End-to-End Dengue Forecasting Platform: ML, LSTM/Transformer Models, and Cloud Deployment

Aichu Tan
Capstone Project, Big Data Analytics Master Program
San Diego State University
December 2025

Project Links

- **Project Website:** https://aichutan.github.io/Dengue_Taiwan_Forecast/
- **ArcGIS Spatiotemporal Dashboard:** https://experience.arcgis.com/experience/1eebab4280a549d294e274392d64625f
- **Live Streamlit Forecasting App:** https://dengue-taiwan-forecast.onrender.com/
- **Video Presentation:** https://youtu.be/aZyTACzGaiY

Abstract

Dengue fever remains a recurring public health challenge in southern Taiwan, where climatic seasonality, rapid urbanization, and expanding mosquito habitats contribute to episodic and often severe outbreaks. This study develops a multi-source dengue forecasting framework for Kaohsiung City, Tainan City, and Pingtung County by integrating 15 years of dengue surveillance, meteorological data, rainfall, mosquito indices, and population density. Four forecasting models : Random Forest, XGBoost, Long Short-Term Memory (LSTM), and a hybrid LSTM-Transformer, were evaluated using a unified weekly dataset and time-based validation.

Results show that **tree-based machine-learning models consistently outperformed deep-learning architectures** across all regions. Random Forest and XGBoost accurately captured outbreak timing and magnitude, while LSTM-based models struggled with sparse outbreak histories and underpredicted sharp epidemic peaks. Autocorrelation analysis confirmed strong short-term temporal dependence and multi-week climatic influences, validating the use of lagged predictors. Regional differences were also apparent: Tainan exhibited the most extreme outbreaks, Kaohsiung showed moderate but recurrent surges, and Pingtung experienced smaller, irregular patterns.

The study introduces an interactive Streamlit dashboard and ArcGIS visualization to support real-time exploration of model outputs and spatial epidemiological patterns. Limitations include missing entomological data, sensor errors in climate observations, weekly temporal resolution, city-level aggregation, a five-week project timeline, and the use of free-tier cloud services that limit application speed and scalability.

Overall, the findings demonstrate that **tree-based forecasting models provide a practical and robust approach for short-term dengue prediction in Taiwan**, offering valuable insights for early-warning systems and targeted vector-control strategies. Future work should incorporate higher-resolution data, additional environmental and mobility indicators, advanced spatiotemporal architectures, and more scalable cloud infrastructure to support operational public health deployment.

1. Introduction

Dengue fever is one of the most significant vector-borne diseases in Taiwan, transmitted primarily by *Aedes aegypti* mosquitoes. Over the past two decades, outbreaks have increased in both frequency and magnitude due to climatic variability, rapid urbanization, population growth, and expanding mosquito habitats. Southern Taiwan, particularly Kaohsiung City, Tainan City, and Pingtung County, consistently reports the highest dengue incidence, experiencing major epidemics in 2002, 2014-2015, and 2023. These regions share environmental characteristics that promote dengue transmission, including a subtropical climate, high humidity, substantial rainfall, and dense urban development. As climate extremes intensify, accurately forecasting dengue trends has become essential for public health preparedness and early intervention.

Taiwan's dengue surveillance system relies on case reporting and entomological inspections, including indicators such as the Breteau Index (BI), House Index (HI), and Container Index (CI). While these metrics are valuable, they often lag behind real-time transmission and cannot fully capture the nonlinear, time-dependent relationships between climate, mosquito abundance, and human infections. Conventional statistical models struggle with these complex interactions, underscoring the need for modern data-driven forecasting frameworks.

Recent advances in machine learning and deep learning provide powerful tools for modeling dengue dynamics. Ensemble algorithms such as Random Forest and XGBoost effectively capture nonlinear patterns and offer interpretable insights into influential predictors. Deep learning models, including Long Short-Term Memory (LSTM) networks and Transformer-based architectures, excel at representing long-range temporal dependencies in epidemiological and climate time series. Hybrid approaches that combine LSTM and Transformer components further enhance prediction accuracy by integrating short- and long-term temporal features.

Building on these advancements, this study develops a comprehensive dengue forecasting framework using weather, rainfall, mosquito surveillance, and demographic data from Taiwan's three major endemic regions in the south. Four models : Random Forest, XGBoost, LSTM, and a hybrid LSTM-Transformer, are trained and evaluated under a consistent feature set and time-based validation strategy. The unified weekly dataset also enables identification of key climatic, entomological, and demographic drivers of dengue outbreaks. By comparing model behavior across multiple high-incidence regions, the study aims to support the development of an early-warning system that strengthens outbreak preparedness and vector control efforts in southern Taiwan.

2. Literature Review

Dengue fever is one of the fastest-growing vector-borne diseases globally, driven by urbanization, climate change, and expanding mosquito habitats. Transmission is strongly influenced by environmental and demographic conditions, particularly in warm, humid, densely populated regions where human-mosquito contact is intensified. In Taiwan, dengue occurs in seasonal outbreaks concentrated in the southern cities of Kaohsiung, Tainan, and Pingtung, which consistently record the highest case counts due to their subtropical climate, heavy rainfall, and favorable breeding environments for *Aedes* mosquitoes.

Extensive research demonstrates that climatic factors are major determinants of dengue transmission. Temperature accelerates mosquito development and viral replication; rainfall generates breeding sites and increases larval density; humidity enhances adult mosquito survival; and wind dynamics influence vector dispersal. Entomological indicators such as BI, HI, and CI quantify larval infestation and are routinely used in vector surveillance. Demographic factors, especially population density, also shape transmission risk by influencing mosquito-human interaction rates. Many of these factors exert lagged effects, reinforcing the importance of time-dependent modeling.

Traditional statistical approaches (e.g., ARIMA, SARIMA, Poisson regression) have been widely used for dengue forecasting but struggle with nonlinear and multivariate interactions. Modern machine learning methods such as Random Forest and XGBoost introduce greater flexibility, capturing complex relationships and producing interpretable feature importance estimates. However, they depend on manually engineered lag features and do not inherently model long-range temporal dependencies.

Deep learning models address these limitations. LSTM networks learn temporal patterns directly from sequential data and have outperformed classical models in various dengue forecasting studies. Transformer-based models, originally developed for language processing, leverage self-attention to capture long-range dependencies and have shown strong performance in climate and epidemiological applications. Hybrid architectures combining LSTM and Transformer layers further enhance temporal modeling capabilities.

Despite these advances, several gaps remain in Taiwan-specific dengue forecasting research: (1) few studies compare model performance across multiple high-incidence regions; (2) mosquito indices and demographic factors are rarely integrated into modern ML/DL frameworks; (3) attention-based models and hybrid LSTM-Transformer architectures remain underexplored; and (4) cross-city generalizability has not been systematically evaluated. This study addresses these gaps by developing and comparing four forecasting models across three major hotspots, integrating climatic, entomological, and demographic predictors into a unified multi-city framework.

3. Study Area and Dataset Description

3.1 Study Area

This study examines dengue transmission across all 22 cities and counties in Taiwan, with a primary focus on Kaohsiung City, Tainan City, and Pingtung County, the regions that consistently experience the nation's highest dengue burden. These areas are characterized by high temperatures, heavy monsoon rainfall, persistent humidity, and dense urban environments that support *Aedes aegypti* proliferation. Although forecasting models are trained separately for each hotspot, nationwide dengue, climate, mosquito, and demographic data were initially analyzed to contextualize spatial patterns of transmission risk. Population density was incorporated to capture structural differences in human-mosquito interaction intensity.

To complement quantitative analyses, an ArcGIS Online dashboard was developed to visualize spatiotemporal dengue trends from 2010 to 2025. The dashboard highlights concentrated clusters in southern Taiwan and strong alignment between rainfall, mosquito indices, and outbreak patterns, reinforcing the selection of these three southern regions as priority areas for prediction.

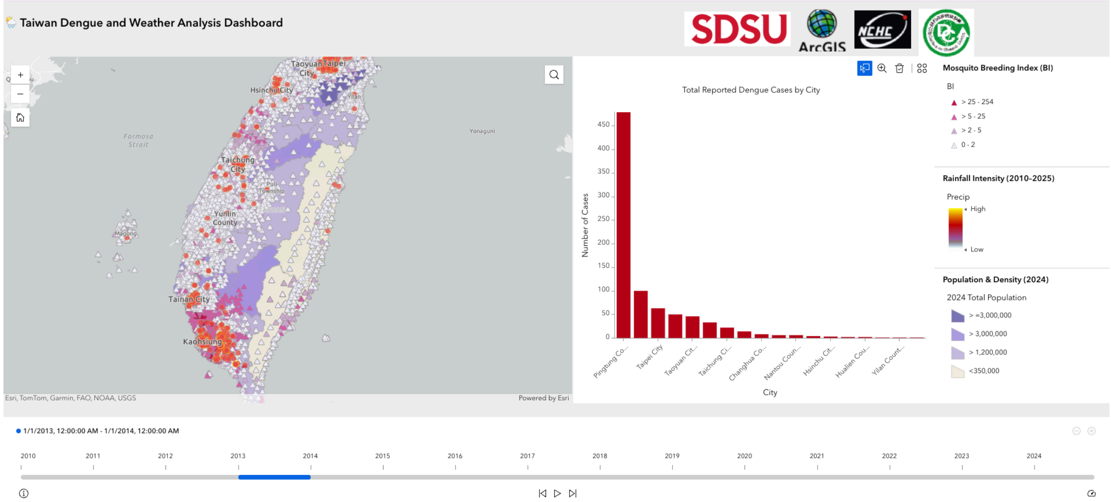


Figure 1. ArcGIS Dashboard for visualizing dengue cases, mosquito breeding indices, rainfall intensity, and city population (2010-2024). The interactive version is available at:

https://experience.arcgis.com/experience/1eebab4280a549d294e274392d64625f

3.2 Dengue Surveillance Data

Daily dengue case records were obtained from the Taiwan Centers for Disease Control (Taiwan CDC) Open Data Portal, covering the period from January 1, 2010 to December 31, 2024. Each record includes onset date, report date, basic demographics, residential location, infection source, and case classification. To ensure accurate modeling of local transmission, only domestic (non-imported) cases were retained.

Daily case counts were aggregated by city, and dates with no reported infections were assigned zero values to maintain temporal continuity. This process generated a complete 15-year city-level incidence time series, which served as the epidemiological foundation of the analysis.

3.3 Dengue Vector Surveillance Data

Mosquito surveillance data were derived from Taiwan CDC's village-level entomological inspections, which monitor larval breeding sites and household infestation. Three standard indices were used:

- **Breteau Index (BI):** number of positive containers per 100 inspected households
- **House Index (HI):** percentage of houses infested with larvae
- **Container Index (CI):** percentage of water-holding containers with larvae

These indices reflect larval density and are widely used to assess dengue vector abundance. Because mosquito inspections are conducted intermittently, BI, HI, and CI values were aggregated to the city level and aligned with the dengue and climate datasets. Missing observations were imputed using forward-fill and backward-fill procedures so that each inspection remained valid until updated. These entomological indicators were included to capture mosquito activity and breeding conditions across the study regions.

3.4 Meteorological and Rainfall Data

Daily meteorological and precipitation data were obtained from the National Center for High-Performance Computing (NCHC), which provides hourly observations from weather stations across Taiwan. Variables included cumulative daily rainfall, mean, minimum, and maximum temperature, relative humidity, atmospheric pressure, and windspeed.

Hourly readings from multiple stations within each city were aggregated using the median to minimize the influence of localized anomalies. Sentinel or unrealistic values (e.g., -99, -999, 9999) were treated as missing, and physical plausibility checks were applied based on accepted ranges for each variable. Temperature records were further validated to ensure logical consistency (Tmin ≤ Tmean ≤ Tmax). These procedures produced a high-quality meteorological dataset suitable for integration into the forecasting framework.

