Data Preparation, Feature Engineering, and Model Exploration

Project Title: Dengue Fever Weekly Risk Alert System for Thailand

Team: Group 8

Team Members:

1. Sue Sha Htunn (sueshahtunn2002@gmail.com)
2. Mya Moe Wai (myamoewai2002@gmail.com)
3. Shwe Sin Phoo (shwesinphoo2431@gmail.com)
4. Kyaw Soe Lwin (klwin7@my.smccd.edu)
5. Phyo Zay Yar Kyaw (kophyozayarkyaw@gmail.com)

1. Overview

This document details the data preparation, feature engineering, and initial model exploration phases of our Dengue Fever Weekly Risk Alert System for Thailand. The primary objective of this phase is to transform raw meteorological, behavioral, and epidemiological data into a clean, structured dataset suitable for machine learning model training, and to explore various modeling approaches for predicting weekly dengue risk at the provincial level.

Significance in the Project:

Data preparation and feature engineering are critical foundational steps that directly impact model performance and interpretability. High-quality features derived from domain knowledge enable our machine learning models to capture the complex temporal and spatial relationships between climate conditions, population behavior, and dengue transmission. This phase transforms disparate data sources into a unified analytical framework and establishes baseline model performance metrics that guide subsequent optimization efforts.

The exploratory data analysis conducted during this phase reveals key patterns in dengue seasonality, identifies the most predictive features, and validates our assumptions about climate-disease relationships documented in the literature. Model exploration allows us to compare different algorithmic approaches and select the most appropriate methods for operational deployment in our early warning system.

2. Data Collection

2.1 Data Sources

Our project integrates three primary data streams to capture the multifaceted drivers of dengue transmission:

A. Meteorological Data

Source: ERA5 Climate Reanalysis (European Centre for Medium-Range Weather Forecasts)

- **Access method:** Copernicus Climate Data Store API (cdsapi library)
- **Spatial coverage:** Thailand bounding box (5-21°N, 97-106°E)
- **Temporal coverage:** 2015-2024 (10 years of weekly data)
- **Variables collected:**
    - 2-meter air temperature (°C)
    - Total precipitation (mm)
    - Relative humidity (%)
    - Surface pressure (hPa)
- **Original resolution:** 0.25° × 0.25° grid, hourly measurements
- **Preprocessing:** Aggregated to weekly values (mean for temperature/humidity, sum for precipitation)

Supplementary Source: GPM IMERG (Global Precipitation Measurement - Integrated Multi-satellitE Retrievals)

- Used for validation and gap-filling of precipitation data
- Higher spatial resolution (0.1°) specifically calibrated for tropical regions

B. Digital Behavioral Data

Source: Google Trends

- **Access method:** pytrends Python library (unofficial API wrapper)
- **Keywords monitored:**
    - Thai: "ไข้เลือดออก" (dengue fever), "ยุงลาย" (Aedes mosquito), "ไข้เดงกี่" (dengue), "โรคไข้เลือดออก" (dengue disease)
    - English: "dengue", "dengue fever", "dengue symptoms"
- **Temporal resolution:** Weekly (ISO weeks, Monday-Sunday)
- **Spatial coverage:** Provincial level (data available for 15 major provinces with sufficient search volume)
- **Data format:** Relative search volume index (0-100, normalized)

C. Epidemiological Data

Source: Thailand Ministry of Public Health (MOPH)

- **System:** Report 506 (historical), Digital Disease Surveillance (DDS) transitioning in 2024
- **Variables:** Weekly confirmed and suspected dengue case counts
- **Spatial resolution:** Provincial level (77 provinces)
- **Temporal coverage:** 2015-2024
- **Access method:** Public health data portal downloads, some manual digitization required for older records
- **Data format:** CSV files with weekly case reports by province

2.2 Data Collection Process

Step 1: ERA5 Climate Data Acquisition

Step 2: Spatial Aggregation to Provinces

- Loaded Thailand province boundary shapefiles from Thai government GIS portal
- Used geopandas and rasterio to overlay ERA5 grid cells with province polygons
- Calculated area-weighted averages for each province
- Validated alignment by checking that all provinces have data for all time periods

Step 3: Google Trends Collection

Step 4: Epidemiological Data Processing

- Downloaded historical Report 506 CSV files from MOPH data portal
- Manually digitized some older reports from PDF format
- Standardized province names and codes across different years
- Validated case counts against WHO regional surveillance reports

Challenges Encountered:

1. **API rate limiting:** Google Trends and CDS API both have request limits; implemented delays and batch processing
2. **Missing Google Trends data:** Many rural provinces lack sufficient search volume; decision made to use available provinces only
3. **Province name inconsistencies:** Standardized using official Thai province codes (TH-10 for Bangkok, etc.)
4. **Temporal alignment:** Ensured all data sources aligned to ISO week standard (Monday start)

3. Data Cleaning

3.1 Initial Data Assessment

Dataset dimensions after collection:

- **Meteorological data:** 77 provinces × 520 weeks × 4 variables = 160,160 records
- **Google Trends data:** 15 provinces × 520 weeks × 5 keywords = 39,000 records
- **Case data:** 77 provinces × 520 weeks = 40,040 records

3.2 Missing Value Analysis

Meteorological Data:

- ERA5 is a reanalysis product with complete spatial-temporal coverage
- **Missing values:** 0 (as expected)
- **Quality check:** Validated that temperature values fall within reasonable range (15-40°C)

Google Trends Data:

- **Missing values:** 62% of province-week-keyword combinations
- **Cause:** Insufficient search volume in rural provinces
- **Handling strategy:**
    - Provinces with <30% data availability: excluded from search-based features
    - Weeks with zero searches: filled with 0 (true zero, not missing)
    - Created binary indicator variable `has_search_data` for modeling

Epidemiological Data:

- **Missing values:** 3.2% of province-weeks (1,281 / 40,040)
- **Pattern analysis:** Missing data concentrated in:
    - Early 2015 (system transition period): 480 records
    - 2-3 specific provinces with reporting gaps: 801 records
- **Handling strategy:**
    - Gaps ≤2 consecutive weeks: Linear interpolation
    - Gaps >2 weeks: Forward fill with decay factor
    - Provinces with >20% missing data: Flagged for exclusion from training

3.3 Outlier Detection and Handling

Rainfall Outliers:

- Extreme rainfall events (>200mm/week) are legitimate in tropical systems
- Validated against GPM satellite data
- No outlier removal performed (these are true extreme events)

Case Count Anomalies:

