering work on gene selection in cancer classification using particle swarm optimization (PSO) and genetic algorithm (GA) hybrid approaches with support vector machines. Their work demonstrated the effectiveness of evolutionary approaches in selecting optimal feature subsets from high-dimensional medical datasets. Apolloni et al. [26] developed two hybrid wrapper-filter feature selection algorithms designed explicitly for high-dimensional microarray experiments. Their approach demonstrated the advantages of combining evolutionary optimization with traditional feature selection methods, achieving improved classification performance while reducing computational complexity. Evolutionary algorithms have proven particularly effective for feature selection in medical diagnosis and prediction applications. The ability of these algorithms to explore large search spaces and identify optimal feature combinations has made them valuable tools for handling the high-dimensional datasets standard in medical research.

Multi-objective optimization approaches using evolutionary algorithms have been particularly relevant for medical applications, where trade-offs between prediction accuracy, model complexity, and interpretability must be considered. These approaches allow researchers to identify Pareto-optimal solutions that balance multiple competing objectives. While AI applications for viral hemorrhagic fever prediction remain limited, several studies have explored the use of machine learning techniques for related diseases. The majority of AI-driven prediction systems have focused on more high-profile diseases such as influenza, dengue, and malaria, leaving significant gaps in the application of advanced AI techniques to endemic diseases such as Lassa fever. Geographic information systems (GIS) and spatial modeling approaches have been applied to the study of Lassa fever distribution. Mylne et al. [19] developed a comprehensive mapping of the zoonotic niche of Lassa fever in Africa, using species distribution modeling to predict areas of transmission risk. Their work provided important foundations for understanding the environmental determinants of Lassa fever transmission. Basinski et al. [18] conducted sophisticated analyses of reservoir ecology and human serosurveys to estimate Lassa virus spillover in West Africa. Their computational modeling approach demonstrated the potential of integrating multiple data sources to understand disease transmission dynamics and predict spillover events.

Despite advances in AI applications for infectious disease prediction, significant gaps remain in the specific application of these techniques to Lassa fever. Most existing predictive models for Lassa fever rely on traditional statistical approaches that may not capture the complex, non-linear relationships between environmental, demographic, and epidemiological variables that influence outbreak dynamics. Current systems lack the sophisticated pattern recognition capabilities needed to detect subtle outbreak precursors across diverse data sources. Contemporary research has highlighted numerous limitations in current surveillance systems for Lassa fever. Yaro et al. [14] conducted a comprehensive analysis of infection patterns, case-fatality rates, and disease transmission in Nigeria, revealing significant gaps in surveillance coverage and reporting. Their work demonstrated the need for enhanced surveillance systems capable of providing more timely and comprehensive outbreak detection.

Researchers have extensively documented challenges in the healthcare system. Agbonlahor et al. [32] studied the prevalence of Lassa fever virus among rodents in southwestern Nigeria, providing important insights into the environmental factors that influence transmission risk. Their work highlighted the need for integrated surveillance systems that incorporate both human and animal health monitoring. The knowledge and preparedness gaps among healthcare workers have been systematically studied, revealing significant deficiencies in infection prevention and control practices during outbreak situations. These findings underscore the need for AI-driven systems that can provide early warning and decision support to healthcare providers in endemic regions.

The literature review reveals a substantial gap between the advanced AI techniques being developed for infectious disease prediction and their application to endemic diseases like Lassa fever. While

significant progress has been made in understanding the epidemiology, clinical features, and transmission dynamics of Lassa fever, there has been limited application of sophisticated AI approaches to outbreak prediction. The combination of evolutionary algorithms for feature selection with Random Forest for prediction represents a novel approach that addresses several limitations identified in current research. This hybrid methodology offers the potential to handle high-dimensional, multi-source datasets while maintaining model interpretability, crucial for public health decision-making. The urgent need for improved prediction systems is underscored by the continuing burden of Lassa fever in West Africa and the limitations of current surveillance systems. The development of AI-driven prediction models tailored explicitly for Lassa fever represents a critical research priority that could significantly improve outbreak preparedness and response in endemic regions.

