

AI-Enabled Early Prediction of Lassa Fever: A Clinical Decision Support System for Endemic Settings

Henry Ohiani Ohize

`ohizeho@custech.edu.ng`

Confluence University of Science and Technology, Osara

Michael Chukwuemeka Etuonuma

Federal University of Technology Minna

Okoeguale Joseph

Institute of Viral and Emergent Pathogens Control and Research

Emmanuel Tanko Umaru

Federal University of Technology Minna

Adeiza James Onumanyi

Next Generation Enterprises and Institutions, Council for Scientific and Industrial Research (CSIR)

James Garba Ambafi

Federal University of Technology Minna

Taliha Abiodun Folorunso

Federal University of Technology Minna

Chibueze Valentine Ikpo

Veritas University

James Agajo

Federal University of Technology Minna

Jane Ekwujo Ohize

East Kent Hospitals University NHS Foundation Trust

Eifediyi Reuben

Institute of Viral and Emergent Pathogens Control and Research

Okogbenin Sylvanus

Institute of Viral and Emergent Pathogens Control and Research

Akpede George

Institute of Viral and Emergent Pathogens Control and Research

Asogun Danny

Institute of Viral and Emergent Pathogens Control and Research

Okokhere Peter

Institute of Viral and Emergent Pathogens Control and Research

Ehigiele Henry Edeko

Institute of Viral and Emergent Pathogens Control and Research

Stephen Ogirima Ohize

International Federation of Red Cross and Red Crescent Societies

Research Article

Keywords: Clinical Decision Support System, Lassa Fever, Machine Learning, Triage, XGBoost, Public Health Informatics, Artificial Intelligence

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11163962/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

AI-Enabled Early Prediction of Lassa Fever: A Clinical Decision Support System for Endemic Settings

Henry Ohiani Ohize^{1,2*}, Michael Chukwuemeka Etuonuma²,
Okoeguale Joseph^{3,4}, Emmanuel Tanko Umaru⁵,
Adeiza James Onumanyi⁶, James Garba Ambafi²,
Taliha Abiodun Folorunso⁷, Chibueze Valentine Ikpo⁸,
James Agajo⁹, Jane Ekwujo Ohize¹⁰, Eifediyi Reuben^{3,4},
Okogbenin Sylvanus^{3,4}, Akpede George³, Asogun Danny³,
Okokhere Peter³, Ehigiele Henry Edeko³, Stephen Ogirima Ohize¹¹

^{1*}Department of Electrical and Electronics Engineering, Confluence University of Science and Technology, Osara, PMB 1040, Kogi, Nigeria .

²Department of Electrical and Electronics Engineering, Federal University of Technology, Minna, PMB 65, Niger State, Nigeria .

³Institute of Viral and Emergent Pathogens Control and Research, Irrua, Edo State, Nigeria .

⁴Department of Obstetrics and Gynaecology, Ambrose Alli University, Ekpoma, Edo State, Nigeria .

⁵Department of Urban and Regional Planning, Federal University of Technology, Minna, PMB 65, Niger, Nigeria .

⁶Next Generation Enterprises and Institutions, Council for Scientific and Industrial Research (CSIR), Pretoria, 0001, South Africa .

⁷Department of Mechatronics Engineering, Federal University of Technology, Minna, PMB 65, Niger, Nigeria .

⁸Department of Software Engineering, Veritas University, Abuja, Nigeria .

⁹Department of Computer Engineering, Federal University of Technology, Minna, PMB 65, Niger State, Nigeria .

¹⁰East Kent Hospital NHS Foundation Trust, Kent, United Kingdom .

¹¹Department of Health and Care, International Federation of the Red Cross and Red Crescent Societies, Yaoundé, Cameroon .

*Corresponding author(s). E-mail(s): ohizeho@custech.edu.ng,
henryohize@futminna.edu.ng;
Contributing authors: michaeletuonuma01@gmail.com;
okoegualejoseph85@gmail.com; eumaru@futminna.edu.ng;
aonumanyi@csir.co.za; ambafi@futminna.edu.ng;
funso.taliha@futminna.edu.ng; ikpoc@veritas.edu.ng;
james.agajo@futminna.edu.ng; janeohize@gmail.com;
gloriagbons1976@gmail.com; okogbenins@yahoo.com;
georgeakpede@yahoo.co.uk; asogun2001@yahoo.com;
okokherep@gmail.com; henryedeko40@gmail.com;
Stephen.ohize@ifrc.org;

Abstract

Lassa fever remains a major public health challenge in West Africa, where early clinical symptoms frequently overlap with other febrile illnesses and laboratory confirmation may be delayed. This study developed and evaluated an AI-enabled Clinical Decision Support System (CDSS) to support early triage of suspected Lassa fever cases using a retrospective clinical surveillance dataset from Irrua Specialist Teaching Hospital, Nigeria, spanning 2015 to 2025, one of the largest longitudinal clinical surveillance datasets for Lassa fever. The final cohort included **31,277** patient encounters, with **3,359** laboratory-confirmed cases (**10.7%** prevalence). To ensure real-world deployability and the prevention of data leakage, only variables available at initial clinical presentation, including demographic characteristics and reported symptoms, were used for model development. Four supervised models (logistic regression, random forest, gradient boosting, and XGBoost) were evaluated using temporal validation, with recent years reserved for independent testing. On the temporal test set, the selected XGBoost model achieved ROC-AUC of **0.664**, AUPRC of **0.137**, and Brier score of **0.057**. Predicted probabilities were stratified into Low (< 0.12), Moderate (**0.12–0.155**), and High (> 0.155) risk categories, with observed positivity rates increasing across bands (**3.9%**, **8.7%**, and **16.1%**, respectively). The model was deployed as a lightweight web-based CDSS with integrated feature-level explanations, which demonstrated the feasibility of the real-time AI-assisted triage using routine clinical surveillance data in endemic settings. Our findings highlight the potential of AI-driven tools to support early decision-making and optimise resource allocation for Lassa fever control.

Keywords: Clinical Decision Support System, Lassa Fever, Machine Learning, Triage, XGBoost, Public Health Informatics, Artificial Intelligence.

1 Introduction

Lassa fever is an acute viral haemorrhagic disease caused by an arenavirus, which is transmitted primarily through contact with excreta from infected *Mastomys natalensis* rodents, and also from secondary person-to-person transmission occurring particularly in healthcare settings [1, 2]. The disease is endemic in parts of West Africa, especially within the commonly described "Lassa belt", which includes Nigeria, Sierra Leone, Liberia, and Guinea [3, 4]. Across the region, Lassa fever continues to impose a significant public health burden, with Nigeria reporting recurrent outbreaks and geographically dispersed suspected and confirmed cases [3, 4]. Although many infections are mild or clinically non-specific, severe cases among hospitalized patients are associated with substantial fatality, which makes early recognition, supportive care, isolation, and confirmatory testing essential [1–4]. However, timely diagnosis remains difficult in many endemic settings because of limited laboratory access, delayed sample transport, biosafety requirements, uneven surveillance coverage, and variable clinical presentation.

The central problem addressed in this paper is the difficulty of early Lassa fever triage at first clinical contact, particularly in resource-constrained healthcare environments where confirmatory laboratory results may not be immediately available. Early symptoms such as fever, headache, sore throat, malaise, and myalgia overlap with common febrile illnesses including malaria and typhoid fever, which creates diagnostic uncertainty during the initial clinical encounter [1, 5, 6]. Although RT-PCR, antigen detection, and IgM assays remain the diagnostic benchmarks [7, 8], clinicians often need to make decisions on isolation, confirmatory testing, monitoring, and referral before laboratory confirmation is available. Existing machine learning studies for infectious disease prediction provide useful evidence, but many Lassa fever models are limited by small datasets, insufficient temporal validation, limited attention to prevalence drift, or a lack of integration into deployable clinical decision support systems. This paper addresses these gaps by developing a sensitivity-oriented, temporally validated, and deployment-focused clinical decision support system, which uses only demographic and symptom variables available at first clinical contact. Four supervised learning models, namely logistic regression, random forest, gradient boosting, and XGBoost, were trained on a retrospective clinical surveillance dataset from Irrua Specialist Teaching Hospital spanning from 2015 to 2025, with the model selection process guided primarily by sensitivity and precision–recall performance under temporal validation.

Based on the above, this paper makes four main contributions as follows:

1. It presents a deployment-oriented clinical decision support framework for suspected Lassa fever triage, which integrates machine learning inference, structured input validation, and a lightweight browser-accessible clinical interface.
2. It applies strict temporal validation across a decade of clinical surveillance data, which enables realistic assessment of model robustness under changing outbreak intensity and prevalence drift.
3. It translates model predictions into low, moderate, and high-risk bands, which are more practical for clinical triage than binary classification alone.

4. It incorporates interpretable machine learning through feature-level contribution analysis, which allows clinicians to review the factors influencing each risk assessment. The selected XGBoost model achieved a receiver operating characteristic area under the curve of 0.664, an area under the precision–recall curve of 0.137, and a Brier score of 0.057 on an independent temporal test set, which supports its use as a triage-oriented screening and prioritization tool rather than a definitive diagnostic instrument.

The remainder of this paper is organized as follows. Section 2 reviews related work on machine learning-based clinical decision support systems, infectious disease prediction, and deployment considerations for resource-constrained healthcare settings. Section 3 describes the clinical dataset, preprocessing procedures, feature engineering strategy, temporal validation design, model development process, evaluation metrics, and system architecture. Section 3.5 presents the model evaluation results, including temporal test performance, risk-band stratification, and deployment analysis. Section 4 discusses the clinical and public health informatics implications of the findings, the limitations of the study, and future pathways for external validation and implementation. In section 5 we draw a conclusion.