3.5 Population Density Data

Population density was incorporated as a demographic factor representing structural differences in human exposure and human-mosquito interaction potential. Annual population counts for each city or county were obtained from Taiwan’s Ministry of the Interior (MOI) and divided by administrative land area to compute density (people per km^2). Because population changes gradually, yearly density values were merged with the weekly dengue dataset by matching both city and year. This allowed demographic context to be consistently reflected across all forecasting models.

3.6 Data Integration and Preprocessing

Dengue case counts, meteorological variables, mosquito indices (BI, HI, CI), and population density were integrated into a unified weekly dataset for Kaohsiung, Tainan, and Pingtung. Daily records were aggregated to align with epidemiological reporting cycles and reduce short-term variability. Missing dengue counts were treated as true zeros, whereas gaps in meteorological data were imputed using time-based interpolation. Mosquito indices were imputed using forward- and backward-fill methods to account for intermittent survey schedules.

After merging all variables by city and epidemiological week, continuous predictors were normalized using z-score standardization for machine-learning models and Min-Max scaling for deep-learning architectures. Finally, lag features were generated (1-12 weeks) to represent the delayed effects of climate conditions and mosquito abundance on dengue transmission. The resulting multi-city weekly dataset formed the input for all subsequent modeling and comparative analysis.

4. Methodology and Database Management

This study employs an end-to-end analytical workflow that integrates multi-source dengue, climate, entomological, and demographic datasets into a reproducible forecasting system. The overall pipeline is shown in **Figure 2**, which outlines the major stages of data ingestion, preprocessing, model development, cloud-based storage, and interactive deployment.

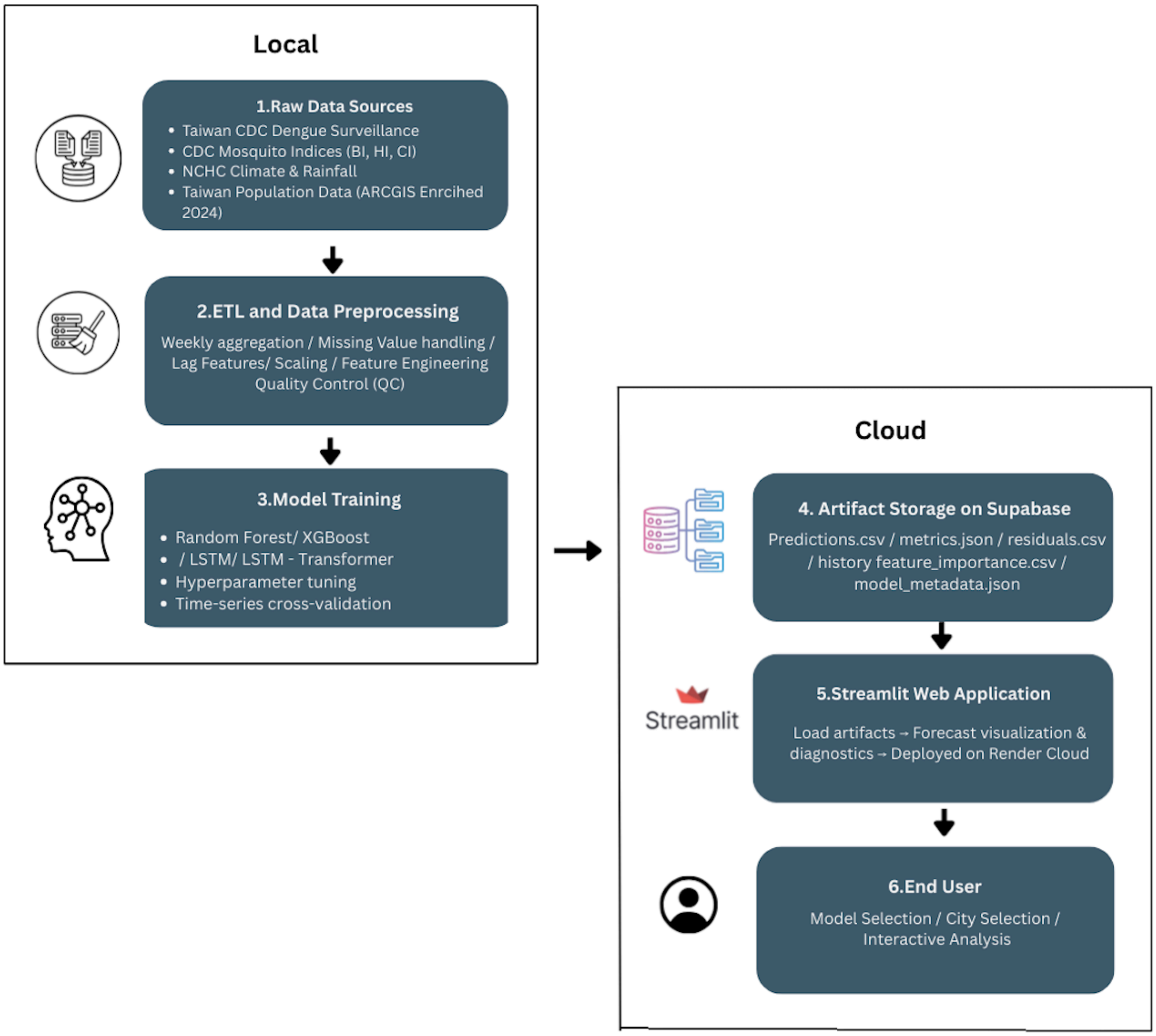

Local
1.Raw Data Sources
Taiwan CDC Dengue Surveillance
CDC Mosquito Indices (BI, HI, CI)
NCHC Climate & Rainfall
Taiwan Population Data (ARCGIS Enrcihed 2024)
2.ETL and Data Preprocessing
Weekly aggregation / Missing Value handling / Lag Features/ Scaling / Feature Engineering Quality Control (QC)
3.Model Training
Random Forest/ XGBoost
/ LSTM/ LSTM - Transformer
Hyperparameter tuning
Time-series cross-validation
Cloud
4. Artifact Storage on Supabase
Predictions.csv / metrics.json / residuals.csv / history feature_importance.csv / model_metadata.json
Streamlit
5.Streamlit Web Application
Load artifacts → Forecast visualization & diagnostics → Deployed on Render Cloud
6.End User
Model Selection / City Selection / Interactive Analysis

Figure 2. End-to-end analytical pipeline for dengue forecasting. The workflow integrates four major components:

(1) raw data ingestion from Taiwan CDC, CDC mosquito surveillance, NCHC climate datasets, and population density sources;

(2) local ETL and preprocessing, including weekly aggregation, missing value handling, lag feature engineering, scaling, and quality control;

(3) local model training using Random Forest, XGBoost, LSTM, and hybrid LSTM Transformer architectures with hyperparameter tuning and time-based cross-validation; and

(4) cloud-based artifact storage on Supabase, enabling downstream visualization and diagnostics through a Streamlit application deployed on Render Cloud.

This pipeline illustrates the complete process from raw data to model deployment and interactive end-user analysis.

4.1 Data Acquisition and Management

This study utilized four primary data sources: dengue surveillance records from the Taiwan Centers for Disease Control (2010-2024), entomological indices from CDC field inspections (Breteau Index, House Index, and Container Index), meteorological and rainfall observations from the National Center for High-Performance Computing (NCHC), and population density statistics obtained from the Ministry of the Interior and enriched through ArcGIS. All raw datasets were initially collected, inspected, and processed locally in Python to ensure data quality and consistency before modeling. Following model development, analytical artifacts, including weekly model predictions, performance metrics, residual errors, feature importance rankings, and model configuration metadata, were exported to Supabase, a cloud-hosted PostgreSQL platform used to centralize and manage project outputs. Supabase stores key files such as *predictions.csv*, *metrics.json* (containing MAE, RMSE, MAPE, R^2, and training history), *residuals.csv*, *feature_importance.csv*, and *model_metadata.json*. Centralizing these outputs in a relational cloud database ensures reproducibility, supports version control, and enables efficient retrieval for visualization within the deployed web application. This architecture effectively separates computationally intensive model training, performed locally, from scalable cloud-based storage and lightweight user interaction.

4.2 ETL and Data Preprocessing

A structured extract-transform-load (ETL) pipeline was implemented to standardize and merge all datasets prior to model development. Daily dengue cases, meteorological variables, mosquito indices, and population counts were aggregated to the weekly city level to align with epidemiological reporting cycles and reduce short-term fluctuations. Missing dengue case counts were treated as true zeros, while gaps in meteorological variables were imputed using time-based interpolation within each city to preserve continuity. Because mosquito indices are collected intermittently, BI, HI, and CI values were imputed using forward-fill and backward-fill procedures. Population density values, which change gradually over time, were matched by city and year and then merged into the weekly dataset. Quality-control procedures were applied to remove unrealistic sentinel values (-99, -999, 9999) and to enforce physical plausibility checks for temperature consistency ($T_{min} \leq T_{mean} \leq T_{max}$) and acceptable meteorological ranges. Continuous predictors were then standardized using z-score normalization for machine-learning models and Min-Max scaling for deep-learning models. This fully cleaned and harmonized dataset formed the foundation for subsequent feature engineering and model development.

4.3 Feature Engineering and Temporal Modeling

Because dengue transmission is strongly influenced by delayed climatic and entomological processes, feature engineering played a central role in the modeling framework. To capture short- and long-term dependencies for the machine learning models, lagged variables were generated for rainfall, temperature, humidity, windspeed, and mosquito indices. Random Forest and XGBoost incorporated lag intervals of 1, 2, 4, 8, 10, 11, 12, and 15 weeks, enabling the models to learn delayed responses such as post-rainfall breeding surges, temperature-driven changes in mosquito development, and gradual shifts in transmission intensity. For deep learning models, sequential input structures were used to represent temporal dynamics directly. The first architecture, an LSTM model, employed a 24-week sliding window of lagged predictors, allowing the network to learn sequential patterns through recurrent memory. The second, a hybrid LSTM-Transformer model, combined an initial LSTM layer to encode short-term temporal structure with a Transformer encoder block designed to capture long-range dependencies through multi-head self-attention and residual normalization. This hybrid design allowed the system to integrate fine-grained weekly variations with broader seasonal and climatic trends.

To prevent temporal leakage, the dataset was chronologically partitioned into training (2010-2017), validation (2018-2021), and testing (2022-2024) subsets. The validation period was used to guide model selection, tune hyperparameters, and apply early stopping procedures.

4.4 Model Development and Evaluation

Four forecasting models were developed and evaluated in this study: Random Forest, XGBoost, a Long Short-Term Memory (LSTM) network with lagged features, and a hybrid LSTM-Transformer architecture. The tree-based models (Random Forest and XGBoost) were trained on engineered lag predictors that captured 1-15-week delayed climatic, entomological, and epidemiological effects. Random Forest models were optimized using a time-series-aware RandomizedSearchCV procedure that varied the number of estimators (300-700), tree depth (none, 15, 25), node-splitting thresholds, and feature-subsampling ratios (0.7, 1.0, or $\sqrt{p}$), with bootstrap sampling enabled to enhance robustness under outbreak variability. XGBoost models were tuned using a structured grid search that explored tree depth (2-8), learning rates (0.01-0.07), the number of boosting rounds (500-800), and subsampling and column-sampling ratios (0.8-1.0). These hyperparameters jointly regulate model complexity, learning speed, and regularization, allowing each model to adapt to the distinct outbreak dynamics of Kaohsiung, Tainan, and Pingtung. The best-performing configuration for each city was selected based on validation RMSE and MAE from the 2018-2021 window.

