End-to-End Dengue Forecasting in Southern Taiwan Using Machine Learning, Deep Learning, 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 and increasingly severe public health threat in southern Taiwan, where climatic seasonality, rapid urbanization, and expanding mosquito habitats create favorable conditions for transmission. This project develops an end-to-end dengue forecasting framework for Kaohsiung, Tainan, and Pingtung by integrating 15 years of dengue surveillance, meteorological data, rainfall, mosquito indices (BI, HI, CI), and population density into a unified weekly dataset. Four models (Random Forest, XGBoost, an LSTM network with engineered lag features, and a hybrid LSTM-Transformer) were trained using time-based validation and evaluated on the 2022-2024 test period.

Across all cities, tree-based models substantially outperformed deep-learning architectures. Random Forest and XGBoost captured outbreak timing and magnitude with higher accuracy, demonstrating robustness in both high-incidence (Kaohsiung, Tainan) and low-count settings (Pingtung). LSTM-based models struggled with sparse outbreak histories and produced overly smoothed predictions, underscoring

the data-intensive nature of sequential neural models. Autocorrelation findings further validated the importance of strong short-term and mid-range temporal dependence in dengue dynamics.

To enhance usability and reproducibility, all model artifacts were stored in Supabase and integrated into an interactive Streamlit dashboard deployed on Render Cloud, complemented by an ArcGIS dashboard for spatial visualization. These tools provide practical early-warning insights for public health planning and establish a reproducible foundation for future dengue forecasting systems.

1. Introduction

Dengue fever is one of the most persistent and escalating public health challenges in Taiwan, particularly in the southern regions of Kaohsiung City, Tainan City, and Pingtung County. These areas are characterized by warm temperatures, high humidity, intense monsoon rainfall, and dense urban environments that support Aedes aegypti, the primary dengue vector. Over the past two decades, the frequency and severity of dengue outbreaks have increased, with major epidemics occurring in 2002, 2014-2015, and most recently in 2023. As climate extremes intensify and mosquito habitats expand, public health agencies increasingly require proactive forecasting tools to support early intervention and outbreak preparedness.

Traditional dengue surveillance in Taiwan relies on confirmed case reporting, environmental inspections, and monitoring of mosquito indices such as the Breteau Index (BI), House Index (HI), and Container Index (CI). While essential, these systems often lag behind actual transmission dynamics and do not fully capture the nonlinear and delayed relationships between dengue incidence, climate variability, and vector abundance. These limitations highlight the need for more advanced, data-driven forecasting approaches.

Recent advances in machine learning (ML) and deep learning (DL) offer promising alternatives. Tree-based methods such as Random Forest and XGBoost are well-suited for tabular epidemiological data, capable of modeling nonlinear relationships and providing interpretable feature importance. Deep learning architectures,including Long Short-Term Memory (LSTM) networks and Transformer-based attention models,have demonstrated strong performance in capturing temporal dependencies in complex datasets and have been applied to epidemic modeling in various contexts.

This study develops an end-to-end dengue forecasting framework for southern Taiwan using 15 years of multi-source data, including dengue surveillance, meteorological variables, rainfall, mosquito indices, and population density. Four forecasting models : Random Forest, XGBoost, LSTM, and a hybrid LSTM-Transformer,are compared using a unified weekly dataset and time-based validation. The framework incorporates cloud-based artifact storage using Supabase and an interactive Streamlit dashboard, enabling real-time visualization and comparative diagnostics.

The goal of this project is to evaluate the relative strengths of ML and DL approaches under Taiwan’s outbreak-sparse epidemiological conditions and to develop a practical, interpretable

tool to support early-warning systems, vector-control strategies, and public health decision-making.

2. Literature Review

Dengue transmission is strongly shaped by environmental and climatic conditions, and a substantial body of research has documented the influence of temperature, rainfall, humidity, and mosquito abundance on outbreak dynamics. Warmer temperatures accelerate mosquito development and shorten the viral extrinsic incubation period, while rainfall expands breeding sites and increases larval density. High humidity enhances adult mosquito survival, and windspeed affects vector dispersal patterns. These climatic effects often operate with multi-week delays, motivating the use of lagged predictors in forecasting models. Entomological indices such as the Breteau Index (BI), House Index (HI), and Container Index (CI) are widely recognized indicators of larval habitat density and have been shown to correlate strongly with dengue transmission risk, though their irregular collection schedules pose challenges for modeling.

Traditional time-series models such as ARIMA, SARIMA, and Poisson regression have been applied to dengue forecasting in various countries, including Taiwan, but they often struggle to capture the nonlinear, multi-factor interactions inherent in dengue epidemiology. Machine learning approaches have therefore gained prominence. Random Forest and XGBoost, in particular, have been demonstrated to offer high predictive accuracy in dengue studies across Southeast Asia, Latin America, and Taiwan. These tree-based models can accommodate complex lag structures, handle noisy or missing data, and provide interpretable feature importance scores. Studies in Taiwan, such as Kuo et al. and Su et al., have shown that incorporating climatic and mosquito indices significantly improves forecast performance, with recent case counts consistently emerging as the strongest predictors.

Deep learning methods have also been explored, particularly Long Short-Term Memory (LSTM) networks, which are designed to capture long-range temporal dependencies absent from traditional models. Prior research in Singapore, mainland China, and Taiwan has shown that LSTMs can model seasonal dengue patterns when sufficient high-frequency data are available. More advanced architectures incorporate attention mechanisms or Transformer encoders to selectively weight influential time steps. For example, Jiao et al. (2020), published in the *International Journal of Data Science and Analytics* (Springer Nature), demonstrated that an attention-enhanced LSTM model using human mobility data improved spatiotemporal epidemic forecasting. However, the performance of deep learning models is highly dependent on data richness and outbreak frequency. In regions like Taiwan,where large dengue outbreaks are infrequent and the time series is highly skewed, LSTM-based architectures may struggle to learn abrupt epidemic transitions without incorporating additional behavioral or mobility data sources.

Despite substantial progress in dengue modeling, several gaps remain in Taiwan's forecasting literature. Most studies focus on a single city, often Kaohsiung or Tainan, limiting understanding of regional variation across the country's major hotspots. Few studies train and evaluate models across multiple cities using standardized preprocessing and validation procedures. Mosquito indices, although operationally important for vector control agencies, are rarely integrated into machine learning or deep learning architectures. Attention-based and Transformer-based forecasting models remain largely unexplored in Taiwan, despite promising results in global studies. Finally, previous work has not typically paired forecasting models with cloud-based dashboards for operational use, limiting practical deployment.

This project contributes to the existing literature by integrating 15 years of dengue, climate, rainfall, mosquito, and demographic data into a unified multi-city forecasting framework and systematically comparing four modeling approaches (Random Forest, XGBoost, LSTM, and a hybrid LSTM-Transformer) across Kaohsiung, Tainan, and Pingtung. Through rigorous temporal validation and reproducible cloud deployment, the study advances both methodological understanding and practical applications of dengue forecasting in Taiwan.

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.

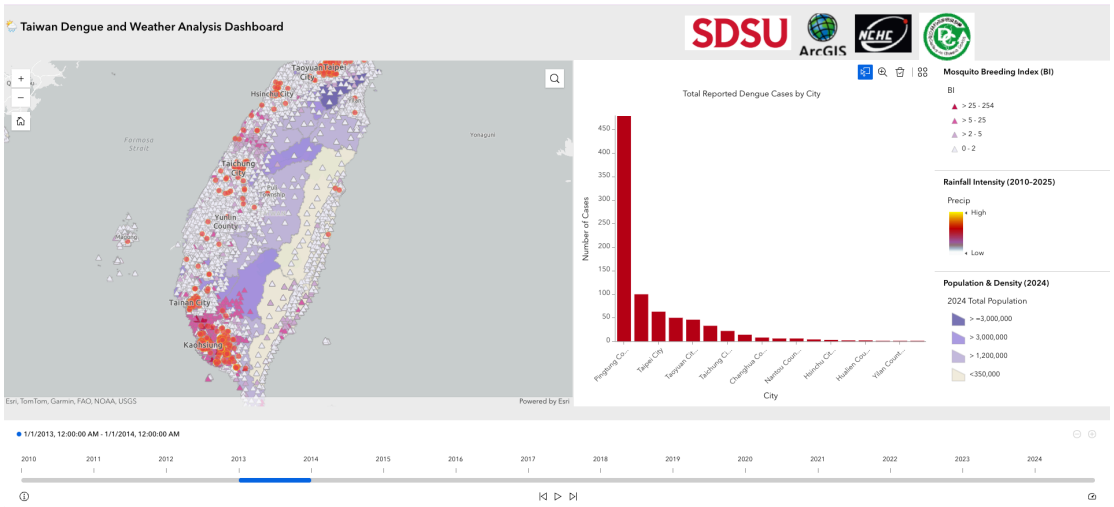


Figure 1. ArcGIS Dashboard for visualizing dengue cases, mosquito breeding indices, rainfall intensity, and city population (2010-2024). The interactive version is available at:

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 ($T_{min} \leq T_{mean} \leq T_{max}$). 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.