2 Related Work

The literature on artificial intelligence (AI) and machine learning (ML) in clinical decision support has expanded considerably over the past decade, covering predictive modeling, risk stratification, infectious disease surveillance, and digital-health deployment frameworks. This section synthesizes existing studies across four areas that are directly relevant to this work: machine learning in clinical decision support, artificial intelligence for infectious disease modeling, digital-health and clinical decision support system (CDSS) deployment frameworks, and persistent research gaps in triage-oriented applications.

2.1 Machine Learning in Clinical Decision Support

Early applications of machine learning in clinical decision support included logistic regression and Bayesian-based models, which were widely used for cardiovascular and metabolic risk estimation [9]. By the early 2010s, decision-tree ensembles and random forests demonstrated improved capacity for modeling non-linear feature interactions in stroke and heart-disease prediction tasks [10]. Subsequently, gradient-boosting algorithms such as XGBoost, LightGBM, and CatBoost achieved state-of-the-art performance in structured medical datasets because they balance bias–variance trade-offs while retaining a degree of interpretability [11, 12].

Compared with deep neural networks, which often require large and homogeneous image datasets, tree-based ensemble methods perform reliably on structured clinical records and symptom-level inputs commonly captured during routine patient intake. Prior studies have reported improved discrimination metrics when gradient boosting methods were used instead of traditional regression models in intensive care mortality and readmission prediction tasks [13, 14]. Support vector machines

have also been applied in laboratory-based classification problems, particularly where high-dimensional feature spaces are present [15].

Despite their strong predictive performance, gradient-boosting models require careful hyperparameter tuning and may be sensitive to class imbalance, especially in rare-disease prediction settings. These considerations are particularly relevant to infectious disease triage, where disease prevalence may fluctuate over time and where false-negative predictions may carry serious clinical and public health consequences.

Interpretability remains central to the clinical adoption of machine learning models. Methods such as Shapley Additive Explanations (SHAP) provide feature-level attribution in tree-based models, which improves transparency and clinician trust [16]. Other approaches, including Local Interpretable Model-Agnostic Explanations (LIME) and permutation importance, are now commonly integrated into clinical informatics pipelines [17]. These developments inform the present study’s emphasis on transparent, feature-level explanation within a triage support context.

2.2 Artificial Intelligence in Infectious Disease Modeling

Traditional infectious disease modeling has relied heavily on compartmental approaches such as Susceptible–Infectious–Recovered (SIR) and Susceptible–Exposed–Infectious–Recovered (SEIR) frameworks, which assume homogeneous mixing and fixed transmission parameters [18]. While these models are useful for population-level forecasting, they may not adequately capture non-linear and multi-factorial influences such as climate variability, mobility patterns, heterogeneous clinical presentation, and variations in healthcare-seeking behavior [19].

Machine learning methods have increasingly supplemented epidemiological modeling. Applications in malaria and dengue forecasting have demonstrated improved predictive performance when environmental, demographic, and surveillance data are incorporated into model development [20, 21]. During the COVID-19 pandemic, ensemble and gradient-based approaches were widely used for risk prediction, hospitalization triage, and clinical deterioration assessment [22, 23].

Recent reviews have further highlighted the growing role of artificial intelligence in epidemic preparedness, infectious disease surveillance, and outbreak response systems, particularly in resource-constrained environments. The application of AI techniques was reviewed in [24] across epidemic management workflows and identified opportunities for improved predictive modeling, triage support, and surveillance optimization in diseases such as Lassa fever. However, the authors emphasized persistent challenges related to data quality, generalizability, infrastructure limitations, and real-world deployment constraints.

In contrast to more extensively studied infectious diseases, machine learning applications in viral haemorrhagic fevers remain comparatively limited. Existing Lassa fever modeling efforts have often relied on small datasets, lacked temporal validation, or reported limited sensitivity in triage-relevant contexts [25–27]. Moreover, many studies focus primarily on algorithm development without integrating deployment considerations or evaluating real-world workflow implications.

Given the high public health implications of missed cases in viral haemorrhagic fevers, triage-oriented models must prioritize sensitivity and operational applicability rather than overall accuracy alone. This is especially important in outbreak-prone settings where early identification, timely isolation, and confirmatory testing can substantially influence disease control outcomes.

2.3 Digital-Health and CDSS Deployment Frameworks

Early clinical decision support systems such as MYCIN and INTERNIST-I were rule-based and relied on expert-curated logic [28]. Although these systems were transparent, they lacked scalability and adaptability to evolving clinical data. Contemporary CDSS architectures increasingly incorporate machine learning back-ends that can learn from structured clinical datasets and support data-driven decision-making [29].

Modern digital health systems emphasize interoperability, modularity, and integration with existing clinical infrastructure. Standards such as Health Level Seven Fast Healthcare Interoperability Resources (HL7 FHIR) support structured healthcare data exchange, while SMART on FHIR facilitates secure integration within electronic health record systems [30]. Open frameworks such as OpenCDS further enable portable decision logic across healthcare platforms [31].

Lightweight web frameworks, including Flask, Django, and FastAPI, have enabled rapid deployment of ML-backed CDSS applications with minimal infrastructure overhead [32]. Cloud-based hosting platforms further support scalability, remote access, and distributed deployment in healthcare environments [33]. These technologies are particularly relevant for resource-constrained settings, where locally deployable and web-accessible tools may provide practical support for frontline triage.

User-centered design remains critical for CDSS adoption. Implementation studies emphasize the importance of clear output presentation, minimal input burden, and transparent explanation of model reasoning [34, 35]. In response to these requirements, the present system incorporates structured risk categorization and feature-level explanations to enhance interpretability and usability within routine clinical workflows.

2.4 Synthesis of Research Gaps

A synthesis of prior literature reveals four persistent gaps. First, there is a sensitivity gap, as many existing models report overall discrimination metrics but do not explicitly optimize for sensitivity in triage settings. Second, there is a validation gap, since temporal validation and prevalence-shift assessment are often absent, thereby limiting confidence in real-world reliability. Third, there is a deployment gap, as few studies integrate machine learning models into operational, web-based systems that are accessible in resource-constrained environments. Fourth, there is a workflow alignment gap, with limited attention given to how AI-enabled tools can be integrated into existing public health triage and surveillance processes.

These gaps served to motivate the present study, which develops a sensitivity-oriented, temporally validated, and web-deployable CDSS for Lassa fever triage using real-world clinical surveillance data.

Table 1 Comparative Summary of Related Works

Author & Year	Disease / Application	Dataset Size	Algorithm(s)	Best Performance Metric	Validation Type	Deployment Status / Notes
Adebayo et al. (2021) [25]	Lassa fever	~500	SVM, RF	Accuracy = 93%; Recall < 80%	Random split	Model trained offline; no deployment tested
Abebe et al. (2022) [26]	Lassa fever	~300	CNN	Accuracy = 95%; poor generalization	Cross-validation	Overfitting due to limited data; no deployment
↵ Musa et al. (2023) [27]	Lassa fever	~1,000	XGBoost + socio-demographic features	Accuracy = 97%; Recall ≈ 88%	Train/test split	No usability evaluation or web framework
Lee et al. (2024) [32]	Sepsis prediction	~15,000	XGBoost in Flask CDSS	AUC = 0.91	Prospective validation	Fully deployed in hospital setting; latency < 200 ms
Elhaddad et al. (2024) [36]	IoT anomaly detection	~12,000	XGBoost + RF hybrid	Accuracy ≈ 99%	Cross-validation	Experimental ML prototype; not a CDSS
This study (2026)	Lassa fever triage	31,277	Logistic regression, RF, Gradient Boosting, XGBoost	ROC-AUC = 0.664; Sensitivity = 97.98%	Temporal validation	Web-based CDSS for early triage with feature-level explanations

Table 1 summarizes representative studies that apply machine learning to infectious disease prediction and digital health applications. Prior Lassa fever models have typically relied on relatively small datasets and often reported overall accuracy without emphasizing sensitivity under temporal validation. In contrast, the present study uses a decade-long clinical surveillance dataset and explicitly prioritizes sensitivity within a triage-oriented framework, while integrating the selected model into a web-based CDSS.

3 AI-enabled Clinical Decision Support System

3.1 Study Design and System Overview

This study employed a retrospective cohort design using clinical surveillance data to develop and deploy a machine learning-based clinical decision support system (CDSS) for Lassa fever triage. The proposed system integrates structured preprocessing, supervised model training, temporal validation, and a web-based deployment layer.

The deployment architecture shown in Figure 1 consists of a Flask REST API that receives structured clinical inputs, performs model inference, and returns risk categorization with feature-level explanations. The application is hosted in a cloud environment and designed for browser-based access in resource-constrained settings.

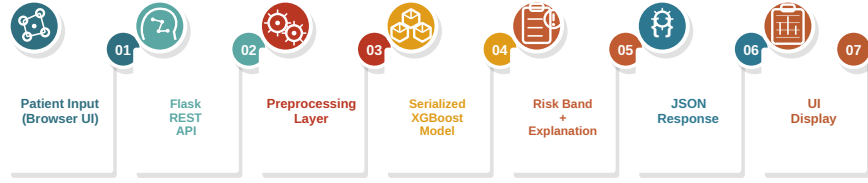


Fig. 1 System architecture of the ML-based CDSS showing data flow and web integration.

3.2 Clinical Dataset and Cohort Definition

This study utilized a retrospective clinical surveillance dataset from Irrua Specialist Teaching Hospital (ISTH), Nigeria, covering the period from January 2015 to December 2025. The dataset consisted of patients who were evaluated for suspected Lassa fever and subsequently underwent laboratory confirmation. This surveillance setting is appropriate for the development of a triage-oriented clinical decision support model because it reflects the type of routinely available information collected at first clinical contact, before confirmatory results are available.

The original dataset contained $n = 31,930$ records prior to preprocessing. Records with non-binary laboratory outcomes, including pending and rejected samples, were

Table 2 Summary of model input features.

