Journal of e-Science Letters
Original Research Article | ISSN 3041-5454

Machine Learning versus Statistical Benchmarks for One-Week-Ahead Forecasting and Seasonally Adjusted EWMA Alerting in Nigerian Lassa Fever Surveillance

Oluwapelumi Alabi*[1], Yusuf O. Afolabi[1], Matthew A. Ogunsan[2], and Oluwatoyin A. Ediwin[1]

[1]*Department of Mathematics and Statistics, Rufus Giwa Polytechnic, Owo, Ondo State, Nigeria*
[2]*Ekiti State Polytechnic, Isan-Ekiti, Ekiti State, Nigeria*

ABSTRACT

The number of cases of Lassa fever reported in Nigeria is overdispersed, autocorrelated and highly seasonal, and claims of better forecasting and outbreak alerting based on machine learning (ML) should be evaluated against appropriate statistical benchmarks. We examine 236 weekly national counts of confirmed cases and deaths from the Nigeria Centre for Disease Control (NCDC) (2021 week 9 to 2025 week 36, 4,849 cases, 765 deaths). Alert weeks are determined using an EWMA chart ($\lambda = 0.2$) of Pearson residuals of a seasonally based negative binomial baseline, and the control-limit multiplier is determined by simulation to have an average run length of approximately 100 weeks ($L = 2.24$). Nine one-step-ahead forecasting methods, five ML-type methods and four statistical benchmarks were evaluated and compared using rolling-origin evaluation over 140 weeks of three dry seasons based on MAE, RMSE, MASE, Diebold–Mariano tests and block-bootstrap intervals. The negative binomial model performed best in terms of MAE (7.94 vs. 8.95 for the benchmark), but this result was not robust to multiplicity adjustments; the linear regression model had the lowest MAE (8.95 vs. 13.3), which was not significantly different from the benchmark. A logistic model with the EWMA statistic alone (AUC 0.92) performed as well as or better than a random forest (AUC 0.88) and a neural network (AUC 0.77) for predicting next week's alert status. Sophisticated ML added very little to simple statistical baselines for this series.

Keywords: *Lassa fever; disease surveillance; EWMA control chart; negative binomial regression; rolling-origin evaluation; forecast comparison; random forest.*

1. Introduction

Lassa fever is an acute viral haemorrhagic disease caused by Lassa virus, an arenavirus transmitted by the multimammate rat (Mastomys natalensis) to humans primarily via contact with the excreta of the rats and secondarily via infected body fluids [1] Nigeria has the highest reported burden, with 20-34 cases reported in the 37 states of the country from 2018 to 2023, and epidemics occurring during the dry season [1] [2] [3] The case fatality rate for confirmed cases has been reported to be approximately 13.5-18.3% (Eneh et al., 2025).

Citation: Oluwapelumi Alabi, Yusuf O. Afolabi, Matthew A. Ogunsan, and Oluwatoyin A. Ediwin (2026). Machine Learning versus Statistical Benchmarks for One-Week-Ahead Forecasting and Seasonally Adjusted EWMA Alerting in Nigerian Lassa Fever Surveillance. *Journal of e-Science Letters.*
DOI: https://doi.org/10.51470/eSL.2026.7.3.13

Received: 10 July 2026
Revised: 12 August 2026
Accepted: 08 September 2026
Available: October 15 2026

Corresponding Authors: **Oluwapelumi Alabi**
Email: **oluwapelumialabi361@gmail.com**

© 2026 by the authors. The license of Journal of e-Science Letters. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/)

The Nigeria Centre for Disease Control (NCDC) is therefore a practical priority to provide timely warning of the annual epidemic, with weekly situation reports being the primary source of public national counts.

Descriptive or seasonal and driver-focused analyses of Lassa fever in Nigeria have been published [4][5]. Other diseases have seen reviews and comparisons that tree ensembles and neural networks perform well or better than classical time-series models [6][7][8]. Boosted trees were also found to be significantly better than count regressions for weekly mortality (R^2 of 0.85 vs below 0.20) for Lassa fever by [9], while other studies use multi-source data to apply ML to outbreak progression and early warning [10][11]. However, these comparisons may be misleading for surveillance series. The counts are overdispersed, and highly autocorrelated, and the naive forecast is the last week's value, which is difficult to outperform [12]. If a control chart is used on the raw series without modelling seasonality, it will only indicate a dry season. High sensitivity does not necessarily mean that ML models are good for early warning when trained to predict an alert label based on the same case counts used as inputs.