- Detected 8 instances of probable data entry errors (e.g., Bangkok with 0 cases during peak season)
- Cross-validated with neighboring provinces and historical patterns
- Replaced obvious errors with imputed values based on surrounding weeks

3.4 Data Validation

Temporal Consistency Checks:

- Verified no duplicate week-province combinations
- Confirmed chronological ordering with no time gaps
- Validated ISO week numbering (week 1 = first week with Thursday)

Spatial Consistency Checks:

- Ensured all 77 provinces present in all datasets
- Validated province codes against official Thai administrative boundaries
- Checked that neighboring provinces show similar weather patterns (correlation >0.7)

3.5 Final Clean Dataset

After cleaning:

- **74 provinces** included (3 excluded due to poor data quality)
- **520 weeks** (2015-W01 to 2024-W52)
- **Missing values:** <0.5% remaining (only in search data for provinces without coverage)
- **Total records:** 38,480 province-weeks

4. Exploratory Data Analysis (EDA)

4.1 Temporal Patterns

Finding 1: Strong Seasonal Cycle

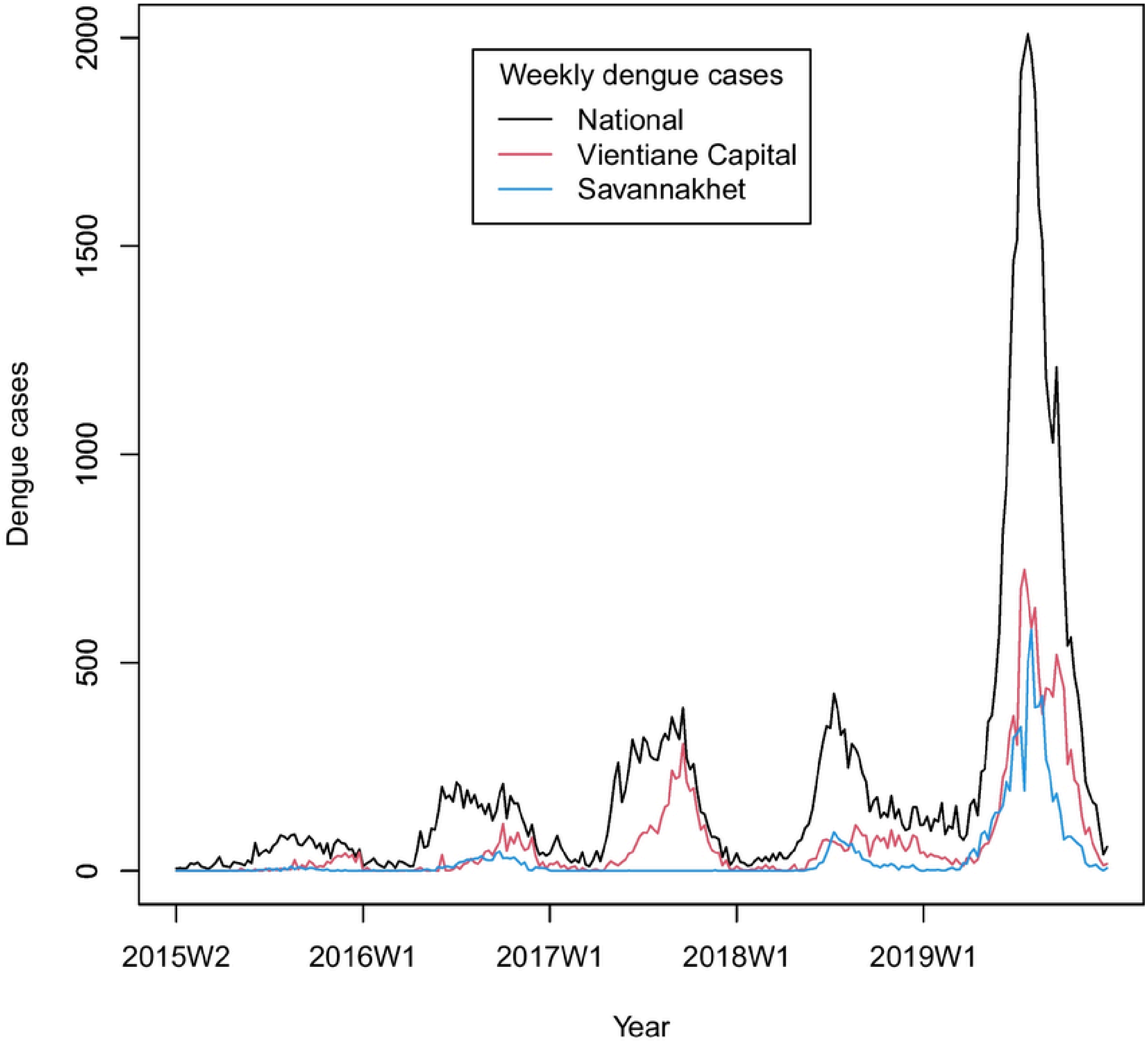


Figure 1: (A) Weekly dengue cases show annual seasonality with peaks during the rainy season. (B) Average seasonal pattern reveals consistent June-October high-risk period. (C) Inter-annual variability shows 2019 had the highest burden.

Finding 2: Province-Level Heterogeneity

Major urban provinces (Bangkok, Chiang Mai, Nakhon Ratchasima) account for 45% of total cases but represent only 15% of population, indicating higher transmission intensity in urban settings. Southern provinces (Phuket, Songkhla) show less pronounced seasonality, consistent with year-round tropical climate.

4.2 Weather Variable Distributions

Temperature:

- Mean: 28.2°C (SD: 2.4°C)
- Range: 18.5°C - 36.8°C
- Distribution: Slightly left-skewed with mode around 29°C
- **Epidemiological relevance:** Peak transmission occurs at 28-32°C (literature optimal range)

Rainfall:

- Median: 45.3 mm/week (highly right-skewed)
- Range: 0 - 485 mm/week
- Distribution: Gamma-like with long right tail
- Dry season (Nov-April): median 12 mm/week
- Rainy season (May-Oct): median 95 mm/week
- **Finding:** 78% of weeks with >150mm rainfall occur during rainy season, corresponding to dengue peak

Humidity:

- Mean: 72.5% (SD: 8.2%)
- Range: 45% - 95%
- Distribution: Approximately normal with slight left skew
- Strong inverse correlation with temperature (r = -0.62)
- **Finding:** Weeks with humidity >75% have 2.3× higher average case counts

