

Temporal Modeling of Weekly Dengue Cases in Marikina City: A Time-Respecting Comparison of Classical, Machine Learning, and Deep Temporal Models

Kal-el John C. Bautista

Mapúa University - Makati

Luis Arnold N. Respecio

lanrespecio@mymail.mapua.edu.ph

Mapúa University - Makati

Cyril Troy C. Vinuya

Mapúa University - Makati

John Paul Q. Tomas

Mapúa University - Makati

Karl C. Ablola

Mapúa University - Makati

Bonifacio T. Doma

Mapúa University - Makati

Research Article

Keywords: dengue forecasting, public health surveillance, time series, ARIMA, LSTM, machine learning, temporal validation, Marikina City

Posted Date: August 31st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10307931/v1>

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

Additional Declarations: No competing interests reported.

Abstract

Background

Short-horizon dengue forecasting can support local preparedness, but model comparisons may be weakened by random splitting, information leakage, limited baseline testing, and insufficient out-of-time evaluation. This study developed a surveillance-only framework for forecasting weekly dengue incidence in Marikina City two weeks ahead.

Methods

Weekly records combined Marikina-filtered Department of Health–Epidemiology Bureau data from 2016–2020 with Marikina Epidemiology Surveillance Unit data from 2023–2025. The years 2021–2022 were excluded because the available records contained zero and three reported cases, respectively. Eight trained models—ARIMA, SARIMA, ETS, Decision Tree, Random Forest, Gradient Boosting, LSTM, and Temporal CNN—were compared with naïve and moving-average reference baselines. Target-year splits used 2024 for internal validation and model selection and 2025 as an out-of-time temporal holdout. Earlier observed 2025 case counts were permitted as causal inputs for forecasting later 2025 targets when they occurred before the relevant forecast origin. Mean absolute error was the primary metric, supported by root mean square error, coefficient of determination, feature importance, and incidence-class error analysis.

Results

The cleaned dataset contained 417 weekly observations and 8,263 reported cases. ARIMA was the strongest trained model during 2024 internal validation, although the two-week moving average and naïve baselines produced slightly lower errors. In the 2025 temporal holdout, LSTM achieved the most favorable metrics, with an MAE of 7.951, RMSE of 10.651, and R^2 of 0.698, followed by ARIMA with an MAE of 8.652, RMSE of 11.686, and R^2 of 0.636. Both achieved lower MAE than the naïve baseline. Tree-based models relied primarily on current incidence and two-week rolling summaries. All evaluated methods tended to underpredict outbreak-level weeks.

Conclusions

Surveillance-only forecasting was feasible under a leakage-safe temporal design. LSTM produced the strongest 2025 holdout metrics, while ARIMA remained an interpretable and computationally efficient alternative. Operational use should retain simple baselines, communicate uncertainty, and prioritize improved detection of rapid outbreak escalation.

1 BACKGROUND

Dengue remains a persistent public health concern in tropical and subtropical settings, where seasonal increases can rapidly strain clinical services, vector-control operations, and local surveillance teams [1, 3]. In densely populated urban areas, short-term increases in reported cases can affect staffing, bed capacity, diagnostics, medical supplies, and the timing of community interventions. City-level forecasts are therefore

potentially more actionable than national or regional summaries because local government units make many operational decisions at the city and barangay levels.

Routine dengue surveillance produces chronologically ordered counts that contain temporal dependence. Recent incidence may represent ongoing transmission chains, reporting momentum, and the early development of outbreak waves [15, 16]. Classical time-series models capture this dependence directly through autoregressive, moving-average, trend, and seasonal components [4, 8, 19]. Machine-learning approaches instead transform the series into lag, rolling, change, trend, and calendar predictors, while deep temporal models learn representations from ordered case sequences [11–14, 17].

Forecasting credibility depends as much on experimental design as on algorithm choice. Random train-test splits violate temporal order, and preprocessing performed before the forecast cutoff may leak future information into model inputs. Time-respecting validation, causal feature construction, and comparison with simple persistence baselines are necessary to determine whether a trained model offers genuine predictive value [5–7]. These safeguards are particularly important for weekly dengue data, where adjacent observations are strongly correlated and naive forecasts can be difficult to beat.

Existing dengue studies have demonstrated the value of autoregressive methods, tree ensembles, and recurrent neural networks, but performance varies across locations, forecast horizons, covariates, and evaluation designs [5, 13, 17, 20]. Local surveillance agencies may also lack consistently available meteorological, mobility, socioeconomic, or environmental variables. A surveillance-only design based on case counts and reporting dates can therefore provide a practical, reproducible baseline for data-constrained settings, although it cannot explain causal transmission mechanisms.

This study developed and evaluated a leakage-safe, two-week-ahead forecasting framework for weekly dengue incidence in Marikina City. The objectives were to: (1) characterize trend, seasonality, autocorrelation, and lag behavior; (2) compare classical, machine-learning, and deep temporal models; (3) determine whether trained models improve on naive and moving-average baselines; and (4) identify the temporal predictors most strongly used by tree-based models. The principal contribution is a complete target-year evaluation design with 2024 reserved for internal model development and 2025 retained as a full unseen external holdout.

2 METHODS

2.1 Study Design and Setting

The study used a quantitative predictive design focused on weekly dengue incidence in Marikina City, Philippines. The task was univariate, two-week-ahead regression: information observed at forecast origin t was used to predict the weekly dengue case count at $t + 2$. The study evaluated predictive accuracy rather than causal effects. No individual-level or personally identifiable health records were analyzed.

2.2 Data Sources and Eligibility

The unified dataset combined two surveillance sources. Department of Health-Epidemiology Bureau records were filtered to retain only Marikina City observations from 2016–2020 [2]. More recent weekly records for 2023–2025 were obtained from the Marikina Epidemiology Surveillance Unit (MESU). The available 2021 record

contained zero cases and the 2022 record contained only three cases; both years were excluded because they were too sparse to support reliable lag, rolling-window, and sequence modeling.