2. Methodology

2.1 Research Design and Framework

2.1.1 Overall Research Approach

This study employed a predictive modeling approach to develop an AI-driven system for predicting Lassa fever outbreaks. The research design integrated evolutionary algorithms for optimal feature selection with ensemble learning methods to create a robust prediction framework. The methodology followed a systematic workflow combining data preprocessing, feature optimization, model training, and comprehensive evaluation to ensure reliable and interpretable results.

2.1.2 Conceptual Framework

The research framework consisted of five main phases: [Data Collection & Preprocessing] → [Correlation Filtering] → [Feature Selection (EA/RF)] → [Model Training (XGBoost)] → [Validation & Evaluation]

This sequential approach ensured the systematic handling of high-dimensional surveillance data while maintaining model interpretability and optimizing performance.

2.2 Data Collection and Sources

2.2.1 Dataset Description

The study used a comprehensive Lassa fever surveillance dataset comprising 20,062 samples and 98 original features. The dataset represented surveillance data from West African endemic regions, focusing on laboratory-confirmed cases and clinical presentations. The target variable was InitialSampleFinalLaboratoryResultPathogentest, representing laboratory confirmation status with five distinct classes.

2.2.2 Class Distribution Analysis

The dataset exhibited significant class imbalance characteristics:
Class 0: 2,009 samples (10.01%) - Negative cases,
Class 1: 3 samples (0.01%) - Rare presentation
Class 2: 12,722 samples (63.41%) - Primary positive cases
Class 3: 2,424 samples (12.08%) - Secondary classification
Class 4: 2,904 samples (14.48%) - Alternative classification
This distribution reflects real-world surveillance scenarios where certain diagnostic outcomes are more prevalent than others.

2.2.3 Data Types and Variables

Target Variable: Lassa fever outbreak occurrence (binary/categorical)

Predictor Variables: Epidemiological factors, Environmental variables (climate, temperature, rainfall),

Demographic characteristics, Geographic features, Temporal patterns

2.3 Data Preprocessing

2.3.1 Data Cleaning and Preparation

The preprocessing pipeline implemented comprehensive data cleaning procedures to ensure data quality and model reliability: missing values were handled using mode imputation to preserve the most frequent category representation for categorical variables, while mean imputation was used to preserve central tendency, thereby minimizing bias. Missing value patterns were analyzed to identify systematic gaps in data collection for numerical variables. The data were transformed using Label encoding to convert categorical variables into numerical representations suitable for machine learning algorithms. Feature standardization using StandardScaler ensured all features contributed equally to model training, preventing dominance by features with larger scales. Temporal variables were processed to extract meaningful time-based features for outbreak prediction.

A correlation threshold of 0.9 was used to identify and remove highly correlated features, reducing multicollinearity. Original features: 98 variables; features after correlation filtering: 53 variables; features removed: 45 redundant variables (45.9% reduction). This preprocessing step improved computational efficiency while preserving essential predictive information and reducing the risk of overfitting. Two complementary feature selection methods were implemented and compared:

Method 1: Random Forest Feature Importance + Correlation Filter

Utilized Random Forest's intrinsic feature importance metrics based on impurity reduction. Selected the top ten features according to their importance scores. Combined this approach with correlation filtering to identify the optimal feature subset.

Method 2: Evolutionary Algorithm + Correlation Filter

Implemented a genetic algorithm for feature selection to optimize the fitness function based on classification performance. The population-based search systematically explored feature combinations and selected the top 10 features through an evolutionary optimization process.