These problems can be addressed with principled tools from the statistical literature on outbreak detection: seasonally adjusted regression baselines for overdispersed counts [13][14], control charts [15][16], and rigorous evaluation of prospective detection [17].

This paper applies these tools to a national Lassa fever series and poses a narrow question: can ML add measurably to statistical benchmarks, after seasonality, overdispersion and honest out-of-sample evaluation are done properly? Our contributions are (i) a seasonally adjusted EWMA with a limit that is calibrated by simulation for negative binomial counts, (ii) a rolling-origin comparison of nine forecasting methods, including ARIMA, seasonal naïve and negative binomial benchmarks, with formal tests and bootstrap intervals, (iii) a prospective alert-prediction task with benchmarks that reveals how much of the apparent skill is the persistence of the EWMA signal, and (iv) a documented data-quality audit of the public series. The rest of the paper is organized as follows: Section 2 presents the data and quality control, Section 3 presents the methods, Section 4 presents the results, and Section 5 discusses limitations and implications.

2. Methodology

2.1 Source of Data

Confirmed Lassa fever cases and deaths were manually extracted from NCDC situation reports.
(https://ncdc.gov.ng/diseases/sitreps) every week. The file is transcribed for 236 epidemiological weeks (2021 week 9 to 2025 week 36). Counts are based on the reporting week, not the onset of symptoms, and are not reflective of revisions to previous weeks. The series were screened for missing values (2.1% of the series) for weeks of 2021 (weeks 21, 23, 24, 30 and 32) were linearly interpolated between adjacent counts; the impact of this was not evaluated separately as all missing weeks were in the low-transmission season with counts below 13.

2.1 Design of Notation and Evaluation

Let *y_t* denote the number of confirmed cases in week *t, d_t* denote the number of deaths in week *t* and *w_t* denote the epidemiological week, *w_t = 1, ... ,p,* in which the cases in week *t* are confirmed. All forecasts are one week ahead, i.e., the forecast for week *t - 1.* The evaluation of models is based on rolling origin [18]. The first T_0 = 96 weeks (2021 week 9 to 2022 week 52) form the initial training window. Afterwards the origin advances by 4 weeks, and at each origin each model is re-fitted on all data up to that point and generates one-step-ahead forecasts (based on the actual observed lags) for the next 4 weeks. The evaluation period is therefore 140 weeks (from 2023 week 1 to 2025 week 36), with 47 weeks in the dry season (weeks 49-52 and 1-13). Hyperparameters were selected only once, at the first window, and then remained constant throughout the evaluation period, meaning that they were not tuned based on data from the evaluation period.

2.2 Seasonal Negative Binomial Baseline.

The negative binomial (NB2) regression model with two annual harmonics is used to model expected weekly counts when there is no unusual activity:

$$y_t \sim NB(\mu_t, \theta), Var(y_t) = \mu_t + \frac{\mu_t^2}{\theta}, (1)$$

$$log\ \mu_t = \beta^0 + \sum_{k=1}^{2}\left[\beta_{1k}\sin\left(\frac{2\pi k\, w_t}{p}\right) + \beta_{2k}\cos\left(\frac{2\pi k\, w_t}{p}\right)\right] \quad (2)$$

Were P is period (52 for weekly data), *k = 1, ... , K* hamonics *(k = 1 or 2)* is usually enough). The parameters are estimated using maximum likelihood (plus a small ridge penalty on the harmonic coefficients for numerical stability).

The fit is in-sample for weeks within the training window, and the baseline μ_t and θ are prospective for each evaluation week since they are based on the most recent origin. The quantity monitored is then the Pearson residuals

$$r_t = \frac{y_t - \mu_t}{\sqrt{\mu_t + \frac{\mu_t^2}{\theta}}}.$$

2.3 EWMA chart and the calibration of the control limit

The EWMA statistic is based on the Pearson residuals,

$$Z_t = \lambda r_t + (1 - \lambda)Z_{t-1}, Z^0 = 0, \lambda = 0.2, (3)$$

A week t is an alert week if Z_t is above the one-sided upper limit.

$$UCL = L\sqrt{\frac{\lambda}{(2 - \lambda)}}. (4)$$

Since r_t is a standardised residual, the variance of z_t is approximately equal to the unit variance (4) in the steady state and is used throughout. The multiplier L is not taken from normal theory. Rather, it is calibrated by Monte Carlo simulation: counts are simulated from the fitted baseline (1)–(2) with a random starting week of the season, the EWMA is applied to the true Pearson residuals of the simulated counts, and L is selected to make the average run length to the first false signal (ARL_0) close to 100 weeks, or about 2 years (4,000 replications, horizon 1,200 weeks). This is the reason for the skewness of low counts. It does not consider estimation error of the baseline, so the actual ARL_0 will be slightly different. Sensitivity to $\lambda \in \{0.1, 0.2, 0.3\}$ and $ARL_0 \in \{52, 100, 200\}$ is reported in Section 3.2. Alert weeks should be interpreted as "excess-incidence alerts", not as independently confirmed outbreaks; there is no external outbreak declaration to validate alert weeks.