Deep learning models were designed to learn temporal structure directly from sequential inputs. The LSTM (lag) model first generated 1-2-week lag features for dengue cases, climate variables, and mosquito indices, followed by log transformation of the target variable and z-score standardization of all inputs. The data were then reshaped into 24-week sliding windows, producing supervised sequences for training and evaluation. The LSTM architecture consisted of stacked recurrent layers (128 and 64 units), an optional dropout layer to reduce overfitting, a 32-unit dense layer with ReLU activation, and a final linear output node. Training used the Adam optimizer, MSE loss, and early stopping with a 30-epoch patience. Sample weighting was applied to emphasize outbreak weeks, improving the model's sensitivity to periods of rapid case growth.

The hybrid LSTM-Transformer model combined an LSTM encoder with a Transformer encoder block to capture both short- and long-range temporal dependencies. A compact hyperparameter search varied the input window length (4-16 weeks), number of attention heads (2 or 4), and feed-forward dimension (64), while keeping the learning rate and dropout fixed. Early stopping on the 2018-2021 validation period ensured that the final architecture balanced representational capacity with the constraints of limited outbreak data.

After hyperparameter tuning, all models were retrained on the combined 2010-2021 training and validation set and assessed on the unseen 2022-2024 test period. Performance evaluation used MSE, RMSE, MAE, MAPE, and R^2. Diagnostic analyses-including predicted-versus-actual curves, scatter plots, temporal residual traces, and LSTM training-validation loss trajectories-were used to assess model fit, temporal stability, and potential bias across outbreak and non-outbreak periods. For the tree-based models, feature-importance rankings provided additional interpretability by identifying the most influential climatic and entomological drivers of dengue transmission.

4.5 Deployment and Interactive Visualization

To facilitate real-time visualization and comparison of model outputs, an interactive Streamlit web application was developed. The application retrieves predictions, residuals, metrics, and model metadata directly from the Supabase cloud database, enabling users to interact with results without running local computations. Through the interface, users can select specific cities and forecasting models, visualize weekly predictions and associated uncertainty intervals, explore feature importance rankings, and examine residual patterns across time. The system was deployed on Render Cloud, ensuring a lightweight, fast, and publicly accessible user experience while keeping all computationally intensive training processes offline.

The interactive forecasting dashboard is publicly accessible via the Streamlit web application: https://dengue-taiwan-forecast.onrender.com/. The interface allows users to select cities and models, visualize predictions, explore feature importance, and examine residual diagnostics.

4.6 End-User Interaction

The final forecasting system allows researchers, students, and public health practitioners to compare model performance across cities, explore climatic and entomological drivers of outbreaks, and evaluate prediction reliability. By centralizing all model artifacts in a cloud database and delivering them through a modern interactive interface, the platform supports transparent, reproducible, and operationally useful dengue forecasting for decision support and outbreak preparedness.

5. Results

5.1 Exploratory Data Analysis

5.1.1 Kaohsiung City

Exploratory analysis of Kaohsiung City (2010-2024) reveals a pattern of low baseline dengue activity punctuated by major outbreaks (Figure 3), most notably the 2014-2015 epidemic with weekly cases exceeding 2,000. A smaller resurgence occurred in 2023-2024. These episodic, climate-amplified spikes highlight Kaohsiung's vulnerability to large outbreaks.

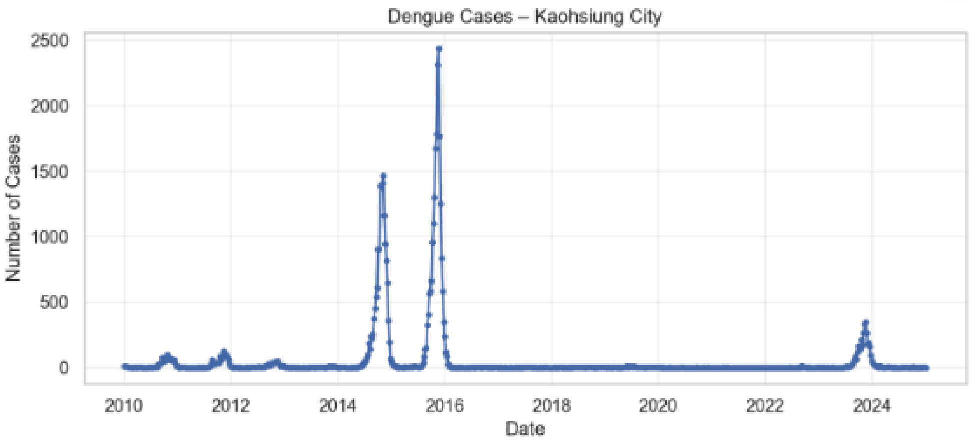


Figure 3. Weekly dengue cases in Kaohsiung City (2010-2024)

Meteorological and entomological variables exhibit clear seasonal cycles (Figure 4). Temperatures peak in July-September, humidity remains high throughout summer, and rainfall increases sharply from May to October, conditions aligned with elevated mosquito indices (BI, HI, CI). Monthly log-scale case distributions (Figure 5) show dengue transmission concentrated between August and November. These patterns underscore the strong climatic seasonality driving dengue risk in Kaohsiung.

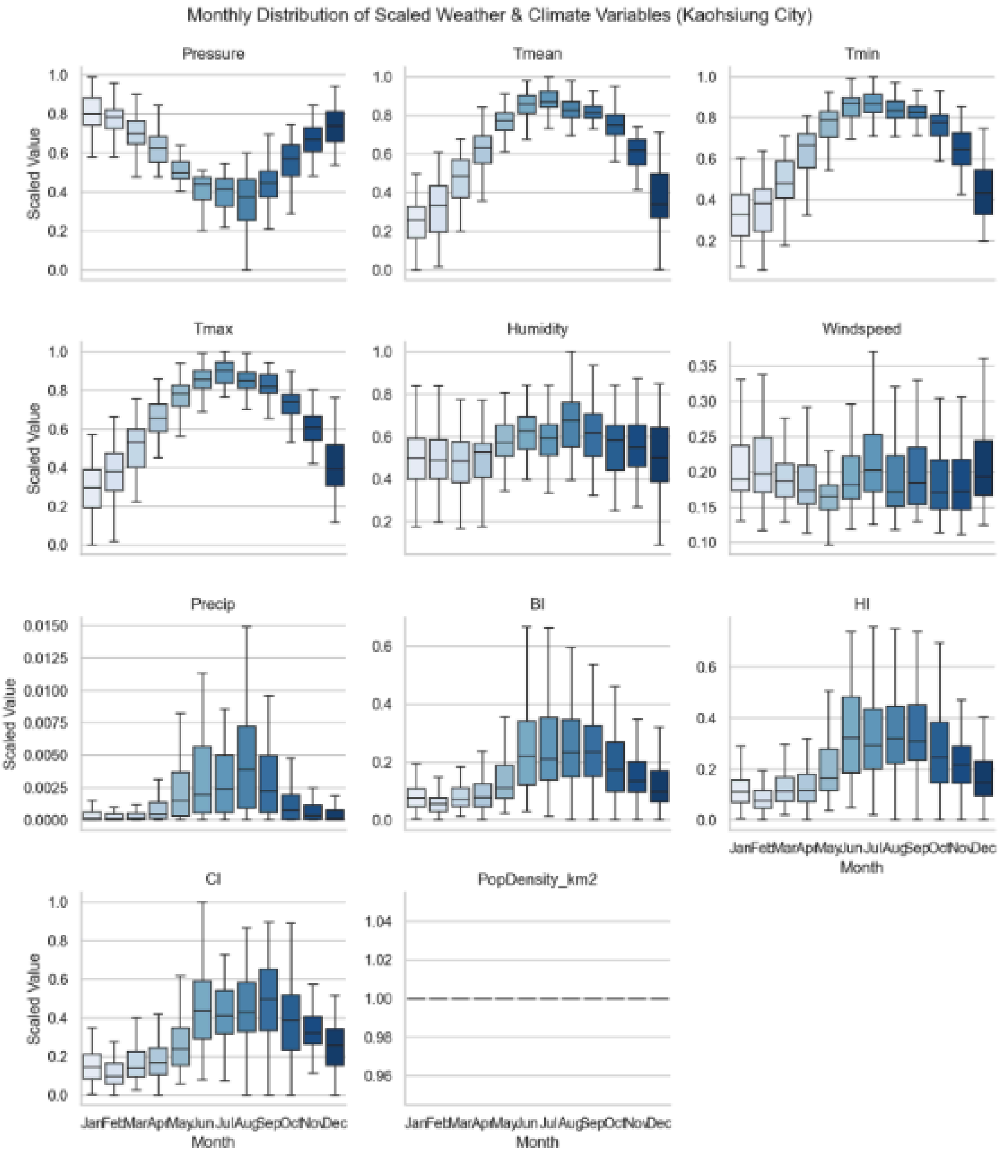


Figure 4. Monthly climate and mosquito indices in Kaohsiung City.

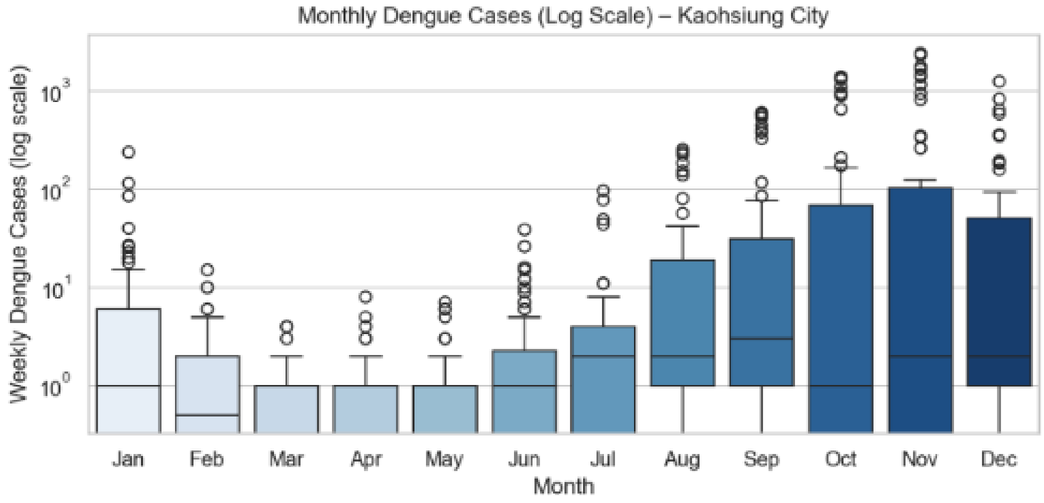


Figure 5. Monthly log-scale dengue cases in Kaohsiung City.

5.1.2 Tainan City

Tainan shows long periods of low transmission interrupted by major epidemics (Figure 6), including a large outbreak in 2015 (peaking above 3,500 weekly cases) and another in 2023-2024. These sharp, high-amplitude surges underscore Tainan's sensitivity to climate-driven dengue amplification.

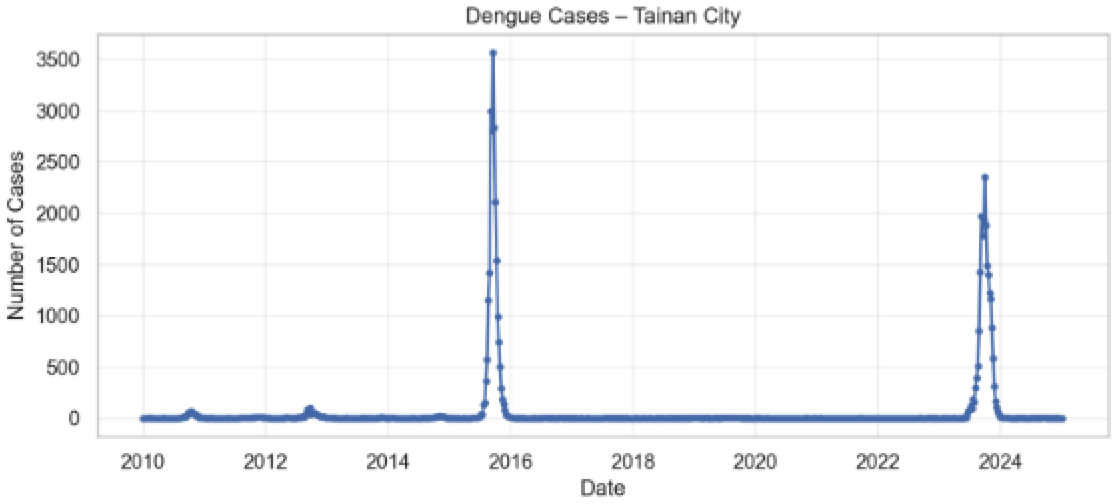


Figure 6. Weekly dengue cases in Tainan City (2010-2024)

Seasonal meteorological and entomological patterns mimic those of Kaohsiung (Figure 7): summer peaks in temperature and humidity, reduced atmospheric pressure, and rainfall increases from May to October. Correspondingly, BI, HI, and CI rise sharply during summer.

