Original Research Article

Four-Week Dengue Outbreak Risk Classification Using Structured Public Health Data

Gautama Narula

Inspirit AI, 335 Bryant Street, Palo Alto, California 94301, United States

ABSTRACT

Background: This study evaluates a supervised classification approach for estimating short-horizon dengue outbreak risk using weekly country-level case-count data. Methods: Weekly country-level dengue case counts were extracted from the OpenDengue national dataset and merged with country-level demographic, health-system, and Water, Sanitation and Hygiene (WASH) covariates. A location-week was labeled positive when the following four weeks met a rule-based elevated-risk definition: forward incidence of at least 100 cases per 100,000 population, or forward cases at least 50% above the prior four-week baseline with at least 20 absolute cases. All predictors were restricted to current or prior information. A temporal split was used: training observations occurred before January 1, 2020; validation observations occurred during 2020; and test observations occurred from January 1, 2021 through the final available observation. RandomForestClassifier and GradientBoostingClassifier models were implemented in scikit-learn and calibrated with Platt scaling on the validation set. Results: After preprocessing, the modeling dataset contained 19,456 country-week observations from 47 countries covering 1924-04-06 through 2022-12-25. The final model was RandomForestClassifier. On the held-out temporal test set (n=3,703; event rate 14.0%), the model achieved AUROC 0.860 (95% CI 0.844-0.875), AUPRC 0.549 (95% CI 0.511-0.590), Brier score 0.090, ECE 0.038, and precision@top-10% 0.589. Conclusions: This study reports a 4-week dengue outbreak-risk classification analysis using temporally separated evaluation and calibrated probabilities. The findings support the feasibility of a leakage-aware, calibrated 4-week probability-estimation pipeline for dengue surveillance, with a 1-100 display scale intended only for prioritization. The study does not claim validation for COVID-19 or influenza without appropriate disease-specific outcome data.

Keywords: dengue; outbreak prediction; machine learning; public health surveillance; calibration; risk scoring

INTRODUCTION

Public health surveillance increasingly depends on timely digital and structured data streams. Digital epidemiology can provide local and time-sensitive signals about disease dynamics, but these signals require careful validation before they can support operational decisions (1). Short-horizon epidemiological models can be useful planning tools, yet COVID-19 forecasting work has also shown that apparently strong model fits may fail when reporting practices, testing behavior, or disease

Corresponding author: Gautama Narula, E-mail: gnarul123@gmail.com.

Accepted June 26, 2026
https://doi.org/10.70251/HYJR2348.43743748

regimes change (2, 3).

This study predicts whether a country-week will be followed by an elevated dengue outbreak-risk period within four weeks. Dengue was selected because the available data included country-week dengue case counts suitable for supervised labeling.

The study objective was to build and evaluate a leakage-aware classifier that maps structured surveillance and contextual variables to a calibrated probability, optionally displayed on a 1-100 scale for operational prioritization. The 1-100 score is a direct display transformation of probability rather than an additional predictive contribution. The main contributions are a transparent 4-week outbreak label, a temporal validation design, calibrated probability estimates, and interpretable feature-importance summaries. Natural Language Processing (NLP) and SHapley Additive exPlanations (SHAP) were not used in the final reported experiment because the project archive did not contain a reproducible NLP or SHAP pipeline. Claims based on those components were removed rather than inferred.

METHODS AND MATERIALS

Study design and data sources

This was a retrospective supervised machine-learning study at country-week granularity. Dengue case data came from the OpenDengue national extract. OpenDengue is a standardized global database of dengue case counts compiled from public sources, and the version 1.2 publication describes 102 countries and records spanning 1924 to 2023 (4). Country-level contextual variables were drawn from the World Bank-style demographic file in the project archive and WASH indicators were drawn from WHO/UNICEF Joint Monitoring Programme datasets, which provide estimates for drinking water, sanitation, and hygiene indicators (5, 6).

The final analysis used weekly records with case_definition_standardised equal to Total and T_res equal to Week. Rows were linked to population and country-level covariates after country-name normalization. Country-weeks without a usable population denominator were excluded because the primary outbreak definition required incidence per 100,000 population.