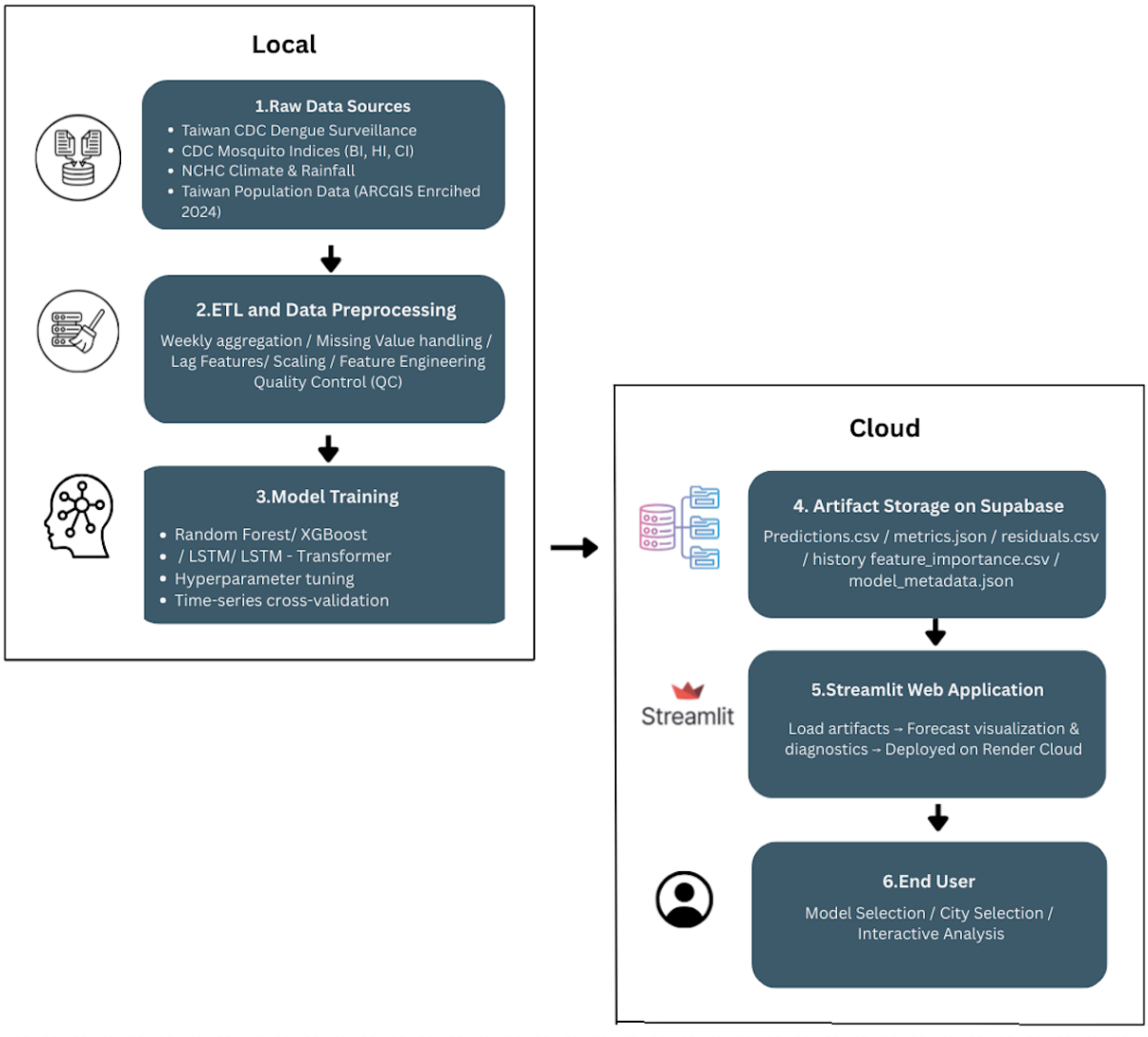


Figure 2. End-to-end analytical pipeline for dengue forecasting. The workflow integrates four major components:
(1) raw data ingestion from Taiwan CDC, CDC mosquito surveillance, NCHC climate datasets, and population density sources;
(2) local ETL and preprocessing, including weekly aggregation, missing 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 (Tmin ≤ Tmean ≤ Tmax) 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.

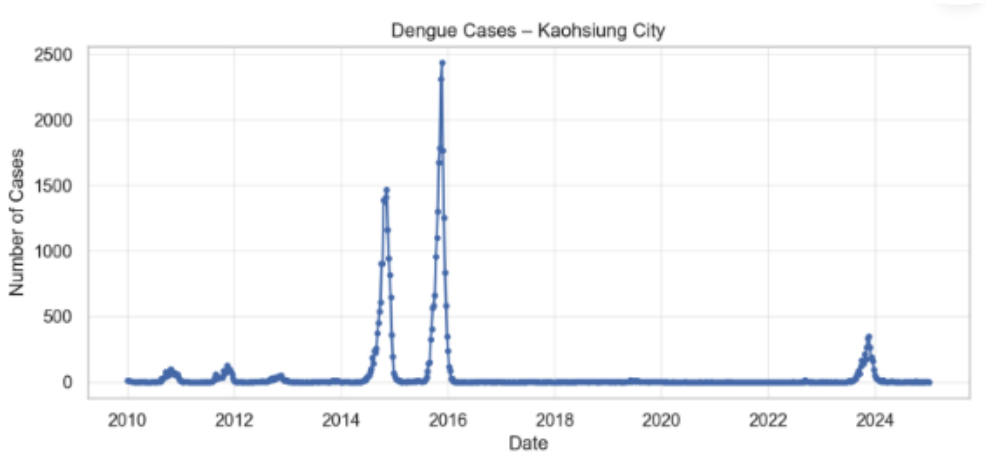


Figure 3. Weekly dengue cases in Kaohsiung City (2010-2024)

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.