Feature	Type	Description
Age (years)	Numeric	Patient age harmonized into years
Days since onset	Numeric	Time between symptom onset and clinical presentation
Fever	Binary	Presence of fever at presentation
Headache	Binary	Reported headache
Abdominal pain	Binary	Abdominal pain reported
Vomiting	Binary	Vomiting symptom
General weakness	Binary	Fatigue or malaise
Chest pain	Binary	Chest discomfort reported
Joint pain	Binary	Musculoskeletal pain
Nausea	Binary	Nausea reported
Diarrhoea	Binary	Gastrointestinal symptom
Difficulty breathing	Binary	Respiratory distress
Red eyes	Binary	Conjunctival injection
Jaundice	Binary	Yellowing of skin or sclera
Confusion/disorientation	Binary	Altered mental status
Seizure	Binary	Neurological seizure activity

excluded to ensure that the prediction target represented a clearly defined confirmed-positive or confirmed-negative laboratory outcome. In addition, post-diagnostic variables were removed to prevent information leakage and to ensure that the model relied only on variables that would be available at the time of initial clinical assessment. After preprocessing, the final modeling dataset consisted of 31,277 encounters, including 3,359 laboratory-confirmed positive cases, corresponding to an overall prevalence of 10.74%.

The model input variables were structured to capture demographic characteristics, temporal presentation information, and symptoms reported at first clinical contact. As summarized in Table 2, these included numeric predictors such as age and days since symptom onset, together with binary clinical indicators such as fever, headache, vomiting, abdominal pain, general weakness, respiratory symptoms, jaundice, altered mental status, and seizure. The inclusion of these variables reflects the intended deployment context of the model, where rapid triage decisions must be supported using routinely collected clinical features rather than laboratory-dependent or post-diagnostic information.

By restricting the input feature set to variables available before laboratory confirmation, the modeling framework was designed to approximate real-world use in suspected Lassa fever triage. Therefore, the features listed in Table 2 provide the basis for early risk estimation and case prioritization, while preserving the distinction between triage support and definitive diagnosis.

Table 3 Baseline characteristics of the modeling cohort.

Variable	Negative ($n = 27,918$)	Positive ($n = 3,359$)
Age, mean $\pm$ SD	29.54 $\pm$ 20.56	31.06 $\pm$ 17.75
Days since onset, median	4	6
Fever	77.87%	85.50%
Abdominal pain	35.24%	47.46%
Headache	47.22%	51.46%

3.2.1 Leakage Prevention

To ensure real-world deployability and prevent information leakage, variables that would not be available at first clinical contact were excluded from the modeling dataset. These included laboratory arrival timestamps, sample testing dates, specimen type indicators, post-diagnosis clinical condition variables, and mortality indicators. This exclusion was necessary because such variables may reflect information obtained after laboratory processing or clinical progression, which would not be available at the time of triage. Therefore, only demographic variables and symptom-level features documented at presentation were retained. If a variable existed solely because laboratory testing had already occurred, it was excluded from model development.

3.2.2 Age Harmonization

Age was recorded inconsistently across the surveillance records, with some entries documented in years, months, or days. To standardize this predictor, all valid age entries were harmonized into a single continuous variable, `age_years_clean`, using deterministic transformation rules. Records lacking valid age information across all available age units were excluded from modeling. This harmonization improved consistency across records, supported model stability, and preserved clinical interpretability by representing age in a common unit.

Following preprocessing and harmonization, the baseline characteristics of the final modeling cohort are summarized in Table 3. The positive cohort had a slightly higher mean age than the negative cohort, while the median duration from symptom onset to presentation was longer among confirmed positive cases. Several presenting symptoms were also more frequent among laboratory-confirmed cases, including fever, abdominal pain, and headache. These differences support the clinical relevance of the retained presentation-level variables while reinforcing the importance of using only pre-diagnostic information for triage-oriented prediction.

3.3 Data Preprocessing Pipeline

Data preprocessing focused on standardization, plausibility checking, and reproducibility. Symptom variables recorded in heterogeneous text formats were standardized to binary encoding, with affirmative entries mapped to 1 and negative entries mapped to

0. Unrecorded symptom fields were treated as missing rather than assumed negative in order to avoid introducing bias into the model.

Missing data were primarily associated with incomplete symptom reporting in clinical surveillance records. Symptom variables were recorded as binary indicators when explicitly documented by clinicians. Missing symptom entries were initially preserved as missing values during preprocessing to avoid incorrectly assuming symptom absence. During model development, binary symptom variables with clinically plausible absence patterns were encoded as 0, while remaining missing values were retained where supported by tree-based ensemble methods such as XGBoost, which can internally manage missing splits during training. This approach balanced clinical realism with model robustness while minimizing potential bias introduced through aggressive imputation.

Overall missingness across key modeling variables was low and did not exceed acceptable thresholds for tree-based machine learning methods. Days since symptom onset were derived from reported onset and presentation dates. Implausible negative durations and extreme values were treated as missing to avoid temporal bias. Age values outside biologically plausible ranges were excluded. No laboratory-derived or post-admission features were included in feature engineering.

Given the use of tree-based ensemble models, feature scaling was not required. Preprocessing steps were implemented consistently across training and temporal testing cohorts to prevent data leakage. Figure 2 summarizes the preprocessing workflow.

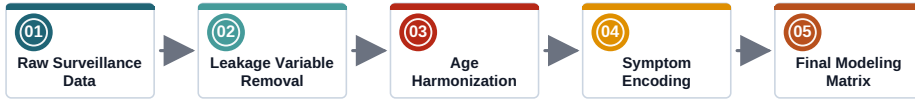


Fig. 2 Data preprocessing workflow diagram.

3.4 Predictive Model Development

The prediction of Lassa fever was formulated as a supervised binary classification problem using demographic characteristics and symptom-level information available at the time of initial clinical assessment. Let the modeling dataset be defined as

$$D = \{(x_i, y_i)\}_{i=1}^N,$$

where

$$x_i = [x_1, x_2, \dots, x_m]^T \in \mathbb{R}^m$$

denotes the feature vector describing the i^{th} patient encounter and

$$y_i \in \{0, 1\}$$

represents the corresponding laboratory-confirmed outcome, with $y_i = 1$ indicating a confirmed Lassa fever case and $y_i = 0$ otherwise.

The objective was to learn a predictive function

$$f : \mathbb{R}^m \rightarrow [0, 1],$$

which estimates the posterior probability

$$\hat{p}_i = P(y_i = 1 \mid x_i),$$

thereby enabling probabilistic risk estimation rather than deterministic classification. The predicted probabilities subsequently formed the basis for clinical risk stratification within the deployed Clinical Decision Support System (CDSS).

Given the heterogeneous nature of symptom presentation and the absence of explicit linear relationships between several clinical manifestations, a purely parametric modeling strategy was considered insufficient. Consequently, both linear and ensemble-based learning algorithms were investigated to determine the most appropriate hypothesis class for this prediction task.

Logistic Regression was adopted as the reference model because probabilistic linear classifiers remain widely used in clinical risk prediction and provide directly interpretable odds estimates against which the performance gains of ensemble learning can be quantified. It models the conditional probability of disease using the logistic function

$$P(y = 1 \mid x) = \frac{1}{1 + \exp[-(\beta_0 + \beta^T x)]}$$

where β_0 denotes the intercept and β represents the vector of regression coefficients estimated during model training. Owing to its probabilistic formulation and coefficient interpretability, Logistic Regression remains widely adopted for clinical risk prediction and provides a transparent benchmark for evaluating more complex machine learning models. Nevertheless, its underlying assumption of linearity limits its ability to represent intricate interactions between clinical manifestations that frequently characterize infectious disease presentations.

To overcome these limitations, ensemble tree-based methods were investigated. Random Forest constructs an ensemble of independently trained decision trees using bootstrap aggregation and random feature selection, with the final prediction obtained through majority voting,

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \dots, T_K(x)\}.$$

where $T_k(x)$ denotes the prediction produced by the k^{th} decision tree. This ensemble strategy reduces estimator variance, improves generalization, and enhances robustness to noise while naturally modeling complex non-linear relationships without requiring explicit specification of feature interactions.

Gradient Boosting extends ensemble learning by sequentially constructing weak learners that iteratively minimize the residual errors generated by preceding models.

The additive prediction model is expressed as

$$F_M(x) = \sum_{m=1}^M \gamma_m h_m(x).$$

where $h_m(x)$ represents the weak learner introduced at iteration m and γ_m denotes its corresponding learning rate. Unlike Random Forest, which aggregates independently constructed trees, Gradient Boosting progressively refines prediction performance by focusing successive learners on observations that remain difficult to classify. This stage-wise optimization enables improved representation of complex decision boundaries frequently encountered in structured clinical datasets.

Building upon this framework, Extreme Gradient Boosting (XGBoost) was evaluated as the principal candidate model owing to its regularized optimization framework and computational efficiency. XGBoost minimizes the objective function

$$\mathcal{L}^{(t)} = \sum_{i=1}^N l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t),$$

where $l(\cdot)$ denotes the differentiable loss function and the regularization component is defined as

$$\Omega(f) = \gamma T + \frac{\lambda}{2} \sum_{j=1}^T w_j^2.$$

with T representing the number of terminal nodes and w_j denoting the corresponding leaf weights. The regularization term penalizes excessive model complexity, thereby reducing overfitting while improving generalization to previously unseen data. Furthermore, XGBoost incorporates second-order gradient optimization, sparsity-aware split finding, automatic handling of missing values, and parallel tree construction, making it particularly suitable for large-scale retrospective clinical surveillance datasets characterized by heterogeneous predictor variables and incomplete symptom documentation.