Monthly log-scale cases (Figure 8) show dengue activity beginning in June and peaking from August to November, with sharper increases than Kaohsiung.

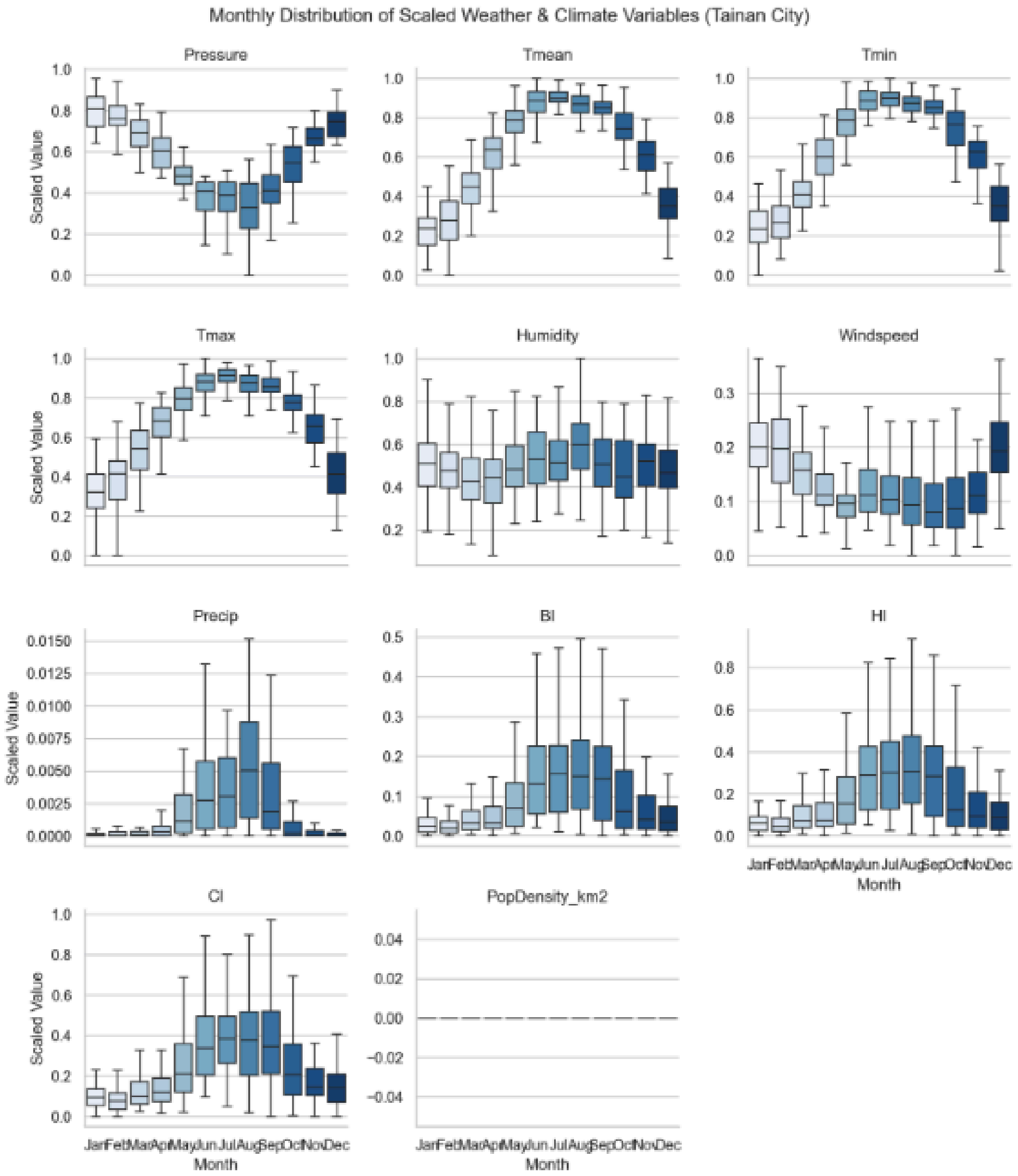


Figure 7. Monthly climate and mosquito indices in Tainan City.

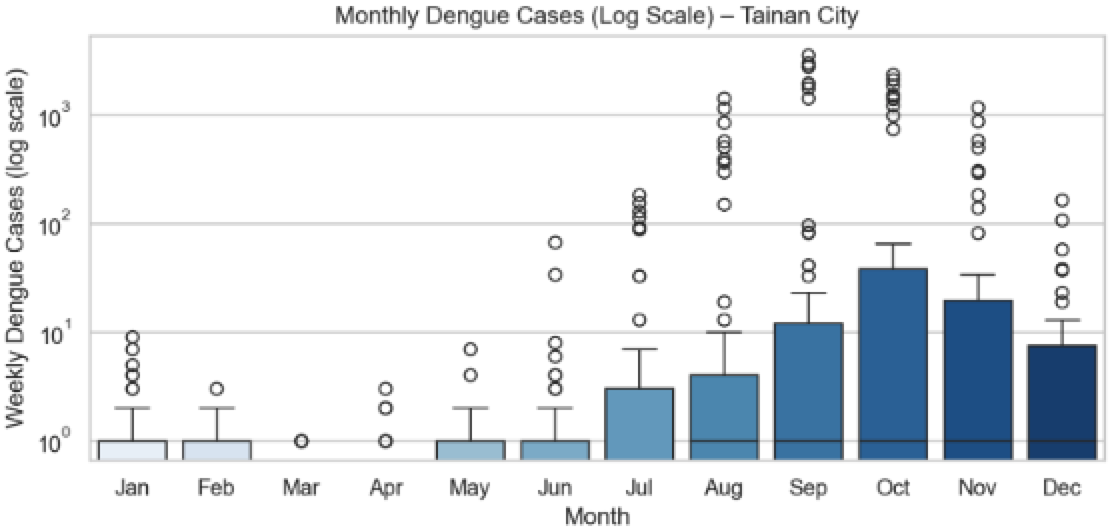


Figure 8. Monthly log-scale dengue cases in Tainan City.

5.1.3 Pingtung County

Pingtung exhibits smaller and more irregular outbreaks than Kaohsiung or Tainan (Figure 9), with peak weekly cases typically between 30-50 during active years (2011-2012, 2014-2016, 2023-2024). Despite suitable climate, lower population density and fewer urban breeding sites moderate outbreak size.

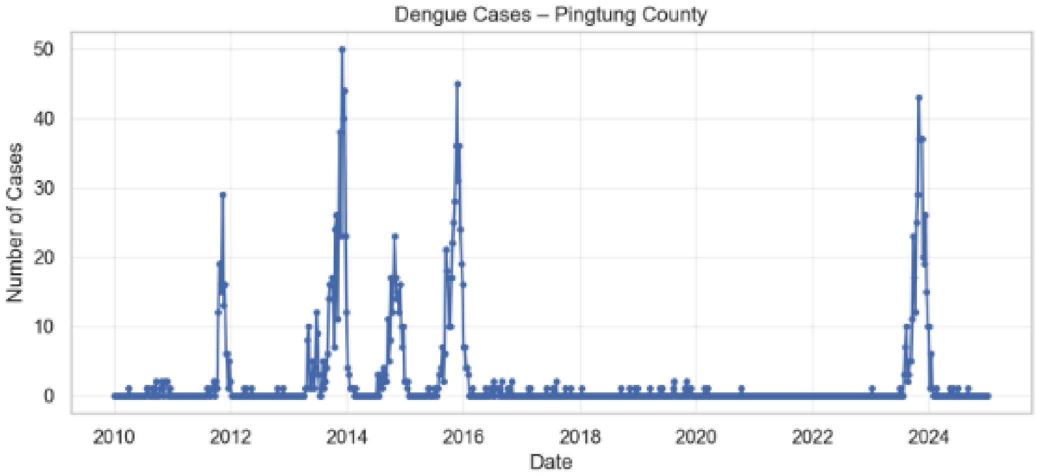


Figure 9. Weekly dengue cases in Pingtung County (2010-2024)

Seasonal climate patterns (Figure 10) include temperature peaks in July-September, high humidity in summer, and strong monsoon rainfall from June to October. BI, HI, and CI peak in the same seasonal window but at lower magnitudes than in the other two cities. Monthly log-scale case distributions (Figure 11) show dengue activity concentrated from October to December, slightly later than in Kaohsiung and Tainan.

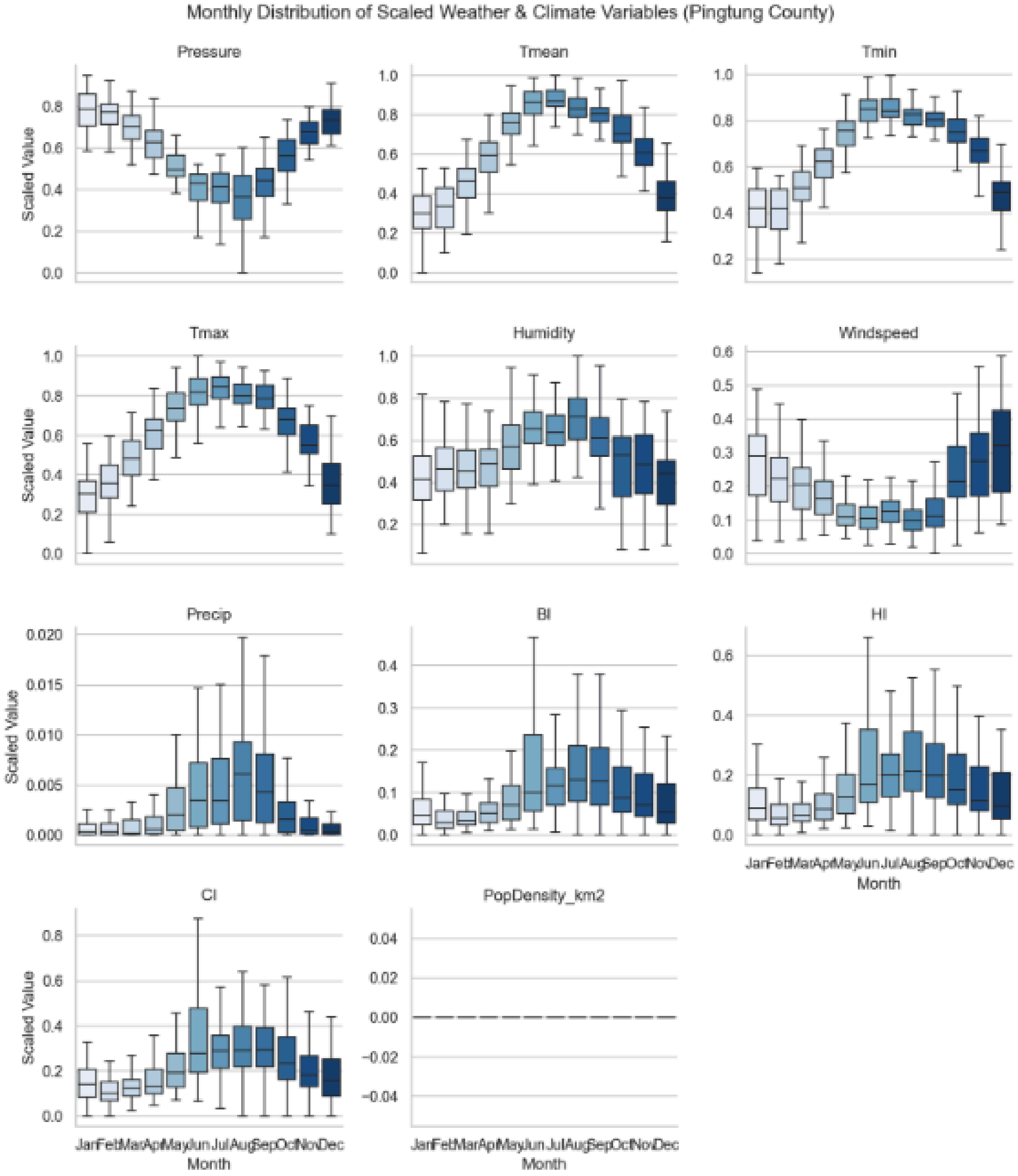


Figure 10. Monthly climate and mosquito indices in Pingtung County.