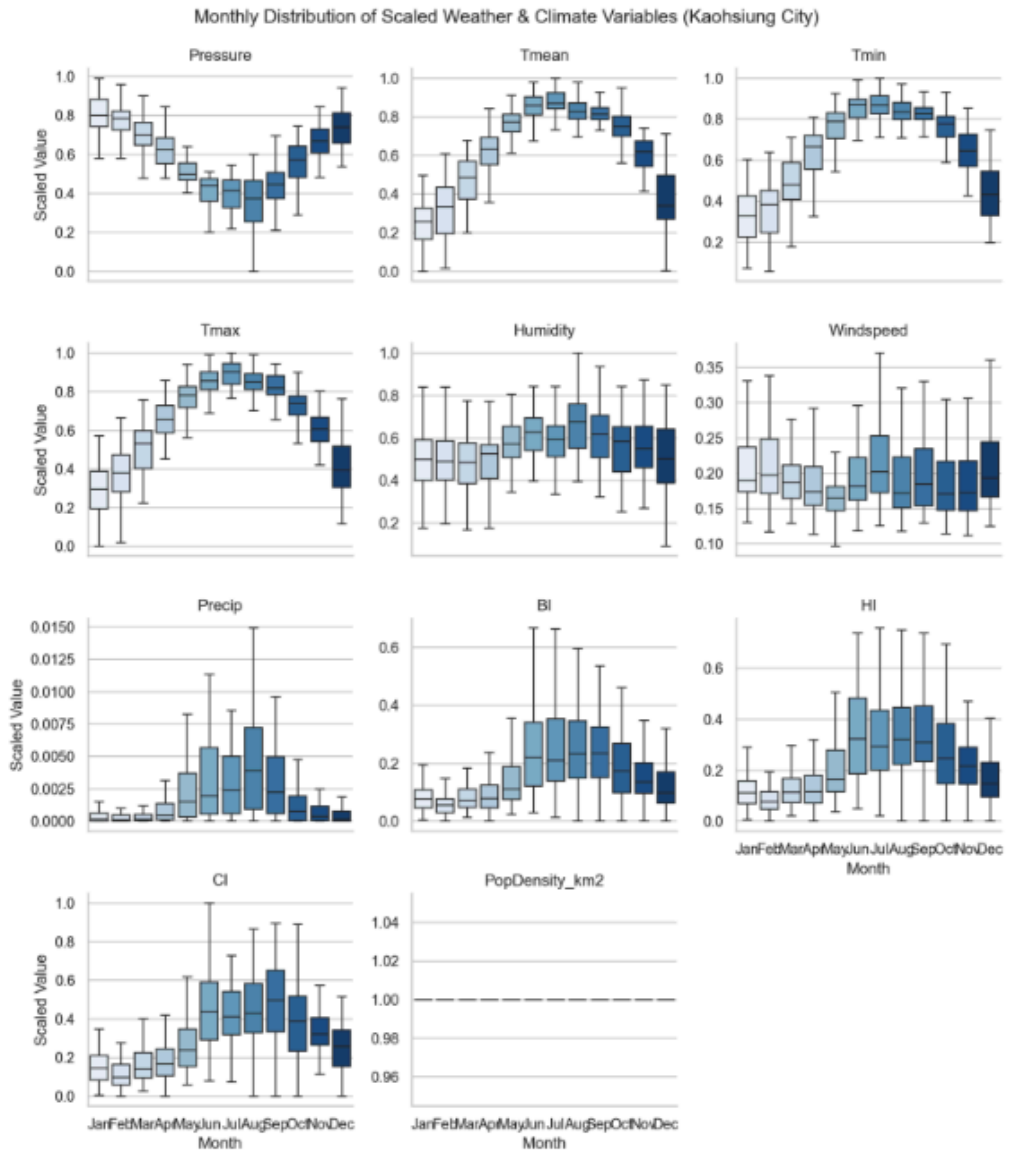


Figure 4. Monthly climate and mosquito indices in Kaohsiung City.

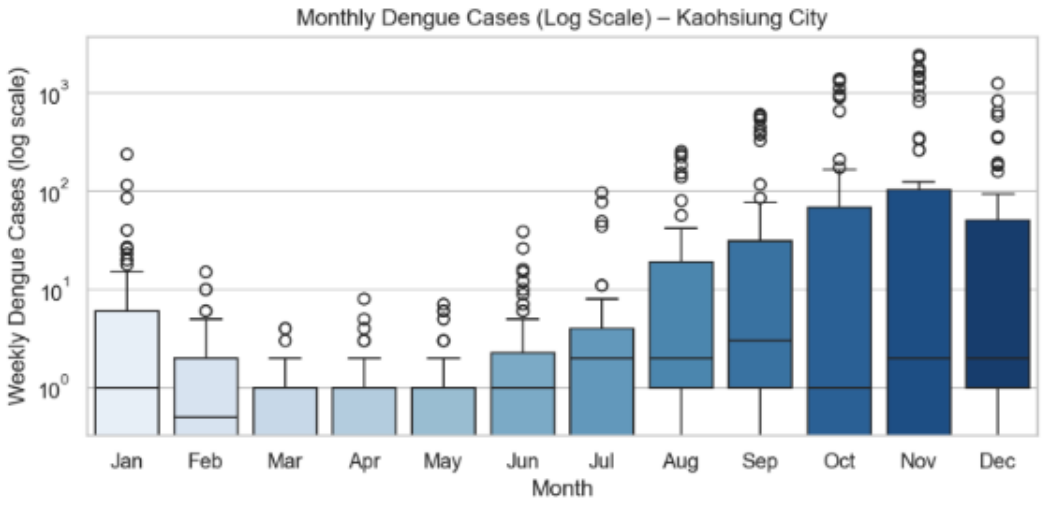


Figure 5. Monthly log-scale dengue cases in Kaohsiung City.

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.

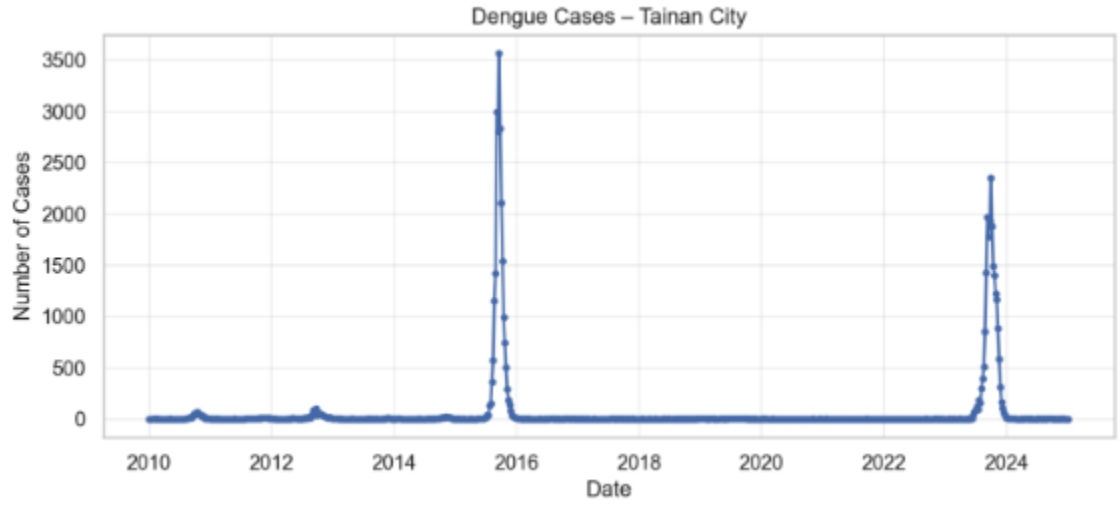


Figure 6. Weekly dengue cases in Tainan City (2010-2024)

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.