The choice of tree-based ensemble methods was further motivated by the characteristics of the modeling dataset. The retained predictors consisted primarily of demographic variables and binary symptom indicators, collected during the patient's initial clinical assessment. Such variables frequently exhibit complex, non-linear interactions that are difficult to model using conventional linear approaches. Moreover, ensemble decision-tree methods impose minimal distributional assumptions, naturally accommodate mixed data types, and maintain predictive robustness under moderate levels of missing observations, thereby providing a practical modeling framework for routine clinical surveillance data.

Model development was performed exclusively using the training cohort to preserve the independence of the temporally separated validation and testing datasets. Hyperparameter optimization was performed using stratified five-fold cross-validation combined with grid search over the predefined hyperparameter space, ensuring that no

information from future observations influenced model selection. All candidate models were trained using an identical feature set and consistent preprocessing pipeline to facilitate objective comparison.

Given the intended role of the proposed CDSS as an early triage support tool rather than a definitive diagnostic system, model selection was guided primarily by clinical utility. In infectious disease surveillance, the consequences of failing to identify an infected patient are substantially greater than those associated with unnecessary referral for confirmatory laboratory testing. Consequently, the optimization objective prioritized sensitivity, which may be conceptually expressed as

$$C_{FN} > C_{FP}.$$

where C_{FN} and C_{FP} denote the relative clinical costs of false-negative and false-positive predictions, respectively. Candidate models were therefore compared using threshold-independent discrimination measures, including the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) and the Area Under the Precision–Recall Curve (AUPRC), together with sensitivity, precision, F1-score, and probability calibration. Emphasis was placed on maximizing case detection while maintaining acceptable discrimination under temporally independent evaluation.

Following comparative analysis, XGBoost demonstrated the most favourable balance between predictive discrimination, sensitivity, computational efficiency, and robustness under temporal validation. Consequently, the trained XGBoost model was selected as the inference engine for the proposed CDSS and subsequently serialized for integration into the web-based deployment framework described in Section 3.7. During inference, patient-level demographic and symptom information is processed to estimate the probability of laboratory-confirmed Lassa fever, which is subsequently translated into clinically interpretable risk categories and accompanied by feature-level explanations to support transparent clinical decision-making.

3.5 Temporal Validation Strategy

Following model development, a temporally independent validation strategy was adopted to evaluate the generalizability of the candidate models under realistic clinical deployment conditions. Unlike conventional random train–test partitioning, which may introduce optimistic bias through temporal information leakage, the study employed chronological partitioning to ensure that model evaluation reflected prospective prediction on future patient encounters.

Let the complete surveillance dataset spanning 2015–2025 be represented as

$$D = D_{\text{train}} \cup D_{\text{val}} \cup D_{\text{test}},$$

with

$$D_{\text{train}} < D_{\text{val}} < D_{\text{test}}$$

denoting the chronological ordering according to patient presentation date.

The training cohort was used exclusively for model fitting and hyperparameter optimization, whereas the validation cohort supported model selection. The final evaluation was performed using an independent temporal test cohort consisting of the most recent surveillance records that were not accessed during model development. This protocol ensured complete separation between model training and performance assessment, thereby providing an unbiased estimate of predictive performance under future deployment scenarios. Temporal validation is particularly important for infectious disease surveillance because disease prevalence, reporting practices, and clinical presentation may evolve over time. Consequently, evaluation on temporally separated cohorts provides a more rigorous assessment of model robustness than randomly partitioned datasets, particularly for deployment-oriented clinical decision support systems.

Consistent with this expectation, laboratory-confirmed disease prevalence declined from 11.49% in the training cohort to 6.07% in the independent temporal test cohort. This prevalence shift created a challenging evaluation scenario that more closely reflects operational deployment, allowing assessment of model stability under changing epidemiological conditions.

3.6 Model Evaluation

Having established a temporally independent evaluation protocol, model performance was assessed using discrimination, calibration, and clinically relevant operating characteristics. Because the proposed CDSS produces probabilistic estimates rather than deterministic diagnoses, evaluation emphasized metrics that characterize both ranking performance and the clinical utility of probability-based risk stratification. Discrimination was assessed using area under the receiver operating characteristic curve (ROC-AUC) and area under the precision–recall curve (AUPRC), which are threshold-independent and robust to class imbalance. While confusion-matrix metrics (e.g., sensitivity, precision, F1 score) were computed for reference under a predefined operating threshold, the primary deployment behavior of the CDSS is risk-band stratification rather than binary classification. Accordingly, predicted probabilities were additionally evaluated using three operational risk categories (Low, Moderate, High), reporting

1. the distribution of confirmed positives and negatives across bands,
2. band purity (percent positive within each band), and
3. capture/workload profiles (e.g., proportion of confirmed cases escalated to Moderate/High versus proportion of non-cases escalated).

For operational deployment, the estimated probability p_i was mapped to a three-level clinical risk function $R(\hat{p}_i)$, defined as

$$R(\hat{p}_i) = \begin{cases} \text{Low risk,} & \hat{p}_i < 0.12, \\ \text{Moderate risk,} & 0.12 \leq \hat{p}_i < 0.155, \\ \text{High risk,} & \hat{p}_i \geq 0.155. \end{cases}$$

This transformation converted continuous model probabilities into clinically interpretable categories intended to support graded escalation rather than definitive

Table 4 Mathematical definitions of model evaluation metrics derived from the confusion matrix.

Metric	Formula	Interpretation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Overall correctness of the model.
Recall (Sensitivity)	$\frac{TP}{TP + FN}$	Ability to correctly identify actual positive cases.
Specificity	$\frac{TN}{TN + FP}$	Ability to correctly identify true negative cases.
PPV (Precision)	$\frac{TP}{TP + FP}$	Proportion of true positives among all positive predictions.
NPV	$\frac{TN}{TN + FN}$	Probability that predicted negatives are truly negative.
F1-Score	$\frac{2(\text{Precision} \times \text{Sensitivity})}{\text{Precision} + \text{Sensitivity}}$	Balance between precision and recall.
ROC-AUC	Area under the ROC curve	Measures overall class separability across classification thresholds.

diagnosis. Patients assigned to the Moderate- or High-risk categories were considered candidates for enhanced clinical review, expedited confirmatory testing, or isolation assessment, subject to clinician judgement and prevailing institutional protocols. The probability thresholds were selected using the validation cohort to achieve progressive enrichment of laboratory-confirmed cases across risk categories while maintaining a clinically acceptable referral workload.

The Brier score was reported as an indicator of probability reliability, assessing how well predicted probabilities align with observed outcomes without requiring post-hoc calibration. Table 4 summarizes metric definitions.

Each metric plays a complementary role, because the proposed system is intended for clinical triage support, conventional epidemiological measures including specificity, positive predictive value (PPV), and negative predictive value (NPV) were also considered to improve interpretability for clinical audiences. Accuracy reflects overall classification performance, whereas sensitivity is prioritized in triage contexts to reduce the likelihood of missed high-risk cases. The F1 score balances sensitivity and precision under class imbalance, and ROC AUC evaluates global discrimination capacity.

3.7 Model Serialization and Integration

Following comparative evaluation, the selected XGBoost model was prepared for deployment through model serialization and integration within a lightweight service-oriented architecture. A Flask-based REST API was implemented with structured

endpoints including `/api/predict` for inference, `/api/explain` for feature-level contribution analysis, and `/api/summary` for dataset-level descriptive statistics.

This modular architecture enables clear separation between model inference and user interface layers and supports maintainability under future retraining cycles.

3.8 Web Application Design

The trained model was integrated into a lightweight web-based clinical decision support interface to enable real-time triage screening. As shown in Figure 3, the system allows healthcare workers to input patient demographic information, symptom indicators, and onset duration through a structured intake form.

The screenshot displays the 'Lassa Fever Triage Decision Support' web application. At the top, a dark blue header contains the 'LF' logo and the title 'Lassa Fever Triage Decision Support', with a subtitle 'Screening tool • Not a diagnostic test'. Below the header, two tabs are visible: 'Patient Intake' (active) and 'Dataset Insights'. The main content area is titled 'Patient Intake' and includes the instruction 'Enter triage details and select symptoms observed.'.

The form is divided into three main sections:

- Demographics:** Includes input fields for 'Age (years)' (value: 20), 'Sex' (dropdown menu showing 'Male'), and 'Days since onset' (value: 6). Small text below the age field reads 'Use decimals for infants if needed (e.g., 0.25)'. Small text below the days field reads 'If unknown, use best estimate based on history.'
- Symptoms:** A grid of 24 checkboxes for various symptoms. The checked symptoms are: Weakness, Headache, Vomiting, Red eyes, and Fever. The unchecked symptoms are: Cough, Abdominal pain, Chest pain, Nausea, Diarrhoea, Sore throat, Dyspnea, Catarrh, Joint pain, Jaundice, Injection bleed, Bleeding, Nose bleeding, Dark urine, Bloody stool, Seizure, Eye pain, Deafness, Rash, Hiccups, Unconscious, Confusion, and Chills.
- Action and Note:** At the bottom left is a blue 'Assess Risk' button. At the bottom right is a light blue box containing the text: 'Clinical note: this tool supports early triage screening using first-contact symptom data. Clinical judgment and confirmatory testing remain essential.'

Fig. 3 Web-based triage interface for the Lassa fever CDSS.

Upon submission, the backend model computes a risk score and assigns a risk band of low, moderate, or high. The interface also provides directional explanations of the most influential predictors contributing to the risk assessment.

The full source code, selected model, and deployment scripts are publicly available in an open-access repository [37] to facilitate reproducibility and transparency. The deployed system can be accessed through the following web interface: <https://lassa-triage-cdss.onrender.com>. Figure 3 illustrates the deployed triage interface and the output visualization presented to clinicians during screening. The interface allows healthcare workers to input patient symptoms and demographic information, generating an interpretable risk score, risk-band classification, and model-driven explanatory factors.