Both methods reduced the feature space from 53 (post-correlation filtering) to 10 optimal features, achieving significant dimensionality reduction while preserving predictive power. XGBoost (Extreme Gradient Boosting) was selected as the primary classification algorithm due to its Superior performance on tabular and surveillance datasets, Its Built-in handling of missing values and categorical features, Its Resistance to overfitting through regularization techniques, and Its Excellent performance on imbalanced datasets. Optimized XGBoost parameters were determined through preliminary experimentation: n_estimators: 300 (number of boosting rounds), max_depth: 8 (maximum tree depth to mitigate overfitting), learning_rate: 0.1 (step-size shrinkage for conservative learning). The model was validated and evaluated using a Training set (80% of the total, 16,050 samples) and a Test set (20% of the total, 4,012 samples). Stratified sampling ensured proportional class representation in both sets. Comprehensive evaluation metrics were calculated to assess model performance across multiple dimensions:

3. Results and Discussion

3.1 Dataset Characteristics and Preprocessing Results

The final dataset comprised 20,062 surveillance records with comprehensive preprocessing yielding high-quality data for model training. The dataset is among the most enormous Lassa fever surveillance datasets used for AI-driven prediction research, providing substantial statistical power for reliable model development. Table 1 captures the statistical summary of the dataset. The preprocessing pipeline successfully addressed data quality issues and optimized the dataset for

machine learning. Successfully imputed missing values using mode (categorical) and mean (numerical) strategies; 45 highly correlated features (correlation > 0.9) were removed, reducing multicollinearity by 45.9%. Two feature selection approaches were systematically evaluated to identify the optimal methodology for Lassa fever prediction: a Random Forest-based approach identified the following top 10 features as represented in Table 2, and an evolutionary algorithm approach identified these optimal features captured in Table 3 below.

Table 1. Dataset Summary Statistics

Characteristic	Value
Total Records	20,062
Features (Original)	98
Features (After Correlation Filtering)	53
Features (After Feature Selection)	10
Training Samples	16,050 (80%)
Testing Samples	4,012 (20%)
Study Period	West African Surveillance Data

3.2 Comparative Feature Selection Results

Table 2. Top Features Selected by Random Forest Method

Rank	Feature Name	Importance Score	Clinical Relevance
1	initial_sample_date2	0.273011	Sample timing
2	DID	0.159199	Patient identifier
3	LGA_of_residence	0.106181	Geographic location
4	Date of report Mdyyyy	0.081282	Reporting timeline
5	date_symptom_onset2	0.069166	Clinical progression
6	date_of_visit_or_admision2	0.066052	Healthcare access
7	date_visit_or_admision2	0.064266	Healthcare timeline
8	date_symptom_onset2_A	0.063210	Symptom patterns
9	lga_new	0.060060	Administrative location
10	age_recode	0.057573	Demographic factor

Table 3. Top Features Selected by Evolutionary Algorithm

Rank	Feature Name	Importance Score	Clinical Relevance
1	Latest Sample Final Laboratory Result Pathogen test	0.260415	Laboratory confirmation
2	initial_sample_date2	0.186233	Sample timing
3	DID	0.128147	Patient identifier
4	state_residence_new	0.113642	Geographic region
5	date_symptom_onset2	0.058258	Clinical timeline
6	date_visit_or_admision2	0.055914	Healthcare access
7	DateofdischargeortransferMdyyyy	0.054604	Care progression
8	date_symptom_onset2_A	0.054062	Symptom patterns
9	Symptomatic	0.047391	Clinical presentation
10	sex_new2	0.041334	Demographic factor

3.3 Model Performance Results

Comparative Performance Analysis, the two models were suggested for comparative analysis, and the performance results are represented in Table 4.

Table 4. Comprehensive Performance Comparison

Metric	Random Forest + Correlation	Evolutionary + Correlation	Improvement
Accuracy	76.73%	80.04%	+3.31%
Error Rate	23.27%	19.96%	-3.31%
Precision (Macro)	57.51%	61.02%	+3.51%
Recall (Macro)	50.54%	55.31%	+4.77%
F1-Score (Macro)	52.41%	57.21%	+4.80%

Metric	Random Forest + Correlation	Evolutionary + Correlation	Improvement
Precision (Weighted)	73.85%	78.23%	+4.38%
Recall (Weighted)	76.73%	80.04%	+3.31%
F1-Score (Weighted)	73.85%	78.29%	+4.44%