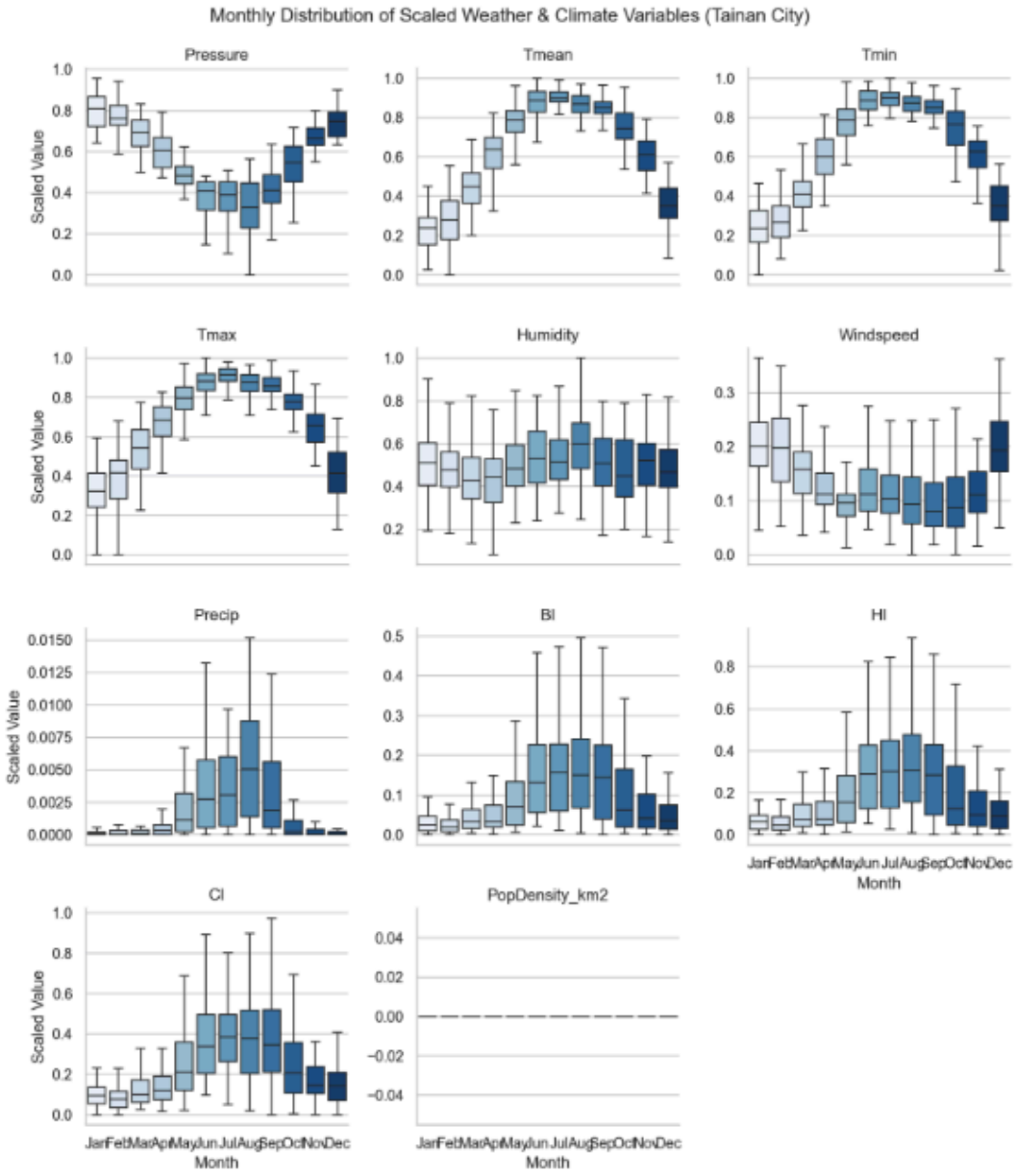


Figure 7. Monthly climate and mosquito indices in Tainan City.

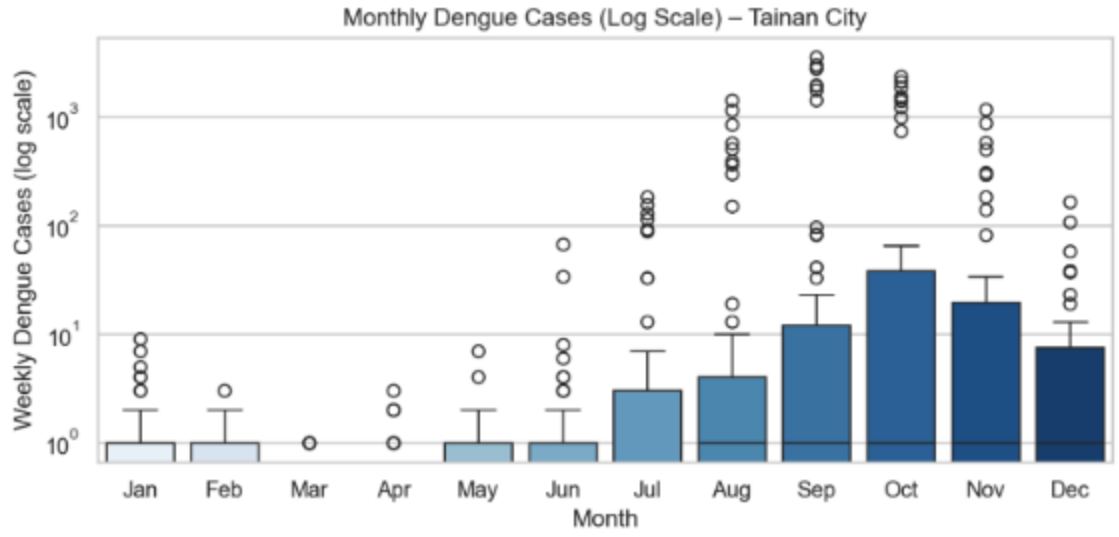


Figure 8. Monthly log-scale dengue cases in Tainan City.

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.

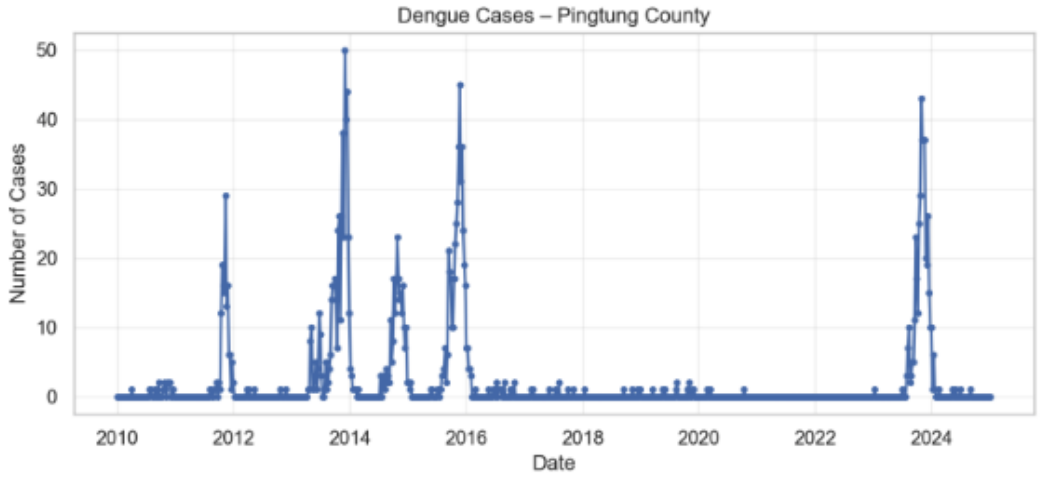


Figure 9. Weekly dengue cases in Pingtung County (2010-2024)

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.