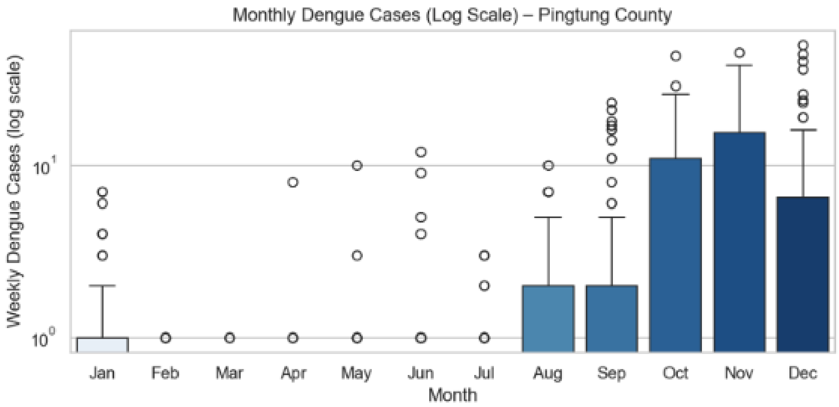


Figure 11. Monthly log-scale dengue cases in Pingtung County.

5.2 Autocorrelation and Partial Autocorrelation Analysis

5.2.1 Kaohsiung City

Kaohsiung's ACF (Figure 12) shows strong autocorrelation through ~10-12 weeks, indicating persistent multi-week transmission momentum. The PACF displays dominant spikes at lags 1-3, reflecting strong short-term autoregressive structure. These findings support the use of multi-week lag features and sequential models.

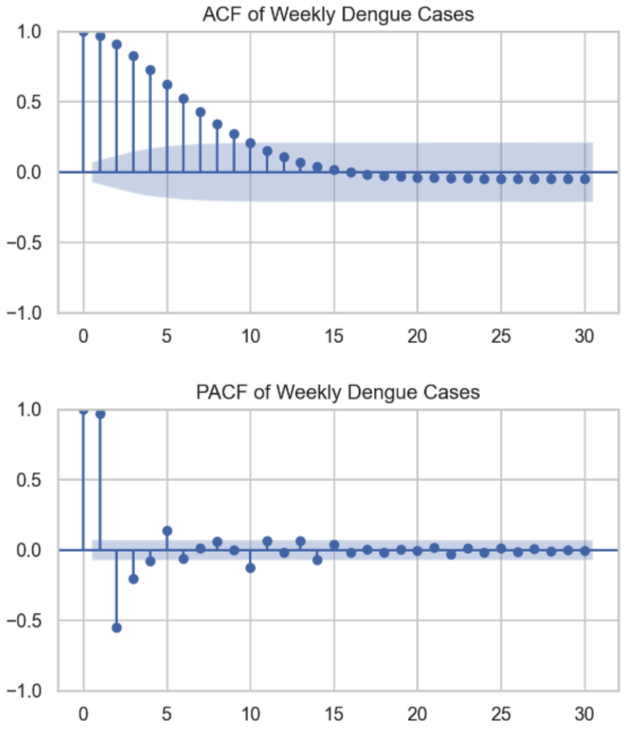


Figure 12. ACF and PACF of weekly dengue cases in Kaohsiung City. The ACF shows significant autocorrelation up to about 10 weeks, while the PACF indicates strong short-term effects at lags 1-3.

5.2.2 Tainan City

Tainan's ACF (Figure 13) mirrors Kaohsiung but with slightly more persistent early-lag autocorrelation, consistent with its rapid outbreak escalation. The PACF again shows strong effects at lags 1-3, confirming the importance of both short-term memory and mid-range temporal dependencies for forecasting.

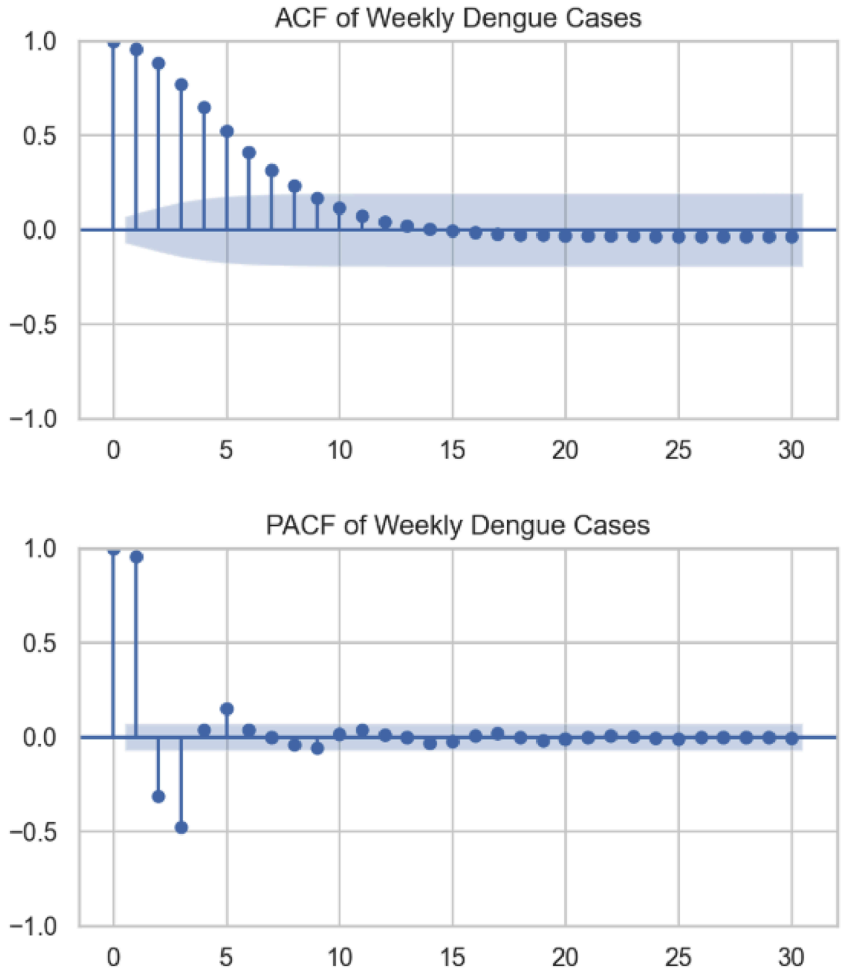


Figure 13. ACF and PACF for Tainan City

5.2.3 Pingtung County

Pingtung's ACF (Figure 14) shows significant but weaker autocorrelation extending ~8-10 weeks. The PACF indicates strongest direct dependence at lag 1 and minor influence at lag 2, with negative values at later lags typical of low-count, irregular series. These results emphasize short-term temporal effects and limited long-range dependency.

