Dengue Fever Weekly Risk Alert System for Thailand

Team Members - Group 8

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. Project Overview

Thailand faces persistent challenges with dengue fever, a mosquito-borne disease that affects thousands annually and places a significant burden on healthcare systems. Despite existing surveillance mechanisms, current response strategies often lag behind disease spread, resulting in reactive rather than proactive interventions.

Our project develops a Dengue Fever Weekly Risk Alert System that leverages machine learning to predict high-risk districts and provinces across Thailand one week in advance. By integrating multi-source data including meteorological conditions (rainfall, temperature, humidity), behavioral signals (Google search trends for dengue-related keywords in Thai and English), and historical case data, our system generates actionable weekly risk scores and classifications for each administrative area.

The system directly supports **SDG 3 (Good Health and Well-Being)** by enabling early intervention strategies such as larval control, community awareness campaigns, and school sanitation drives before dengue cases spike. It aligns with **SDG 11 (Sustainable Cities and Communities)** by providing city planners and local health departments with data-driven insights for vector control resource allocation. Additionally, it addresses **SDG 13 (Climate Action)** by explicitly linking climate patterns to disease transmission, supporting climate-health adaptation planning.

The potential impact is substantial: early warnings can reduce case numbers, prevent hospital overcrowding, save lives, and significantly decrease healthcare costs. By empowering local health teams with clear, interpretable predictions and explanations, our tool transforms dengue surveillance from reactive reporting to proactive prevention.

2. Objectives

Our project aims to achieve the following specific objectives:

- **Develop accurate predictive models** that classify dengue risk (Low/Medium/High) for each district or province in Thailand on a weekly basis, achieving a minimum AUROC of 0.75 and F1-score of 0.70 for high-risk prediction.

- **Integrate diverse data sources** including satellite-derived weather data (ERA5, GPM, TRMM), Google Trends search behavior, and historical dengue case records from Thai Ministry of Public Health into a unified analytical framework.

- **Create an intuitive dashboard** featuring interactive maps, top-K high-risk district rankings, and plain-language explanations of risk drivers to support decision-making by public health officials with varying technical backgrounds.

- **Implement model interpretability** using SHAP (SHapley Additive exPlanations) to provide transparent, explainable insights into why specific areas are flagged as high-risk each week.

- **Enable early intervention** by providing actionable alerts at least one week before anticipated dengue spikes, allowing time for preventive measures such as vector control operations, public awareness campaigns, and healthcare resource preparation.

- **Establish a reproducible pipeline** that can be updated weekly with new data, ensuring the system remains operationally viable and can be maintained by public health agencies long-term.

3. Background

Dengue fever is a significant public health challenge in Thailand, with all 77 provinces reporting cases annually. The disease exhibits strong seasonal patterns influenced by climatic factors, with transmission rates peaking during and after the rainy season when mosquito breeding conditions are optimal. Traditional surveillance relies on weekly case reporting (Report 506) from healthcare facilities to the Ministry of Public Health, but this represents a lagging indicator cases are counted only after patients seek treatment, by which time local transmission is already established.

Existing dengue control efforts include larval surveys, fogging operations, and public education campaigns. However, these interventions are often deployed reactively after case numbers rise, limiting their effectiveness. While Thailand has modernized its disease surveillance infrastructure with the Digital Disease Surveillance (DDS) system launched in 2024, predictive capabilities remain underdeveloped at the local level.

Recent research demonstrates that dengue transmission is highly predictable using machine learning approaches that combine meteorological variables with novel data sources. Studies have shown that temperature, rainfall, and humidity create windows of opportunity for mosquito population growth and viral replication. Furthermore, internet search behavior, specifically Google Trends queries for dengue-related terms provides a real-time signal of community awareness and concern that often precedes official case reporting by 1-2 weeks.

A machine learning approach is particularly beneficial for this problem because: (1) the relationships between weather variables, search behavior, and dengue risk are complex and non-linear, making them difficult to capture with simple statistical models; (2) ML algorithms can learn spatiotemporal patterns and province-specific risk profiles from historical data; (3) ensemble methods like XGBoost can handle mixed data types and missing values common in public health datasets; and (4) modern interpretability tools like SHAP enable transparent decision-making, crucial for public health applications where trust and understanding are essential.

Our system builds upon successful implementations in other Southeast Asian contexts while being specifically tailored to Thailand's administrative structure, data availability, and public health infrastructure. By providing weekly, district-level predictions with clear explanations, we fill a critical gap between national surveillance systems and local operational needs.