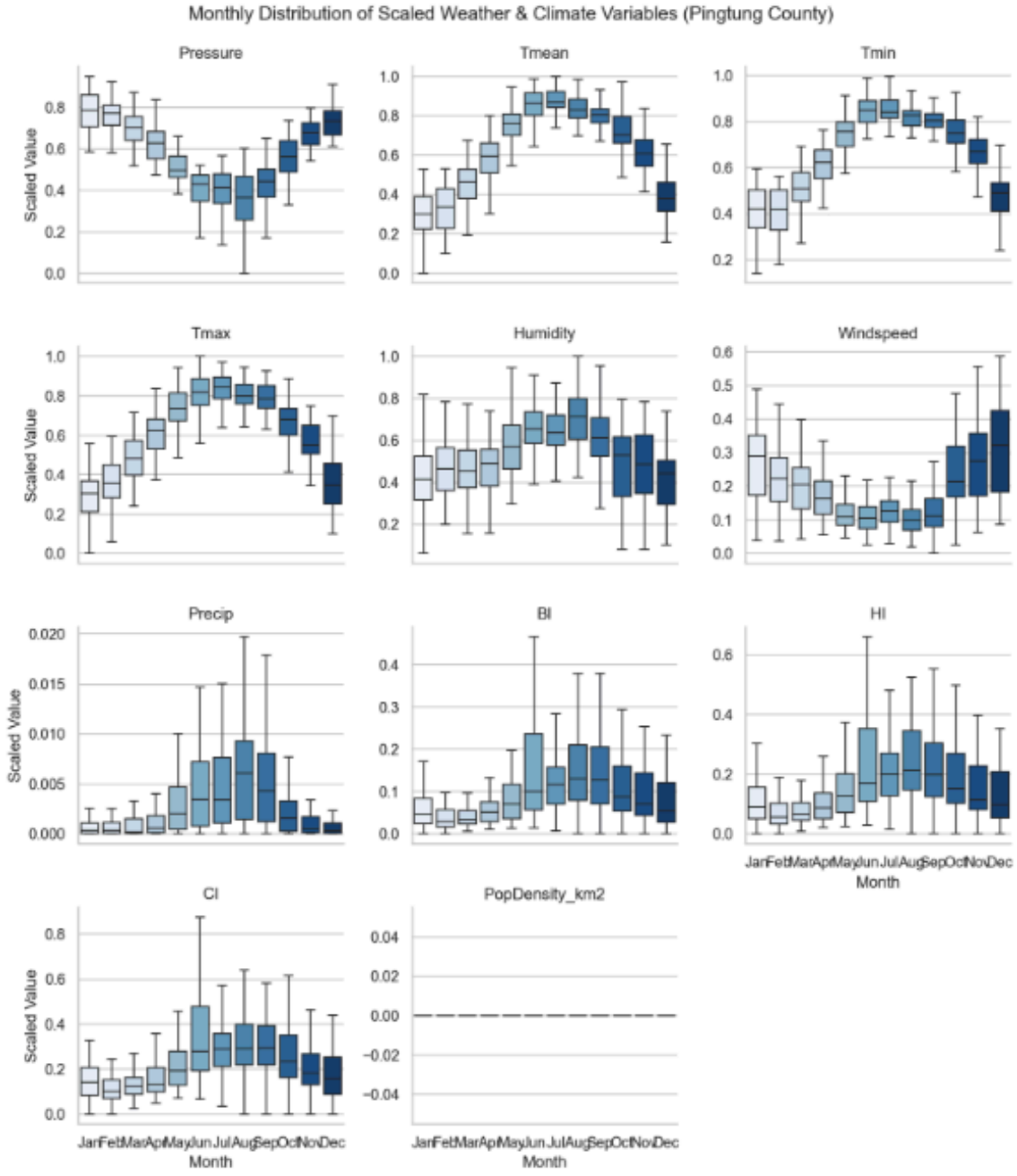


Figure 10. Monthly climate and mosquito indices in Pingtung County.

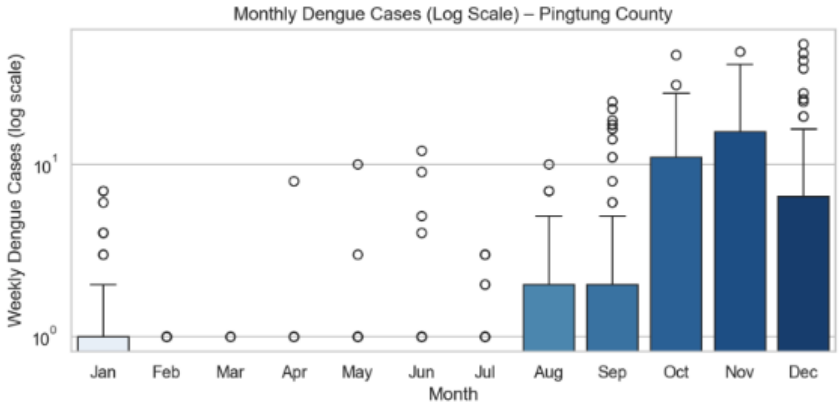
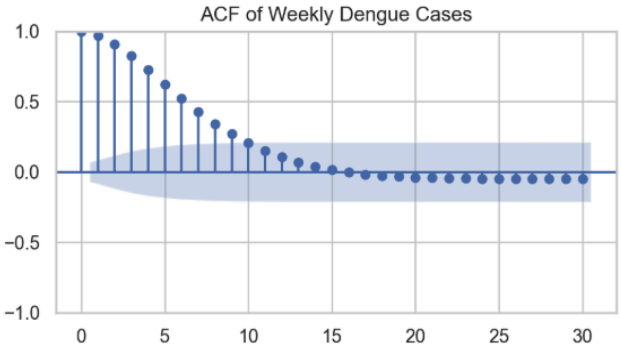


Figure 11. Monthly log-scale dengue cases in Pingtung County.

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.



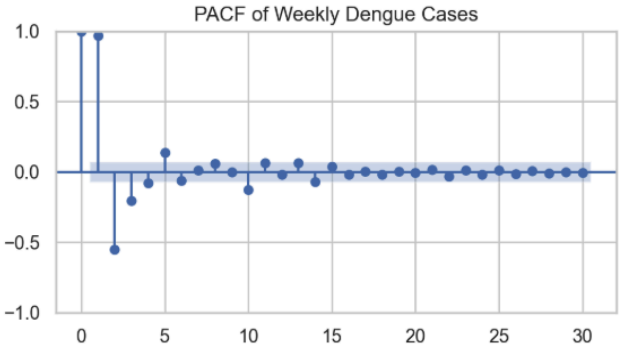


Figure 12. ACF and PACF of weekly dengue cases in Kaohsiung City. The ACF shows significant autocorrelation up to about 10 weeks, while the PACF indicates strong short-term effects at lags 1-3.

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.

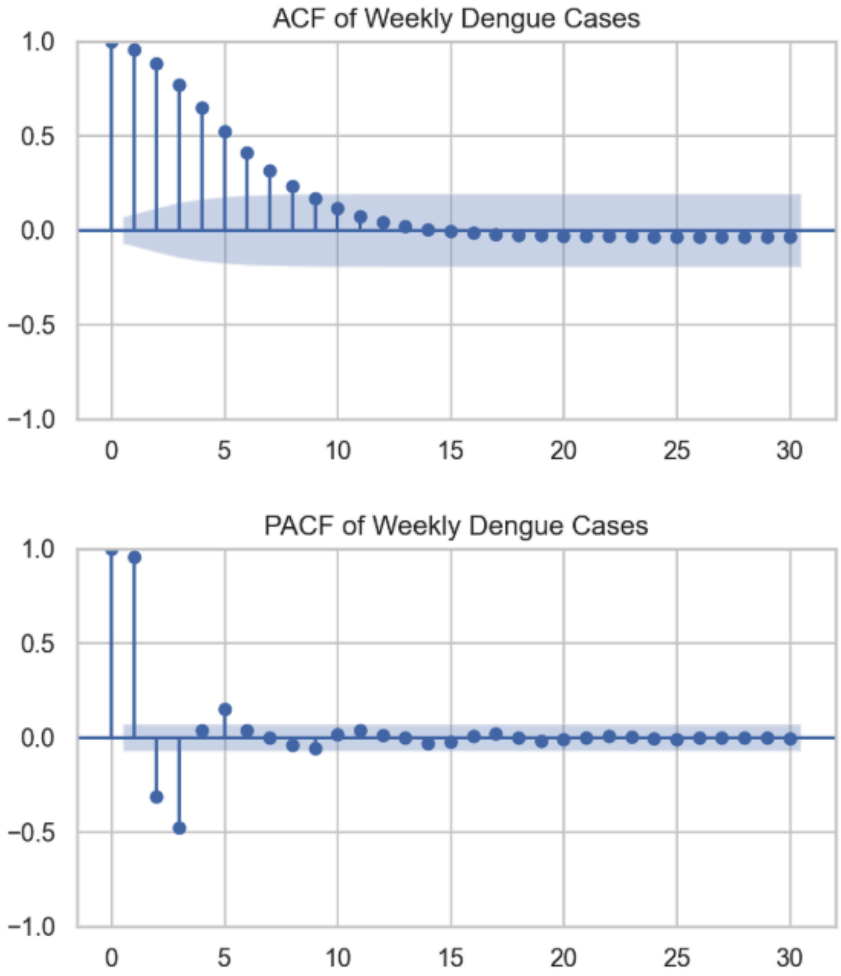


Figure 13. ACF and PACF for Tainan City

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.