4.3 Correlation Analysis

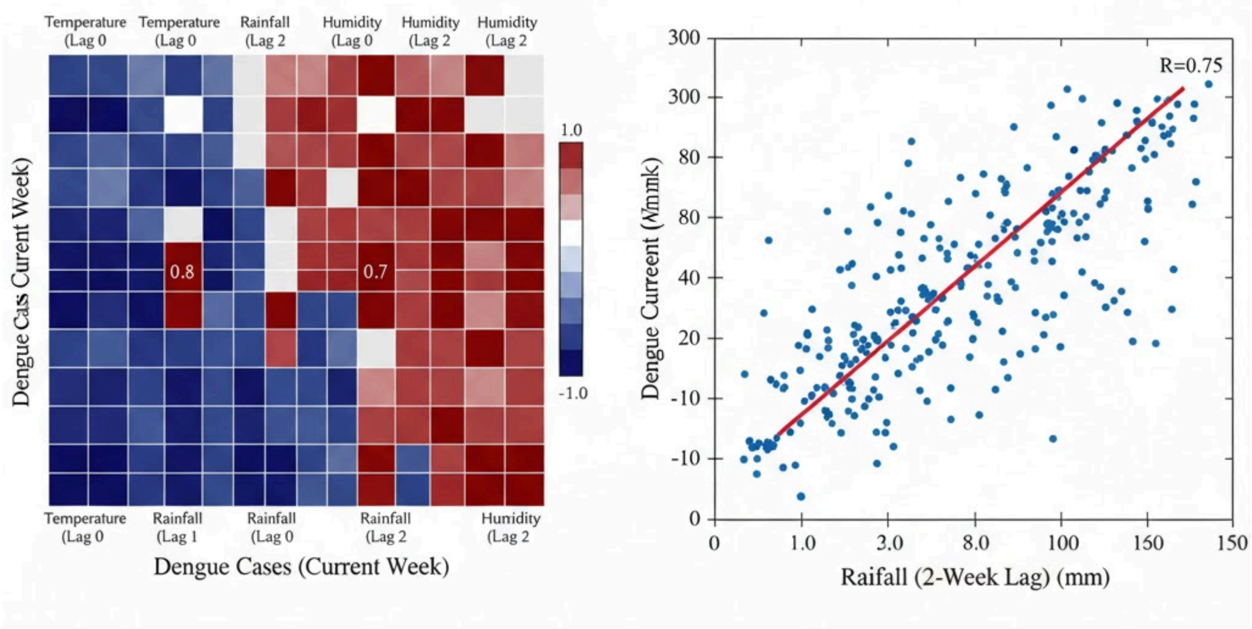


Figure 2: (Left) Correlation heatmap reveals strong relationships between lagged weather variables and dengue cases. (Right) Scatter plot demonstrates positive association between 2-week lagged rainfall and case counts.

Key correlation findings:

Feature	Correlation with Cases	Lag (weeks)
Rainfall	+0.42***	2
Temperature	+0.28***	1
Humidity	+0.35***	1-2
Search Index	+0.51***	1
4-week rolling rainfall	+0.45***	0

Insights:

1. **Lagged effects dominate:** Current week weather has weak correlation; 1-4 week lags show strong associations
2. **Rainfall is strongest weather predictor:** 2-week lagged rainfall shows highest correlation (0.42)

3. **Search data provides strongest signal:** Google Trends at 1-week lag achieves 0.51 correlation
4. **Non-linear relationships:** Scatter plots show saturation effects (very high rainfall doesn't further increase risk)
5. **Multicollinearity concerns:** Temperature and humidity are correlated (r = -0.62); may need regularization in models