Outcome definition

For each country-week, the binary outcome represented whether an elevated-risk period occurred during the next four weeks. A positive label was assigned when either 1) the forward four-week incidence was at least 100 dengue cases per 100,000 population, or 2) the forward four-week case count was at least 50% higher than the prior four-week baseline and included at least 20 cases. The prior baseline used only weeks before the prediction point. Forward cases were used only to construct the outcome label and were excluded from all predictor variables.

Predictor variables and leakage prevention

Predictors included current dengue cases and incidence, 1-, 2-, 4-, 8-, and 12-week lagged cases and incidence, rolling means and rolling sums over prior 4-, 8-, and 12-week windows, calendar month, ISO week of year, population, population density, health-system and demographic covariates, and WASH indicators. Target-derived and future variables were explicitly excluded, including future four-week cases, future incidence, and the outbreak label. Missing predictor values were imputed using the median within the training pipeline.

Preprocessing and model fitting were performed inside scikit-learn pipelines so that imputation and scaling were learned from training data and then applied to validation and test data (7). The leakage audit file generated with the revised code lists the exact predictors used in the final model.

Train/validation/test split

A forecasting-style temporal split was used. Training observations included all eligible records before January 1, 2020 (n=13,340; event rate 17.6%). The validation set included observations from January 1, 2020 to December 31, 2020 (n=2,413; event rate 15.9%). The held-out test set included observations from January 1, 2021 through the final available observation (n=3,703; event rate 14.0%). This design prevented random mixing of future observations into training.

Models, calibration, and software

Two scikit-learn classifiers were evaluated: RandomForestClassifier and GradientBoostingClassifier. The RandomForestClassifier used 150 trees, maximum depth of 8, minimum leaf size of 12, and balanced subsampling class weights. The GradientBoostingClassifier used 120 estimators, learning rate 0.05, maximum depth of 2, and minimum leaf size of 10. Random seeds were fixed at 42. The analysis was run in Python with pandas, numpy, matplotlib, joblib, and scikit-learn 1.8.0.

Post-hoc calibration was performed with Platt scaling fitted on validation-set predicted probabilities only (8).

Calibrated probabilities were mapped to a 1-100 display score using risk score = 1 + 99 x calibrated probability. The score is a monotonic linear transformation of the calibrated probability and therefore adds no predictive information. It is retained only as an operational display aid: public-health users may sort country-weeks into a surveillance-review queue or allocate limited review capacity by score, while the calibrated probability remains the primary quantitative estimate. For transparency, all discrimination and calibration analyses are reported using probabilities rather than the score.

Evaluation and uncertainty

Performance was evaluated on the held-out temporal test set using the Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), Brier score, Expected Calibration Error (ECE), precision at the top 10% of predicted risk, and precision at a 0.5 probability threshold. Brier score was included because it evaluates probabilistic accuracy (9). AUPRC and precision@top-10% were emphasized because outbreak labels were imbalanced and operational surveillance often prioritizes the highest-risk locations (10). Nonparametric bootstrap intervals with 500 resamples were computed for AUROC, AUPRC, and Brier score.

RESULTS

Dataset construction

The raw dengue extract contained 31,032 rows. Filtering to weekly total case records yielded 20,944 rows before the population-denominator filter. After country matching, population filtering, feature construction, and removal of rows without sufficient prior or forward windows, the final modeling dataset included 19,456 observations from 47 countries. The final date range was 1924-04-06 to 2022-12-25. The model used 41 predictor variables after the leakage audit.

Model performance

RandomForestClassifier had the best AUPRC on the temporal test set and was selected as the final model. The results are shown in Table 1. The evaluation no longer reports R2 or mean absolute error because the revised experiment is a binary outbreak-classification task, not a vaccination regression task.

Operational interpretation of the risk score: The 1-100 score is not a separate model output or a new predictive measure. It is a display representation of the calibrated probability, intended only to make priority ordering easier in a surveillance workflow. For example, higher scores identify country-weeks that could be reviewed first when public-health attention is limited; model performance is evaluated solely with calibrated probabilities and the metrics reported in Table 1.

Uncertainty estimates

For the final RandomForestClassifier, the AUROC was 0.860 with a 95% bootstrap confidence interval of 0.844 to 0.875. The AUPRC was 0.549 with a 95% confidence interval of 0.511 to 0.590. The Brier score was 0.090 with a 95% confidence interval of 0.085 to 0.096.