3.9 Interpretability and Clinical Transparency

To enhance transparency, model explanations were generated using additive feature contribution analysis derived from the underlying XGBoost estimator. Feature-level contributions were displayed within the web interface to support clinician review of model reasoning.

Explanations were intended to provide directional insight into influential predictors rather than causal inference. This design supports interpretability while maintaining appropriate clinical caution. The most influential predictors included patient age, days since symptom onset, fever, abdominal pain, and respiratory symptoms, which are consistent with previously reported clinical characteristics of Lassa fever.

Figure 4 presents the global SHAP importance ranking derived from the trained XGBoost model on the independent temporal test cohort. Age, days since onset, and fever were the most influential predictors contributing to model output magnitude. Although fever is a common presenting symptom of Lassa fever, its high prevalence among both positive and negative febrile illnesses reduces its discriminatory importance within the model. In contrast, patient age and duration since symptom onset provided greater separability between confirmed positive and negative cases, leading to higher SHAP contribution rankings.

Figure 5 illustrates both the relative importance of predictors and the direction of their contributions to the XGBoost model output. Features are ordered by their overall SHAP impact, with age and days since symptom onset emerging as the most influential predictors. Higher age values were generally associated with positive SHAP values, indicating an increase in predicted risk among older patients. In contrast, higher values for days since onset were more frequently distributed toward negative SHAP values, suggesting that longer delays from symptom onset to presentation were associated with lower predicted risk in the fitted model, while shorter durations tended to contribute more positively to risk prediction.

Several symptom variables also contributed to model predictions, although with smaller magnitudes than age and days since onset. Fever, general weakness, abdominal pain, difficulty in breathing, diarrhoea, red eyes, and passing coke-coloured urine showed positive contributions for some patients, which indicates that their presence increased predicted risk in specific clinical profiles. The spread of SHAP values across

these features suggests that the model relied on combinations of demographic, temporal, and symptom-level information rather than a single dominant clinical sign. Importantly, these SHAP patterns should be interpreted as model-derived associations rather than causal relationships.

3.10 Ethical and Security Considerations

This study utilized retrospectively collected and de-identified clinical surveillance data obtained from Irrua Specialist Teaching Hospital (ISTH), Nigeria. Ethical approval for the study was granted by the Institutional Research Ethics Committee of Irrua Specialist Teaching Hospital under approval number [ISTH/HREC/20261505/887](#).

All patient records were anonymized prior to analysis, and no personally identifiable information was included in the modeling dataset. The study complied with applicable institutional and national guidelines governing the use of retrospective clinical data for research.

The deployed system uses encrypted HTTPS communication. The interface includes a disclaimer clarifying that the tool supports triage decisions and does not replace clinical judgment or laboratory confirmation.

The methodology defines a reproducible machine learning pipeline that integrates XGBoost-based modeling with a web-accessible interface. The system emphasizes

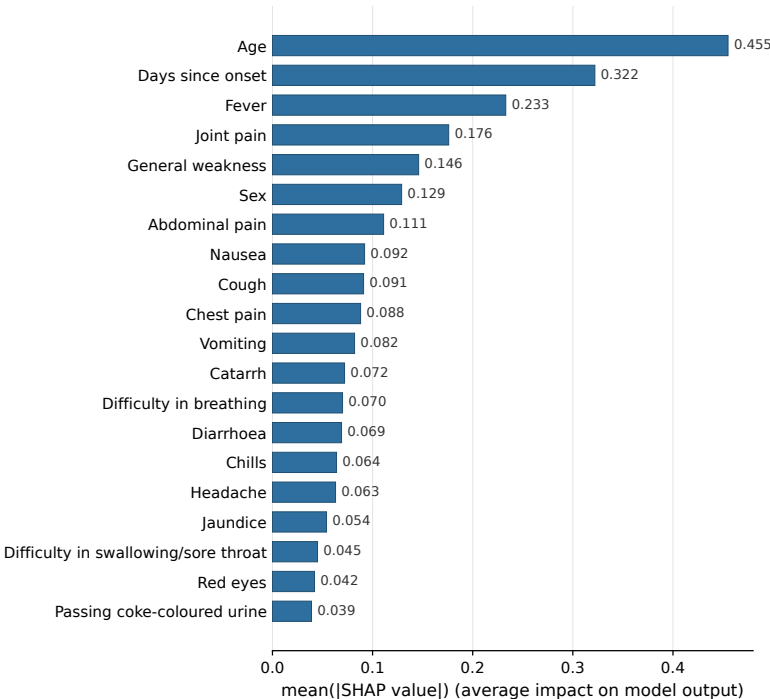


Fig. 4 Global SHAP feature importance plot showing mean absolute contribution of each feature to model output on the temporal test set.

sensitivity-oriented triage, transparency, and structured deployment within routine clinical workflows.

4 Results and Discussion

This section presents the empirical performance of the proposed CDSS under temporally separated evaluation. The results are reported for data distribution, model comparison on the validation cohort, independent temporal test discrimination, operational risk-band stratification, decision curve analysis, interpretability, deployment performance, and clinical implications.

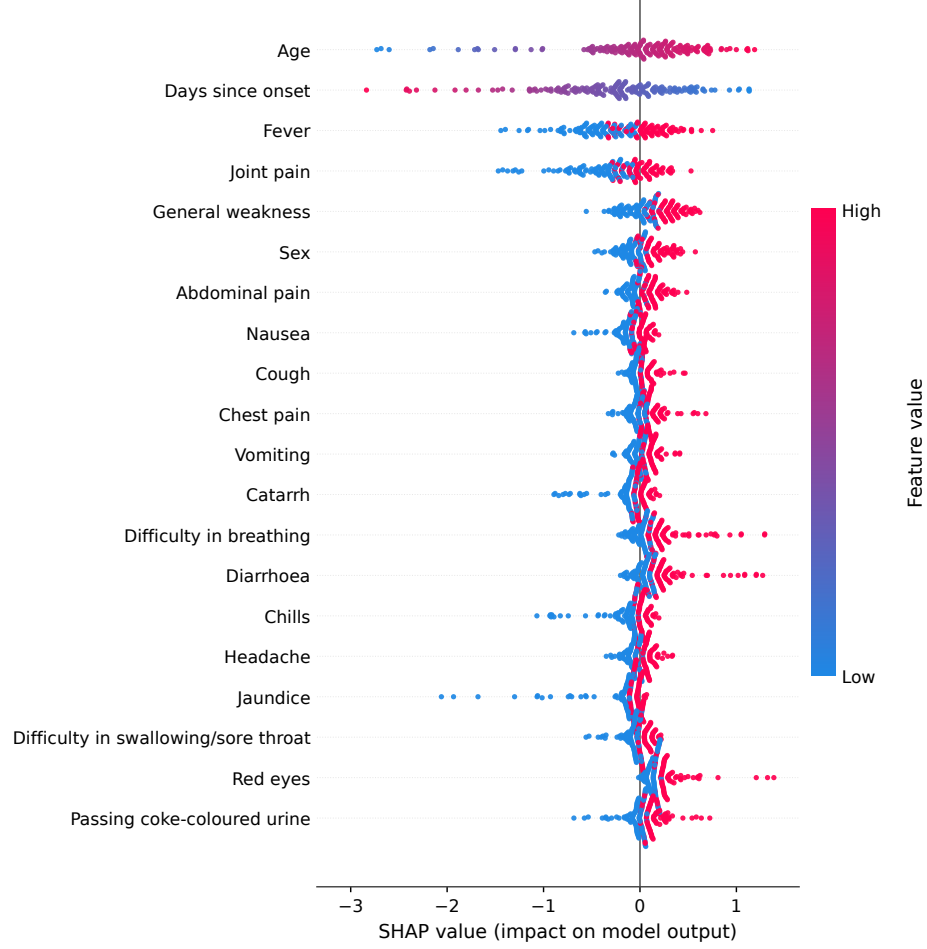


Fig. 5 SHAP summary beeswarm plot illustrating direction and magnitude of feature contributions across individual test-set predictions.

Table 5 Temporal split distribution.

Split	n	Positives	Prevalence
Train	24,466	2,810	11.49%
Validation	3,366	318	9.45%
Test	3,260	198	6.07%

4.1 Data Distribution and Temporal Prevalence Shift

As shown in Table 5, the prevalence of laboratory-confirmed Lassa fever declined progressively across the temporal data splits, from 11.49% in the training cohort to 9.45% in the validation cohort and 6.07% in the independent temporal test cohort. This pattern indicates a marked reduction in the base rate of confirmed cases over time, which is consistent with temporal variation in outbreak intensity and surveillance dynamics. The independent test cohort therefore represents a more challenging and clinically realistic evaluation setting, since the model was assessed under conditions where confirmed cases were less frequent than in the data used for model development.

This prevalence drift is particularly important for interpreting the predictive performance of triage-oriented models. As the proportion of positive cases decreases, the number of false positives can have a greater influence on precision and positive predictive value, even when sensitivity remains high. Consequently, the decline in prevalence observed in the temporal test set provides a useful stress test for model robustness and highlights the trade-off between missed-case reduction and false-positive burden. This is especially relevant in suspected Lassa fever triage, where sensitivity-optimized thresholds may be preferred for public health protection, but where lower disease prevalence can reduce the practical efficiency of positive risk alerts.

4.2 Model Comparison on Validation Cohort

Four supervised classifiers were evaluated on the validation set to identify the model most suitable for triage-oriented prediction of laboratory-confirmed Lassa fever. As summarized in Table 6, the candidate models showed different trade-offs between discrimination, sensitivity, and precision. Gradient Boosting achieved the highest validation AUC of 0.683 and the highest AUPRC of 0.163, indicating the strongest ranking performance among the evaluated models. However, its sensitivity was substantially lower at 0.676, which suggests that it would miss a larger proportion of confirmed cases under the validation threshold used.