4. Methodology

Our methodology employs a supervised machine learning classification approach optimized for tabular time-series data. We prioritize tree-based ensemble methods over deep learning due to their superior performance on structured data, faster training times, lower data requirements, and inherent interpretability.

Core Machine Learning Pipeline:

- **Baseline Model:** Logistic Regression for simple, interpretable benchmarking
- **Primary Models:** XGBoost and LightGBM (gradient boosting frameworks known for handling tabular data excellently)
- **Backup Model:** Random Forest for ensemble diversity

Target Variable Construction:

We create a binary or multi-class risk label for each district-week observation. High-risk weeks are defined as those where case counts exceed the 80th percentile of historical distribution for that district, or where cases surpass a dynamic threshold based on rolling averages. This approach accounts for regional baseline differences in dengue incidence.

Feature Engineering:

- **Temporal lags:** Weather variables at t-1, t-2, t-3, t-4 weeks prior (reflecting mosquito lifecycle)
- **Rolling statistics:** 2-week and 4-week moving averages/sums of rainfall and temperature
- **Wet-spell indicators:** Count of consecutive days with rainfall above threshold
- **Seasonality features:** Week-of-year, month indicators, Thai holiday flags
- **Search trends:** Google Trends relative search volume for Thai keywords ("ไข้เลือดออก", "ยุงลาย") and English terms ("dengue", "dengue fever")
- **Climate indices:** ENSO indicators (ONI/MEI) if correlation analysis supports inclusion

Training and Validation Strategy:

We implement walk-forward cross-validation that respects temporal ordering to prevent data leakage. The dataset is split chronologically, with models trained on historical data and validated on subsequent time periods. Province-aware splitting ensures that model evaluation reflects real-world deployment where predictions are made for all provinces simultaneously.

Model Interpretability:

SHAP (SHapley Additive exPlanations) values are computed for each prediction to identify the top 3 contributing features. These are translated into plain language explanations such as "High risk due to: heavy rainfall past 2 weeks, elevated humidity, increased dengue searches."

Evaluation Metrics:

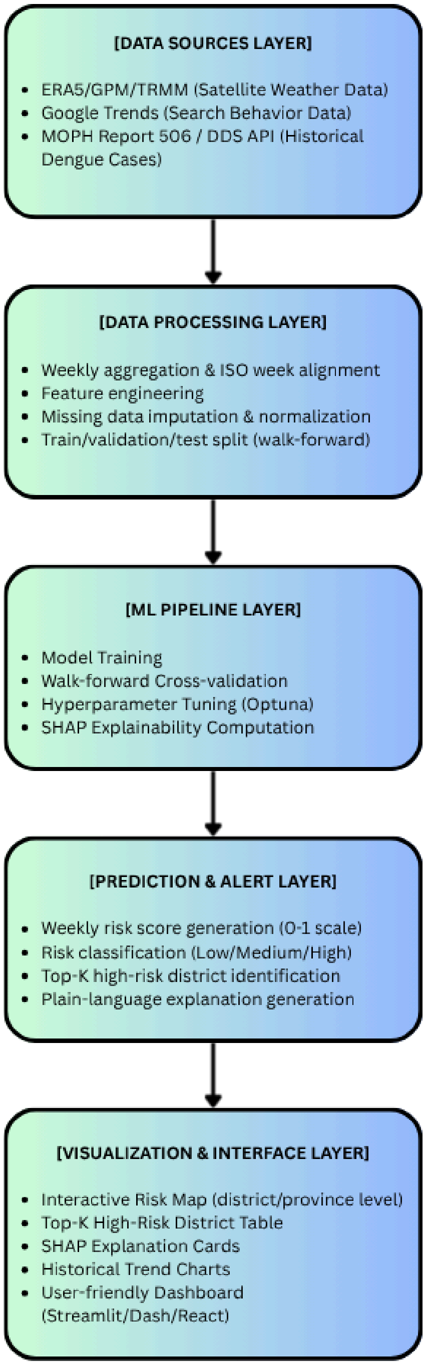
- **AUROC and PR-AUC:** Overall discrimination ability
- **F1-score:** Balance between precision and recall for high-risk class
- **Brier score:** Calibration of probability estimates
- **Top-K hit rate:** Percentage of actual outbreak districts captured in top-K predictions

Stretch Goals (if time permits):