The cleaned series contained 417 weekly records and 8,263 reported cases. The training and development period included 313 weeks from 2016–2020 and 2023, with 4,999 cases. The 2024 internal-validation year contained 52 weeks and 1,867 cases. The 2025 external holdout contained 52 weeks and 1,397 cases. All three sets had no remaining duplicate dates or missing case values. The 2025 holdout included 15 low, 14 moderate, 16 high, and seven very high or outbreak-level weeks, providing a broader test than a partial-year holdout. Table 1 summarizes the resulting dataset structure and data quality by period.

Table 1
Dataset structure and data quality

Period	Years	Weeks	Cases	Use
Historical segment 1	2016–2020	261	3,468	Training/development
Excluded interval	2021–2022	-	0 and 3	Excluded for insufficient reporting
Historical segment 2	2023	52	1,531	Training/development
Internal validation	2024	52	1,867	Model and hyperparameter selection
External holdout	2025	52	1,397	Final unseen evaluation
Note. The training/development total was 313 weeks and 4,999 cases. Dashes indicate that the excluded records were not treated as a modeling segment.				

2.3 Data Preparation and Temporal Segmentation

Dates were converted to a standard datetime representation, case counts were converted to numeric form, invalid dates and missing targets were removed, negative values were bounded at zero, and records were sorted chronologically. The median interval between observations was checked to verify reporting frequency; the uploaded unified dataset was already weekly, so no additional aggregation was applied. Outbreak peaks were retained because they represented epidemiologically meaningful events rather than errors.

A gap longer than 21 days defined a new continuous segment. This prevented lags, rolling statistics, expanding seasonal means, and deep-learning sequences from crossing the excluded 2021–2022 interval. Consequently, the last observation in 2020 was not treated as immediately adjacent to the first observation in 2023. Classical-model input was handled differently: the available observations before each cutoff were chronologically concatenated without explicit missing calendar positions. This did not introduce future targets, but it may have affected estimates of autocorrelation and seasonality and is treated as a limitation.

2.4 Exploratory Temporal Analysis and Feature Engineering

The autocorrelation function (ACF), partial autocorrelation function (PACF), monthly and weekly seasonal profiles, and Seasonal-Trend decomposition using Loess (STL) were used to examine short-range persistence and annual structure [4, 19]. Because the strongest dependence occurred at early lags and annual recurrence was irregular, the machine-learning feature set emphasized recent history rather than direct 26- or 52-week lags.

For Decision Tree, Random Forest, and Gradient Boosting, each forecast origin included the current case count; lags at 1, 2, 4, 8, and 12 weeks; rolling means, sums, and standard deviations over 2, 4, 8, and 12 weeks; exponentially weighted means with spans of 4 and 8 weeks; one- and two-week differences; a bounded one-week percentage change; four- and eight-week local trend slopes; sine and cosine encodings of epidemiological week and month; a segment-local trend index; and expanding historical means for the same week and month. All transformations used only information available at or before the forecast origin. The supervised target was the case count two weeks later.

LSTM and Temporal CNN inputs were constructed separately as segment-aware 12-week case sequences paired with the target at $t + 2$. ARIMA, SARIMA, and ETS operated on the chronological case-count series and did not use the engineered feature matrix.

2.5 Forecasting Models

Five reference baselines were calculated without training: a naive forecast equal to the current observed case count and moving averages over the previous 2, 4, 8, and 12 weeks. Baselines were used only to establish the minimum practical standard and were not candidates for final trained-model selection.

The classical models were ARIMA, SARIMA, and exponential smoothing (ETS). Candidate ARIMA orders (1,0,0), (2,0,0), and (3,0,0) were compared by 2024 rolling-origin MAE; ARIMA(3,0,0) was selected. SARIMA used order (1,0,0) with seasonal order (1,0,0,52). ETS used additive trend and additive seasonality with a 52-week period. ARIMA was repeatedly refitted at forecast origins, whereas SARIMA and ETS generated static full-horizon forecasts after fitting before the evaluation year [8–10].

The tree-based regressors were Decision Tree, Random Forest, and Gradient Boosting. Candidate settings varied tree depth, leaf size, number of estimators, feature subsampling, learning rate, and subsampling ratio. Each candidate was evaluated with raw and log1p-transformed targets, and the configuration with the lowest 2024 MAE was selected. Random Forest reduces variance through bootstrap aggregation [11], while Gradient Boosting sequentially corrects residual error [12].

The LSTM used one recurrent layer with 32 hidden units followed by a linear output layer. The Temporal CNN used two one-dimensional convolutional layers with 32 channels per layer, kernel size 3, ReLU activations, adaptive average pooling, and a linear output. It is described as a Temporal CNN rather than a canonical dilated residual TCN because the implementation did not include explicit dilation or residual TCN blocks [14, 21]. Both neural models used training-derived standardization, Adam optimization with learning rate 0.001 and weight decay 0.0001, SmoothL1 loss, a maximum of 150 epochs, and early stopping with patience 12. Table 2 summarizes the input, selected configuration, and evaluation procedure for each trained model.

Table 2
Trained models and evaluation procedures

Model	Input and selected setup	Evaluation procedure
ARIMA	Weekly series; selected order (3,0,0)	Rolling-origin refitting in 2024 and 2025
SARIMA	Weekly series; (1,0,0)(1,0,0,52)	Static full-horizon forecast
ETS	Additive trend and seasonality; period 52	Static full-horizon forecast
Decision Tree	Engineered features; depth 3; leaf 5; log target	Target-year validation and refit through 2024
Random Forest	200 trees; depth 5; max features 0.7; leaf 5	Target-year validation and refit through 2024
Gradient Boosting	100 trees; learning rate 0.05; depth 2	Target-year validation and refit through 2024
LSTM	12-week sequence; 32 hidden units	Early stopping on 2024; final fit through 2024
Temporal CNN	12-week sequence; two Conv1D layers	Early stopping on 2024; final fit through 2024
Note. Different forecast-generation procedures were retained by model family and are considered when interpreting direct comparisons.		