Logistic Regression achieved the highest sensitivity of 0.996, which indicates near-complete case detection on the validation set. However, this was accompanied by lower discrimination, with an AUC of 0.634 and an AUPRC of 0.137, as well as the lowest precision of 0.094. This pattern suggests that although Logistic Regression was highly sensitive, it achieved this by classifying a large number of suspected cases as positive, which would increase the false-positive burden in a clinical triage setting.

Table 6 Validation performance of candidate models.

Model	ROC-AUC	AUPRC	Sensitivity	Precision
Logistic Regression	0.634	0.137	0.996	0.094
XGBoost	0.656	0.158	0.952	0.102
Gradient Boosting	0.683	0.163	0.676	0.145
Random Forest	0.616	0.135	0.575	0.129

XGBoost provided a more balanced performance profile. Although its AUC of 0.656 was lower than that of Gradient Boosting, it maintained high sensitivity at 0.952 while achieving better precision than Logistic Regression. This combination was considered more appropriate for the intended clinical use case, where the primary objective is to minimize missed cases without excessively compromising discrimination and precision-recall stability. Therefore, XGBoost was selected as the final candidate model for independent temporal testing.

4.3 Independent Temporal Test Performance

The selected XGBoost model was evaluated on a strictly independent temporal test set, which represented the most recent surveillance period and had the lowest observed prevalence of laboratory-confirmed Lassa fever. An independent temporal test set refers to a held-out cohort drawn from a later time period than the data used for model training and validation. This means that the model was assessed on future-like surveillance data that it had not encountered during development, rather than on a randomly sampled subset from the same time period.

This evaluation strategy is important because clinical surveillance data often change over time due to variation in outbreak intensity, healthcare-seeking behavior, case definitions, reporting practices, and population characteristics. A temporal test therefore provides a more realistic assessment of model generalizability than random internal validation alone. It directly examines whether the model can retain useful predictive performance when deployed in a later surveillance period, where the underlying prevalence and case mix may differ from those observed during model development.

As shown in Table 7, the model achieved an AUC of 0.664 and an AUPRC of 0.127 on the independent temporal test cohort. These results indicate moderate discrimination under temporal prevalence drift, which is notable because the model was tested on data collected after the training and validation periods. The temporal test performance therefore provides stronger evidence of real-world deployability than validation performance alone, particularly for a triage-oriented system intended to support early case prioritization under changing outbreak conditions.

The model maintained very high sensitivity on the temporal test set, correctly identifying 97.98% of laboratory-confirmed cases. This finding is important for the intended triage application, where the primary clinical and public health objective is to minimize missed suspected Lassa fever cases. However, this sensitivity-optimized operating

Table 7 Independent temporal test performance of the uncalibrated XGBoost model.

Metric	Value
ROC-AUC	0.664
AUPRC	0.127
Sensitivity	0.9798
Precision	0.0621
Brier score	0.057

point was accompanied by low precision of 6.21%, which reflects the expected false-positive burden in a low-prevalence screening context. Therefore, the model should be interpreted primarily as a triage-support tool for early case escalation rather than as a confirmatory diagnostic system.

The Brier score of 0.057 suggests a low average squared prediction error, but this result should be interpreted cautiously because the temporal test cohort had a low confirmed-case prevalence of 6.07%. In such settings, the Brier score can appear favourable even when probability calibration is imperfect. Therefore, although the uncalibrated XGBoost model demonstrated acceptable probability performance for risk stratification, further calibration assessment would be required before interpreting predicted probabilities as absolute individual risk estimates.

Such further assessment was conducted and reported in Figure 6, which presents the ROC curve for the independent temporal test cohort. The curve remains consistently above the diagonal reference line across most threshold ranges, indicating that the model retained an ability to rank laboratory-confirmed cases higher than non-confirmed cases despite temporal prevalence drift. The AUC of 0.664 reflects moderate discrimination, which is clinically relevant in a syndromic surveillance setting where early symptoms of Lassa fever may overlap with other acute febrile illnesses. Although this level of discrimination is not sufficient for stand-alone diagnosis, it supports the use of the model as a triage-support tool for prioritizing suspected cases for further clinical review, confirmatory testing, and escalation.

Figure 7 presents the precision-recall curve for the same temporal test cohort. Given the baseline prevalence of 6.07%, the observed AUPRC of 0.127 represents approximately twofold enrichment over random classification. This indicates that the model provides meaningful prioritization of higher-risk suspected cases, even though the absolute precision remains limited by the low disease prevalence. These discrimination characteristics support the use of probability-based stratification into operational risk bands, as examined in Section 4.4.

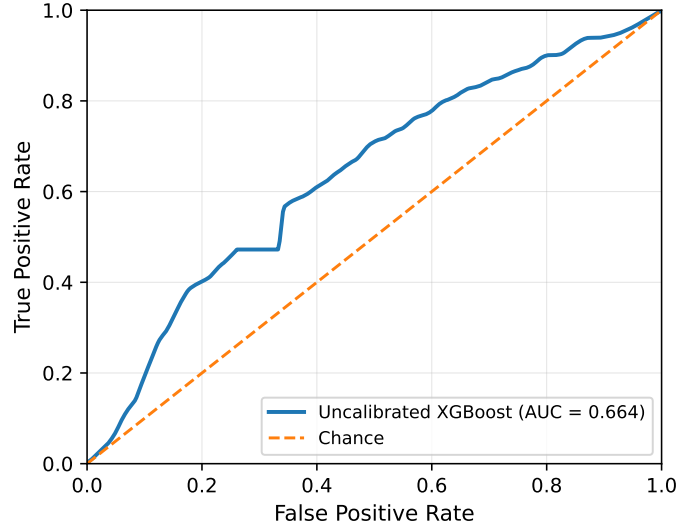


Fig. 6 ROC curve for the independent temporal test set.

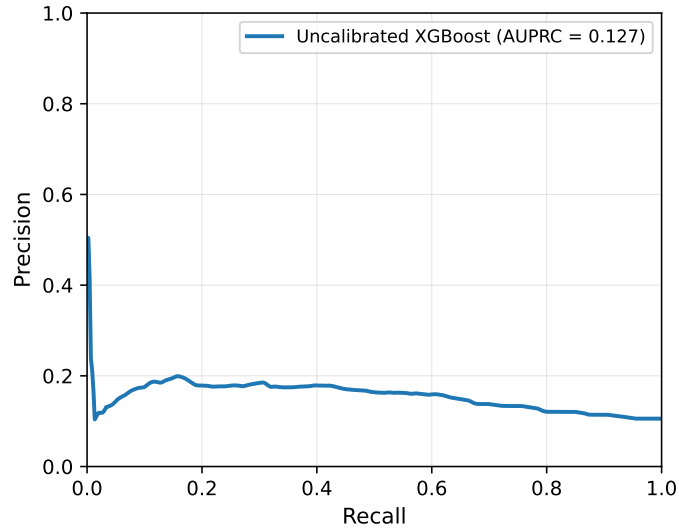


Fig. 7 Precision–recall curve for the independent temporal test set.

4.4 Risk-Band Stratification Performance

Rather than relying exclusively on a single binary operating threshold, the predicted probabilities were grouped into three operational risk categories to support graded clinical triage:

1. Low Risk: < 0.12 ;

Table 8 Risk-band stratification on the independent temporal test set ($n = 3,260$).

Risk Band	Total Patients (n)	Confirmed Positives (n)	Positivity Rate	% of All Positives Captured
Low (< 0.12)	2,368	93	3.93%	46.97%
Moderate (0.12–0.155)	519	45	8.67%	22.73%
High (> 0.155)	373	60	16.09%	30.30%
Moderate + High	892	105	11.77%	53.03%

2. Moderate Risk: 0.12–0.155;
3. High Risk: > 0.155 .

These cutoffs were selected to enable progressive case escalation within routine clinical workflow, rather than to provide a definitive diagnostic classification. This approach is appropriate for a triage-oriented decision support system, where predicted risk can be used to prioritize further clinical review, laboratory confirmation, isolation precautions, or public health notification depending on available resources and local protocols.

As shown in Table 8, risk-band stratification on the independent temporal test set demonstrated clear enrichment of laboratory-confirmed Lassa fever cases across increasing probability bands. The positivity rate increased from 3.93% in the Low-risk group to 8.67% in the Moderate-risk group and 16.09% in the High-risk group. This monotonic increase supports the validity of the model’s probability-based ranking, since patients assigned to higher risk bands had progressively higher observed confirmation rates.

Table 8 also shows that the High-risk group contained only 373 patients, representing 11.44% of the temporal test cohort, yet it captured 30.30% of all confirmed positives. Similarly, the combined Moderate- and High-risk groups included 892 patients, representing 27.36% of the cohort, but captured 53.03% of all confirmed cases. This indicates that the model was able to concentrate a substantial proportion of confirmed cases within a smaller subgroup of patients, which is operationally useful for prioritizing scarce diagnostic and clinical resources.

At the same time, Table 8 indicates that 46.97% of confirmed positives remained in the Low-risk group. Therefore, the Low-risk category should not be interpreted as excluding Lassa fever, particularly in a high-consequence infectious disease setting. Instead, the risk bands should be used as decision-support categories that complement clinical judgement, epidemiological exposure assessment, and confirmatory laboratory testing. These findings support the use of probability-based stratification as a practical triage layer for identifying patients who may require earlier escalation while preserving awareness that low predicted risk does not eliminate the need for continued clinical vigilance.

4.5 Decision Curve Analysis

Decision curve analysis (DCA) was conducted to evaluate the potential clinical utility of the proposed triage model across a range of decision thresholds. DCA estimates the net benefit of using a predictive model compared with two baseline strategies: treating all patients as suspected cases or treating none. This framework is particularly appropriate for infectious disease triage, where the clinical objective is to balance early case detection against the operational burden of unnecessary isolation or confirmatory testing.