2.4 Forecasting Methods

A comparison of nine one-step-ahead methods was performed. The models are: (i) naïve persistence A comparison of nine one-step-ahead methods was performed. The models are: (i) naïve persistence, $\hat{y}_t = y_{t-1}$; (ii) seasonal naïve, $\hat{y}_t = y_{t-52}$; (iii) an ARIMA(p, d, q) model for log *(1 + y_t)* with order determined by AIC on the initial window, estimated by conditional sum of squares (our own implementation) and back-transformed with truncation at zero; and (iv) an NB – AR model, i.e. NB2 regression with log *(1 + y_{t-k}), k = 1, ..., 4,* plus a single annual harmonic pair, whose mean is the point forecast and whose NB quantiles yield prediction intervals. The ML-type methods include multiple linear regression (MLR), support vector regression with a radial basis kernel (SVR; [19]), random forest (RF; [20]), gradient boosting (GB; [21]) and a feed-forward neural network (MLP) with ReLU units and L2 penalty averaged over five random initialisations.The eight predictors all ML-type methods employ are lags $y_{t-1}, ..., y_{t-4}$ and the mean and standard deviation of the previous four weeks, and *sin ($2\pi w_t/P$)* and *cos ($2\pi w_t/P$).* There is no linear trend term, and so tree ensembles (which do not extrapolate) and linear models are given the same information. The standardisation of the predictors for SVR and the MLP is done using only the moments within the training window.

The hyperparameters were selected using the expanding window cross-validation method (three folds of 13 validation weeks each) within the initial window, by minimising RMSE on the small grids in Appendix A. The selected values were: *SVR C = 1, γ = 0.05, ε = 1;* RF maximum depth 3, minimum leaf size 3 (300 trees); GB 100 trees of depth 2, learning rate 0.05, subsampling 0.8; MLP hidden layers (16, 8), penalty 1.

2.5 Alert-Status Prediction

Alert prediction is formulated prospectively: given the eight forecasting predictors and the previous week's EWMA statistic Z_{t-1} and alert indicator a_{t-1}, predict $a_t = 1$ using only the information available up to week t − 1. The classifiers are logistic regression (ridge penalty $C = 0.3$), random forest (300 trees, depth 4, minimum leaf 3) and a single hidden layer neural network (8 units, 5 initialisations averaged, ridge penalty C = 0.1, refitted at each origin with labels observed at that point). Two benchmarks are provided to indicate the extent to which the skill is due to the EWMA state itself: persistence of alert status (a_t predicted by a_{t-1}) and logistic regression on Z_{t-1} alone. Probability of ≥ 0.5 is considered as an alert (class-prior thresholds were not tuned due to small number of alert weeks). Correct prediction of the first alert week of an episode (onset detection) is reported separately, as onsets cannot be detected by construction.

2.6 Performance Measures and Inference

The forecasts are compared using mean absolute error (MAE), root mean squared error (RMSE), mean absolute scaled error (MASE; [22]), and RMSE relative to the naïve forecast. The mean absolute percentage error (MAPE) is not used because it is undefined if counts are zero. The Diebold–Mariano test [23] is used to test differences in predictive accuracy in each of the eight comparisons, with squared and absolute error loss, and using a Newey–West variance with four lags and the small-sample correction of [24]. The 95% percentile intervals for the RMSE, MASE and RMSE ratio are presented with uncertainty, calculated using a moving-block bootstrap (block length 8, 3,000 resamples; [25]). The nominal 90% prediction interval for the NB-AR model is reported.

For alert prediction we report the area under the ROC curve (AUC, with block-bootstrap interval), Brier score, sensitivity, specificity and positive predictive value with exact Clopper–Pearson 95% intervals [26], which are optimistic in light of the run structure of alerts.