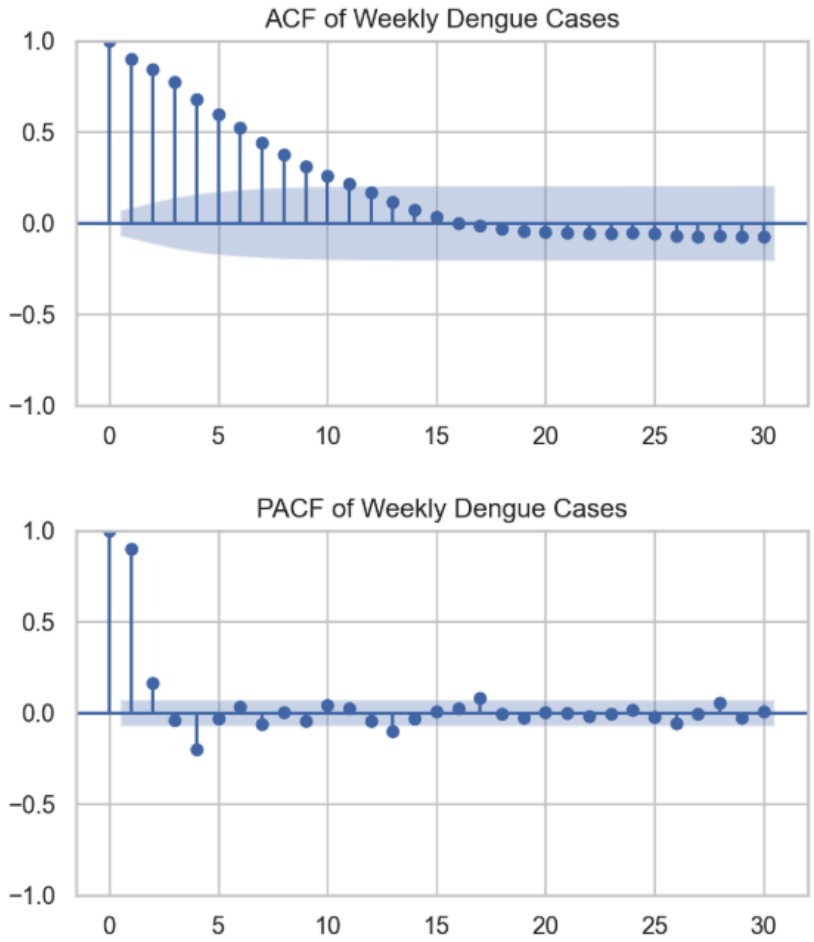


Figure 14. ACF and PACF for Pingtung County.

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.

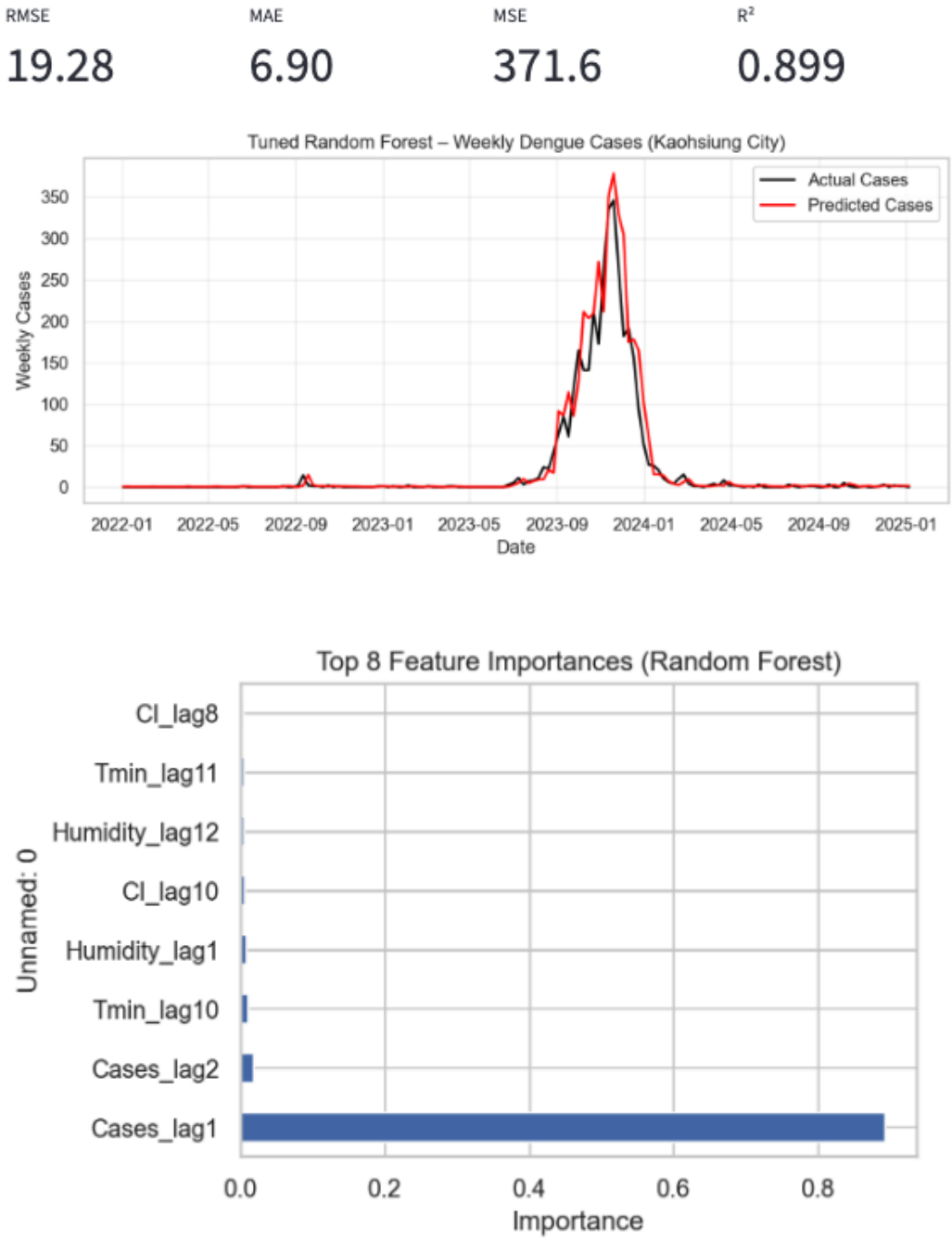


Figure 15. RF predictions and top 8 feature importances for Kaohsiung.

XGBoost

XGBoost performed well (RMSE = 25.53; MAE = 8.15; R^2 = 0.822) but tended to underpredict the peak weeks. Cases_lag1 remained the strongest predictor, followed by temperature and humidity lags.