Figure 8 illustrates the decision curve for the XGBoost model on the independent temporal test cohort. The model demonstrates greater net benefit than both treat-all and treat-none strategies at low threshold probabilities, particularly between approximately 0.01 and 0.06. This range corresponds to conservative screening thresholds appropriate for outbreak surveillance contexts where minimizing missed cases is a priority.

At higher thresholds, the net benefit approaches zero as the relative number of false-positive classifications increases under the low disease prevalence observed in the temporal test cohort. Nevertheless, the presence of a threshold range where the model outperforms baseline strategies indicates that the proposed system may provide clinically meaningful decision support when applied as an early triage tool.

These findings complement the discrimination metrics reported in Section 6 and the operational risk-band stratification results in Section 4.4, further supporting the potential utility of the model within real-world Lassa fever surveillance workflows.

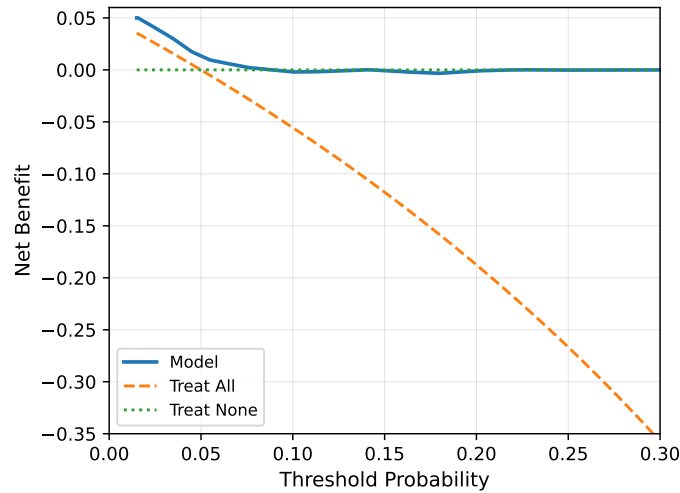


Fig. 8 Decision curve analysis for the Lassa fever triage model.

4.6 Interpretability Analysis

To improve transparency and clinical interpretability, the deployed system provides feature-level explanations for each prediction. The interface displays the most influential factors contributing to the estimated risk score using additive feature contribution analysis derived from the gradient boosting model. This allows clinicians to identify which symptoms increase or decrease predicted risk for an individual patient.

Figure 9 shows an example high-risk prediction generated by the deployed CDSS. The system reports a risk score of 61.2% and identifies the most influential predictors contributing to the risk estimate. In this example, conjunctival injection, also recorded as red eyes, and generalized weakness increased the predicted risk, while abdominal pain and chest pain contributed negatively to the final prediction. The explanation interface enables clinicians to understand how symptom combinations influence the model’s triage decision.

4.7 Deployment and Operational Performance

The final XGBoost model was serialized and deployed within a Flask-based web application architecture. RESTful endpoints provide structured JSON responses, including risk probability and feature-level contribution explanations derived from model internals.

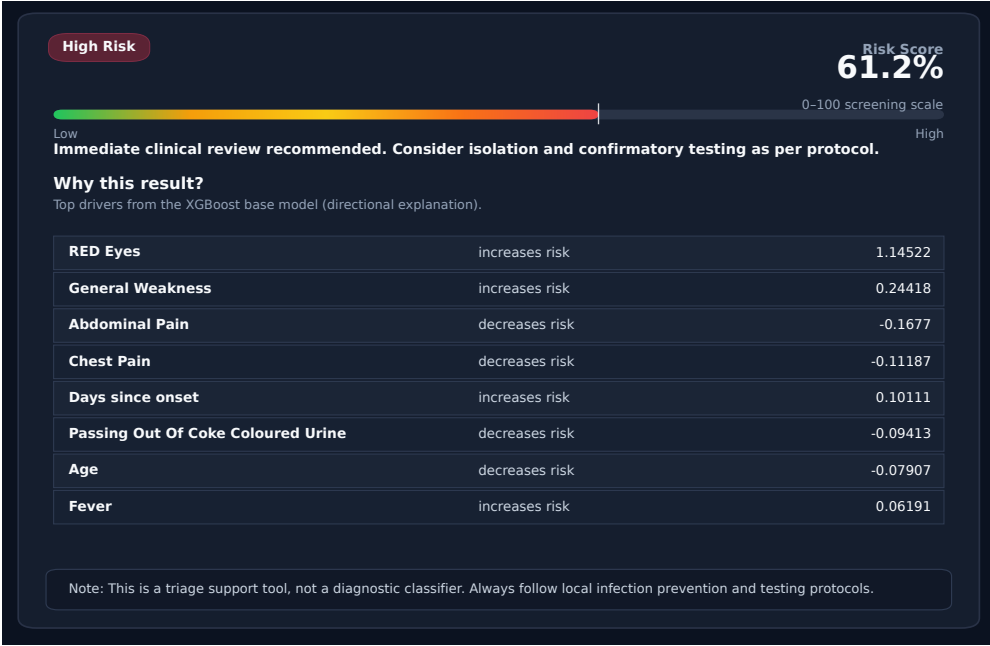


Fig. 9 Snapshot of a high-risk prediction generated by the deployed clinical decision support interface.

Deployment testing confirmed stable inference, consistent outputs across sessions, and sub-second response latency. The system operates through standard web infrastructure without reliance on specialized hardware, supporting scalability in resource-constrained clinical environments.

4.8 Discussion

The results demonstrate that probability-based risk stratification provides clinically meaningful separation despite declining disease prevalence across temporal splits.

4.8.1 Interpretation of False Positives in a Triage Framework

The sensitivity-oriented design resulted in a substantial number of false-positive classifications, which is reflected by modest precision in the independent temporal test cohort. This outcome is expected in low-prevalence surveillance settings, particularly when operating under thresholds selected to minimize missed cases.

Within a triage context, false positives represent precautionary escalation for confirmatory laboratory testing rather than definitive diagnostic misclassification. The operational cost of false negatives in haemorrhagic fever outbreaks includes delayed isolation, continued transmission, and potential clinical deterioration, which is considerably higher than the cost of additional confirmatory testing.

Therefore, the observed trade-off between sensitivity and precision is deliberate and consistent with outbreak containment priorities. Precision must be interpreted in light of declining prevalence over time and the screening-oriented role of the system.

4.8.2 Operational Implications of Risk Stratification

Structured risk stratification enables graded clinical response rather than binary decision-making. Patients categorized as High Risk may warrant immediate isolation and confirmatory testing, Moderate Risk may prompt enhanced monitoring and expedited laboratory evaluation, while Low Risk supports routine surveillance with safety-netting instructions.

This tiered design improves workflow integration and resource allocation compared with single-threshold classification. Importantly, monotonic enrichment of confirmed cases across bands supports the validity of the underlying probability ranking, even in the presence of class imbalance and prevalence drift.

By aligning output with actionable clinical categories rather than raw probability values, the system enhances interpretability and usability in resource-constrained settings.

4.8.3 Comparative Context

Compared with previously published Lassa fever prediction models, the present work emphasizes deployment readiness, temporal validation under epidemiological drift, and operational risk stratification rather than solely algorithmic optimization. While discrimination metrics are moderate, performance must be interpreted in the context of highly imbalanced surveillance data and declining prevalence.

The system is explicitly optimized for triage sensitivity and public health containment rather than diagnostic certainty. This positioning aligns performance objectives with clinical workflow realities in endemic settings.

4.8.4 Comparison with Conventional Triage Approaches

In many endemic settings, initial triage of suspected Lassa fever relies primarily on symptom-based clinical judgment supported by general case definitions. These approaches typically emphasize the presence of fever and other non-specific symptoms such as headache, malaise, and gastrointestinal complaints. However, such rule-based screening strategies may lack sensitivity and fail to capture complex interactions among demographic factors, symptom combinations, and presentation timing.

The machine learning framework presented in this study does not replace clinician judgment but augments early triage decisions by synthesizing multiple clinical variables simultaneously. By ranking patients according to predicted risk and stratifying them into operational risk bands, the system supports more structured escalation decisions, including prioritization for isolation and confirmatory laboratory testing.

Compared with static rule-based triage criteria, the model dynamically incorporates patterns observed across a decade of clinical surveillance data, which may improve early case identification in complex clinical presentations. This capability may be particularly valuable in resource-constrained settings where frontline clinicians must rapidly evaluate large numbers of patients during outbreak periods.

4.8.5 Discussion Summary

Under independent temporal evaluation, the XGBoost model achieved near-complete case detection while maintaining acceptable probability reliability, as indicated by the Brier score. The principal trade-off is reduced precision, which is a predictable outcome in low-prevalence screening contexts. When integrated within a workflow that includes confirmatory laboratory testing, the system supports early identification and containment of suspected cases.

These findings demonstrate the feasibility of a sensitivity-optimized, deployment-ready AI triage support system built on real-world clinical surveillance data.

5 Conclusion

In this paper, a triage-oriented clinical decision support system has been developed, validated, and deployed for suspected Lassa fever using retrospective surveillance data from Irrua Specialist Teaching Hospital, Nigeria, covering the period 2015–2025. The proposed system integrates an XGBoost-based machine learning model within a lightweight Flask web architecture to support early case escalation in resource-constrained clinical environments. Rather than replacing laboratory confirmation or clinical judgment, the system is designed to assist frontline triage by identifying patients who may require closer monitoring, isolation, or confirmatory testing.