2.6 Chronological Validation

Observations were assigned according to the year of the forecast target rather than the year of the forecast origin. Targets occurring in 2023 or earlier formed the training and development set, targets occurring in 2024 formed the internal-validation set, and targets occurring in 2025 formed the out-of-time temporal holdout. This target-year assignment prevented forecasts originating near the end of one year from placing a target from the following year into an earlier split.

The 2025 target values were excluded from ARIMA-order selection, machine-learning hyperparameter tuning, and initial neural-model validation. During the 2025 temporal holdout evaluation, earlier observed 2025 case counts could be used as historical inputs for later 2025 forecasts only when those observations occurred before the relevant forecast origin. Therefore, the 2025 targets remained held out from model development while forecasts were allowed to use causally available surveillance history.

For the neural models, the 2024 validation period was used to select the architecture, training configuration, and number of training epochs. After these settings were fixed, the selected models were refitted using all eligible sequences through 2024 for the predetermined number of epochs without monitoring the 2025 holdout. The 2025 targets were then used only for final temporal evaluation.

2.7 Performance Measures and Interpretation

Mean absolute error was treated as the primary operational metric because it directly represented the average weekly case-count error. Root mean square error was used to identify models with larger individual forecasting errors, while R^2 provided supplementary information regarding variance explanation. Model conclusions were

based primarily on MAE and RMSE rather than on a custom percentage-accuracy measure. This evaluation approach follows established principles for assessing point forecasts using proper scoring rules [18].

Native impurity-based feature importance was extracted from the selected tree-based models. These scores describe how strongly predictors were used within each fitted model but do not represent causal relationships or permutation-based importance. Errors during the 2025 temporal holdout were also summarized according to incidence level.

All analyses were performed in Python using pandas and NumPy for data handling, statsmodels for classical models, scikit-learn for tree regressors and evaluation, and PyTorch for neural models. Fixed random seeds and explicit configurations were used where supported.

3 RESULTS

3.1 Data Characteristics and Temporal Structure

Annual totals varied substantially. The highest observed total was 1,867 cases in 2024, followed by 1,531 in 2023 and 1,397 in 2025. Weekly incidence included low-transmission periods and abrupt peaks, indicating that the series was not governed by a uniform annual trajectory. The feature-generation process produced 292 model-ready supervised observations: 189 with targets through 2023, 51 for 2024, and 52 for 2025. The deep-learning sequence dataset retained 52 targets in each evaluation year. Figure 1 shows the resulting weekly incidence series.

Figure description: A weekly line plot covering 2016–2020 and 2023–2025 shows several irregular peaks, a gap for excluded 2021–2022 data, high activity in 2023–2024, and a lower but still variable full-year pattern in 2025.

The ACF showed strong positive dependence at early lags, declining gradually across the following weeks. The PACF concentrated its strongest direct effects at lags 1 and 2. A weaker secondary ACF rise appeared near the annual range, but it was less pronounced than short-range persistence. These findings supported the reduced lag set and the emphasis on current incidence and rolling summaries. Figure 2 presents the ACF and PACF diagnostics described above.

Figure description: Panel a shows high short-lag autocorrelation that gradually declines and a smaller rise near the annual range. Panel b shows the largest partial autocorrelation at the first two lags, with smaller values thereafter. Both diagnostics use the concatenated observed sequence after excluding 2021–2022.

STL decomposition of the latest continuous segment identified a changing trend, recurring fluctuations, and substantial residual spikes. The seasonal component was present but not sufficiently stable to justify treating annual recurrence as the dominant signal. The residual component remained large during high-incidence periods, suggesting that outbreak peaks could not be explained by smooth trend and seasonality alone. Figure 3 shows the STL decomposition described above.

Figure description: Four aligned panels display the observed weekly series, long-term trend, seasonal component, and residual component. The seasonal pattern is visible but irregular, while several high-incidence spikes remain in the residuals.

3.2 Internal Validation in 2024

ARIMA was the strongest of the eight trained models in the 2024 internal-validation period, with MAE 11.789, RMSE 15.028, and R-squared 0.680. Gradient Boosting ranked second by MAE, followed by LSTM and Random Forest. Temporal CNN was weaker, and the static SARIMA and ETS forecasts produced negative R-squared values. Table 3 reports the full 2024 internal-validation ranking.

Table 3
Internal-validation performance in 2024

Rank	Model	n	MAE	RMSE	R ²
1	ARIMA	51	11.789	15.028	0.680
2	Gradient Boosting	51	13.285	17.619	0.561
3	LSTM	52	14.031	18.114	0.531
4	Random Forest	51	14.265	18.572	0.512
5	Decision Tree	51	14.961	20.172	0.424
6	Temporal CNN	52	18.818	22.281	0.291
7	ETS	51	25.966	31.247	-0.381
8	SARIMA	51	29.234	38.951	-1.146
Note. ARIMA, tree models, and baselines had 51 feature-aligned targets; deep models used 52 sequence targets. Baselines were not ranked as trained models.					

The reference baselines were slightly stronger than ARIMA internally. The two-week moving average achieved MAE 11.480, RMSE 14.801, and R-squared 0.690, while the naive current-case baseline achieved MAE 11.510. Thus, ARIMA was the best trained model but did not demonstrate added value beyond the strongest simple recent-history benchmarks during model development. Figure 4 plots the corresponding 2024 internal-validation forecasts.

Figure description: Observed weekly dengue counts are compared with the top trained models and reference baselines. The curves follow the broad annual decline and later increase, but differ around abrupt weekly peaks.

3.3 External Validation in 2025

The external holdout produced a different ranking. LSTM achieved the lowest MAE and RMSE and the highest R-squared: MAE 7.951, RMSE 10.651, and R-squared 0.698. ARIMA ranked second with MAE 8.652, RMSE 11.686, and R-squared 0.636. Decision Tree and Random Forest provided moderate performance, while Gradient Boosting and Temporal CNN generalized less strongly. Static SARIMA and ETS again produced negative R-squared values. Table 4 reports the full 2025 external-holdout ranking.