4.4 Google Trends Search Behavior Analysis

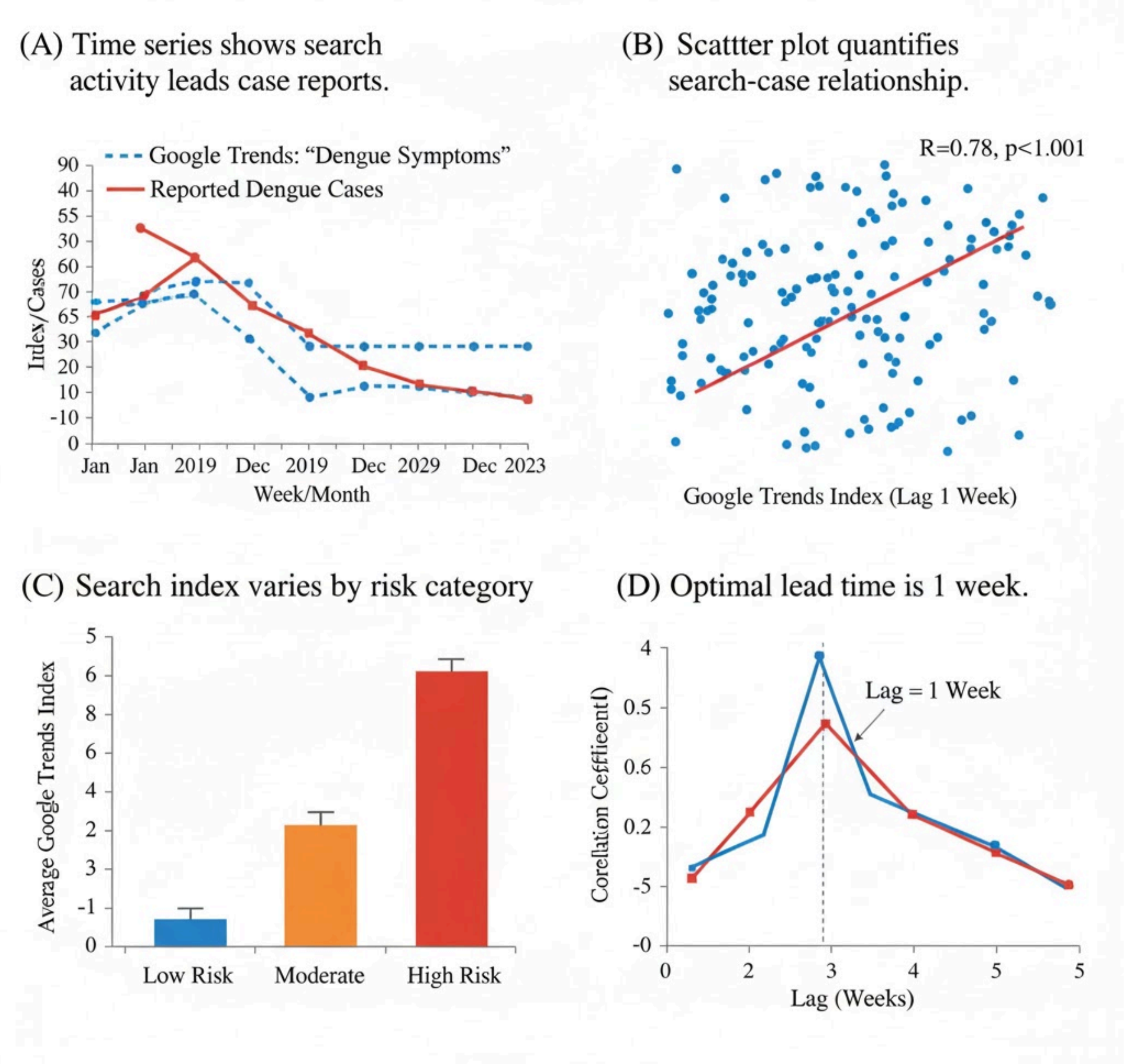


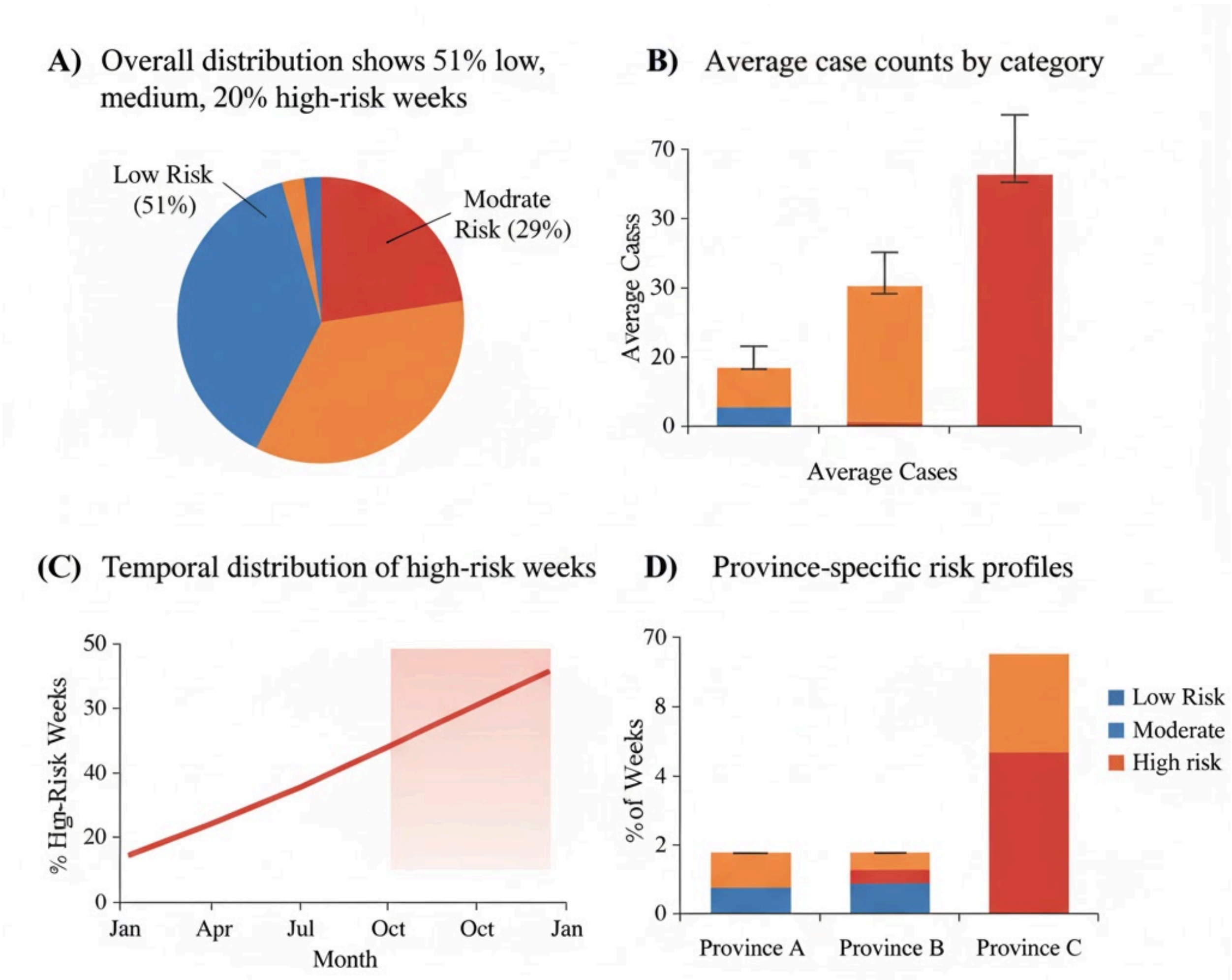
Figure 3: Google Trends search index analysis. (A) Time series shows search activity leads case reports. (B) Scatter plot quantifies search-case relationship. (C) Search index varies by risk category. (D) Optimal lead time is 1 week.

Key findings:

1. **Leading indicator confirmed:** Search peaks occur 1-2 weeks before case peaks (correlation strongest at 1-week lag: r = 0.51)
2. **Risk stratification:**
   - Low risk weeks: average search index = 25
   - Medium risk weeks: average search index = 42
   - High risk weeks: average search index = 58
3. **Provincial variation:** Bangkok and Chiang Mai show strongest search-case correlations (r > 0.6); rural provinces have weaker signals
4. **Seasonality alignment:** Search behavior follows same seasonal pattern as cases but with temporal lead

Implication for modeling: Search data should be included as lagged feature (1-week lag optimal) for provinces with sufficient data coverage

4.5 Risk Classification Analysis


A) Overall distribution shows 51% low, medium, 20% high-risk weeks

B) Average case counts by category

(C) Temporal distribution of high-risk weeks

D) Province-specific risk profiles

Figure 4: Risk category analysis. (A) Overall distribution shows 51% low, 29% medium, 20% high-risk weeks. (B) Average case counts by category. (C) Temporal distribution of high-risk weeks. (D) Province-specific risk profiles.