Discrimination and Calibration Performance

Figures 1-3 provide the visual counterparts to the held-out temporal test metrics. The receiver operating characteristic curve shows ranking performance

Table 1. *Held-out temporal test-set performance for 4-week dengue outbreak classification. AUROC and AUPRC summarize discrimination; Brier score and expected calibration error summarize probabilistic calibration; the final two rows report precision for operational review rules. Values are reported separately for the two evaluated classifiers.*

Metric	Gradient boosting classifier	Random forest classifier
AUROC	0.850	0.860
AUPRC	0.532	0.549
Brier score	0.093	0.090
Expected calibration error	0.040	0.038
Precision among highest-risk 10%	0.541	0.589
Precision at probability threshold 0.50	0.871	0.682

across thresholds (Figure 1), the precision-recall curve emphasizes performance under the 14.0% event rate (Figure 2), and the calibration curve compares predicted probabilities with observed outbreak frequencies (Figure 3). Together, these figures contextualize the AUROC, AUPRC, Brier score, and expected calibration error reported above.

Feature-importance Analysis

Feature importance was evaluated as part of the Results rather than as a separate contribution. The most important predictors in the final model were incidence_per_100k, cases_rollmean_12w, cases_rollsum_8w, cases_rollsum_12w, cases, cases_rollmean_8w, incidence_lag_8w, incidence_lag_2w, cases_lag_8w, and weekofyear. These variables mainly represented current incidence, recent rolling case burden, lagged incidence, and seasonality. Figure 4 displays the fitted RandomForestClassifier feature-importance ranking. The importances are descriptive model summaries and should not be interpreted causally. SHAP analyses are not reported because the project archive did not contain generated SHAP outputs.

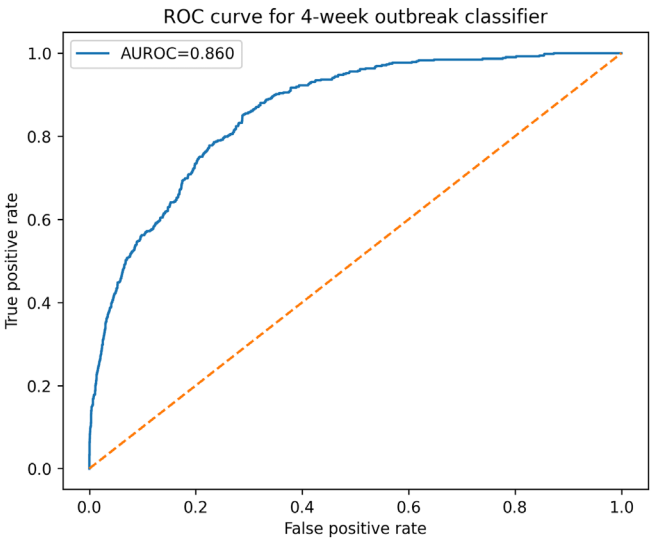

Figure 1. *Receiver operating characteristic curve for the final 4-week dengue outbreak classifier on the held-out temporal test set. The curve shows sensitivity against 1-specificity across decision thresholds; the dashed diagonal line represents random ranking. The observed AUROC is 0.860.*

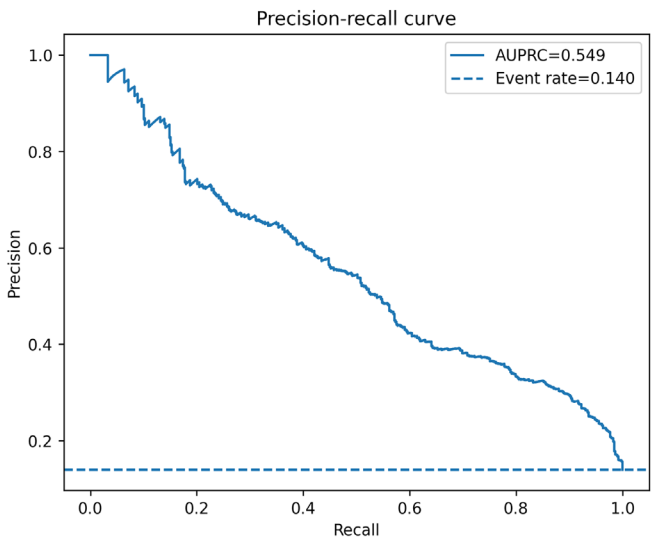

Figure 2. *Precision-recall curve for the final classifier on the held-out temporal test set. Precision is the proportion of predicted positive observations that were true outbreak events, and recall is the proportion of all outbreak events identified. The dashed horizontal line marks the 14.0% test-set event rate, the expected precision under random ranking.*