Table 4
External-holdout performance in 2025

Rank	Model	Family	n	MAE	RMSE	R ²
1	LSTM	Deep	52	7.951	10.651	0.698
2	ARIMA	Classical	52	8.652	11.686	0.636
3	Decision Tree	Machine learning	52	10.841	14.286	0.456
4	Random Forest	Machine learning	52	10.869	15.208	0.383
5	Gradient Boosting	Machine learning	52	12.715	17.415	0.191
6	Temporal CNN	Deep	52	15.279	18.398	0.098
7	SARIMA	Classical	52	17.873	24.753	-0.634
8	ETS	Classical	52	26.794	31.844	-1.704
Note. The table ranks only trained models. The 2025 targets were excluded from model selection and hyperparameter tuning.						

Both leading trained models improved on the external baselines by MAE. The naive current-case baseline achieved MAE 8.846, and the two-week moving-average baseline achieved MAE 9.298. LSTM reduced MAE by 0.895 cases relative to the naive baseline; ARIMA reduced it by 0.194 cases. The improvement was numerically clear for LSTM and small for ARIMA, and no formal forecast-error significance test was conducted. Table 5 compares the leading trained models with the external baselines.

Table 5
Leading models versus external baselines

Method	Type	MAE	Mean weekly accuracy (%)
LSTM	Trained model	52	7.951
ARIMA	Trained model	52	8.652
Naïve current cases	Reference baseline	52	8.846
Moving average 2	Reference baseline	52	9.298
Note. No formal paired forecast-error test was conducted; therefore, these results indicate numerical rather than statistically confirmed superiority. Figure 5 shows the corresponding observed and forecast series for these methods.			

Figure description: Observed 2025 weekly cases are shown with ARIMA, LSTM, naive, and moving-average forecasts. The selected methods capture the overall decline, later-year rise, and final decrease, but smooth or miss some abrupt peaks.

3.4 Temporal Predictor Importance

The selected Decision Tree relied primarily on the current observed case count, which received an importance score of 0.727. Random Forest assigned nearly equal importance to the two-week rolling sum (0.289) and rolling mean (0.284), followed by current cases (0.092). Gradient Boosting ranked the two-week rolling mean first

(0.339), the two-week rolling sum second (0.182), and current cases third (0.100). Historical same-week and same-month means contributed secondary seasonal context. Table 6 lists the highest-ranked predictors for each tree-based model.

Table 6
Highest tree-based temporal predictor scores

Model	Rank	Predictor	Importance
Decision Tree	1	current_cases	0.727
Decision Tree	2	month_hist_mean	0.104
Decision Tree	3	trend_idx	0.097
Decision Tree	4	week_sin	0.072
Random Forest	1	roll_sum_2	0.289
Random Forest	2	roll_mean_2	0.284
Random Forest	3	current_cases	0.092
Random Forest	4	week_hist_mean	0.070
Random Forest	5	month_hist_mean	0.070
Gradient Boosting	1	roll_mean_2	0.339
Gradient Boosting	2	roll_sum_2	0.182
Gradient Boosting	3	current_cases	0.100
Gradient Boosting	4	month_hist_mean	0.083
Gradient Boosting	5	week_hist_mean	0.076
Note. Scores are native impurity-based feature importance and should not be interpreted as causal effects.			

3.5 Error by Incidence Level

LSTM achieved the lowest MAE during moderate-incidence weeks (5.483) and performed similarly to ARIMA during high-incidence weeks (9.186 versus 9.285). The naive baseline was strongest during low-incidence weeks, with MAE 5.333. All selected methods underpredicted very high or outbreak-level weeks. For the seven outbreak-level observations, the actual mean was 60.143 cases, whereas LSTM predicted 45.143 and ARIMA predicted 42.082 on average. The four-week moving average had the lowest outbreak-class MAE among the selected methods (12.571), but remained a benchmark rather than a final model candidate. Table 7 reports the full breakdown of error by incidence class.

Table 7
Error by Incidence Class in the 2025 Temporal Holdout.

Method	Class	n	MAE	Mean predicted
LSTM	Low	15	5.649	8.849
LSTM	Moderate	14	5.483	20.673
LSTM	High	16	9.186	41.801
LSTM	Very high/outbreak	7	14.999	45.143
ARIMA	Low	15	6.473	11.440
ARIMA	Moderate	14	5.558	20.845
ARIMA	High	16	9.285	40.246
ARIMA	Very high/outbreak	7	18.061	42.082
Naive	Low	15	5.333	8.733
Naive	Moderate	14	6.571	20.786
Naive	High	16	11.875	44.875
Naive	Very high/outbreak	7	14.000	48.143
Moving average 4	Low	15	8.283	12.283
Moving average 4	Moderate	14	9.786	22.929
Moving average 4	High	16	13.172	45.172
Moving average 4	Very high/outbreak	7	12.571	47.571
Note. Mean actual values were 5.400, 19.357, 39.000, and 60.143 cases for low, moderate, high, and very high/outbreak classes, respectively. Baselines are shown only for comparison.				

4 DISCUSSION

4.1 Principal Findings

The study produced three main findings. First, recent surveillance history carried the strongest forecasting signal. Short-lag autocorrelation, the dominance of current cases and two-week rolling summaries, and the competitive naive baselines all indicate that weekly dengue incidence in Marikina exhibits substantial short-term persistence. Second, the best-performing trained model depended on the evaluation stage: ARIMA was strongest during 2024 model development, while LSTM produced the most favorable metrics on the full unseen 2025 holdout. Third, all methods were less reliable during outbreak-level weeks, where forecasts systematically underestimated the highest observed counts.