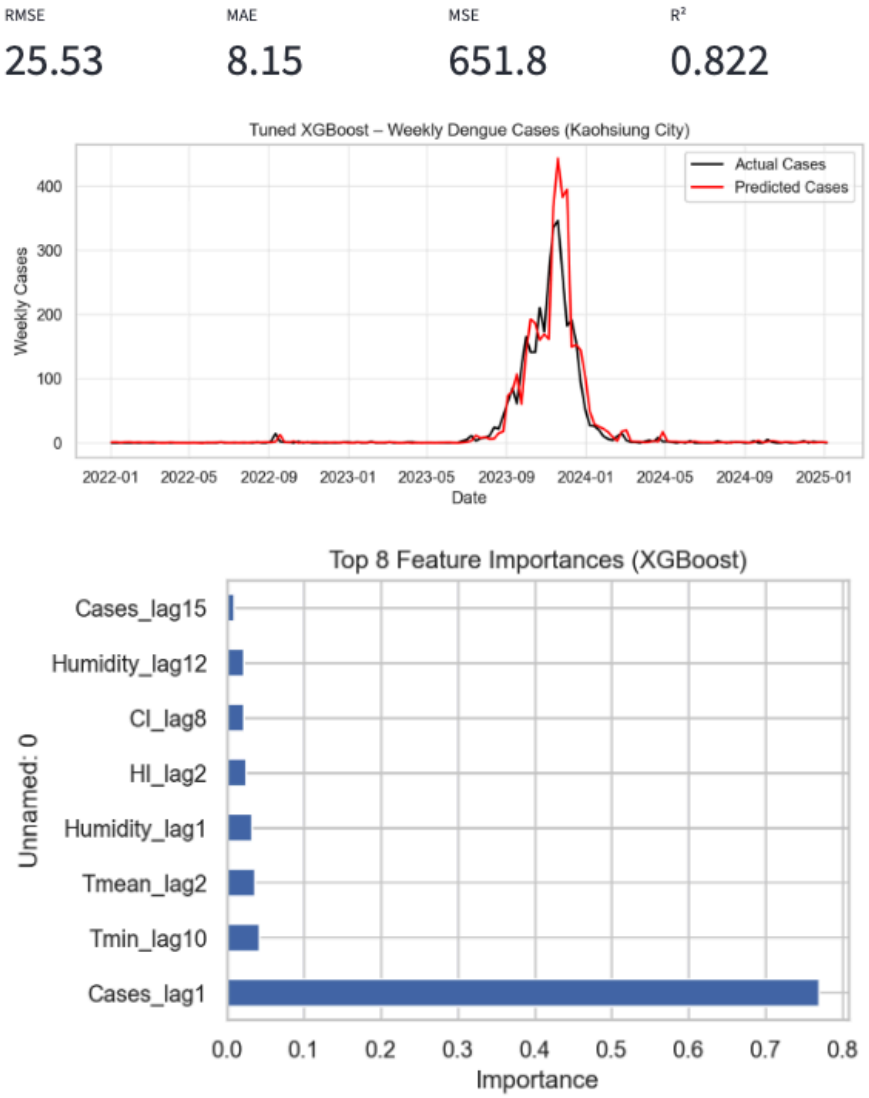
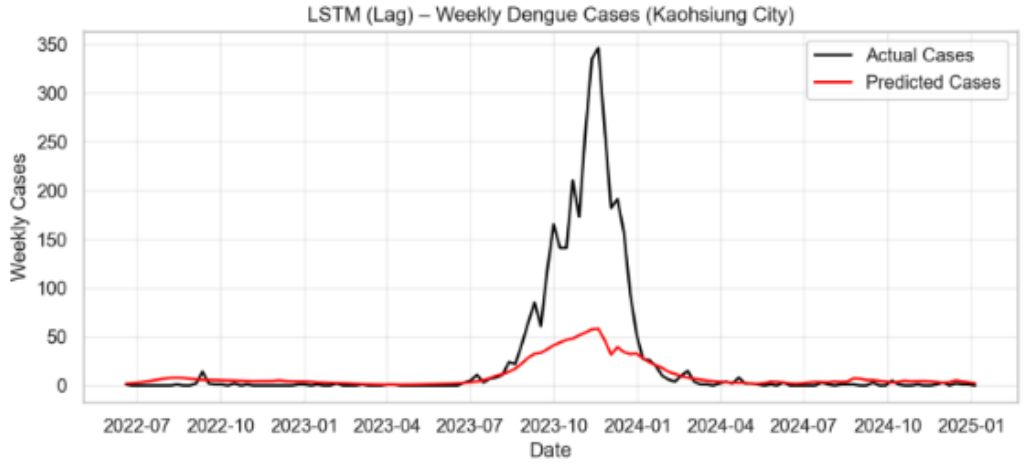


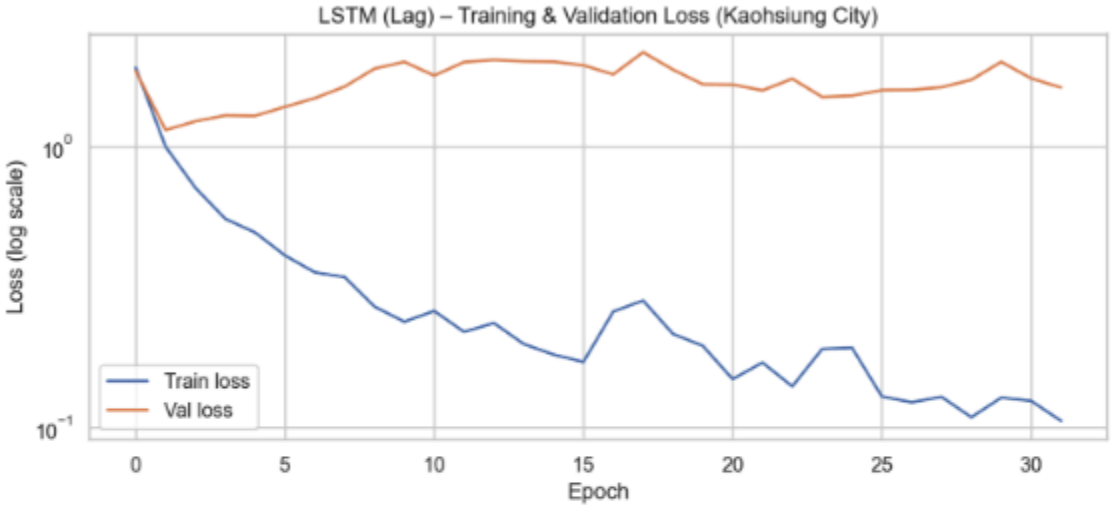
Figure 16. XGBoost predictions and top 8 feature importances.

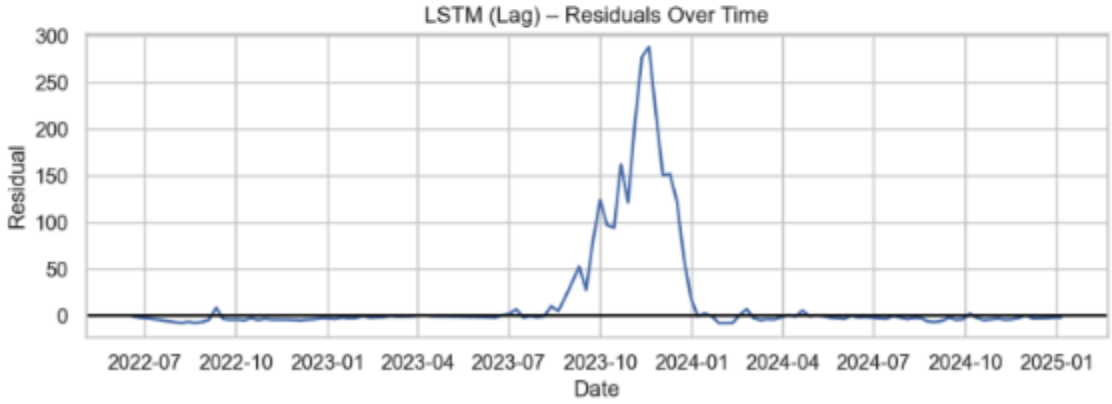
LSTM

LSTM performed poorly (RMSE = 54.83; MAE = 20.01; R^2 = 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