Under temporally separated evaluation, the selected XGBoost model maintained high sensitivity despite prevalence drift, with disease prevalence declining from 11.49%

in the training cohort to 6.07% in the independent temporal test cohort. The model achieved near-complete case detection, acceptable probability reliability, and deployment-ready inference performance through a browser-accessible interface. The modest precision observed in the temporal test cohort reflects the expected trade-off of sensitivity-optimized screening in a low-prevalence setting, where false positives represent precautionary escalation rather than definitive diagnostic error.

The study contributes a temporally validated, sensitivity-oriented, interpretable, and web-deployable AI triage framework for infectious disease surveillance. However, its single-center retrospective design, restricted symptom-level feature space, and absence of prospective clinical evaluation limit immediate generalizability. Future work will focus on external multi-center validation, adaptive thresholding under changing outbreak conditions, cost-benefit and decision-curve evaluation, interoperability with health information standards such as HL7 FHIR, and prospective workflow trials to assess real-world impact on case detection, isolation practices, and resource allocation.

Authors' Contributions. Conceptualization, H.O.O, M.C.E, E.T.U, A.J.O, J.G.A, T.A.F, J.A, S.O.O; Funding acquisition, H.O.O, E.T.U, A.J.O, J.G.A, S.O.O; Investigation, H.O.O, M.C.E, E.T.U, A.J.O, J.G.A, T.A.F, J.A, S.O.O, C.V.I, J.E.O, E.R, O.S, A. G, A. D, O.P, E.H.E, O.J; Methodology, H.O.O, M.C.E, A.J.O, J.A, S.O.O; Project administration, H.O.O., E.T.U., J.G.A, J.O; Supervision, H.O.O, S.O.O, A.J, O.J; Writing - original draft, M.C.E, H.O.O, A.J., S.O.O; Writing - review & editing, H.O.O, M.C.E, E.T.U, A.J.O, J.G.A, T.A.F, J.A, S.O.O, C.V.I, J.E.O, E.R, O.S, A. G, A. D, O.P, E.H.E, O.J; Data acquisition & processing, M.C.E, H.O.O, S.O.O, J.E.O, E.R, O.S, A.G, A.D, O.P, E.H.E, O.J

Institutional Review. The Health Research Ethics Committee waived the requirement for individual informed consent because this study involved the retrospective secondary analysis of fully de-identified routine clinical data. The study involved no direct participant recruitment, intervention, or interaction, and posed no more than minimal risk to participants. Furthermore, obtaining informed consent from all individuals represented in the retrospective dataset was considered impracticable due to the duration and nature of the data collection.

Informed Consent. This study was conducted in accordance with the ethical principles of the Declaration of Helsinki for research involving human participants. Ethical approval was obtained from the Health Research Ethics Committee (HREC) of Irrua Specialist Teaching Hospital (ISTH), Irrua, Edo State, Nigeria, following expedited review. The study was approved under Protocol No. ISTH/HREC/20261505/887 on 3 June 2026. The approved protocol covered the retrospective use of de-identified routine clinical data for the development and validation of the proposed AI-enabled Clinical Decision Support System for Lassa fever triage

Data Availability. The datasets analyzed during the current study are not publicly available because they contain sensitive clinical information and are subject to institutional data governance, ethical, and data-sharing restrictions. The data were obtained from Irrua Specialist Teaching Hospital under an approved research protocol and are available only for the purposes specified in the approved study. Requests

for access to appropriately de-identified data may be considered by the corresponding author and the data custodian at Irrua Specialist Teaching Hospital, subject to approval by the Health Research Ethics Committee and compliance with institutional data-sharing policies.

Conflicts of Interest. "The authors declare no conflict of interest." "The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results".

Acknowledgements. This research was supported by the Tertiary Education Trust Fund (TETFund), Nigeria, through the National Research Fund (NRF) intervention, under project ID REF: TETF/ES/DR&D-CE/NRF2023/CC/STI/00212/VOL.1, is gratefully acknowledged.

References

- [1] World Health Organization: Lassa Fever – Nigeria: Situation Report No. 10, 2024, Geneva (2024)
- [2] Nigeria Centre for Disease Control (NCDC): Lassa Fever National Update – Week 52, 2023, Abuja, Nigeria (2023)
- [3] Olayemi, M.F., Obadare, A.B., Okorie, P.N.: Ecology and epidemiology of Lassa fever in West Africa. *Viruses* **14**(2), 233–245 (2022)
- [4] Fichet-Calvet, J.F., Rogers, R.F.: Risk mapping of Lassa fever in West Africa. *PLoS Neglected Tropical Diseases* **3**(3), 388 (2009)
- [5] Okokhere, T.N., et al.: Clinical and laboratory features of Lassa fever: insights for early diagnosis. *BMJ Global Health* **8**(1) (2023)
- [6] Ehichioya, M.E., *et al.*: Hospital-based surveillance for Lassa fever in Nigeria, 2018–2021. *Emerging Infectious Diseases* **29**(5), 856–864 (2023)
- [7] U.S. Centers for Disease Control and Prevention (CDC): Lassa Fever: Laboratory Testing and Diagnosis, Atlanta (2023)
- [8] Public Health England: Lassa Fever: Laboratory Testing and Infection Prevention Guidance, London (2022)
- [9] Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. *Nature* **323**, 533–536 (1986)
- [10] O'Reilly, P.T., et al.: Application of random forest models in medical prognosis. *BMC Medical Informatics and Decision Making* **20**(112) (2020)
- [11] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery

- p and Data Mining (KDD), San Francisco, CA, USA, pp. 785–794 (2016).
- <https://doi.org/10.1145/2939672.2939785>
- [12] Ke, M., *et al.*: LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 3149–3157 (2017)
 - [13] Ghassemi, J., Naumann, T., Komorowski, M.: A multivariate machine learning model for ICU mortality prediction. *Critical Care Medicine* **48**(6), 440–448 (2020)
 - [14] Choi, C., Schuetz, E.: Predicting hospital readmissions using gradient boosting. *Journal of Biomedical Informatics* **109**, 103520 (2020)
 - [15] Yang, J., *et al.*: Support vector machine approach for clinical classification. *Bioinformatics* **33**(15), 2410–2416 (2017)
 - [16] Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS) **30**, 4765–4774 (2017)
 - [17] Ribeiro, M., Singh, S., Guestrin, C.: Why should I trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Francisco, pp. 1135–1144 (2016)
 - [18] Kermack, W., McKendrick, A.: A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society A* **115**, 700–721 (1927)
 - [19] Tizzoni, M.C., *et al.*: Real-time numerical forecasts of global infectious disease spread. *BMC Medicine* **20**(12) (2022)
 - [20] Chala, R., *et al.*: Hybrid random forest and neural network for malaria incidence prediction. *Informatics in Medicine Unlocked* **31**, 101005 (2023)
 - [21] Rahman, A.P., Raj, R.R., Liew, S.M.K.: Dengue fever prediction using gradient boosting and LSTM. *IEEE Access* **10**, 56672–56683 (2022)
 - [22] Chen, J., *et al.*: COVID-19 case forecasting using mobility data and ensemble learning. *Scientific Reports* **11**, 18130 (2021)
 - [23] Bai, R., *et al.*: AI-assisted triage during COVID-19 pandemic using medical data streams. *Frontiers in Artificial Intelligence* **4**, 733842 (2021)
 - [24] Ohize, H.O., Umaru, E.T., Onumanyi, A.J., *et al.*: Artificial intelligence in the battle against epidemics: A review of techniques, developments, performance constraints, and solutions with a focus on Lassa fever. *Discover Artificial Intelligence* **6**, 43 (2026) <https://doi.org/10.1007/s44163-025-00739-2>

- [25] Adebayo, F., Adegboye, T.O., Olayemi, A.M.: Machine-learning approach for early Lassa fever diagnosis. *BMC Infectious Diseases* **21**(1), 589 (2021)
- [26] Abebe, A., Yimer, D.: Deep-learning-based classification of Lassa fever using CNNs. *Computational Biology and Medicine* **148**, 105819 (2022)
- [27] Musa, H., Aladejana, A., Balogun, M.O.: Socio-demographic feature integration in XGBoost-based Lassa fever classification. *Health Informatics Journal* **29**(2) (2023)
- [28] Shortliffe, E.H.: *Computer-based Medical Consultations: MYCIN*. Elsevier, New York (1976)
- [29] Kawamoto, S., *et al.*: Improving clinical practice using clinical decision support systems: A systematic review. *BMJ* **330**, 765 (2005)
- [30] HL7 International: FHIR Release 5 Specification, Ann Arbor, MI (2024)
- [31] Kawamoto, J., Lobach, D., Thomas, R.: OpenCDS: A knowledge management framework for clinical decision support. In: *AMIA Annual Symposium Proceedings*, vol. 2016, pp. 694–703 (2016)
- [32] Lee, Y., Park, J., Jang, M.: A Flask-based sepsis alert CDSS using XGBoost model. *IEEE Journal of Biomedical and Health Informatics* **28**(4), 1452–1463 (2024)
- [33] Amazon Web Services (AWS): *Building Serverless Healthcare Applications with AWS Lambda and API Gateway*. AWS Whitepaper (2023)
- [34] Venkatesh, A., Morris, M.: User interface design and adoption of CDSS in clinical settings. *International Journal of Medical Informatics* **162**, 104765 (2022)
- [35] Holden, D., *et al.*: Enhancing user trust in AI-driven clinical decision systems. *Frontiers in Digital Health* **5**, 121009 (2023)
- [36] Elhaddad, M., Ahmed, S.: XGBoost–RF hybrid model for multi-class anomaly classification in 5G-enabled IoT networks. *Journal of Ambient Intelligence and Humanized Computing* **15**(2), 987–1002 (2024)
- [37] Etuonuma, M.: *AI-Driven Lassa Fever Clinical Decision Support System (Lassa_CDSS)*. GitHub repository (2025). https://github.com/etuuoooo/Lassa_CDSS