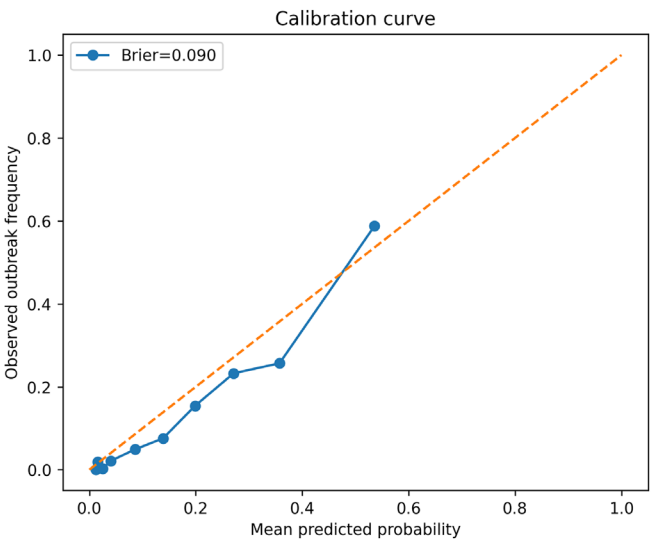

Figure 3. *Calibration curve for the final classifier on the held-out temporal test set. Points compare mean predicted outbreak probability with observed outbreak frequency across probability bins; the diagonal line represents perfect calibration. The final model had a Brier score of 0.090 and expected calibration error of 0.038.*

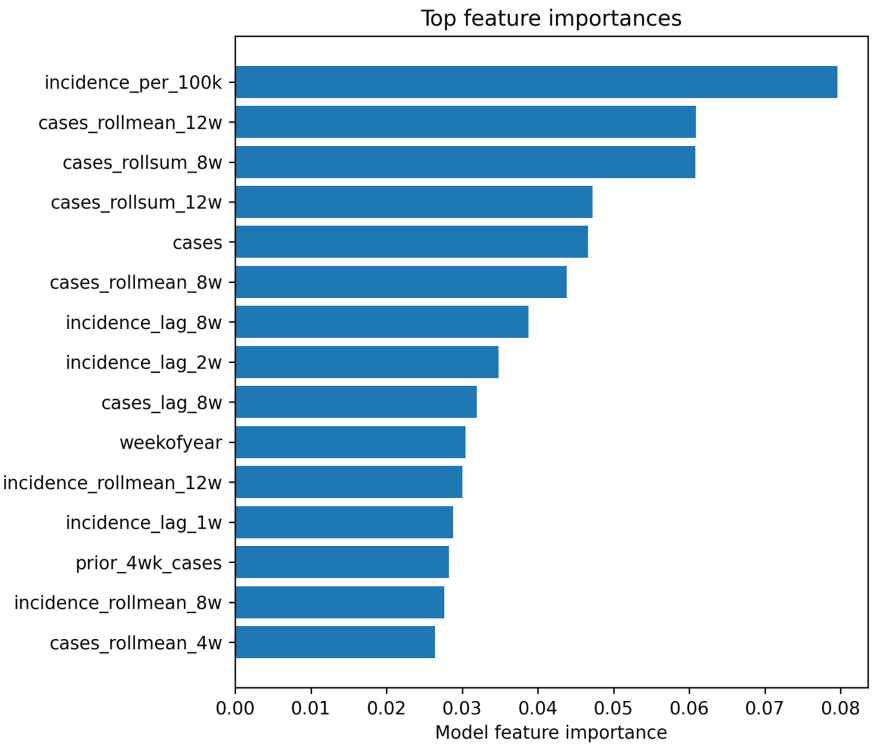

Figure 4. *Feature-importance ranking from the final RandomForestClassifier fitted on the training data and evaluated on the held-out temporal test set. Larger values indicate variables used more strongly in fitted model splits. Current and recent dengue-burden variables dominate the ranking; these associations are descriptive rather than causal.*

DISCUSSION

This study demonstrates a coherent 4-week outbreak-risk classification pipeline. The outcome is a binary forward-looking outbreak-risk label, and performance is evaluated using classification and calibration metrics. The conclusions are limited to dengue and the dataset analyzed. The 1-100 risk score is not treated as a novel predictive output; it is only a transparent display transformation of the calibrated probability for prioritization.