3 Results and Discussions

3.1 Descriptive analysis

The series has 4,849 confirmed cases and 765 deaths over 236 weeks, with a case fatality rate of 15.8%. The number of cases per week is highly skewed and overdispersed relative to a Poisson distribution, with 20.5 cases per week (SD 24.9, range 0-137), skewness 2.06, and a variance-to-mean ratio of 30.1. The lag-1 autocorrelation is 0.88. Weekly deaths average 3.24 (SD 4.71, maximum 23). The 73 dry-season weeks (weeks 49–52 and 1–13, 30.9% of the series) contain 72.4% of all cases (Figure 1). The highest weekly peak (137 cases) was recorded in 2023, while the annual CFR in 2025 (to week 36) was 18.4% compared to 13.7% in 2022, and there was no formal trend test conducted for the two years. Annual totals are not directly comparable as 2021 starts at week 9 and 2025 ends at week 36.

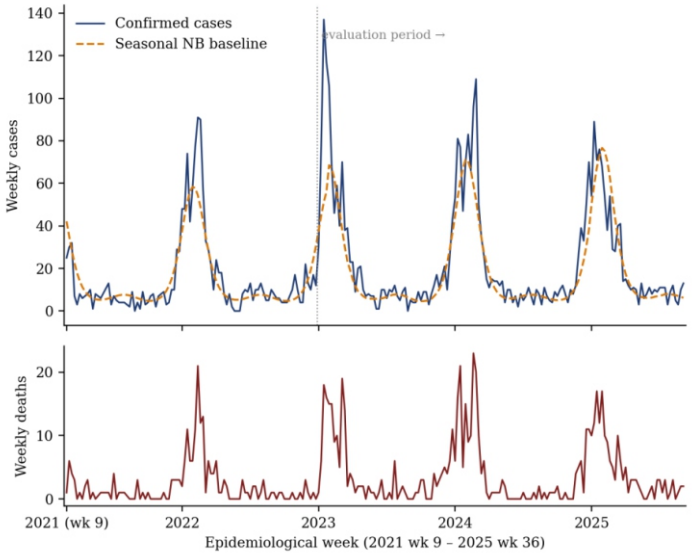


Figure 1. Weekly confirmed cases (top, with the seasonal negative binomial baseline; the dotted line marks the start of the evaluation period) and deaths (bottom), Nigeria, 2021 week 9 to 2025 week 36.

Table 1: Annual summary of confirmed cases, deaths, case fatality rate (CFR, with exact 95% interval) and number of EWMA alert weeks

Year	Weeks	Cases	Mean/week	Peak	Deaths	CFR % (95% CI)	Alert weeks
2021 (wk 9–52)	44	377	8.6	32	51	13.5 (10.2–17.4)	0
2022	52	1,025	19.7	91	140	13.7 (11.6–15.9)	2
2023	52	1,265	24.3	137	210	16.6 (14.6–18.8)	10
2024	52	1,312	25.2	109	204	15.5 (13.6–17.6)	8
2025 (wk 1–36)	36	870	24.2	89	160	18.4 (15.9–21.1)	3
Total	236	4,849	20.5	137	765	15.8	23

3.2 Seasonal EWMA Alerts

Calibration resulted in L = 2.24 (simulated ARL_0 99.7 weeks), which is equivalent to a limit of 0.747 on the scale of Z_t in 8 episodes, 23 of 236 weeks (9.7%) were flagged as alerts, and 21 of 140 evaluation weeks (15.0%) were alerts, with 7 episode onsets (Figure 2). There were 11 alert weeks during the Dec–Mar season, and 12 weeks in other seasons, including a six-week period from April–May 2023 (weeks 16–21). The longest episode was for seven weeks from week 49 of 2024 to week 3 of 2025. The baseline for April-May 2023 was only approximately 6-9 cases per week (observed 7-21), which meant that any excursions above this very low baseline were considered excess incidence. This is likely because of a baseline that is based on a short history of low counts during the off-season of 2021, and it demonstrates that alert definitions are dependent on the reference period.

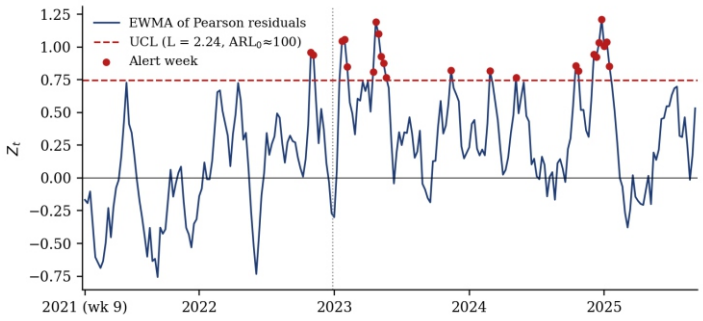


Figure 2. EWMA chart (λ = 0.2) of Pearson residuals from the seasonal negative binomial baseline. Red points exceed the control limit calibrated to an in-control ARL of about 100 weeks.