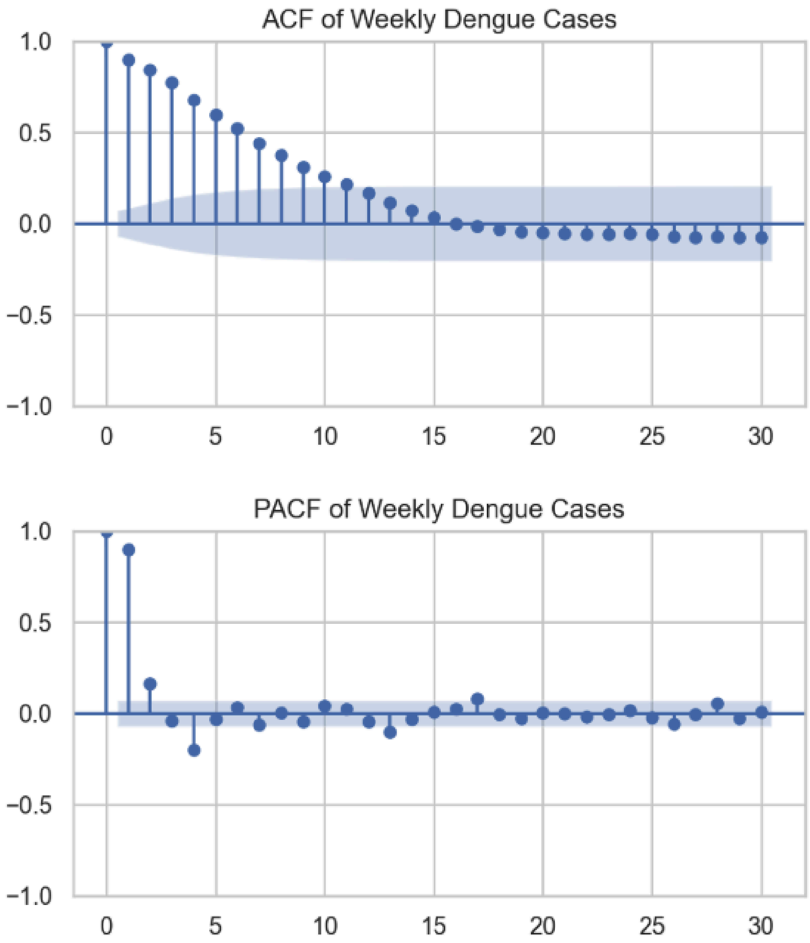


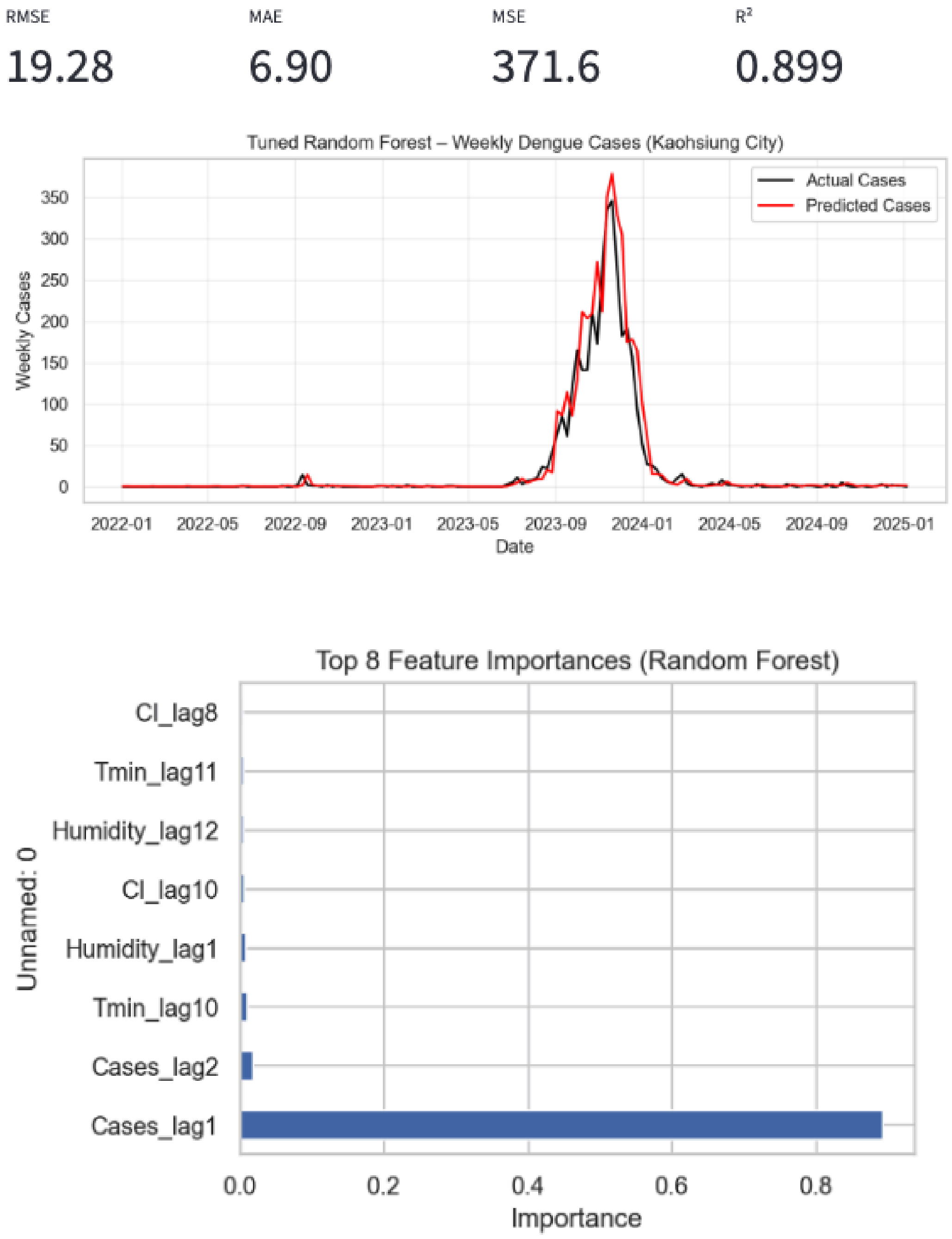
Figure 14. ACF and PACF for Pingtung County.

5.3 Model Performance - Kaohsiung City

Model performance for Kaohsiung (2022-2024) varied substantially across the four forecasting approaches.

Random Forest

Random Forest achieved the best performance (RMSE = 19.28; MAE = 6.90; $R^2 = 0.899$), accurately capturing the 2023-2024 outbreak shape and magnitude. Feature importance was dominated by recent case lags, with climatic and entomological lags also contributing.

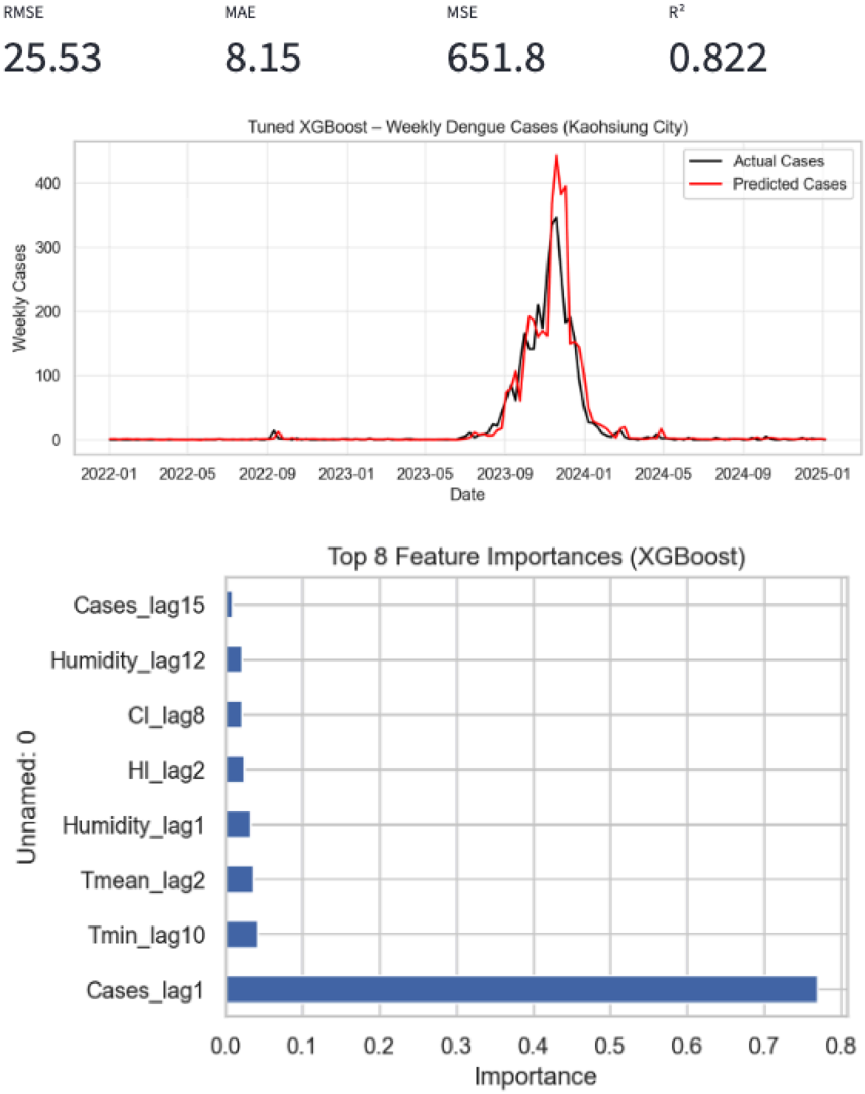


RMSE	MAE	MSE	R²
19.28	6.90	371.6	0.899

Figure 15. RF predictions and top 8 feature importances for Kaohsiung.

XGBoost

XGBoost performed well (RMSE = 25.53; MAE = 8.15; R² = 0.822) but tended to underpredict the peak weeks. Cases_lag1 remained the strongest predictor, followed by temperature and humidity lags.



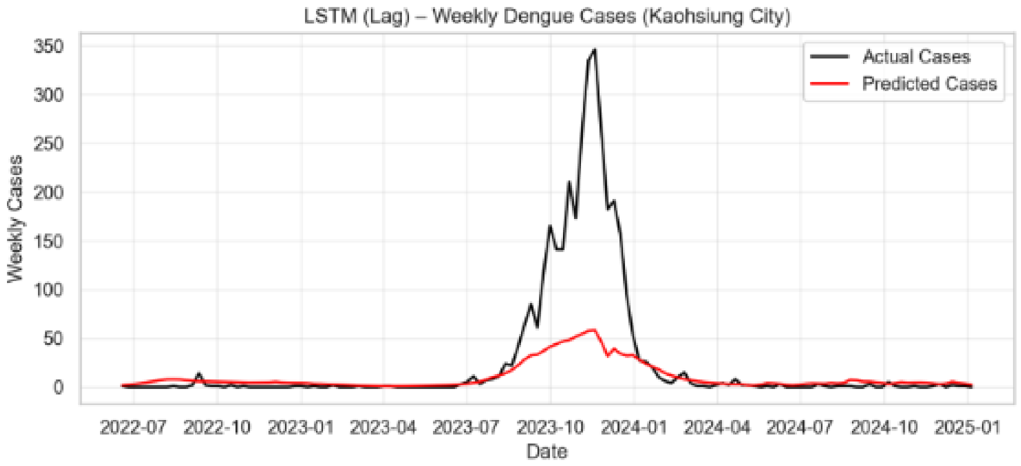
RMSE	MAE	MSE	R²
25.53	8.15	651.8	0.822

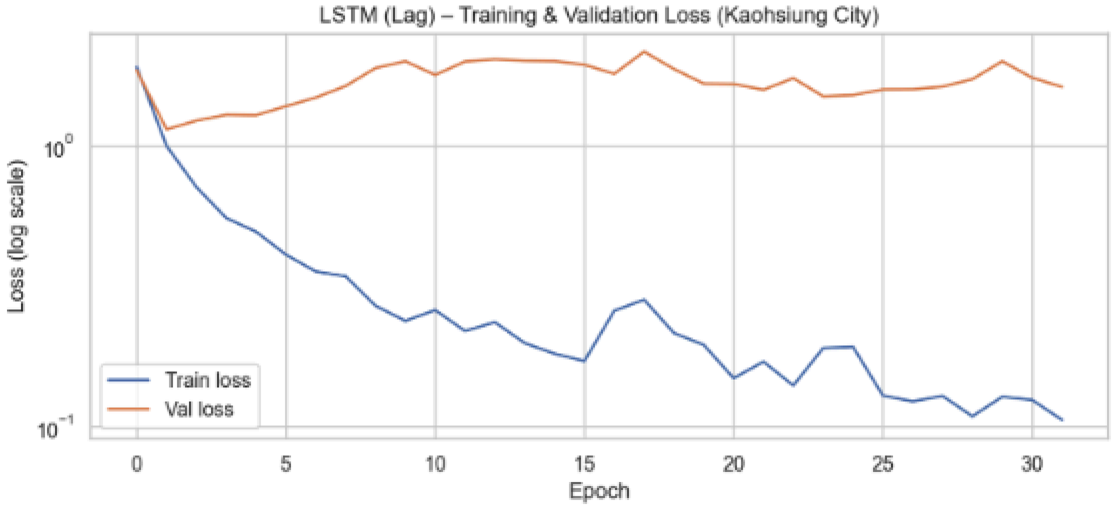
Figure 16. XGBoost predictions and top 8 feature importances.

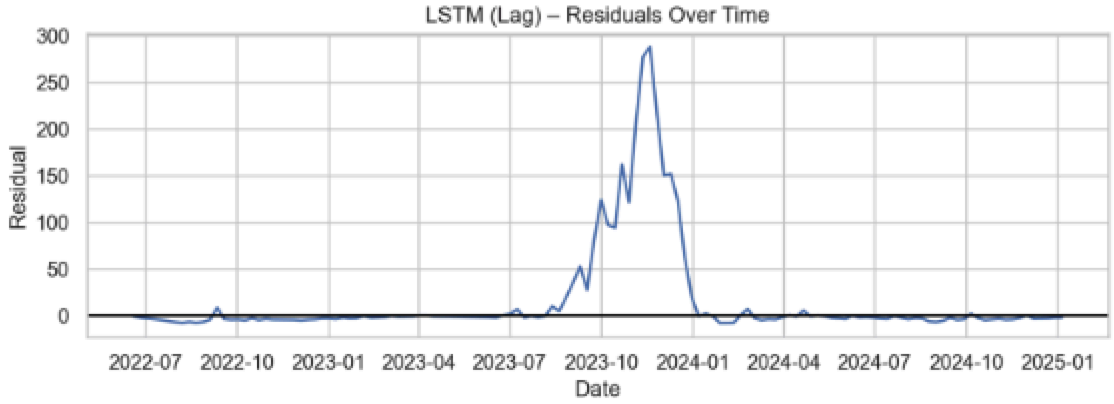
LSTM

LSTM performed poorly (RMSE = 54.83; MAE = 20.01; R² = 0.288), generating overly smooth predictions and underestimating the outbreak peak. Training/validation divergence indicated overfitting.