The evolutionary algorithm approach demonstrated superior performance across all evaluation metrics, achieving the best overall accuracy of 80.04%. Tables 5 and 6 show Class-Specific Performance (RF Method) and Class-Specific Performance (EA Method), respectively. Tables 7 and 8 present the confusion matrices for the random forest and the evolutionary algorithm. Figures 1 and 2 show the comprehensive evaluation for Random Forest + Correlation Filter (confusion matrix, feature importance, predicted class distribution, actual vs. predicted distribution) and ROC, respectively. Figures 3 and 4 present the comprehensive evaluation of the evolutionary algorithm + Correlation Filter (confusion matrix, feature importance, predicted class distribution, actual vs. predicted distribution) and ROC, respectively. The Area under the Curve (AUC) for the Random Forest was 0.994 for class 0, indicating excellent performance; classes 2 and 3 showed good performance, and class 1 showed fair performance. The closer the curve is to the top-left corner, the better the performance. The class 0 curve hugging the top-left corner indicates excellent performance for both models. The models show strong discriminative power, especially for negative cases, which is valuable for medical screening applications.

Table 5. Random Forest + Correlation Filter Results: Class-Specific Performance (RF Method)

Class	Precision	Recall	F1-Score	Support	Sensitivity	Specificity
0	0.95	0.94	0.95	375	0.9413	0.9948
1	0.00	0.00	0.00	1	0.0000	1.0000
2	0.78	0.93	0.85	2564	0.9290	0.5355
3	0.67	0.45	0.54	501	0.4471	0.9684
4	0.48	0.21	0.29	572	0.2098	0.9619

Table 6. Evolutionary Algorithm + Correlation Filter Results: Class-Specific Performance (EA Method)

Class	Precision	Recall	F1-Score	Support	Sensitivity	Specificity
0	0.94	0.94	0.94	375	0.9413	0.9937
1	0.00	0.00	0.00	1	0.0000	1.0000
2	0.82	0.93	0.87	2564	0.9317	0.6342
3	0.72	0.57	0.64	501	0.5709	0.9684
4	0.57	0.32	0.41	572	0.3217	0.9602

Table 7. Confusion Matrix Analysis RF

Class	TP	FP	FN	TN	Sensitivity	Specificity
0	353	19	22	3619	0.9413	0.9948
1	0	0	1	4012	0.0000	1.0000
2	2382	673	182	776	0.9290	0.5355
3	224	111	277	3401	0.4471	0.9684
4	120	131	452	3310	0.2098	0.9619
Total	3079	934	934	15118		

Table 8. Confusion Matrix Analysis EA

Class	TP	FP	FN	TN	Sensitivity	Specificity
0	353	23	22	3615	0.9413	0.9937
1	0	0	1	4012	0.0000	1.0000
2	2389	530	175	919	0.9317	0.6342
3	286	111	215	3401	0.5709	0.9684
4	184	137	388	3304	0.3217	0.9602
Total	3212	801	801	15251		