Table 2: Sensitivity of the alert definition to the smoothing constant λ and the target in-control ARL_0 (L recalibrated for each setting)

λ	Target ARL_0	L	Alert weeks (all)	Evaluation	
				Alert weeks	Onsets
0.1	52	1.60	76	69	13
0.1	100	1.86	59	54	12
0.1	200	2.28	34	32	9
0.2	52	1.80	42	36	11
0.2	100	2.24	23	21	7
0.2	200	2.66	13	11	3
0.3	52	2.04	26	20	7
0.3	100	2.48	16	13	5
0.3	200	2.90	6	5	3

The number of alert weeks is sensitive to the design choices, from 5 to 69 evaluation weeks; reported alert counts should therefore always be quoted with their chart settings.

3.3 Forecasting Performance

Table 3 presents the out-of-sample performance of the forecasts. The differences between the better methods are relatively small. The NB-AR model (13.34) and MLR (RMSE 13.29) have slightly lower RMSE than naïve persistence (13.90), while random forest (14.15) and gradient boosting (13.96) are not distinguishable from it. The MLP (17.08), ARIMA (16.49), seasonal naïve (19.56) and SVR (24.09) are worse. The naïve forecast is the only method that does not outperform the others in terms of squared-error loss (Diebold–Mariano $p \geq 0.32$ for all methods except the naïve forecast), and the bootstrap intervals for the RMSE ratio of all methods except the seasonal naïve and SVR contain 1. The NB-AR model has the lowest MAE (7.94 vs 8.95; raw $p = 0.032$), but this result is not statistically significant after Holm adjustment (adjusted $p = 0.26$) when compared to the other models. The differences between RF, GB and MLR and naïve persistence are within sampling noise, with MAE ranging from 8.36 to 8.41. The seasonal naïve forecast is far from being as good as persistence ($p = 0.039$), as the size of the peak of each year varies from the previous year.

Table 3: One-step-ahead forecast accuracy, rolling-origin evaluation (n = 140 weeks, 2023 week 1 to 2025 week 36). RMSE and MASE are shown with 95% block-bootstrap intervals; the last two columns give raw Diebold–Mariano p-values against naïve persistence (squared and absolute loss)

Method	MAE	RMSE (95% CI)	MASE (95% CI)	RMSE ratio to naïve	DM p (sq.)	DM p (abs.)
Naïve persistence	8.95	13.90 (8.4–16.7)	1.54 (0.97–2.00)	1.00	–	–
Seasonal naïve (lag 52)	11.49	19.56 (11.1–24.8)	1.97 (1.15–2.76)	1.41	0.039	0.035
ARIMA(2,0,0) on log(1+y)	9.41	16.49 (8.3–20.1)	1.62 (0.91–2.14)	1.19	0.249	0.549
NB_ AR + harmonics	7.94	13.34 (7.0–16.2)	1.36 (0.83–1.75)	0.96	0.634	0.032
Multiple linear regression	8.41	13.29 (7.9–16.2)	1.44 (0.94–1.87)	0.96	0.325	0.241
Support vector regression	12.70	24.09 (10.7–30.9)	2.18 (0.99–3.13)	1.73	0.064	0.063
Random forest	8.36	14.15 (7.6–17.2)	1.44 (0.83–1.87)	1.02	0.844	0.289
Gradient boosting	8.38	13.96 (7.7–17.1)	1.44 (0.84–1.94)	1.00	0.960	0.335
Neural network (MLP)	9.30	17.08 (8.0–22.0)	1.60 (0.85–2.22)	1.23	0.301	0.741

There is a significant difference in accuracy among seasons (Table 4). The RMSE is four to ten times higher in the dry-season weeks than in the rest of the year, and it is in these weeks that the methods differ: MLR and NB-AR model slightly reduce the RMSE for the dry season (21.7 and 22.2 respectively) below the naïve value (22.9), while SVR and MLP are much worse (41.2 and 28.9 respectively), with the same pattern of under-prediction of peaks higher than those in the training window. The tree ensembles, SVR and the MLP are best outside the dry season (RMSE 4.0–4.1 compared to 5.0 for the naïve forecast), but the errors are small in absolute terms. The importance of the random forest (Appendix A) is focused on the previous week's count (89%), which means that it acts as a persistence model to a large extent.