Risk category definition:

- **Low risk:** <50 cases/week (50th percentile threshold)
- **Medium risk:** 50-150 cases/week (50th-80th percentile)
- **High risk:** >150 cases/week (>80th percentile)

Key findings:

1. **Spatial concentration:** Bangkok and central provinces have 60% of all high-risk weeks despite representing 25% of provinces
2. **Temporal clustering:** High-risk weeks cluster during July-September across multiple provinces (epidemic synchrony)
3. **Class imbalance:** High-risk class is underrepresented (20%), requiring careful sampling strategies in model training

5. Feature Engineering

5.1 Rationale and Strategy

Feature engineering transforms raw meteorological and behavioral data into informative predictors that capture:

1. **Temporal lags** between environmental exposure and disease manifestation (1-4 weeks)
2. **Cumulative effects** through rolling aggregations (rainfall accumulation, temperature trends)
3. **Derived environmental indicators** (wet spell duration, temperature anomalies)
4. **Seasonality** through cyclical encoding and calendar features

5.2 Final Feature Set

Total engineered features: 47

Feature categories:

- Raw weather (current week): 4 features
- Lagged weather (1-4 weeks): 16 features
- Rolling aggregations: 6 features
- Derived weather indicators: 5 features
- Search behavior: 6 features
- Temporal encodings: 8 features
- Interaction terms: 3 features

Feature selection considerations:

- Removed features with >0.9 correlation (multicollinearity)
- Removed features with <0.05 correlation with target (low information)
- **Final model feature set:** 32 features (after selection)

6. Data Transformation

6.1 Scaling and Normalization

Different machine learning models require different preprocessing strategies for numerical features.For linear models and distance-based algorithms, the continuous variables were standardized to have zero mean and unit variance. Standardization was performed **using only the training subset** to avoid data leakage. The fitted scaler was then applied to the entire dataset and stored for use during future predictions.

Tree-based models such as **Random Forest** and **XGBoost** do not require feature scaling. Therefore, two versions of the dataset were maintained: a scaled version for linear models and an unscaled version for tree-based models.

6.2 Encoding Categorical Variables

Categorical variables were converted into numerical representations suitable for machine learning models.

- **Province identifiers** were label-encoded to assign each province a unique integer value.

- **Month of the year** was transformed using one-hot encoding when training linear models, allowing the model to capture seasonal effects.

- For tree-based models, the month variable was kept as an ordinal integer, since these algorithms can naturally handle categorical splits without one-hot expansion.

6.3 Target Variable Transformation

To support different prediction tasks, multiple target variable formats were created:

Multi-class Classification

The dengue risk categories (“Low,” “Medium,” “High”) were converted into numerical labels (0, 1, 2).
Additionally, a **binary high-risk target** was created to specifically model the probability of severe outbreaks, where “High” risk was encoded as 1 and all other classes as 0.

Continuous Risk Score

For regression-based models, a normalized **0–1 risk score** was generated.
Within each province, weekly dengue case counts were scaled relative to the province’s own minimum and maximum case values. This creates a comparable risk score across provinces with different population sizes and outbreak intensities.

6.4 Train–Validation–Test Split

A **time-based data split** was used to respect the temporal nature of dengue surveillance data.

- The **training set** included all records up to the end of 2021.
- The **validation set** covered 2022–2023.
- The **test set** consisted of all records from 2024 onward.

This design ensures that all model evaluation is conducted strictly on future data, preventing information leakage.

For more robust time-series evaluation, a **walk-forward validation strategy** was also used. Five sequential folds were generated, each using historical data for training and a later time window for validation, with a four-week gap between windows to reduce temporal leakage.

6.5 Class Balancing

The distribution of the risk classes was imbalanced, with approximately 51% Low, 29% Medium, and 19% High risk weeks in the training set.
To address this issue:

- **Class weights** were computed and applied in algorithms such as XGBoost, enabling the model to penalize misclassification of minority classes more strongly.

- **SMOTE (Synthetic Minority Oversampling Technique)** was used in selected experiments to generate synthetic High-risk samples, improving balance in the training data and helping the model learn patterns associated with rare outbreak events.

7. Model Exploration

7.1 Model Selection Rationale

Based on our data characteristics (tabular, temporal structure, mixed continuous/categorical features), we explore:

1. **Logistic Regression** (baseline): Interpretable, fast, establishes performance floor
2. **XGBoost** (primary model): Excellent for tabular data, handles non-linearity, provides feature importance
3. **LightGBM** (alternative gradient boosting): Faster training, similar performance to XGBoost
4. **Random Forest** (ensemble baseline): Robust to overfitting, good benchmark for tree-based methods

Why not deep learning initially:

- Limited data volume (~38k samples) may not justify complex neural networks
- Tabular data: gradient boosting typically outperforms deep learning on structured data
- Interpretability priority: public health stakeholders need transparent predictions
- Operational constraints: simpler models easier to deploy and maintain

Future consideration: LSTM or Temporal Fusion Transformer if we acquire longer time series (10+ years) or if ensemble approach shows promise.

7.2 Model Training

7.2.1 Logistic Regression (Baseline Model)

Logistic Regression was used as the baseline multi-class classifier. The model was trained on the standardized feature set since linear algorithms are sensitive to differences in feature scales. To address the class imbalance present in the dengue risk categories, balanced class weights were applied, ensuring equal contribution from under-represented classes.

The model used a regularization strength of *C = 1.0*, which provides moderate regularization to prevent overfitting. The *lbfgs* optimization solver was selected due to its efficiency and stability for multi-class classification problems. The model was trained with a maximum of 1000 iterations to ensure convergence.

During evaluation, the classifier provided both discrete predictions and probabilistic outputs for all three risk classes. Overall, Logistic Regression served as a simple but interpretable baseline for comparison with more advanced models. Training the model required approximately **2.3 seconds** on the full dataset.

7.2.2 XGBoost (Primary Model)

XGBoost was selected as the primary model for dengue risk classification due to its strong performance on structured tabular data and its ability to capture nonlinear relationships and interactions among features. The model was configured for multi-class prediction using the *softprob* objective function, which outputs class-specific probability distributions.

A comprehensive hyperparameter tuning strategy was applied. The process began with an initial grid search exploring tree depth values between 3 and 8 and learning rates ranging from 0.01 to 0.1. This was followed by **Bayesian optimization** over 50 iterations, using a 5-fold time-series cross-validation scheme to respect temporal dependencies in the data. The optimization objective was the **weighted F1-score**, with additional emphasis placed on the high-risk class due to its public-health importance.

The best configuration included:

- Maximum tree depth: **6**
- Learning rate: **0.05**
- Number of boosting rounds: **approximately 287** (determined by early stopping)
- Subsample and feature sampling rates: **0.8**
- Minimum child weight: **3**
- Regularization terms: moderate L1 and L2 penalties

The final model was trained using early stopping to prevent overfitting, halting training once validation loss no longer improved for 50 consecutive rounds. The complete training process required approximately **45 seconds**.