The results reinforce why simple baselines must remain part of infectious-disease forecasting evaluation. In 2024, the two-week moving average and naive forecast marginally outperformed the strongest trained model. A comparison restricted to complex models would have overstated the value of ARIMA. In 2025, LSTM and ARIMA

both exceeded the baselines by MAE, indicating that their benefit became visible under the external annual pattern. This change across years demonstrates that a single internal holdout is insufficient to establish universal superiority.

4.2 Model Behavior and Relation to Prior Work

ARIMA performed consistently well because the series showed strong short-range dependence, a structure that autoregressive models represent directly [8, 19]. Its rolling-origin procedure also allowed the model to incorporate newly observed history before later forecasts. In contrast, the fixed seasonal SARIMA and ETS configurations performed poorly and produced negative R-squared values. The STL result provides a plausible explanation: annual variation existed, but it was irregular and accompanied by large residual spikes, so a fixed 52-week seasonal form did not adequately represent the evaluation years.

LSTM produced the strongest 2025 metrics despite the limited city-level sample. Recurrent networks are designed to represent ordered dependencies and have previously been applied to dengue forecasting [13, 20]. The compact 32-unit architecture, 12-week sequence, SmoothL1 loss, and early stopping may have limited excessive fitting. However, the study did not inspect learned representations or conduct a statistical test of the LSTM-ARIMA forecast difference. The result should therefore be described as numerical superiority on this holdout, not proof that LSTM is generally superior.

The tree-based predictor rankings were consistent with broader dengue forecasting work in which recent incidence and lagged summaries provide substantial signal [17]. Seasonal history remained useful but secondary. This supports a surveillance-only framework for local health units that do not have stable access to synchronized weather or mobility data. At the same time, surveillance-only inputs are unlikely to fully capture sudden outbreak acceleration caused by changing vector density, rainfall, population movement, reporting behavior, or interventions.

4.3 Public Health Implications

A two-week horizon is operationally relevant for local preparedness. Forecasts could support scheduling of surveillance review, barangay-level communication, vector-control planning, staffing, and inventory checks. The results do not justify autonomous deployment or direct clinical decision-making. A practical tool should display observed counts, baseline and trained-model forecasts, uncertainty intervals, data recency, and warnings when the input history falls outside the training range. Human review by MESU and local health personnel remains essential.

ARIMA may be preferable where transparency, low computation, and frequent refitting are priorities. LSTM may be retained as a complementary model where the technical capacity exists to manage scaling, retraining, and monitoring. An ensemble or model-switching strategy could be investigated, but it should be evaluated prospectively and compared with the same baselines. Because outbreak weeks were underpredicted, alerts should not rely on point forecasts alone; threshold-sensitive monitoring and conservative escalation rules may be more appropriate.

4.4 Strengths and Limitations

The main strengths were strict target-year separation, a complete unseen 2025 holdout, causal feature construction, explicit reference baselines, comparison across eight trained models, manual verification of R-

squared, and incidence-stratified error analysis. The study also avoided removing outbreak peaks and distinguished trained models from baselines in final model assessment.

Several limitations affect interpretation. The 2021–2022 gap reduced continuity and the classical models did not explicitly preserve the missing weekly calendar positions. The dataset contained only one city and one complete external year. Model families used different forecast-generation procedures, and the deep-model internal set contained 52 observations whereas feature-based models used 51. The final neural fit reused 2024 sequences for early-stopping monitoring. Hyperparameter search was limited, formal residual diagnostics and prediction intervals were not completed, and the numerical advantage of LSTM over ARIMA was not tested statistically.

The surveillance-only design excluded meteorological, environmental, mobility, demographic, intervention, and healthcare-utilization variables. This improves reproducibility but limits outbreak sensitivity and causal interpretation. Native tree importance may be biased toward predictors that offer more split opportunities and does not establish causal influence. Finally, reporting delays or revisions in operational surveillance data could reduce real-time accuracy compared with the finalized records used in retrospective evaluation.

4.5 Future Work

Future studies should extend the holdout design to additional complete years and conduct prospective weekly evaluation. Classical models should be fitted within continuous segments or with explicit missing calendar positions. Neural models should use a fully nested validation procedure in which early stopping is independent of the final training set. Forecast intervals, calibration, Diebold-Mariano or related error comparisons, and residual diagnostics should accompany point metrics.

Outbreak performance may improve through count-aware or asymmetric loss functions, hurdle or regime-switching models, quantile forecasting, oversampling of high-incidence sequences, and carefully synchronized rainfall, temperature, vector, mobility, or intervention variables. Transfer learning from larger Philippine surveillance datasets may also help deep models, provided that geographic transfer, target-year leakage, and source-target overlap are controlled.

5 CONCLUSIONS

A surveillance-only, leakage-safe framework can produce meaningful two-week-ahead dengue forecasts for Marikina City using weekly case counts and reporting dates. ARIMA was the strongest trained model during the 2024 internal-validation stage but did not outperform the leading simple baselines. On the complete 2025 out-of-time temporal holdout, LSTM achieved the lowest MAE and RMSE and the highest R-squared, while ARIMA remained a close, interpretable, and computationally efficient alternative. Both models improved on the naive baseline by MAE.

Recent incidence, the two-week rolling mean, and the two-week rolling sum were the most influential tree-based predictors. Seasonal context contributed but did not dominate. The major unresolved limitation was systematic underprediction of very high or outbreak-level weeks. Any operational implementation should therefore combine trained forecasts with simple baselines, uncertainty communication, threshold monitoring, and public health review rather than relying on a single point estimate.

Abbreviations

ACF
autocorrelation function
ARIMA
autoregressive integrated moving average
CNN
convolutional neural network
DOH-EB
Department of Health-Epidemiology Bureau
ETS
error-trend-seasonality exponential smoothing
GB
Gradient Boosting
LSTM
long short-term memory
MAE
mean absolute error
MESU
Marikina Epidemiology Surveillance Unit
PACF
partial autocorrelation function
RF
Random Forest
RMSE
root mean square error
 R^2
coefficient of determination
SARIMA
seasonal autoregressive integrated moving average
STL
Seasonal-Trend decomposition using Loess.