Table 4: Forecast error by season: dry season (weeks 49–52 and 1–13; 47 evaluation weeks) and remainder of the year (93 weeks)

Method	RMSE dry	MAE dry	RMSE other	MAE other
Naïve persistence	22.9	18.9	5.0	3.9
Seasonal naïve (lag 52)	32.9	25.5	5.4	4.4
ARIMA(2,0,0) on log(1+y)	27.7	20.9	4.6	3.6
Negative binomial AR + harmonics	22.2	16.6	4.4	3.5
Multiple linear regression	21.7	16.8	5.2	4.2
Support vector regression	41.2	31.5	4.0	3.2
Random forest	23.8	18.6	4.0	3.2
Gradient boosting	23.4	18.7	4.0	3.2
Neural network (MLP)	28.9	21.3	4.1	3.2

The NB-AR model had a 90% prediction interval that covered 95.7% of the evaluation weeks (95.7% during the dry season and 95.7% during the rest of the year), and a mean width of 34.8 cases. The width of the intervals indicates that a one week-ahead interval is wide compared to the typical off-season counts, and coverage above the nominal level indicates that the intervals are conservative. The point forecasts are compared to the observed series in Figure 3

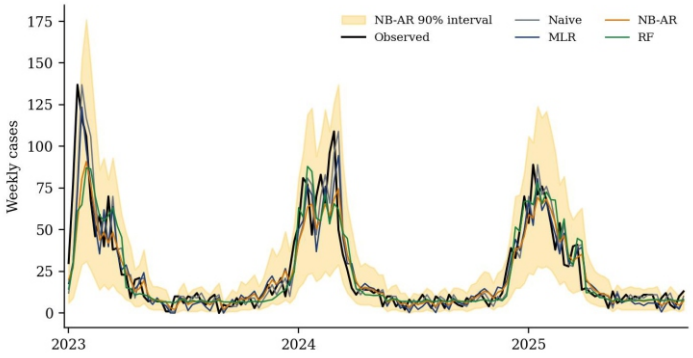


Figure 3. Observed weekly cases and one-step-ahead forecasts from naïve persistence, MLR, NB-AR and random forest over the evaluation period, with the NB-AR 90% prediction interval.

3.4 Alert status prediction

Table 5 shows the prospective prediction of next week's alert status over the 140 evaluation weeks, 21 of which were alert weeks and 7 were episode onsets. By construction, the previous week's alert status has a sensitivity of 67% and a specificity of 94% and will not detect any of the 7 onsets. The logistic regression model based on the EWMA statistic alone has the highest AUC (0.925, 95% CI 0.82–0.97) and identifies 5 of 7 onsets (positive predictive value 38%) at the cost of 31 false-alarm weeks. Random forest has a lower AUC (0.882), but the best Brier score (0.086) and specificity (96%) and at the 0.5 threshold it detects 43% of alert weeks and 0 of 7 onsets. The full-feature logistic model (AUC 0.834) and the neural network (0.767) are not better than the EWMA-only model. The case-history predictors are the least important and account for the remaining 17% of the importance of the random forest (Figure 4). Bootstrap intervals for AUC overlap for all classifiers, so no ranking should be taken as given. Early-origin fits are limited by the two weeks of alert time in the initial training window

Table 5. Prospective one-week-ahead alert prediction (140 evaluation weeks; 21 alert weeks; 7 onsets). AUC and Brier score are threshold-free; sensitivity, specificity and PPV use a 0.5 probability threshold. Intervals: block bootstrap (AUC), exact binomial (others)

Method	AUC (95% CI)	Brier	Sensitivity % (CI)	Specificity % (CI)	PPV % (CI)	Onsets found
Persistence of alert status	0.804	0.100	67 (43–85)	94 (88–98)	67 (43–85)	0/7
Logistic, EWMA statistic only	0.925 (0.82–0.97)	0.148	90 (70–99)	74 (65–82)	38 (25–53)	5/7
Logistic, full feature set	0.834 (0.67–0.95)	0.117	67 (43–85)	88 (81–93)	50 (31–69)	1/7
Random forest	0.882 (0.78–0.96)	0.086	43 (22–66)	96 (90–99)	64 (35–87)	0/7
Neural network (MLP)	0.767 (0.56–0.91)	0.107	24 (8–47)	97 (93–99)	63 (24–91)	0/7