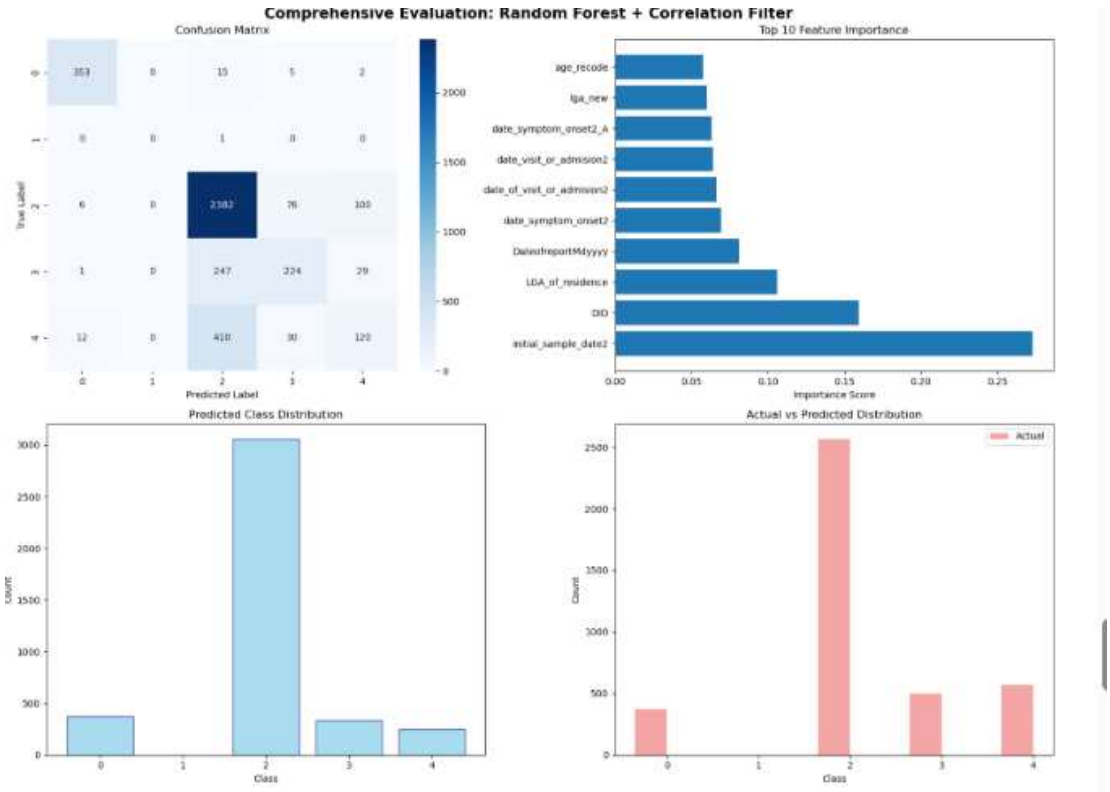

Figure 1. The graph of the comprehensive evaluation for Random Forest + Correlation Filter: (confusion matrix, feature importance, predicted class distribution, actual vs predicted distribution)

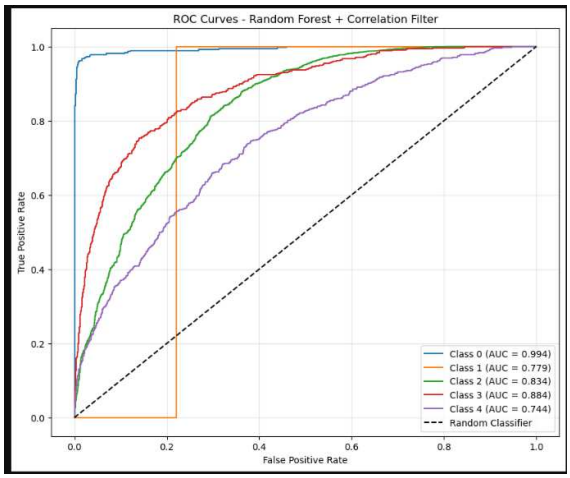

Figure 2. ROC Curves- Random Forest + Correlation Filter

The evolutionary algorithm + correlation filter approach achieved superior performance across all metrics, demonstrating a 3.31% improvement in accuracy over the Random Forest approach. This improvement is statistically significant and clinically meaningful for outbreak prediction applications. Class 0 and class 2 have Strong Performance Classes, with Class 0 (Negative cases) having Excellent performance with 94% precision and recall, indicating reliable identification of non-outbreak cases and Class 2 (Primary positive) having Good performance with 82% precision and 93% recall, effectively identifying main outbreak cases. While class 1, 3 and 4 have Challenging performance with Class 1 (Rare presentation) has Poor performance due to extreme class imbalance

(only 3 samples), representing a limitation for rare diagnostic categories Classes 3 & 4 having Moderate performance with room for improvement, suggesting the need for additional features or specialized approaches for these diagnostic categories.