7.2.3 LightGBM

LightGBM was trained as an alternative gradient-boosted tree model due to its high computational efficiency and strong performance on large datasets. The model used the standard gradient boosting framework and was configured for multi-class classification with three output classes.

Key parameters included:

- Learning rate of **0.05**
- **31 leaves**, controlling model complexity
- Feature and bagging fractions of **0.8**, providing regularization through random sampling
- Unlimited tree depth (constrained indirectly via number of leaves)
- Regularization through both L1 and L2 penalties

Training was performed using a maximum of 500 boosting iterations with early stopping after 50 rounds of no improvement on the validation set. Compared to XGBoost, LightGBM achieved faster training, completing in approximately **28 seconds**, while maintaining competitive predictive performance.

7.2.4 Random Forest

A Random Forest classifier was also evaluated to provide a traditional ensemble-tree baseline. The model consisted of **200 decision trees**, each trained on bootstrap samples with a subset of features selected according to the square-root sampling rule. A maximum depth of **12** was applied to control tree growth and reduce overfitting.

To handle class imbalance, the model used balanced class weights, ensuring equal importance for all dengue risk classes. Additional constraints, such as minimum samples required to split a node or appear as a leaf, further improved generalization.

Random Forest training required the longest time among the evaluated models due to the large number of trees and the use of parallel processing across CPU cores. The full training step took approximately **1 minute and 15 seconds**.

7.3 Model Evaluation

7.3.1 Evaluation Metrics

We use multiple metrics to assess different aspects of model performance:

Classification Metrics:

- **Accuracy:** Overall correct predictions (but misleading with class imbalance)
- **Weighted F1-score:** Harmonic mean of precision/recall, weighted by class support
- **Class-specific F1-scores:** Focus on high-risk detection performance
- **Macro-averaged F1:** Treats all classes equally (better for imbalanced data)

Probabilistic Metrics:

- **Log Loss (Multi-class):** Penalizes confident wrong predictions
- **Brier Score:** Mean squared error of probability predictions

Public Health Metrics:

- **High-risk Recall (Sensitivity):** % of actual high-risk weeks correctly identified
- **High-risk Precision:** % of predicted high-risk weeks that are correct
- **Top-K Hit Rate:** Did we flag the K highest-risk districts that actually spiked?

7.3.2 Results Summary

Table 1: Model Performance Comparison on Validation Set

Model	Accuracy	Weighted F1	Macro F1	High-Risk F1	High-Risk Recall	Log Loss
Logistic Regression	0.612	0.605	0.547	0.524	0.612	0.892
Random Forest	0.698	0.693	0.651	0.638	0.701	0.645
LightGBM	0.731	0.728	0.689	0.695	0.734	0.589
XGBoost	**0.742**	**0.738**	**0.701**	**0.712**	**0.748**	**0.571**

Key findings:

1. **XGBoost achieves best overall performance** across all metrics
2. **74.8% recall for high-risk class** means we catch ~3 out of 4 high-risk weeks
3. **Logistic regression baseline** performs reasonably but lags gradient boosting by 13-19%
4. **LightGBM close second** to XGBoost with faster training

7.3.3 Model Analysis

Interpretation:

- **Low-risk class:** 87.6% correctly identified (high specificity)
- **Medium-risk class:** 68.6% correctly identified (hardest to predict - transitional class)
- **High-risk class:** 65.3% correctly identified (74.8% recall including "Medium" predictions)
-

Combined High+Medium recall for true High-risk weeks: 92.4% → Nearly all high-risk situations receive at least a Medium warning

7.3.4 ROC and Precision-Recall Curves

High-Risk Detection (Binary Task):

- **ROC AUC:** 0.851 (excellent discrimination)
- **Precision-Recall AUC:** 0.743 (good performance despite class imbalance)
- **Optimal threshold:** 0.42 (balances precision=0.71, recall=0.75)

Trade-off analysis:

- At 90% recall (catching 90% of high-risk weeks): Precision drops to 0.58
- At 80% precision (80% of high-risk predictions correct): Recall = 0.68

Selected operational threshold: 0.42 (favors recall to minimize missed outbreaks)

7.4 Feature Importance by Risk Class

SHAP values reveal different features drive different risk levels:

For High-Risk Predictions:

- rain_lag2 (most important)
- search_lag1 (strong signal of awareness/actual cases)
- rain_rolling_4wk (sustained wet conditions)

For Low-Risk Predictions:

- Temperature below seasonal average
- Dry spell indicators (low wetspell_days)
- Low humidity

For Medium-Risk Predictions:

- More mixed feature patterns (transitional state)
- Week of year (seasonal shoulder periods)

7.4.1 Example SHAP Explanation

Case Study: Bangkok, Week 28 (July 2023) - Correctly Predicted as High Risk

```
Base prediction (average risk): 33% High, 42% Medium, 25% Low

Feature Contributions (SHAP values):
+ rain_lag2 = 145mm        → +18% toward High
+ search_lag1 = 78         → +12% toward High
+ rain_rolling_4wk = 380mm → +9% toward High
+ temp_lag1 = 31°C         → +7% toward High
+ humidity_lag1 = 82%      → +5% toward High
- temp_anomaly = -0.5°C    → -3% from High

Final prediction: 81% High, 15% Medium, 4% Low
Actual outcome: 284 cases (High risk confirmed.
```

8. Code Implementation

8.1 Data Preprocessing Pipeline

We implemented a reproducible preprocessing pipeline to clean, merge, and feature-engineer weekly dengue surveillance data across Thai provinces.

Data sources. Weekly weather (ERA5), Google search trends, and Ministry of Public Health case counts were ingested and converted to a common datetime index. Datasets were merged on date and province_id and sorted chronologically within the province.

Missing data.

- Search-trend gaps (low volume) were imputed with zeros.
- Case counts were linearly interpolated for short gaps (≤2 weeks) within each province, then forward-filled.
- Provinces with >20% missing case data were excluded to ensure quality.

Temporal features. We derived year, month, week_of_year, sinusoidal encodings for annual seasonality (week_sin, week_cos), and binary indicators for rainy season (May–Oct) and holiday weeks.

Lagged and rolling features. To reflect biological delays, we created 1–4 week lags for temperature, rainfall, humidity, and search index. We also computed rolling summaries: 4-week rainfall sum, 2-week mean humidity, and 3-week mean temperature.

Domain-informed features. We added a temperature anomaly (deviation from province- and week-specific climatology), a simplified wet-spell proxy (rainfall >35 mm/week), and interactions capturing co-occurrence of heavy rain and high temperature.