Declarations

Ethics approval and consent to participate

The study used aggregated, non-identifiable weekly dengue surveillance counts and did not include individual-level patient records, personally identifiable information, human tissue, or direct participant involvement. Formal informed consent was therefore not applicable. The authors complied with the data-use conditions established by the Marikina Epidemiology Surveillance Unit.

Consent for publication

Not applicable. The manuscript contains no identifiable individual person's data, images, or videos.

Clinical trial number

Not applicable.

Availability of data and materials

The Department of Health-Epidemiology Bureau source records used for the earlier period were obtained from public surveillance data. The MESU records used for 2023-2025 are not publicly redistributed because access is subject to the data provider's conditions. The derived analytical data and code may be available from the corresponding author on reasonable request and with permission from the Marikina Epidemiology Surveillance Unit.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors for this research.

Authors' contributions

KJCB, LANR, and CTCV contributed equally to the conceptualization, data preparation, methodology implementation, analysis, visualization, and drafting of the manuscript. All authors reviewed and approved the final manuscript and accept responsibility for the reported work.

Acknowledgements

The authors thank the Marikina Epidemiology Surveillance Unit (MESU) for collecting and providing the local weekly dengue surveillance dataset used in this study. The authors also acknowledge the Department of Health–Epidemiology Bureau as the source of the earlier surveillance records and Mapúa University for institutional support.

ChatGPT was used to assist with condensation of the original thesis into research-article form, language editing, and manuscript formatting. It was not used to create the surveillance dataset, execute model training, calculate the reported metrics, or determine the scientific conclusions. The authors reviewed and verified all generated content and remain fully responsible for the accuracy and integrity of the manuscript.

CCS CONCEPTS

Applied computing → Health informatics; Computing methodologies → Time series analysis; Computing methodologies → Machine learning; Applied computing → Public health.

References

1. World Health Organization. 2025. Dengue and severe dengue. WHO Fact Sheet. Retrieved August 2025 from <https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue>

2. Department of Health–Epidemiology Bureau. 2016–2020. Marikina City dengue surveillance records. Department of Health, Philippines. Data provided through the Marikina Epidemiology Surveillance Unit.
3. World Health Organization Regional Office for the Western Pacific. 2025. Dengue Situation Updates. WHO WPRO. Retrieved August 2025 from <https://www.who.int/westernpacific/emergencies/surveillance/dengue>
4. Rob J, Hyndman. and George Athanasopoulos. 2021. *Forecasting: Principles and Practice* (3rd ed.). OTexts. Retrieved from <https://otexts.com/fpp3/>
5. Leung XY, Islam RM, Adhami M, Ilic D, McDonald L, Palawaththa S, Diug B, Munshi SU, Md Nazmul K. 2023. A systematic review of dengue outbreak prediction models: Current scenario and future directions. *PLoS Neglected Tropical Diseases* 17, 2 (2023), Article e0010631. <https://doi.org/10.1371/journal.pntd.0010631>
6. Ai Deng. 2023. Time series cross validation: A theoretical result and finite sample performance. *Economics Letters* 233 (2023), Article 111369. <https://doi.org/10.1016/j.econlet.2023.111369>
7. Yang X et al. 2024. Research on information leakage in time series prediction. *Scientific Reports* 14 (2024), Article 4218. <https://doi.org/10.1038/s41598-024-80018-9>
8. George EP, Box GM, Jenkins GC, Reinsel, Ljung GM. *Time Series Analysis: Forecasting and Control*. 5th ed. Hoboken, NJ: Wiley; 2015.
9. Statsmodels. 2024. SARIMAX documentation, version 0.14.4. Retrieved from <https://www.statsmodels.org/stable/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html>
10. Statsmodels. 2025. ETSModel documentation. Retrieved from https://www.statsmodels.org/stable/generated/statsmodels.tsa.exponential_smoothing.ets.ETSModel.html
11. Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001), 5–32. <https://doi.org/10.1023/A:1010933404324>
12. Jerome H, Friedman. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29, 5 (2001), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
13. Imee V, Necesito JMS, Velasco J, Kwak JH, Lee MJ, Lee JS, Kim, Hung Soo K. 2021. Combination of univariate long-short term memory network and wavelet transform for predicting dengue case density in the National Capital Region, the Philippines. *The Southeast Asian Journal of Tropical Medicine and Public Health* 52, 4 (2021), 479–494.
14. Shaojie Bai JZ, Kolter, Koltun V. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv:1803.01271*. <https://doi.org/10.48550/arXiv.1803.01271>
15. Roy M, Anderson, May RM. *Infectious Diseases of Humans: Dynamics and Control*. Oxford, UK: Oxford University Press; 1991.
16. William O, Kermack, McKendrick AG. 1927. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society A* 115, 772 (1927), 700–721. <https://doi.org/10.1098/rspa.1927.0118>
17. Alessandro Sebastianelli D, Spiller R, Carmo et al. 2024. A reproducible ensemble machine learning approach to forecast dengue outbreaks. *Scientific Reports* 14 (2024), Article 3807. <https://doi.org/10.1038/s41598-024-52796-9>
18. Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102, 477 (2007), 359–378. <https://doi.org/10.1198/016214506000001437>
19. Robert H, Shumway, Stoffer DS. *Time Series Analysis and Its Applications: With R Examples*. 5th ed. Cham: Springer; 2025. <https://doi.org/10.1007/978-3-031-70584-7>.