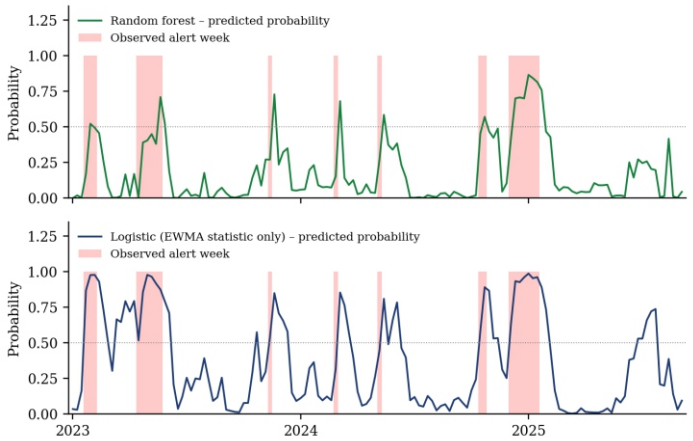


Figure 4. Predicted next-week alert probability from random forest and from logistic regression on the EWMA statistic alone; shaded bars mark observed alert weeks.

5. Discussion

None of the ML methods performed significantly better than the simple statistical benchmarks in the one-week-ahead forecasting of national Lassa fever counts under a leakage-free rolling-origin design with formal inference. This is not in contradiction with the optimistic comparisons found in the literature, and is not due to different targets (e.g., mortality), longer or multi-region series, or weaker baselines (as is often the case in those studies). In this case the naive forecast is already good since the lag-1 autocorrelation of the weekly series is close to 0.88, and most of the more elaborate methods are based primarily on the previous week's count. The NB-AR model provides the smallest errors and, crucially, calibrated-in-principle prediction intervals for counts (which point-forecast ML models do not have).

The reading of the alert-prediction analysis is similar. The benchmark of persistence of alert is a formidable one and the most useful single predictor of next week's alert is the EWMA statistic itself. The addition of case-history predictors to the ML classifiers did not enhance discrimination, and few onsets were detected with the random forest and the neural network at a conventional threshold. Its usefulness is thus that it transforms a noisy count series into an interpretable, calibrated signal, and whether or not ML on top of it adds anything can only be determined after more seasons and, ideally, independent outbreak labels.

This operating threshold was set at 0.5, but a lower threshold would increase sensitivity and increase the number of false alarms, which should be determined in consultation with NCDC based on the relative costs.

6 Conclusion

After a data-quality audit, we used 236 weeks of NCDC data and found that the naïve, MLR and negative binomial forecasts were not distinguishable in terms of RMSE, that the negative binomial model had a nominal advantage in terms of MAE but this did not survive multiplicity adjustment, and that a logistic regression on a seasonally adjusted EWMA statistic predicted next-week alerts as well as random forest or a neural network. Any claim of ML superiority for Lassa fever surveillance should be backed up by comparison with calibrated statistical baselines, and rolling-origin evaluation and inference on differences. NCDC is recommended to pilot the seasonal negative binomial – EWMA signal in addition to the situation reports, and to test it out in the coming seasons and make the weekly series underlying the negative binomial – EWMA signal available in a verified machine-readable format.

References

1. Dalhat, M. M., Olayinka, A., Meremikwu, M. M., et al. (2022). Epidemiological trends of Lassa fever in Nigeria, 2018–2021. *PLoS ONE*, 17, e0279467. https://doi:10.1371/journal.pone.0279467
2. Al-Mustapha, A. I., Adesiyan, I. M., Orum, T. G., et al. (2024). Lassa fever in Nigeria: epidemiology and risk perception. *Scientific Reports*, 14, 27669. https://doi:10.1038/s41598-024-78726-3
3. Eneh, S. C., Obi, C. G., Ikpongifono, U. E., et al. (2025). The resurgence of Lassa fever in Nigeria: economic impact, challenges, and strategic public health interventions. *Frontiers in Public Health*, 13, 1574459. https://doi:10.3389/fpubh.2025.1574459
4. Okoro, O. A., Bamgboye, E., Dan-Nwafor, C., et al. (2020). Descriptive epidemiology of Lassa fever in Nigeria, 2012–2017. *Pan African Medical Journal*, 37, 15. https://doi:10.11604/pamj.2020.37.15.21160
5. Redding, D. W., Gibb, R., Dan-Nwafor, C. C., et al. (2021). Geographical drivers and climate-linked dynamics of Lassa fever in Nigeria. *Nature Communications*, 12, 5759. https://doi:10.1038/s41467-021-25910-y
6. Kane, M. J., Price, N., Scotch, M. and Rabinowitz, P. (2014). Comparison of ARIMA and random forest time series models for prediction of avian influenza H5N1 outbreaks. *BMC Bioinformatics*, 15, 276. https://doi:10.1186/1471-2105-15-276