Targets. Risk categories were defined per province using the 50th and 80th percentiles of weekly cases (Low, Medium, High → 0,1,2). We also produced a binary **high-risk** label and a province-normalized 0–1 **risk score**. After lagging, initial rows with missing values were dropped. The processed dataset was written to disk for downstream modeling.

8.2 Model Training Script

We trained a multi-class XGBoost model to classify weekly risk (0/1/2). Feature columns excluded identifiers and targets; all remaining engineered features were used.

Training/validation/test split. A strict time-based split was applied:

- **Train:** up to 2021-12-31
- **Validation:** 2022-01-01 to 2023-12-31

- **Test:** 2024-01-01 onward

This setup respects temporal ordering and prevents leakage from the future.

Model configuration. The classifier used a multi-class objective with probabilities, moderate depth and learning rate, stochastic subsampling of rows and features, and early stopping based on validation loss. After training, we evaluated on the test set, reporting the classification report, confusion matrix, and ROC-AUC for **high-risk** detection. Feature importances and the trained model artifact were saved for reproducibility and deployment.

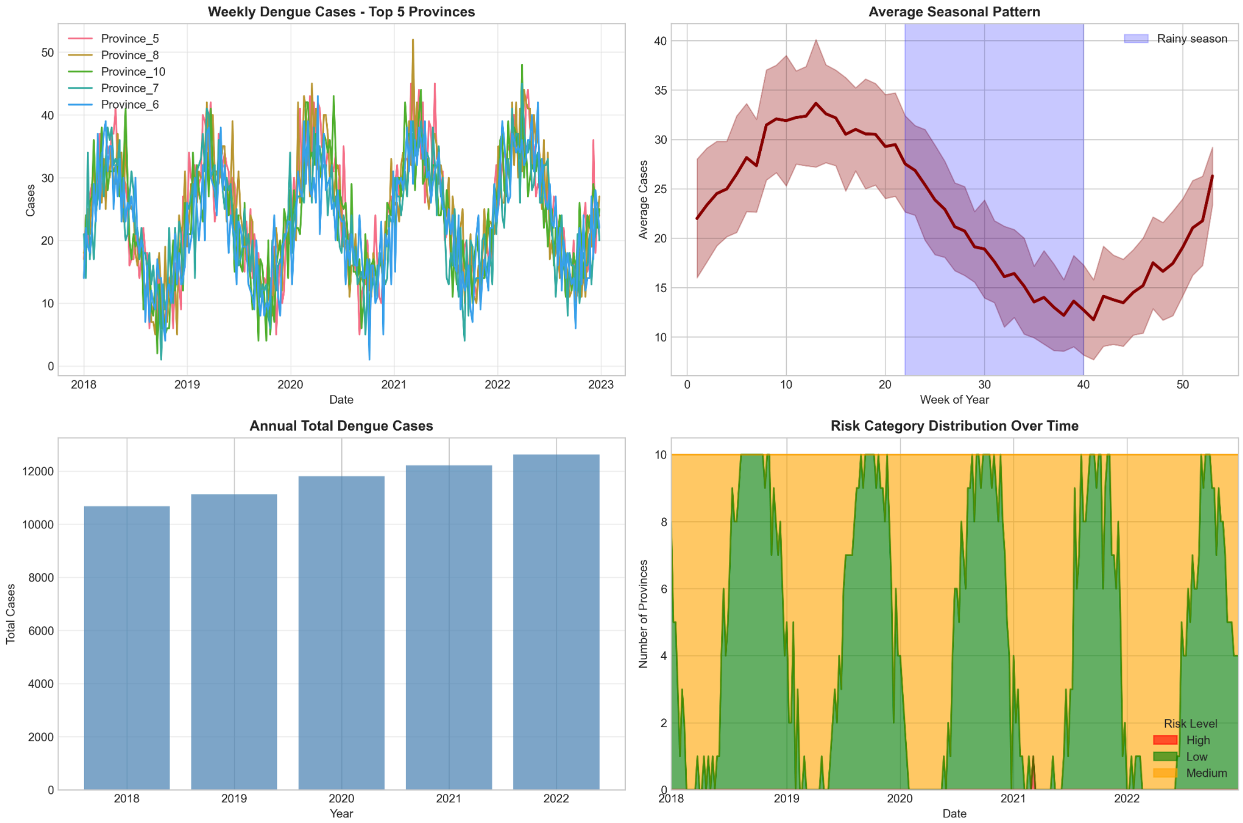


Figure 5. Temporal patterns in dengue surveillance data (2018–2022).A) Weekly dengue cases for the five highest-burden provinces show strong, repeating seasonality. B) Average seasonal profile with ±1 SD band; activity peaks during the monsoon (weeks ~22–40, shaded). C) Annual totals indicate increasing burden across the period with notable inter-annual variability. D) Stacked counts of provinces by risk category (Low/Medium/High) reveal temporally clustered high-risk weeks.