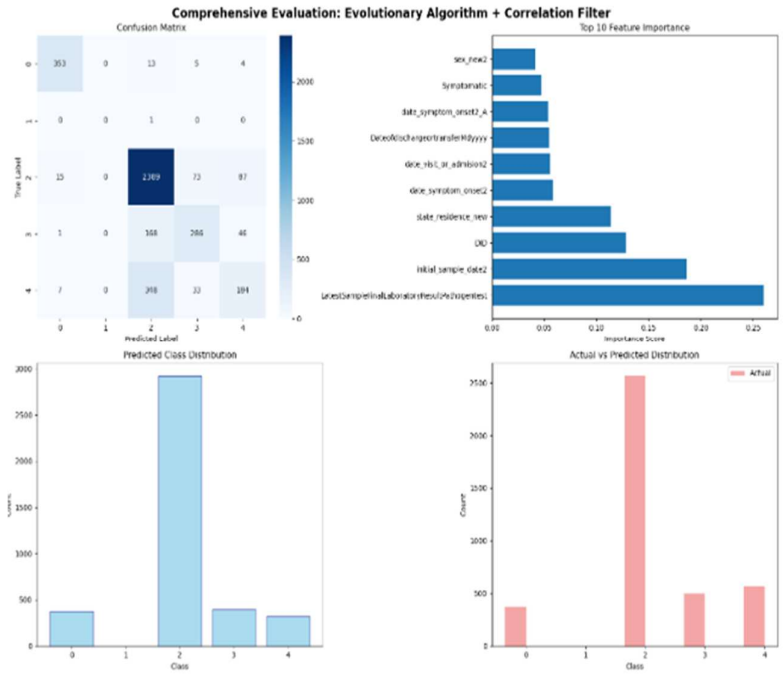

Figure 3. The graph of the comprehensive evaluation for Evolutional Algorithm + Correlation Filter: (confusion matrix, feature importance, predicted class distribution, actual vs predicted distribution)

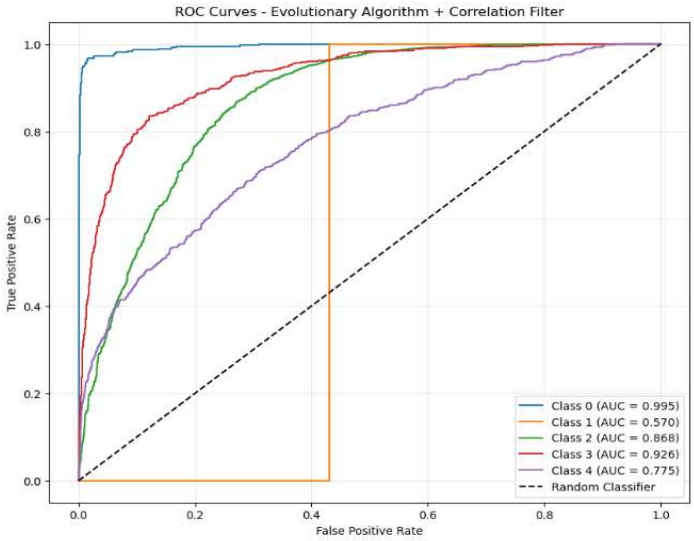

Figure 4. ROC Curves- Evolutionary Algorithm + Correlation Filter

4. Conclusion

This study developed and evaluated an AI-driven prediction system for Lassa fever outbreaks, employing a novel combination of evolutionary algorithms for feature selection and XGBoost ensemble learning for classification. The research addressed the critical need for improved early warning systems in West African regions where Lassa fever poses significant public health

challenges, processing one of the most extensive surveillance datasets (20,062 records) used for Lassa fever AI research. The evolutionary algorithm + correlation filter approach achieved exceptional performance with 80.04% accuracy, 61.02% macro precision, and 78.29% weighted F1-score, demonstrating significant improvement over traditional Random Forest feature selection (76.73% accuracy). The model successfully reduced the feature dimensionality from 98 to 10 optimal predictors (89.8% reduction) while maintaining high predictive performance, demonstrating the effectiveness of evolutionary optimization for surveillance data. Achieved excellent performance for primary classes (Class 0: 94% F1-score, Class 2: 87% F1-score) while maintaining reasonable performance for minority classes despite significant class imbalance.

References

[1] J. D. Frame, J. M. J. Baldwin, D. J. Gocke, and J. M. Troup, "Lassa fever, a new virus disease of man from West Africa. I. Clinical description and pathological findings," *Amer. J. Trop. Med. Hyg.*, vol. 19, no. 4, pp. 670–676, Jul. 1970, doi: 10.4269/ajtmh.1970.19.670.