7. Santangelo, O. E., Gentile, V., Pizzo, S., Giordano, D. and Cedrone, F. (2023). Machine learning and prediction of infectious diseases: a systematic review. *Machine Learning and Knowledge Extraction*, 5, 175–198. https://doi:10.3390/make5010013
8. Dixon, S., Keshavamurthy, R., Farber, D. H., Stevens, A., Pazdernik, K. T. and Charles, L. E. (2022). A comparison of infectious disease forecasting methods across locations, diseases, and time. *Pathogens*, 11, 185. https://doi:10.3390/pathogens11020185
9. Samson, T. K., Akingbade, T. and Orija, J. (2023). Comparative analysis of mortality predictions from Lassa fever in Nigeria: a study using count regression and machine learning methods. *Acadlore Transactions on AI and Machine Learning*, 2(4), 204–211. https://doi:10.56578/ataiml020403
10. Oluwole, A. S. and Nkonyana, T. (2022). Forecasting Lassa fever outbreak progression with machine learning. In *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, 1–5. IEEE. https://doi:10.1109/ICECCME55909.2022.9987787
11. Osaseri, R., Usiobaifo, A. R. and Okhionkpamwonyi, F. O. (2026). AI-driven prediction of Lassa fever using evolutional and random forest: a machine learning approach for enhanced surveillance in West Africa. *NIPES Journal of Energy Technology and Environment*, 8(1), 65–78.
12. Hyndman, R. J. and Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). OTexts, Melbourne.
13. Farrington, C. P., Andrews, N. J., Beale, A. D. and Catchpole, M. A. (1996). A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society A*, 159, 547–563.
14. Noufaily, A., Enki, D. G., Farrington, P., Garthwaite, P., Andrews, N. and Charlett, A. (2013). An improved algorithm for outbreak detection in multiple surveillance systems. *Statistics in Medicine*, 32, 1206–1222.
15. Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, 1, 239–250.
16. Lucas, J. M. and Saccucci, M. S. (1990). Exponentially weighted moving average control schemes: properties and enhancements. *Technometrics*, 32, 1–12.
17. Unkel, S., Farrington, C. P., Garthwaite, P. H., Robertson, C. and Andrews, N. (2012). Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society A*, 175, 49–82.
18. Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16, 437–450.
19. Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14, 199–222.
20. Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
21. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29, 1189–1232.
22. Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22, 679–688.
23. Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13, 253–263.
24. Harvey, D., Leybourne, S. and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13, 281–291.
25. Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17, 1217–1241.
26. Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26, 404–413.

Appendix A Predictors, hyperparameters and variable importance

Predictors common to MLR, SVR, RF, GB and MLP (eight): $y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}$; mean and standard deviation of $y_{t-4}, \ldots, y_{t-1}$; $\sin(2\pi w_t/52)$; $\cos(2\pi w_t/52)$. NB-AR uses $\log(1 + y_{t-k})$, $k = 1, \ldots, 4$, and one harmonic pair.

Table A1: Hyperparameter grids searched on the initial window (selected values in Section 3.4).

Method	Grid
SVR	$C \in \{1, 10, 50\}$; $\gamma \in \{0.05, 0.17, 0.5\}$; $\varepsilon = 1$
Random forest	300 trees; depth $\in \{3, 6\}$; minimum leaf $\in \{1, 3\}$
Gradient boosting	trees $\in \{100, 300\}$; depth $\in \{2, 3\}$; learning rate 0.05; subsample 0.8
MLP	hidden layers $\in \{(8), (16, 8)\}$; L2 penalty $\in \{0.001, 0.1, 1\}$; L-BFGS; 3 seeds in tuning, 5 in evaluation
ARIMA	$p \le 3$, $q \le 2$, $d \in \{0, 1\}$ by AIC (conditional sum of squares)

Table A2: Random forest impurity importance (fit to all data, for illustration only)

Predictor	Regression importance	Alert-classifier importance
lag1	0.887	0.072
lag2	0.011	0.033
lag3	0.004	0.029
lag4	0.009	0.039
mean4	0.006	0.067
sd4	0.025	0.057
sin	0.025	0.042
cos	0.033	0.073
Z_{t-1}	–	0.419
a_{t-1}	–	0.169