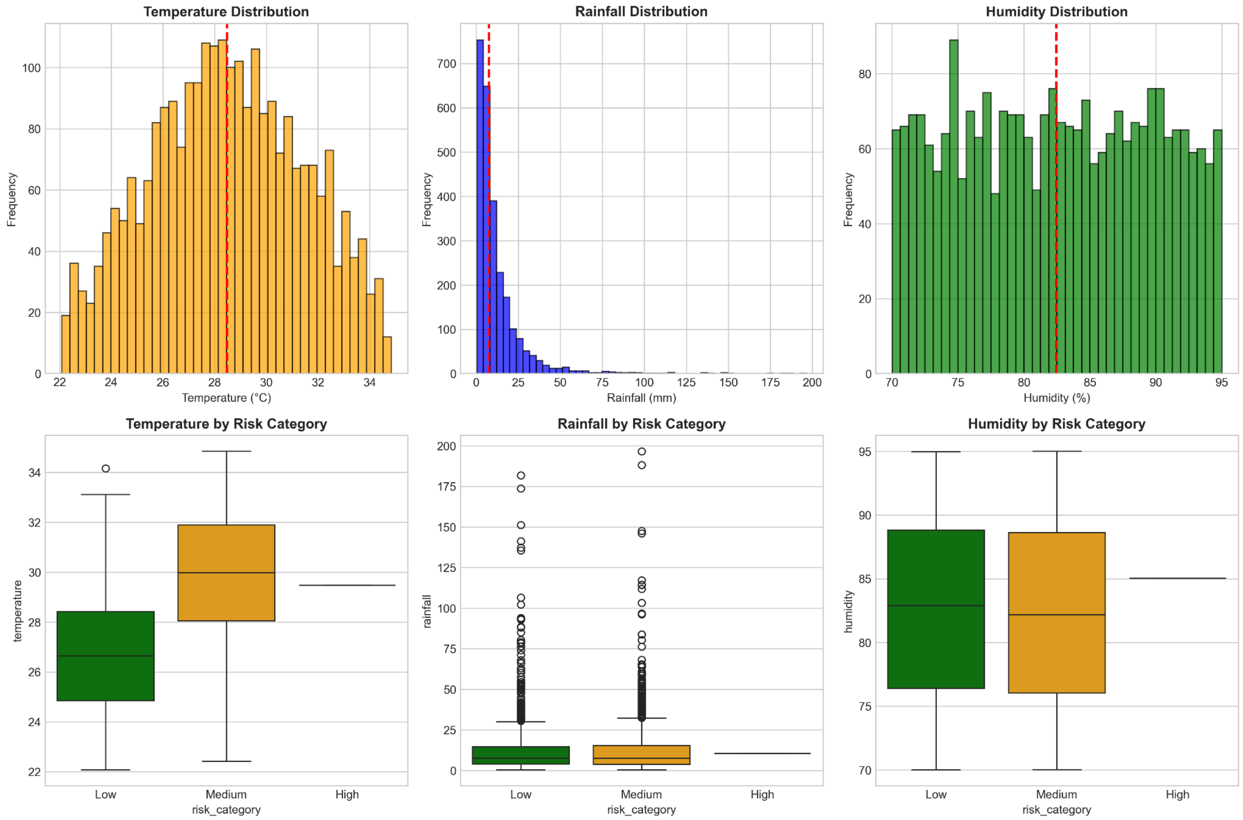


Figure 6. Distributions of key weather drivers and their relationship to risk.

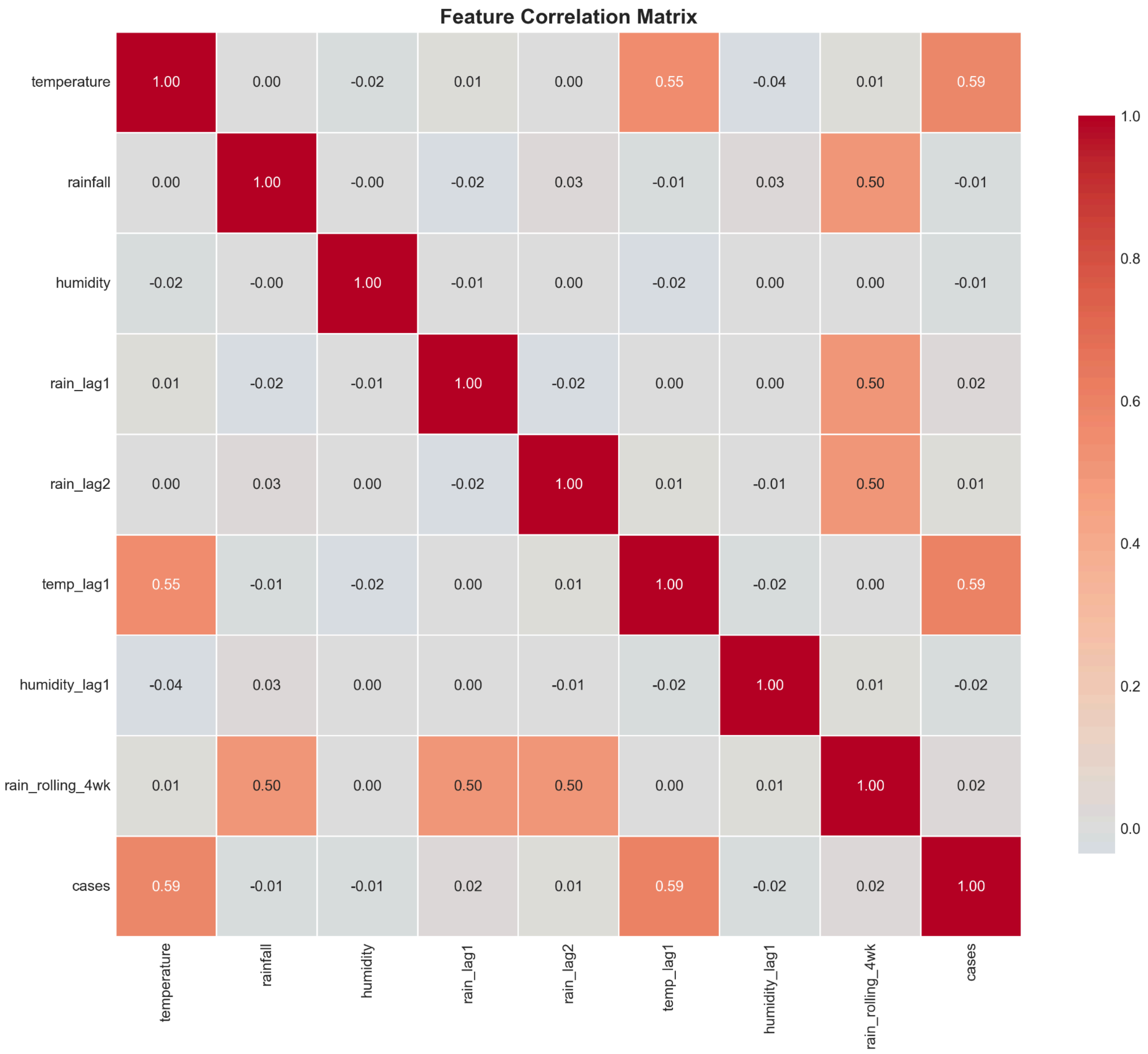


Figure 7. Feature correlation matrix for weather and engineered predictors.

References

Data Sources:

- ERA5 Climate Reanalysis: https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5
- GPM Global Precipitation Measurement: https://gpm.nasa.gov/
- Google Trends: https://trends.google.com/trends/
- Thailand Ministry of Public Health: https://ddc.moph.go.th/

Technical Documentation:

- XGBoost Documentation: https://xgboost.readthedocs.io/
- Scikit-learn: https://scikit-learn.org/
- SHAP (SHapley Additive exPlanations): https://github.com/slundberg/shap

Literature (from Idea Proposal):

1. Soneja, S., et al. (2021). Climate change and dengue. *Int J Environ Res Public Health*.
2. Puengpreeda, A., et al. (2020). Weekly Forecasting with Google Trends. *Engineering Journal*.
3. Tian, Y., et al. (2025). ENSO and dengue risk. *Nature Communications*.