- Implement LSTM or Temporal Fusion Transformer for enhanced sequence modeling
- Incorporate satellite-derived vegetation indices (NDVI) and water body detection
- Develop SMS/LINE notification system for automated alerts to local health departments

5. Architecture Design Diagram

System Architecture Overview



[DATA SOURCES LAYER]

- ERA5/GPM/TRMM (Satellite Weather Data)
- Google Trends (Search Behavior Data)
- MOPH Report 506 / DDS API (Historical Dengue Cases)

[DATA PROCESSING LAYER]

- Weekly aggregation & ISO week alignment
- Feature engineering
- Missing data imputation & normalization
- Train/validation/test split (walk-forward)

[ML PIPELINE LAYER]

- Model Training
- Walk-forward Cross-validation
- Hyperparameter Tuning (Optuna)
- SHAP Explainability Computation

[PREDICTION & ALERT LAYER]

- Weekly risk score generation (0-1 scale)
- Risk classification (Low/Medium/High)
- Top-K high-risk district identification
- Plain-language explanation generation

[VISUALIZATION & INTERFACE LAYER]

- Interactive Risk Map (district/province level)
- Top-K High-Risk District Table
- SHAP Explanation Cards
- Historical Trend Charts
- User-friendly Dashboard (Streamlit/Dash/React)

Component Descriptions:

- **Data Sources Layer:** Ingests raw data from satellite weather services, Google Trends API, and Thai public health databases on a weekly schedule.
- **Data Processing Layer:** Performs ETL operations, creates engineered features, handles missing values, and structures data for ML consumption.
- **ML Pipeline Layer:** Trains ensemble models using walk-forward validation, generates predictions, and computes SHAP explanations for interpretability.
- **Prediction & Alert Layer:** Converts model outputs into actionable alerts, ranks districts by risk, and generates human-readable explanations.
- **Visualization Layer:** Presents insights through an intuitive web dashboard accessible to public health officials, featuring maps, tables, and contextual information.

6. Data Sources

Our system integrates three primary categories of open-access data sources, each contributing unique dimensions to dengue risk assessment. Weather data from ERA5 reanalysis (European Centre for Medium-Range Weather Forecasts), GPM (Global Precipitation Measurement), and TRMM (Tropical Rainfall Measuring Mission) satellites provides daily measurements of rainfall accumulation, temperature, and relative humidity at spatial resolutions suitable for district-level analysis. These are aggregated to weekly temporal resolution to match surveillance reporting cycles. Behavioral signals are captured through Google Trends, which offers relative search volume indices for dengue-related keywords in both Thai ("ไข้เลือดออก" for dengue fever, "ยุงลาย" for Aedes mosquito) and English terms, reflecting public awareness and information-seeking behavior that often precedes official case detection. Historical dengue case data originates from Thailand's Ministry of Public Health Report 506 system and, where available, the newer Digital Disease Surveillance (DDS) API, providing weekly provincial and district-level confirmed case counts spanning 3-10 years. Data is structured in CSV or Parquet format with columns for date_week, administrative identifiers, meteorological variables, search indices, and case counts, totaling approximately 12,000-40,000 rows (77 provinces × 52 weeks × 3-10 years). Preprocessing involves ISO week standardization, interpolation of minor gaps, creation of lagged and rolling features, and normalization to prepare for machine learning model ingestion.

7. Literature Review

Extensive research establishes the predictability of dengue transmission through meteorological and behavioral indicators. Climate variables particularly temperature, rainfall, and humidity consistently emerge as primary drivers of mosquito population dynamics and viral replication rates in Southeast Asian contexts, with warm temperatures and moderate rainfall creating optimal conditions while extreme precipitation can temporarily suppress transmission by flushing larvae. Large-scale climate signals such as ENSO (El Niño-Southern Oscillation) modulate dengue risk regionally, with El Niño-related warming associated with increased epidemic probability and spatiotemporal synchronization across multiple countries, providing useful seasonal forecasting horizons. Internet search behavior has proven valuable as a real-time surveillance supplement, with studies demonstrating that Google Trends queries for dengue-related terms correlate strongly with incidence and improve short-term forecast accuracy, particularly in urban areas with high internet penetration. Thailand-specific research has successfully combined Google Trends data with meteorological variables to predict weekly dengue cases at the provincial level, demonstrating feasibility of the multi-source data integration approach. Additionally, analyses of Thailand's decades-long provincial case series reveal significant spatiotemporal synchronization patterns and traveling waves of dengue transmission, supporting the use of province/district-granular models with temporal awareness rather than purely spatial or temporal approaches alone. Our project builds directly on this evidence base by implementing ensemble machine learning methods that can capture complex non-linear relationships between these diverse predictors while providing the interpretability necessary for operational public health decision-making.