RMSE	MAE	MSE	R^2
54.83	20.01	3006.4	0.288



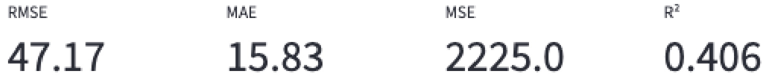




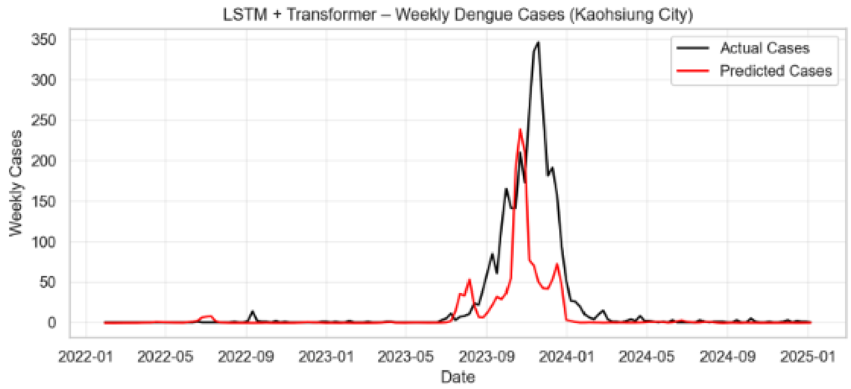
Figures 17-18. LSTM predictions, loss curves, and residuals.

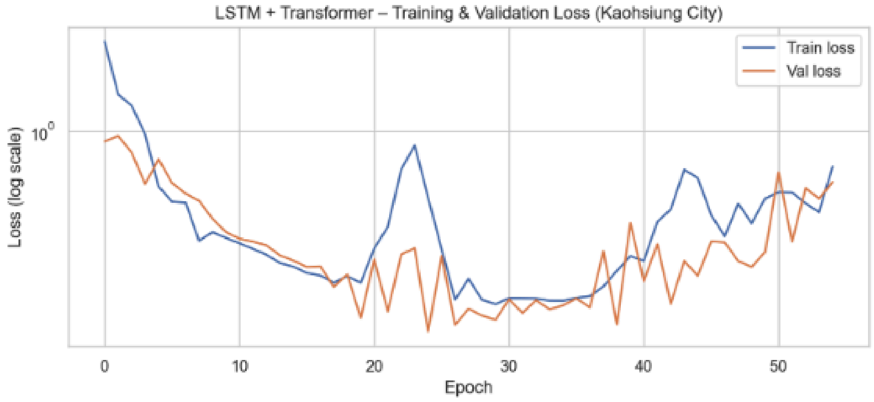
LSTM-Transformer

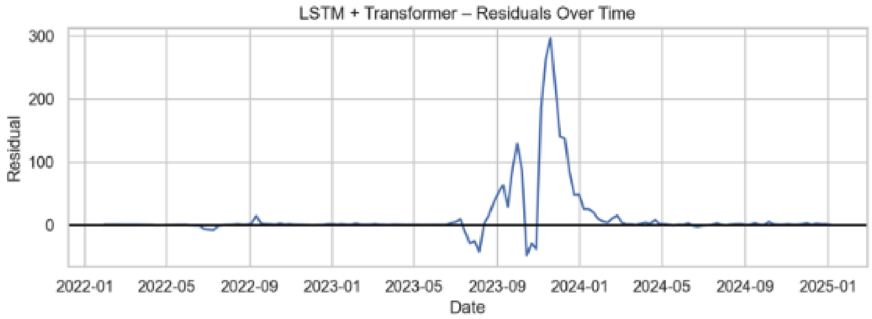
LSTM-Transformer improved timing but continued to underestimate peak magnitude (RMSE = 47.17; MAE = 15.83; R² = 0.406). Training loss showed unstable convergence.



RMSE	MAE	MSE	R²
47.17	15.83	2225.0	0.406







Figures 19-20. LSTM-Transformer predictions, loss curves, and residuals.

Summary - Kaohsiung:

Tree-based models significantly outperformed deep-learning models, reflecting the episodic, nonlinear nature of Kaohsiung's outbreaks and the value of engineered lag features.

5.4 Model Performance - Tainan City

Model behavior in Tainan mirrored Kaohsiung, but the extreme 2015 and 2023-2024 outbreaks amplified performance differences. Random Forest (RMSE = 121.21; R^2 = 0.918) and XGBoost (RMSE = 113.88; R^2 = 0.928) closely tracked the outbreak, with recent case lags and climatic features driving predictions.
LSTM (RMSE = 483.51; R^2 = -0.122) and LSTM-Transformer (RMSE = 469.91; R^2 = -0.115) failed to model the sharp epidemic peak, showing underprediction, unstable validation loss, and large residuals.

Full visualizations for Tainan City are available at:
https://dengue-taiwan-forecast.onrender.com/

5.5 Model Performance - Pingtung County

In Pingtung's lower-incidence setting, Random Forest (RMSE = 2.95; R^2 = 0.856) and XGBoost (RMSE = 2.99; R^2 = 0.851) performed robustly, accurately reflecting outbreak timing and modest peak heights. Feature importance emphasized case lags, supplemented by precipitation and temperature lags.

LSTM (RMSE = 7.97; R^2 = 0.085) and LSTM-Transformer (RMSE = 4.30; R^2 = 0.700) underpredicted outbreak magnitude and showed unstable validation behavior.

Full visualizations for Pingtung County are available at:
https://dengue-taiwan-forecast.onrender.com/

5.6 Model Comparison Across Cities

Across all three regions, Kaohsiung, Tainan, and Pingtung, the model comparison results show a clear and consistent pattern: tree-based models (Random Forest and XGBoost) outperformed the deep-learning models (LSTM and LSTM-Transformer) across every evaluation metric. In Kaohsiung, Random Forest achieved the lowest error (RMSE = 19.28, R^2 = 0.899), followed closely by XGBoost (RMSE = 25.53, R^2 = 0.822), while both LSTM-based models substantially underpredicted the outbreak peak. Tainan exhibited the largest disparities among models due to its extreme 2023-2024 outbreak. XGBoost delivered the strongest performance (RMSE = 113.88, R^2 = 0.928), slightly surpassing Random Forest, whereas LSTM and LSTM-Transformer performed poorly, with very large errors and negative R^2 values, reflecting their difficulty modeling Tainan's steep and highly nonlinear epidemic curves. In Pingtung, where dengue incidence is lower and outbreaks are more irregular, both Random Forest and XGBoost performed exceptionally well (RMSE ≈ 3.0, $R^2 \approx 0.85$), demonstrating strong robustness in low-count settings. The deep-learning models again lagged behind, with LSTM producing the weakest performance and the LSTM-Transformer offering moderate improvement but still falling short of tree-based approaches. Overall, recent case lags consistently emerged as the most influential predictors, while climatic and entomological lags contributed more variably across regions. Collectively, these results indicate that tree-based machine-learning models provide the most accurate and stable dengue forecasts for southern Taiwan, whereas deep-learning architectures struggle under conditions of limited outbreak history, sharp epidemic peaks, and highly skewed case distributions.

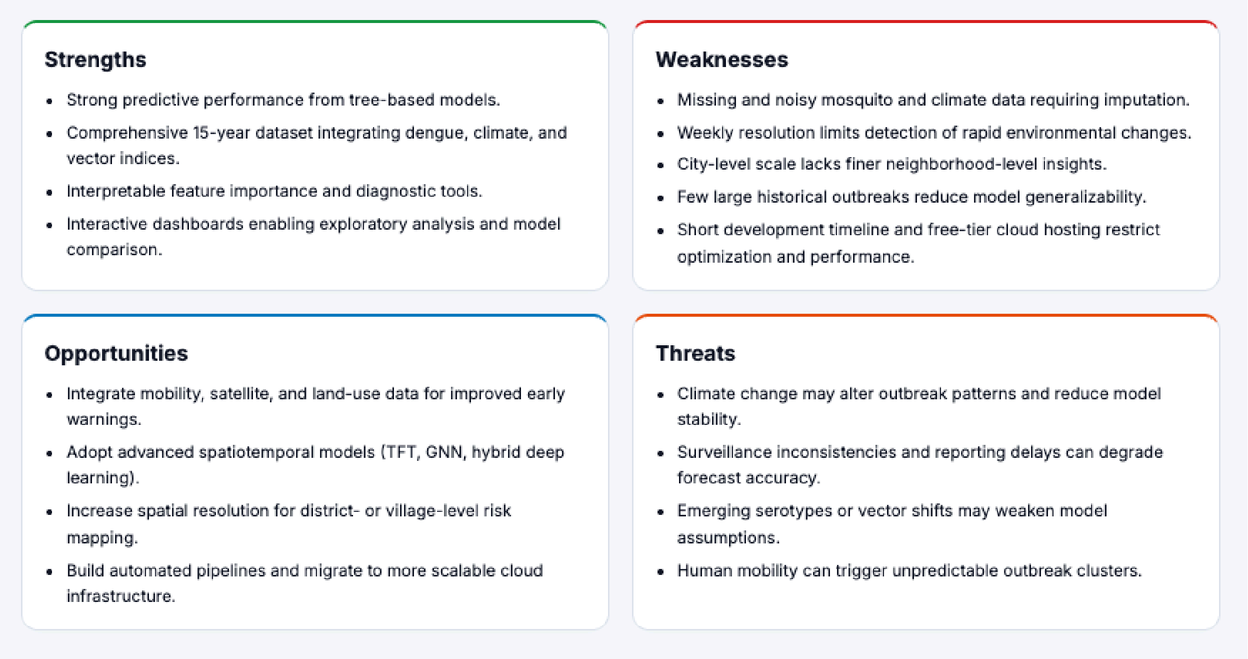
6. Discussion

This study examined dengue transmission dynamics and forecasting performance across Kaohsiung City, Tainan City, and Pingtung County by integrating epidemiological, climatic, and entomological data into a unified analytical framework. Seasonal climatic patterns, previously described in the EDA, were closely aligned with dengue activity in all regions, with transmission consistently peaking during late summer and early autumn. Autocorrelation analysis confirmed strong short-term temporal dependence at lags of one to three weeks and meaningful influence extending up to roughly eight to twelve weeks, underscoring the importance of incorporating multi-week lag features into forecasting models.

Although all three regions share similar seasonal patterns, their outbreak behaviors differed substantially. Tainan exhibited the most intense epidemic surges, Kaohsiung showed moderate recurrent peaks, and Pingtung experienced smaller, irregular outbreaks. These differences likely reflect variation in demographic density, urban structure, mosquito habitat distribution, and microclimatic conditions, emphasizing the need for region-specific forecasting strategies.

Across all cities, the model comparison revealed a consistent pattern: **tree-based machine-learning models (Random Forest and XGBoost) outperformed sequential deep-learning models**. Tree-based models demonstrated strong accuracy, effectively capturing outbreak timing and magnitude and showing robustness even in regions with limited historical outbreaks. Their advantage stems from their ability to leverage engineered lag features and model nonlinear relationships without requiring large amounts of sequential training data. In contrast, LSTM and LSTM-Transformer models struggled with sparse outbreak histories and tended to generate overly smoothed forecasts that underpredicted rapid epidemic escalation. These challenges were most evident in Tainan, where extreme peaks exposed the difficulty of learning complex temporal dynamics from highly skewed time series.

The regional comparison further highlights the importance of aligning forecasting tools with local epidemiological characteristics. While tree-based models remained strong across all regions, deep-learning models were particularly sensitive to outbreak magnitude and data scarcity. These findings suggest that traditional machine-learning approaches may currently be more operationally reliable than deep-learning architectures for dengue forecasting in Taiwan's outbreak-sparse and climate-driven context.



Strengths

- Strong predictive performance from tree-based models.
- Comprehensive 15-year dataset integrating dengue, climate, and vector indices.
- Interpretable feature importance and diagnostic tools.
- Interactive dashboards enabling exploratory analysis and model comparison.

Weaknesses

- Missing and noisy mosquito and climate data requiring imputation.
- Weekly resolution limits detection of rapid environmental changes.
- City-level scale lacks finer neighborhood-level insights.
- Few large historical outbreaks reduce model generalizability.
- Short development timeline and free-tier cloud hosting restrict optimization and performance.