20. Dang AT. 2024. Leveraging climate data for dengue forecasting in Ba Ria Vung Tau Province, Vietnam: An advanced machine learning approach. Tropical Medicine and Infectious Disease 9, 10 (2024), Article 250. <https://doi.org/10.3390/tropicalmed9100250>

21. Mengjiao Wang and Fengtai Qin. 2024. A TCN-Linear hybrid model for chaotic time series forecasting. Entropy 26, 6 (2024), Article 467. <https://doi.org/10.3390/e26060467>

Figures

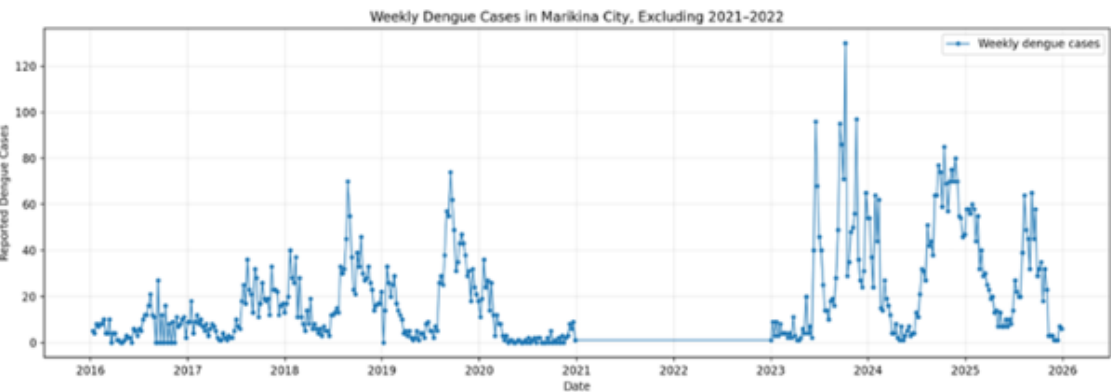


Figure 1

Weekly dengue incidence in Marikina City

Figure description: A weekly line plot covering 2016-2020 and 2023-2025 shows several irregular peaks, a gap for excluded 2021-2022 data, high activity in 2023-2024, and a lower but still variable full-year pattern in 2025.

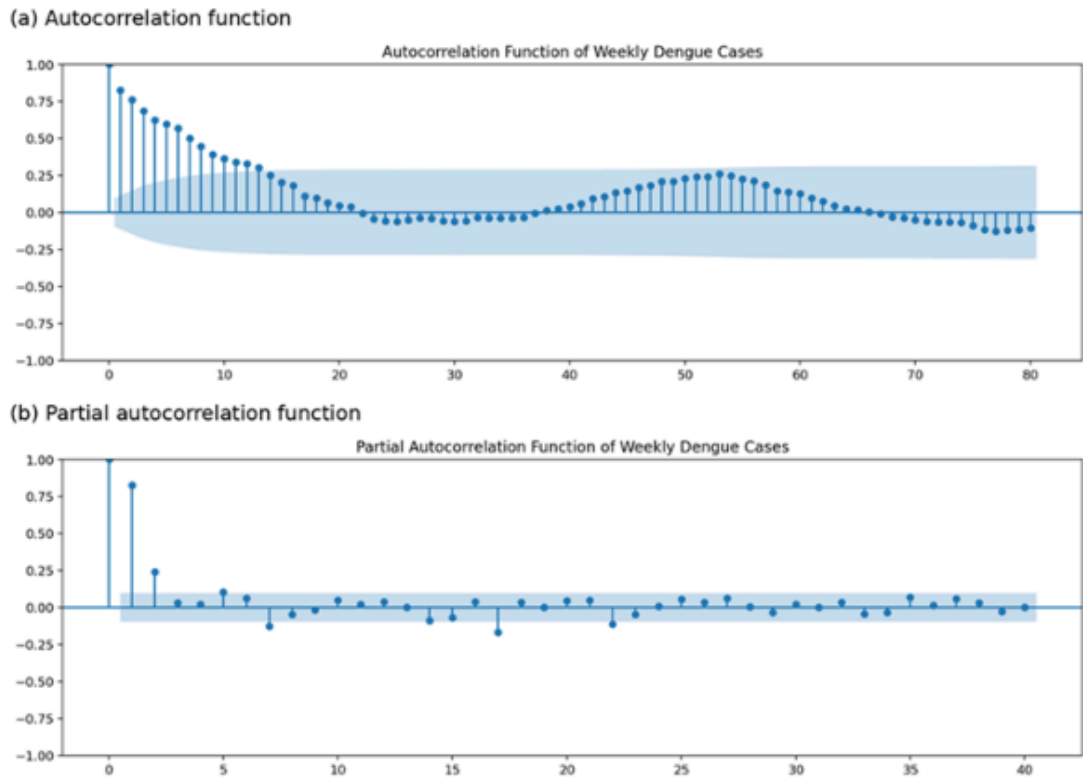


Figure 2

ACF and PACF of weekly dengue incidence

Figure description: Panel a shows high short-lag autocorrelation that gradually declines and a smaller rise near the annual range. Panel b shows the largest partial autocorrelation at the first two lags, with smaller values thereafter. Both diagnostics use the concatenated observed sequence after excluding 2021-2022.

STL decomposition of the latest continuous segment identified a changing trend, recurring fluctuations, and substantial residual spikes. The seasonal component was present but not sufficiently stable to justify treating annual recurrence as the dominant signal. The residual component remained large during high-incidence periods, suggesting that outbreak peaks could not be explained by smooth trend and seasonality alone.