IMPLEMENTATION PLAN

1. Technology Stack

Programming Languages

- Python 3.9+ (primary)
- JavaScript/React (dashboard frontend)
- SQL (data queries)

ML Libraries & Frameworks

- XGBoost 2.0+
- LightGBM 4.0+
- Scikit-learn 1.3+
- SHAP (explainability)
- Optuna (hyperparameter tuning)

Data Processing

- Pandas & NumPy
- PyArrow (Parquet handling)
- xarray (weather data)
- pytrends (Google Trends API)

Visualization & Dashboard

- Plotly/Dash or Streamlit
- Folium (interactive maps)
- Matplotlib/Seaborn (static plots)
- React.js (optional web interface)

Development & Deployment

- Git/GitHub (version control)
- Docker (containerization)
- Jupyter Lab (development)
- GitHub Actions (CI/CD)

Data Storage & APIs

- PostgreSQL/SQLite (structured data)
- AWS S3 or local storage (data lake)
- FastAPI (REST API for predictions)

2. Timeline

Weeks 1-2: Data Collection & Exploration (Nov 4-17)

Tasks:

- Download ERA5/GPM weather data via API
- Collect Google Trends data for Thai & English keywords
- Obtain historical dengue case data from MOPH sources
- Perform exploratory data analysis (EDA)
- Document data quality issues and gaps

Responsible: All team members

- Sue Sha: Weather data collection
- Mya Moe: Search trends collection
- Shwe Sin: Case data collection
- Kyaw Soe & Phyo Zay: EDA

Weeks 3-4: Data Preprocessing & Feature Engineering (Nov 18 - Dec 1)

Tasks:

- Align all datasets to ISO week format
- Handle missing values and outliers
- Create temporal lags (t-1 to t-4 weeks)
- Engineer rolling statistics and wet-spell features
- Normalize and scale features
- Create train/validation/test splits (walk-forward)

Responsible:

- Kyaw Soe: Temporal features
- Phyo Zay: Weather features
- Sue Sha: Data integration

Weeks 5-6: Model Development & Training (Dec 2-15)

Tasks:

- Implement baseline Logistic Regression model
- Develop XGBoost and LightGBM models
- Implement walk-forward cross-validation
- Hyperparameter tuning using Optuna
- Train Random Forest as backup model
- Compare model performance metrics

Responsible:

- Mya Moe: XGBoost/LightGBM development
- Shwe Sin: Baseline & Random Forest
- Kyaw Soe: Validation strategy

Weeks 7-8: Model Evaluation & Interpretability (Dec 16-29)

Tasks:

- Calculate AUROC, PR-AUC, F1-score, Brier score
- Compute Top-K hit rate for high-risk districts
- Implement SHAP explainability analysis
- Generate feature importance visualizations
- Validate model on holdout test set
- Document model performance and limitations

Responsible:

- Phyo Zay: SHAP implementation
- Sue Sha: Metrics calculation
- All: Interpretation

Weeks 9-10: Dashboard Development (Dec 30 - Jan 12)

Tasks:

- Design dashboard UI/UX wireframes
- Implement interactive risk map (Folium/Plotly)
- Create Top-K high-risk district table
- Build explanation cards with SHAP insights
- Add historical trend visualizations
- Integrate all components into cohesive interface

Responsible:

- Shwe Sin: Frontend design
- Mya Moe: Map implementation
- Kyaw Soe: Backend integration

Weeks 11-12: Testing, Documentation & Deployment (Jan 13-26)

Tasks:

- End-to-end system testing
- User acceptance testing with sample stakeholders
- Write comprehensive technical documentation
- Create user guide for public health staff
- Prepare final presentation and report
- Deploy system (local or cloud-based)

Responsible:

- Phyo Zay: Documentation lead
- Sue Sha: Deployment
- All: Testing & presentation preparation

Task Distribution Matrix

Task Category	**Sue Sha**	**Mya Moe**	**Shwe Sin**	**Kyaw Soe**	**Phyo Zay**
Data Collection	Weather	Search Trends	Case Data	EDA	EDA
Feature Engineering	Integration	Support	Support	Temporal	Weather
Model Development	Support	XGB/LGBM	Baseline/RF	Validation	Support
Evaluation	Metrics	Analysis	Support	Support	SHAP
Dashboard	Support	Map	Frontend	Backend	Support
Documentation	Deployment	Testing	Testing	Testing	Lead Writer

3. Milestones

Milestone 1: Data Integration Complete (Week 2)

All three data sources successfully downloaded, cleaned, and merged into a unified dataset with documented quality assessment.