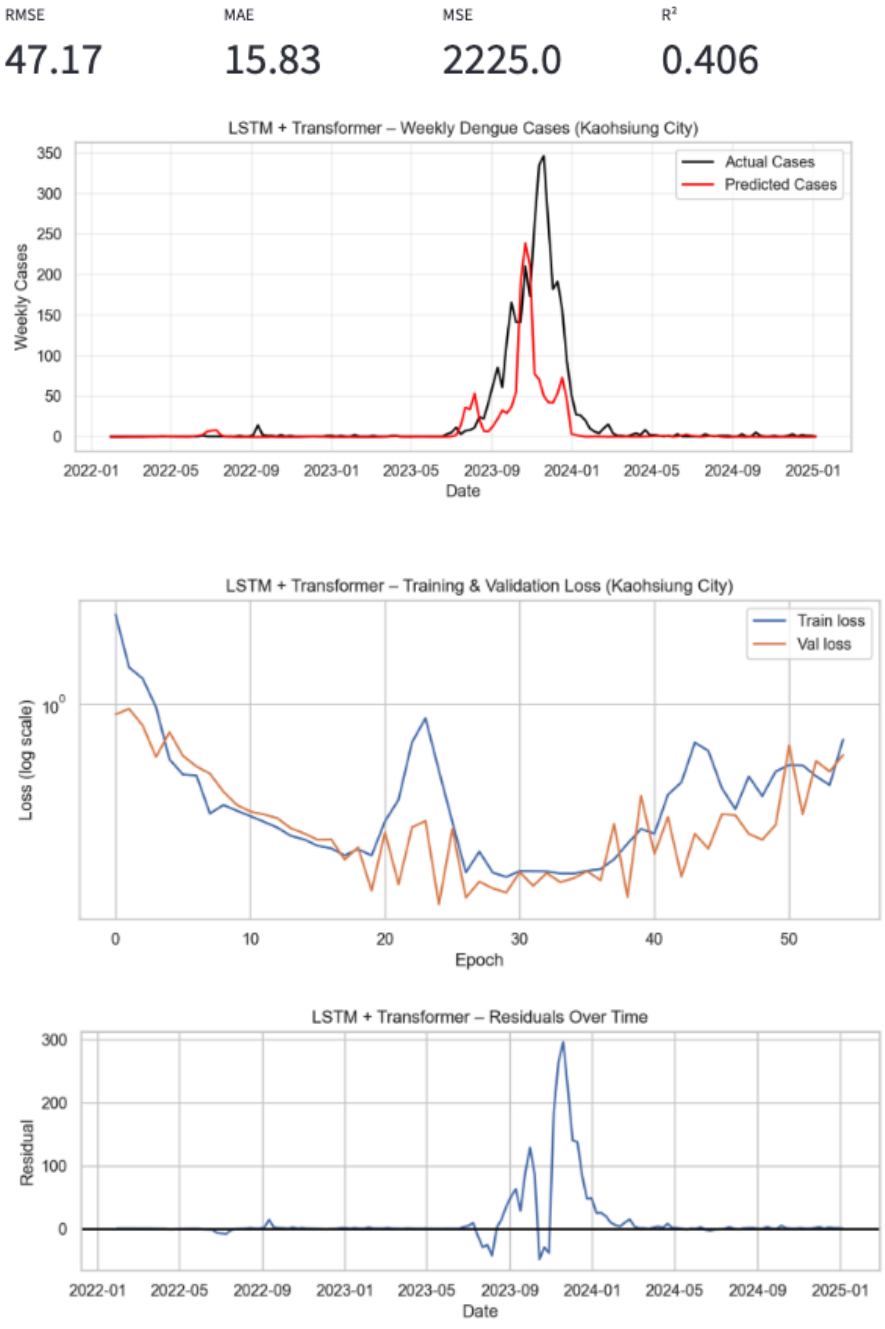




Figures 17-18. LSTM predictions, loss curves, and residuals.

LSTM-Transformer

LSTM-Transformer improved timing but continued to underestimate peak magnitude (RMSE = 47.17; MAE = 15.83; R^2 = 0.406). Training loss showed unstable convergence.



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

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 ≈ 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.

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 comprehensive dengue forecasting framework for southern Taiwan by integrating 15 years of epidemiological, climatic, and entomological data into a unified analytical pipeline. Focusing on Kaohsiung City, Tainan City, and Pingtung County (the regions with the highest national disease burden) the analysis revealed strong seasonal patterns and pronounced temporal dependence in dengue transmission. Autocorrelation findings confirmed that recent dengue activity, along with multi-week climatic and mosquito index patterns, plays a central role in shaping outbreak trajectories. These insights informed the construction of lag-based predictors and sequential models designed to capture the nonlinear, episodic structure of dengue outbreaks in the region.

Across all cities, the model comparison revealed a consistent and robust pattern: **tree-based machine-learning models substantially outperformed both LSTM and LSTM-Transformer architectures**. Random Forest and XGBoost accurately captured outbreak timing and magnitude, demonstrating low error even in low-incidence or highly skewed conditions. Their feature-importance outputs reinforced the dominant influence of recent case lags, supported by selected climatic variables. Conversely, deep-learning models struggled with sparse outbreak histories and frequently produced overly smoothed forecasts that failed to reflect sudden epidemic surges, particularly in Tainan, where outbreak peaks were exceptionally large. These limitations highlight the data-intensive nature of sequential neural models; LSTM-based architectures typically require richer spatiotemporal information, such as Google Trends activity, GPS-based mobility data, or behavioral indicators, to perform well, as shown in mobility-driven epidemic forecasting literature. In Taiwan's outbreak-sparse context, **traditional ML approaches currently offer greater stability, interpretability, and operational readiness**.

A major contribution of this project is the **end-to-end, cloud-enabled architecture** that separates model training from user-facing visualization. All analytical artifacts, including predictions, residuals, performance metrics, and training histories, were stored in a Supabase cloud database and served dynamically to a Streamlit application deployed on Render Cloud. This design promotes reproducibility, scalability, and real-time interaction, enabling public health practitioners, researchers, and students to explore model behavior and climatic or entomological drivers without requiring local computation. The accompanying ArcGIS dashboard further contextualizes spatial transmission patterns and illustrates how rainfall and mosquito indices align with dengue risk across Taiwan.

Despite its strengths, the forecasting framework has several limitations. Weekly temporal resolution masks rapid environmental fluctuations; mosquito indices are measured intermittently; and city-level aggregation limits fine-scale spatial insights. Deep-learning models were constrained by limited outbreak history and computational restrictions inherent to free-tier cloud hosting. These limitations, however, highlight pathways for future research. Incorporating additional datasets, such as mobility, satellite-derived environmental indicators, and high-resolution land-use layers, could strengthen forecasting performance. Advanced spatiotemporal architectures (e.g., Temporal Fusion Transformers, graph neural networks) may

better capture long-range dependencies. Building an automated real-time data pipeline and migrating to more scalable cloud infrastructure would further support operational deployment.

Overall, this study demonstrates that integrating multi-source climatic, entomological, and epidemiological data into a structured forecasting pipeline can produce **practical, interpretable, and region-specific early-warning models** for dengue in Taiwan. The consistent superiority of tree-based models underscores their suitability for near-term public health applications, while deep-learning approaches hold longer-term promise as richer datasets become available. This work establishes a foundation for next-generation early-warning systems and provides actionable insights for vector-control planning and outbreak preparedness across southern Taiwan.

8. Limitations and Future Work

Although the forecasting framework developed in this study provides practical and interpretable insights into dengue risk across southern Taiwan, several limitations must be acknowledged. First, the use of weekly temporal resolution smooths short-term fluctuations in climate and mosquito activity, reducing the model's ability to capture rapid environmental changes that may precede outbreak onset. Future work should incorporate higher-frequency (daily) meteorological and rainfall data to better reflect real-time transmission drivers. Second, mosquito indices such as BI, HI, and CI are collected intermittently and at inconsistent intervals across cities. Although forward- and backward-fill imputation maintains continuity, it cannot fully represent actual changes in vector abundance. Integrating additional entomological inputs, such as automated sensor networks, ovitrap counts, or remote-sensing proxies for breeding sites, would strengthen the accuracy of vector-related predictors. A third limitation is the city-level spatial scale of analysis. Aggregated data obscure neighborhood-level heterogeneity in mosquito habitats, microclimate, human mobility, and socioeconomic factors that influence transmission. Expanding to district- or village-level forecasting could enable more precise, targeted early-warning systems. Deep-learning models in particular were constrained by limited outbreak history, as southern Taiwan experiences relatively few large epidemics. This scarcity reduces a model's ability to learn sharp temporal patterns, contributing to underprediction during major surges. Prior research demonstrates that LSTM and Transformer-based models benefit from richer spatiotemporal features such as Google Trends activity, GPS-based mobility data, and satellite-derived environmental indicators. Incorporating these additional data sources may substantially improve sequential model performance. Finally, the use of free-tier cloud infrastructure limited computational resources, model experimentation depth, and application speed. Migrating to more scalable platforms and implementing automated real-time data ingestion pipelines would enhance operational viability and long-term sustainability.

Looking ahead, future work should focus on integrating higher-resolution temporal and spatial data, incorporating mobility and remote-sensing indicators, exploring advanced architectures such as Temporal Fusion Transformers and graph neural networks, and developing multi-step and probabilistic forecasting capabilities. By addressing these limitations, the forecasting system

can evolve into a fully operational early-warning tool that supports more effective, data-informed vector control and public health decision-making throughout southern Taiwan.

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.