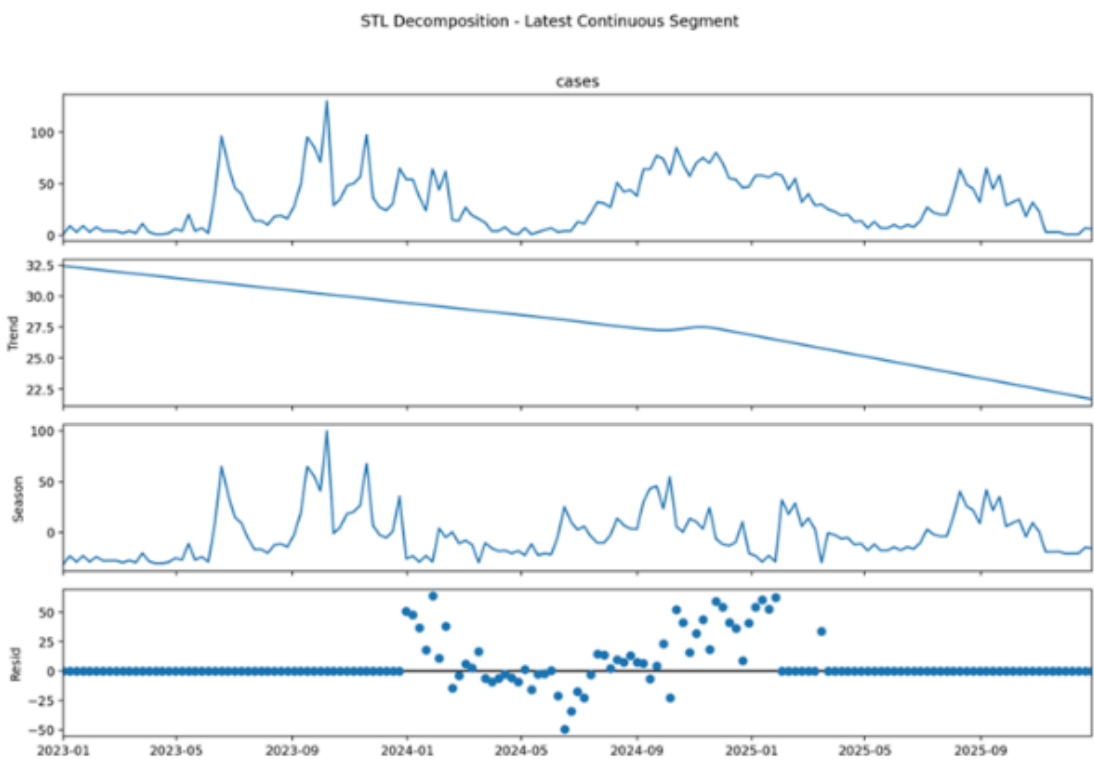


Figure 3

STL decomposition of the latest continuous segment

Figure description: Four aligned panels display the observed weekly series, long-term trend, seasonal component, and residual component. The seasonal pattern is visible but irregular, while several high-incidence spikes remain in the residuals.

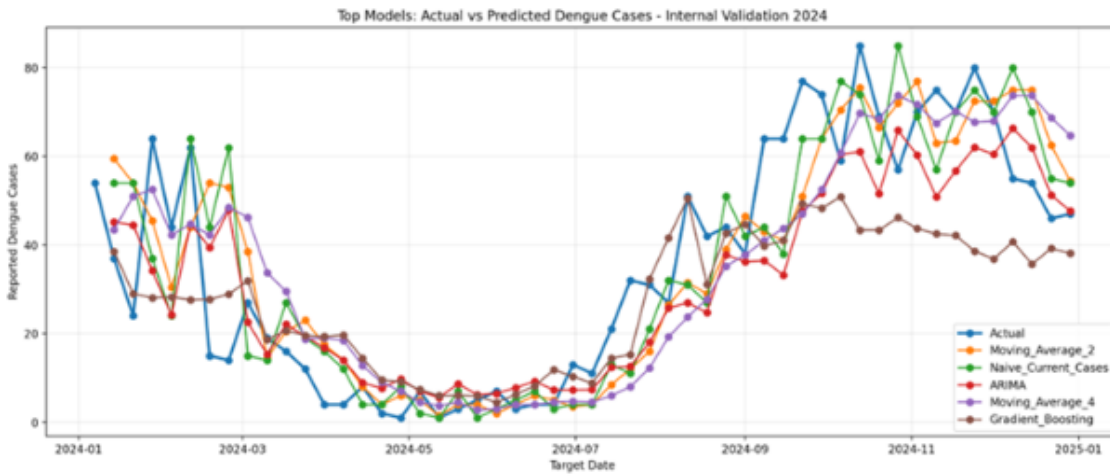


Figure 4

Internal-validation forecasts for 2024

Figure description: Observed weekly dengue counts are compared with the top trained models and reference baselines. The curves follow the broad annual decline and later increase, but differ around abrupt weekly peaks.

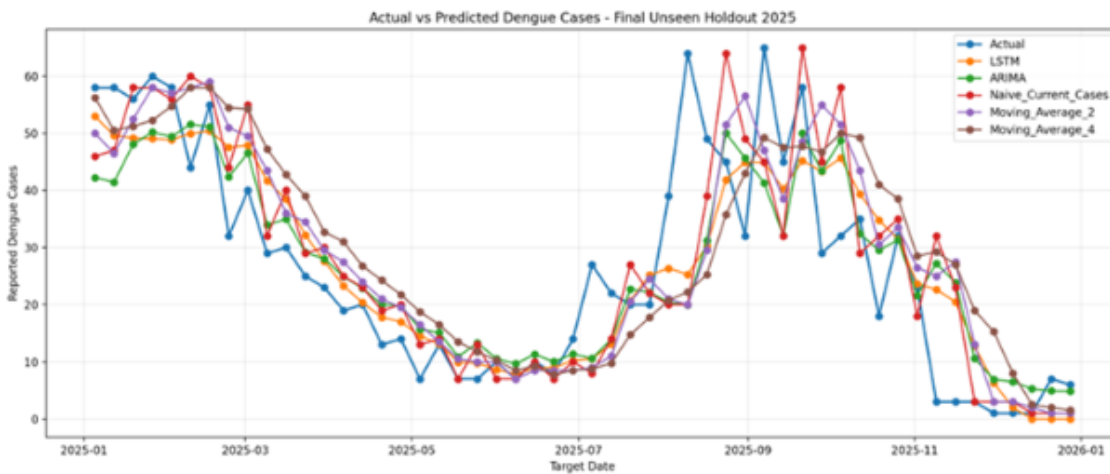


Figure 5

Selected 2025 temporal holdout forecasts.

Figure description: Observed 2025 weekly cases are shown with ARIMA, LSTM, naive, and moving-average forecasts. The selected methods capture the overall decline, later-year rise, and final decrease, but smooth or miss some abrupt peaks.