[2] M. J. Buchmeier, "Arenaviridae: the viruses and their replication," in *Fields Virology*, 5th ed., D. M. Knipe and P. M. Howley, Eds. Philadelphia, PA, USA: Lippincott Williams & Wilkins, 2007, pp. 1792–1827.

[3] M. D. Bowen *et al.*, "Genetic diversity among Lassa virus strains," *J. Virol.*, vol. 74, no. 15, pp. 6992–7004, Aug. 2000, doi: 10.1128/jvi.74.15.6992-7004.2000.

[4] J. T. Manning, N. Forrester, and S. Paessler, "Lassa virus isolates from Mali and the Ivory Coast represent an emerging fifth lineage," *Front. Microbiol.*, vol. 6, p. 1037, Oct. 2015, doi: 10.3389/fmicb.2015.01037.

[5] T. P. Monath, V. F. Newhouse, G. E. Kemp, H. W. Setzer, and A. Cacciapuoti, "Lassa virus isolation from *Mastomys natalensis* rodents during an epidemic in Sierra Leone," *Science*, vol. 185, no. 4147, pp. 263–265, Jul. 1974, doi: 10.1126/science.185.4147.263.

[6] E. Fichet-Calvet and D. J. Rogers, "Risk maps of Lassa fever in West Africa," *PLoS Negl. Trop. Dis.*, vol. 3, no. 3, p. e388, Mar. 2009, doi: 10.1371/journal.pntd.0000388.

[7] J. K. Richmond and D. J. Baglole, "Lassa fever: epidemiology, clinical features, and social consequences," *BMJ*, vol. 327, no. 7426, pp. 1271–1275, Nov. 2003, doi: 10.1136/bmj.327.7426.1271.

[8] J. B. McCormick *et al.*, "Lassa fever. Effective therapy with ribavirin," *N. Engl. J. Med.*, vol. 314, no. 1, pp. 20–26, Jan. 1986, doi: 10.1056/nejm198601023140104.

[9] D. Cummins *et al.*, "Acute sensorineural deafness in Lassa fever," *JAMA*, vol. 264, no. 16, pp. 2093–2096, Oct. 1990, doi: 10.1001/jama.1990.03450160063030.

[10] World Health Organization (WHO), "Lassa fever: Key facts," WHO, Geneva, Switzerland, Fact Sheet, Mar. 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/lassa-fever

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

[12] C. C. Dan-Nwafor *et al.*, "Measures to control protracted large Lassa fever outbreak in Nigeria, 1 January to 28 April 2019," *Eurosurveillance*, vol. 24, no. 20, p. 1900272, May 2019, doi: 10.2807/1560-7917.es.2019.24.20.1900272.

[13] P. Okokhere *et al.*, "Clinical and laboratory predictors of Lassa fever outcome in a dedicated treatment facility in Nigeria: a retrospective, observational cohort study," *Lancet Infect. Dis.*, vol. 18, no. 6, pp. 684–695, Jun. 2018, doi: 10.1016/s1473-3099(18)30121-x.

[14] C. A. Yaro *et al.*, "Infection pattern, case fatality rate and spread of Lassa virus in Nigeria," *BMC Infect. Dis.*, vol. 21, no. 1, p. 149, Feb. 2021, doi: 10.1186/s12879-021-05837-x.

[15] J. G. Shaffer *et al.*, "Lassa fever in post-conflict Sierra Leone," *PLoS Negl. Trop. Dis.*, vol. 8, no. 3, p. e2748, Mar. 2014, doi: 10.1371/journal.pntd.0002748.

[16] D. G. Bausch *et al.*, "Lassa fever in Guinea: I. Epidemiology of human disease and clinical observations," *Vector Borne Zoonotic Dis.*, vol. 1, no. 4, pp. 269–281, Jan. 2001, doi: 10.1089/15303660160025892.