The final model showed useful ranking ability on a temporally later test set, with AUROC 0.860 and AUPRC 0.549. Because the test-set event rate was 14.0%, the AUPRC indicates improvement over random ranking. Precision@top-10% of 0.589 suggests that the highest-ranked observations were enriched for subsequent outbreak events. The Brier score of 0.090 and ECE of 0.038 indicate that calibration was imperfect but substantially reported and measurable.

These results apply to dengue and should not be generalized to COVID-19 or influenza without disease-specific outcome data and independent validation.

Several limitations remain. First, country-level case reporting practices vary over time, and backfills or surveillance changes may affect labels. Second, the model used national-level features and cannot capture subnational heterogeneity. Third, static covariates such as population density and WASH indicators may not reflect rapid local changes. Fourth, although temporal validation was used, a full leave-one-location-out evaluation was not conducted and should be included in future work. Fifth, feature importance is not causal explanation.

Future work should add disease-specific COVID-19 or influenza case outcomes only if consistent location-time case data are available, repeat the full leakage audit for those diseases, report external geographic validation, and include reproducible text-processing steps before reintroducing NLP-derived features.

CONCLUSION

This study developed and evaluated a leakage-aware 4-week dengue outbreak-risk classifier using structured surveillance, demographic, health-system, and WASH data. The study uses a supervised classification design with temporal validation, calibrated probabilities, classification metrics, uncertainty intervals, and figures integrated into the Results. The accompanying 1-100 score is a non-novel display scale derived directly from calibrated probability and is intended only to support prioritization. The work supports the feasibility of a transparent outbreak-prioritization pipeline for dengue surveillance; external validation and disease-specific data are needed before broader operational use.

ACKNOWLEDGEMENTS

The author acknowledges Varsha Sandadi for guidance during the writing process and Eashan Sinha for assistance with code and analysis. Any remaining errors are the responsibility of the author.

FUNDING SOURCES

The author received no specific funding for the conduct of this research or preparation of this manuscript.

CONFLICT OF INTEREST

The author declares no conflicts of interest related to this work.

DATA AND CODE AVAILABILITY

Code and derived analysis outputs are available from the corresponding author upon reasonable request. The analysis relies on the public datasets cited in the References.

REFERENCES

1. Salathe M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, *et al*. Digital epidemiology. *PLoS Comput Biol*. 2012; 8 (7): e1002616. doi:10.1371/journal.pcbi.1002616.
2. Holmdahl I, Buckee C. Wrong but useful - what COVID-19 epidemiologic models can and cannot tell us. *N Engl J Med*. 2020; 383 (4): 303-305. doi:10.1056/NEJMp2016822.
3. Petropoulos F, Makridakis S. Forecasting the novel coronavirus COVID-19. *PLoS One*. 2020; 15 (3): e0231236. doi:10.1371/journal.pone.0231236.
4. Clarke J, Lim A, Gupte P, Pigott DM, van Panhuis WG, Brady OJ. A global dataset of publicly available dengue case count data. *Sci Data*. 2024; 11 (1): 296. doi:10.1038/s41597-024-03120-7.
5. WHO/UNICEF Joint Monitoring Programme for Water Supply, Sanitation and Hygiene. Data: household drinking water, sanitation and hygiene estimates [Internet]. Geneva: WHO/UNICEF. Available from: https://washdata.org/data (accessed on 2026-06-15).
6. World Bank. World Bank Open Data [Internet]. Washington (DC): World Bank. Available from: https://data.worldbank.org/ (accessed on 2026-06-15).
7. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al*. Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011; 12: 2825-2830.
8. Platt JC. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola AJ, Bartlett P, Scholkopf B, Schuurmans D, editors. Advances in large margin classifiers. Cambridge (MA): MIT Press; 1999. p. 61-74.
9. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev.* 1950; 78 (1): 1-3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
10. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning; 2006; Pittsburgh, PA. New York: ACM. 2006; p.233-240. https://doi.org/10.1145/1143844.1143874