Success Criteria: >90% data completeness, aligned time periods, no critical errors

Milestone 2: Feature Engineering Pipeline (Week 4)

Automated pipeline producing all lag features, rolling statistics, seasonality indicators, and normalized datasets ready for model training.

Success Criteria: 50+ engineered features, reproducible pipeline, proper train/test splits

Milestone 3: Baseline Model Performance (Week 6)

Trained XGBoost/LightGBM models achieving minimum performance thresholds on validation set with documented hyperparameters.

Success Criteria: AUROC ≥0.75, F1-score ≥0.70, Top-10 hit rate ≥60%

Milestone 4: SHAP Explainability Implementation (Week 8)

Working SHAP analysis generating top-3 feature explanations for each prediction with human-readable summaries.

Success Criteria: Explanations match model behavior, clear visualizations, <2 second computation time

Milestone 5: MVP Dashboard Launch (Week 10)

Functional web dashboard displaying current week predictions with interactive map, Top-K table, and explanation cards.

Success Criteria: All visualization components working, responsive design, <3 second load time

Milestone 6: Project Delivery (Week 12)

Complete system deployed with full documentation, user guide, final presentation, and handover materials ready for stakeholders.

Success Criteria: All deliverables completed, system operational, successful final presentation

4. Challenges and Mitigations

Challenge 1: Data Quality & Completeness

Issue: Historical case data may have missing weeks, underreporting, or delayed updates. Weather satellite data can have gaps due to sensor issues.

Mitigation: Implement robust missing data imputation strategies (forward-fill for short gaps, interpolation for weather, exclusion for long gaps). Use multiple weather data sources (ERA5 + GPM + TRMM) for redundancy. Document all data quality issues and exclude problematic time periods from training if necessary. Build model uncertainty estimates to flag predictions with low confidence due to data quality.

Challenge 2: Model Overfitting & Generalization

Issue: Limited historical data (3-10 years) increases risk of overfitting. Province-specific patterns may not generalize across regions.

Mitigation: Use rigorous walk-forward cross-validation that simulates real deployment. Apply regularization in tree models (max_depth limits, min_child_weight constraints). Implement province-aware validation splits. Monitor validation curves for overfitting signs. Use ensemble methods (XGBoost + LightGBM + Random Forest) to improve generalization. Start with simpler features before adding complexity.

Challenge 3: Computational Resource Constraints

Issue: Training ensemble models on 77 provinces × 10 years × 52 weeks with hyperparameter tuning can be computationally intensive.

Mitigation: Use efficient implementations (LightGBM for speed, GPU acceleration if available). Implement smart hyperparameter search (Optuna with early stopping, coarse-to-fine grid search). Cache intermediate results. Parallelize province-level model training. Use cloud computing resources (Google Colab, Kaggle kernels) if local resources insufficient. Optimize data storage formats (Parquet for faster I/O).

Challenge 4: Real-Time Data Pipeline

Issue: Operational deployment requires automated weekly data fetching from multiple APIs with different formats and update schedules.

Mitigation: Build modular ETL pipeline with error handling and retry logic. Implement data validation checks to detect anomalies. Create fallback mechanisms (use previous week's weather forecast if satellite data delayed). Schedule automated runs with monitoring alerts. For

MVP, focus on batch processing; real-time automation is a stretch goal. Document manual update procedures as backup.

Challenge 5: Stakeholder Adoption & Trust

Issue: Public health officials may be skeptical of ML predictions or find technical outputs difficult to interpret and act upon.

Mitigation: Prioritize interpretability through SHAP explanations in plain language. Design user-friendly dashboard with minimal jargon. Include uncertainty indicators (confidence scores). Validate predictions against known historical outbreaks to build trust. Create comprehensive user guide with case studies. Offer training sessions for end users. Emphasize that system is decision-support tool, not replacement for human expertise.

Challenge 6: Class Imbalance

Issue: High-risk weeks are rare events (top 20% by definition), creating imbalanced classification problem that can bias models toward predicting "low risk."