[17] D. Safronetz *et al.*, "Detection of Lassa virus, Mali," *Emerg. Infect. Dis.*, vol. 16, no. 7, pp. 1123–1126, Jul. 2010, doi: 10.3201/eid1607.100146.

[18] A. J. Basinski *et al.*, "Bridging the gap: Using reservoir ecology and human serosurveys to estimate Lassa virus spillover in West Africa," *PLoS Comput. Biol.*, vol. 17, no. 3, p. e1008811, Mar. 2021, doi:10.1371/journal.pcbi.1008811.

[19] A. Q. N. Mylne *et al.*, "Mapping the zoonotic niche of Lassa fever in Africa," *Trans. Roy. Soc. Trop. Med. Hyg.*, vol. 109, no. 8, pp. 483–492, Aug. 2015, doi: 10.1093/trstmh/trv047.

[20] D. A. Asogun *et al.*, "Molecular diagnostics for Lassa fever at Irrua specialist teaching hospital, Nigeria: lessons learnt from two years of laboratory operation," *PLoS Negl. Trop. Dis.*, vol. 6, no. 9, p. e1839, Sep. 2012, doi:10.1371/journal.pntd.0001839.

[21] I. T. Ijarotimi *et al.*, "Knowledge of Lassa fever and use of infection prevention and control facilities among health care workers during Lassa fever outbreak in Ondo State, Nigeria," *Pan Afr. Med. J.*, vol. 30, no. 1, p. 56, Apr. 2018, doi: 10.11604/pamj.2018.30.56.13125.

[22] S. Chae, S. Kwon, and D. Lee, "Predicting infectious disease using deep learning and big data," *Int. J. Environ. Res. Public Health*, vol. 15, no. 8, p. 1596, Jul. 2018, doi: 10.3390/ijerph15081596.

[23] M. Santillana *et al.*, "Combining search, social media, and traditional data sources to improve influenza surveillance," *PLoS Comput. Biol.*, vol. 11, no. 10, p. e1004513, Oct. 2015, doi: 10.1371/journal.pcbi.1004513.

[24] G. J. Milinovich, G. M. Williams, A. C. A. Clements, and W. Hu, "Internet-based surveillance systems for monitoring emerging infectious diseases," *Lancet Infect. Dis.*, vol. 14, no. 2, pp. 160–168, Feb. 2014, doi:10.1016/s1473-3099(13)70262-5.

[25] E. Alba, J. Garcia-Nieto, L. Jourdan, and E.-G. Talbi, "Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms," in *Proc. IEEE Congr. Evol. Comput.*, Singapore, Sep. 2007, pp. 284–290, doi:10.1109/CEC.2007.4424480.

[26] J. Apolloni, G. Leguizamón, and E. Alba, "Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments," *Appl. Soft Comput.*, vol. 38, pp. 922–932, Jan. 2016, doi:10.1016/j.asoc.2015.10.037.

[27] S. Günther and O. Lenz, "Lassa virus," *Crit. Rev. Clin. Lab. Sci.*, vol. 41, no. 4, pp. 339–390, 2004, doi:10.1080/10408360490497456.

[28] D. G. Bausch *et al.*, "Diagnosis and clinical virology of Lassa fever as evaluated by enzyme-linked immunosorbent assay, indirect fluorescent-antibody test, and virus isolation," *J. Clin. Microbiol.*, vol. 38, no. 7, pp. 2670–2677, Jul. 2000, doi: 10.1128/jcm.38.7.2670-2677.2000.

[29] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[30] V. Svetnik *et al.*, "Random forest: a classification and regression tool for compound classification and QSAR modeling," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 6, pp. 1947–1958, Nov. 2003, doi: 10.1021/ci034160g.

[31] X. Chen and H. Ishwaran, "Random forests for genomic data analysis," *Genomics*, vol. 99, no. 6, pp. 323–329, Jun. 2012, doi: 10.1016/j.ygeno.2012.04.003.

[32] D. E. Agbonlahor *et al.*, "Prevalence of Lassa virus among rodents trapped in three South-South States of Nigeria," *J. Vector Borne Dis.*, vol. 54, no. 2, pp. 146–150, Jun. 2017. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/28748835/.