Opportunities

- Integrate mobility, satellite, and land-use data for improved early warnings.
- Adopt advanced spatiotemporal models (TFT, GNN, hybrid deep learning).
- Increase spatial resolution for district- or village-level risk mapping.
- Build automated pipelines and migrate to more scalable cloud infrastructure.

Threats

- Climate change may alter outbreak patterns and reduce model stability.
- Surveillance inconsistencies and reporting delays can degrade forecast accuracy.
- Emerging serotypes or vector shifts may weaken model assumptions.
- Human mobility can trigger unpredictable outbreak clusters.

Figure 21. SWOT analysis highlighting the forecasting system's strengths, weaknesses, opportunities for enhancement, and external threats affecting long-term applicability.

To contextualize the performance and practical relevance of the forecasting system, a SWOT analysis was conducted. The system's **strengths** include the strong predictive performance of Random Forest and XGBoost, the integration of a robust 15-year multi-source dataset, interpretable feature importance outputs, and the availability of interactive visual dashboards (ArcGIS and Streamlit) that support real-time exploration. The key **weaknesses** involve data limitations, such as irregular mosquito index collection, sensor errors in climate data, and weekly temporal resolution, as well as the city-level aggregation that limits spatial precision. The system's **opportunities** lie in incorporating additional data sources (mobility, satellite imagery, land-use information), adopting more advanced spatiotemporal architectures, automating real-time pipelines, and scaling to finer spatial units for targeted control. Potential **threats** include climate change, changes in surveillance quality, shifting dengue serotypes or vectors, and unexpected mobility-driven outbreak clusters that may challenge long-term model stability.

Overall, this study demonstrates the value of integrating multi-source environmental and epidemiological data into operational dengue forecasting models. Given their reliability and interpretability, tree-based models remain the most practical tools for short-term forecasting in Taiwan, while deep-learning approaches may require richer, more consistent outbreak histories to achieve comparable performance. These insights can guide future development of early-warning systems and inform targeted vector-control efforts across southern Taiwan.

7. Conclusion

This study developed a multi-source dengue forecasting framework for Kaohsiung City, Tainan City, and Pingtung County by integrating 15 years of dengue surveillance, climatic variables, rainfall, mosquito indices, and demographic information. Through a comparative evaluation of four forecasting approaches (Random Forest, XGBoost, LSTM and a hybrid LSTM-Transformer model) the results demonstrate that **tree-based machine-learning models consistently provide the highest predictive accuracy and the most reliable performance** across all regions. Random Forest and XGBoost effectively captured outbreak timing, magnitude, and seasonal transmission patterns, outperforming sequential deep-learning models that struggled with sharp epidemic peaks, sparse outbreak histories, and highly skewed case distributions.

The analysis also highlights the central role of climatic and entomological drivers in shaping dengue dynamics in southern Taiwan. Strong seasonal cycles, coupled with short-term autocorrelation and multi-week climatic dependencies, underscore the importance of lagged predictors in dengue forecasting. Regional differences in outbreak intensity, particularly Tainan's extreme epidemic surges and Pingtung's smaller, irregular patterns, further emphasize the need for **region-specific modeling strategies** and careful consideration of local transmission environments.

Beyond methodological insights, the forecasting system developed in this study offers practical value for public health. The integration of interpretable feature importance outputs, GIS-based spatial visualization, and a Streamlit-based interactive dashboard provides a foundation for **operational early-warning systems** that can support real-time risk assessment and decision-making. These tools can help health authorities anticipate outbreak onset, allocate vector control resources more efficiently, and plan interventions during high-risk periods.

Despite its strengths, the framework faces limitations related to weekly temporal resolution, intermittent mosquito index data, and city-level spatial granularity. Future work should explore finer spatial scales, incorporate mobility or land-use data, and expand to advanced spatiotemporal deep-learning architectures as more outbreak data become available. Continued improvements in data quality, surveillance consistency, and integration of real-time environmental information will be essential for enhancing forecasting performance.

In summary, this study demonstrates that **data-driven, machine-learning-based forecasting provides a practical and effective approach to supporting dengue preparedness in Taiwan**, with tree-based models showing particular promise for short-term operational use. As climate variability and dengue risk continue to evolve, robust forecasting frameworks such as the one developed here will play an increasingly important role in guiding proactive and targeted public health responses.

8. Limitations and Future Work

This study has several limitations. First, the dataset contains substantial missing values, particularly in mosquito surveillance indices (BI, HI, CI), which required forward- and backward-fill imputation. While necessary to maintain temporal continuity, this may smooth over genuine fluctuations in vector abundance. Meteorological data also included erroneous or missing sensor readings that required interpolation and filtering, introducing uncertainty into climate-related predictors. Second, the use of **weekly temporal resolution** may obscure rapid changes in weather and mosquito activity, reducing the ability of deep-learning models to learn fine-grained outbreak dynamics. Third, the analysis was conducted at the **city level**, limiting spatial granularity and potentially overlooking neighborhood-level variation in environmental conditions, human mobility, and vector habitats that influence dengue transmission. Fourth, the limited number of historically large outbreaks constrained the generalizability of sequential deep-learning models, especially in low-incidence regions such as Pingtung where outbreak patterns are sparse and irregular.

The study was also **completed within a five-week timeline**, which restricted the extent of model tuning, sensitivity analysis, and experimentation with additional forecasting architectures. With a longer development period, the modeling framework could be further optimized and expanded. In addition, the deployment architecture relies on **free-tier cloud services** (Streamlit, Render, and Supabase), which impose constraints on bandwidth, compute resources, and request rates. These limitations can lead to slow retrieval times, delayed loading of model artifacts, or occasional cold-start latency, reducing the responsiveness of the forecasting dashboard for real-time public health use.

Future work should focus on improving data quality and expanding the forecasting framework. Integrating additional data sources, such as human mobility traces, satellite-derived environmental indicators, or land-use information could enhance early detection of outbreak conditions. Increasing spatial resolution to the district or village level would improve risk mapping and support more targeted vector-control interventions. Exploring advanced spatiotemporal architectures, including Temporal Fusion Transformers or Graph Neural Networks, may further improve model performance as more outbreak data become available. Finally, migrating the system to more robust cloud infrastructure and incorporating automated real-time data pipelines would support the development of a fully operational dengue early-warning system for public health agencies.

References

Jiao, S., Wang, Y., Ye, X., Nagahara, L., & Sakurai, T. (2020). *Spatio-temporal epidemic forecasting using mobility data with LSTM networks and attention mechanism.* International Journal of Data Science and Analytics, 10(2), 99-113.

Lin, C. H., Wen, T. H., Teng, H. J., Chang, N. T., & Lin, Y. Y. (2022). *Real-time dengue forecast for outbreak alerts in Southern Taiwan.* PLoS Neglected Tropical Diseases, 16(7), e0010671.

Yeh, D. Y., Leu, J. H., Ye, S., & Cheng, C. H. (2018). *An intelligent autoregressive-distributed lag model: A climate-driven approach for predicting dengue fever incidence in Taiwan cities.* Environmental Research, 164, 311-319.

Chien, L. C., Yu, H. L., & Chuang, T. W. (2020). *Challenges and implications of predicting the spatiotemporal distribution of dengue fever outbreaks in Taiwan.* Scientific Reports, 10, 4863.

Lai, S. C., Ko, H. Y., Lin, Y. L., Chen, T. H., & Wu, H. S. (2017). *Dengue outbreaks and the geographic distribution of dengue vectors in Taiwan: A 20-year epidemiological analysis.* American Journal of Tropical Medicine and Hygiene, 97(3), 1128-1134.

Kuo, C. Y., Yang, W. W., & Su, E. C. (2021). *Improving dengue fever predictions in Taiwan based on feature selection and random forests.* Scientific Reports, 11, 11892.

Hii, Y. L., Zhu, H., Ng, N., Ng, L. C., & Rocklöv, J. (2012). *Forecast of dengue incidence using temperature and rainfall.* PLoS Neglected Tropical Diseases, 6(11), e1908.

Wu, P. C., Lay, J. G., Guo, H. R., Lin, C. Y., Lung, S. C., & Su, H. J. (2007). *Higher temperature and urbanization affect the spatial patterns of dengue fever transmission in subtropical Taiwan.* Science of the Total Environment, 407(7), 222-232.

Learning Experiences and Outcomes Summary

Over the 16-month Master of Business Data Analytics (BDA) program, I have undergone a transformative academic and professional development journey that strengthened my capabilities in data science, machine learning, artificial intelligence, and applied analytics. Entering the program, my goal was to bridge my technical background with advanced analytical skills and gain the confidence to tackle real-world problems using data-driven methods. Through rigorous coursework, hands-on projects, and continuous practical learning, the BDA program equipped me with the technical foundation, analytical mindset, and professional competencies required to excel in modern data-centric roles.

The curriculum provided a strong foundation in statistical reasoning, predictive modeling, and data management. I developed practical proficiency in Python, SQL, and R, working extensively with tools and libraries such as Pandas, NumPy, Scikit-learn, TensorFlow, Keras, and MLflow. Courses in machine learning, applied statistics, data mining, time series forecasting, big data analytics, and database systems deepened my understanding of regression, classification, ensemble learning, clustering, PCA, neural networks, and forecasting models. Additionally, I gained experience in data cleaning, feature engineering, database design, and exploratory data analysis, skills that supported every project I completed. Training in visualization tools such as Tableau, ArcGIS, and Streamlit helped me communicate insights effectively to both technical and non-technical audiences.

A central component of my learning was the sequence of applied projects that allowed me to translate theory into fully functional systems. One of the most impactful experiences was my **Dengue Forecasting Capstone Project,** where I built an end-to-end predictive modeling pipeline integrating epidemiological, climatic, and entomological data across three cities. I trained Random Forest, XGBoost, LSTM, and LSTM-Transformer models on a unified multi-city weekly dataset with engineered lag features and evaluated performance using RMSE, MAE, MSE, and R^2. For deployment, I built an interactive Streamlit dashboard and implemented a cloud architecture in which Supabase served as the PostgreSQL storage layer for predictions, metrics, and model artifacts, while Render Cloud hosted the application backend and handled data retrieval. This project strengthened my skills in MLOps, cloud deployment, and user-facing analytics.

Another significant milestone was the **AI-Powered Plant Disease Detection Project**, where I developed an image-based diagnostic tool using YOLOv8 combined with a LangChain-powered LLM for automated treatment recommendations. I deployed the system on Hugging Face Spaces to support smallholder farmers in Africa. This project enhanced my experience in computer vision, LLM integration, model deployment, cloud platforms, and socially responsible AI. Presenting the project at an international One Health conference strengthened my scientific communication skills and increased my confidence in sharing complex analytical work.

The **Smart Cities project on San Diego Homelessness** further broadened my experience by integrating GIS analysis, census data, PIT counts, and 311 reports into a spatial inequality dashboard. Using ArcGIS Online, Leaflet.js, and HTML/CSS, I created an interactive platform to visualize housing burden, homelessness patterns, and environmental injustice in Downtown San Diego. This project taught me how data analytics can inform policy, planning, and community-focused solutions.

Throughout the BDA program, I also built strong professional competencies. Group projects enhanced my collaboration and communication skills, while frequent presentations and written reports helped me articulate technical findings clearly. I became proficient with GitHub and adopted systematic workflows for version control, reproducible analysis, and end-to-end project management. The program strengthened my ability to think critically, troubleshoot complex analytical problems, and design solutions that balance accuracy, interpretability, and operational feasibility.

Overall, the 16-month BDA program prepared me to excel in data-driven roles by providing a strong technical foundation, extensive hands-on experience, and exposure to real-world applications across domains such as epidemiology, agriculture, and smart cities. I now feel confident pursuing opportunities in machine learning engineering, data science, and applied AI development, and I look forward to contributing to impactful projects that leverage data for meaningful decision-making.