Mitigation: Use appropriate evaluation metrics (PR-AUC, F1-score) that account for imbalance. Apply class weighting in training (scale_pos_weight in XGBoost). Consider SMOTE or other resampling techniques cautiously. Optimize probability thresholds for practical use case (prioritize recall for public health). Evaluate using Top-K hit rate which is imbalance-agnostic.

5. Ethical Considerations

Data Privacy & Patient Confidentiality

We use only aggregated, de-identified dengue case counts at provincial/district level from publicly available government sources. No individual patient data, medical records, or personally identifiable information is collected or processed. Google Trends data is aggregated search volume with no individual user tracking. We commit to never attempting to re-identify individuals from aggregated data and will handle all data in compliance with Thai data protection regulations.

Algorithmic Bias & Fairness

Dengue risk prediction may exhibit spatial bias if certain regions have better data quality, internet penetration (affecting search trends), or healthcare reporting infrastructure. We mitigate this by:

(1) validating model performance separately for urban vs rural provinces;
(2) not penalizing areas with limited search data by making features optional;

(3) using province-specific baselines to avoid disadvantageously historically high-burden areas; and
(4) providing uncertainty indicators when data quality is poor. We acknowledge that predictions are only as good as underlying data and will clearly communicate limitations.

Potential for Harm & Misuse

Inaccurate predictions could lead to either:

(1) false alarms causing unnecessary panic and wasted resources, or
(2) missed outbreaks resulting in preventable illness

We address this by emphasizing that our system is a decision-support tool to complement, not replace, existing surveillance. Predictions should be validated by local health officials and combined with other information sources. We also avoid creating stigma around "high-risk" areas by framing alerts as opportunities for proactive intervention rather than blame. Clear communication emphasizes that high risk reflects environmental conditions, not community failure.

Resource Allocation & Equity

Risk alerts could inadvertently direct resources away from chronically underserved areas toward regions with better prediction accuracy or political visibility. We mitigate this by:

(1) providing equal coverage for all provinces regardless of historical data quality;
(2) highlighting uncertainty in predictions to prevent overconfidence;
(3) recommending baseline prevention measures for all areas, with enhanced measures for predicted high-risk zones; and
(4) advocating for equitable resource distribution in our user documentation.

Transparency & Accountability

We commit to full transparency about our methodology, data sources, model limitations, and performance metrics. SHAP explanations provide interpretation for every prediction. We will publish our code openly and document all assumptions. If our system is adopted operationally, we recommend establishing a feedback loop where public health officials can report prediction accuracy, enabling continuous improvement and accountability for model performance.

Community Impact & Benefit Sharing

This project is designed to benefit Thai communities affected by dengue, not to extract value. We commit to:

(1) making our tool freely available to public health agencies;
(2) training local staff to maintain and update the system;

(3) prioritizing deployment in underserved areas with high disease burden; and
(4) actively seeking feedback from affected communities and health workers to ensure the tool meets real needs.

Success is measured by reduced dengue incidence and improved public health outcomes, not commercial metrics.

6. References

1. Soneja, S., et al. (2021). A review of dengue's historical and future health risk from a changing climate. *International Journal of Environmental Research and Public Health*. https://pmc.ncbi.nlm.nih.gov/articles/PMC8416809/

2. Sugeno, M., et al. (2023). Association between environmental factors and dengue in Lao PDR. *BMC Public Health*. https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-023-17277-0

3. Tian, Y., et al. (2025). Rising dengue risk with increasing ENSO. *Nature Communications*. https://www.nature.com/articles/s41467-025-63655-0

4. van Panhuis, W. G., et al. (2015). Region-wide synchrony & traveling waves of dengue. *Proceedings of the Royal Society B*. https://pmc.ncbi.nlm.nih.gov/articles/PMC4620875/

5. Johansson, M. A., et al. (2009). Multiyear climate variability & dengue ENSO. *PLOS Medicine*. https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000168

6. Li, Z., et al. (2022). Forecasting dengue by integrating Google Earth Engine, AI & GT. *International Journal of Applied Earth Observation and Geoinformation*. https://pmc.ncbi.nlm.nih.gov/articles/PMC9603269/

7. Puengpreeda, A., et al. (2020). Weekly Forecasting Model for DHF in Thailand using Google Trends & meteorology. *Engineering Journal*. https://engj.org/index.php/ej/article/view/3498