Executive Summary

This study develops and evaluates a machine-learning–based early warning model for dengue outbreaks in Malaysia using weekly surveillance data (2011–2024) and weather variables derived from the NASA POWER API. An outbreak was operationally defined as weeks exceeding the 75th percentile of dengue cases over the preceding four weeks to capture short-term surges while maintaining sufficient prevalence for modelling.

To reduce seasonal confounding, dengue case series were detrended using a Generalized Additive Model (negative binomial), and cross-correlation analysis on residuals guided candidate weather lag selection. Models were trained using XGBClassifier with rolling-origin cross-validation. Feature importance was assessed using SHAP, permutation importance, and remove-and-retrain ablation tests with bootstrap confidence intervals.

Results show that lagged dengue cases remain the strongest predictors; however, weather features particularly lagged temperature and rolling precipitation, consistently improved model performance. The final tuned model achieved ROC-AUC $\approx$ 0.73 and PR-AUC $\approx$ 0.60 on a 2023-2024 holdout set. While default probability thresholds yielded low outbreak detection (due to conservative nature), adjusting the decision threshold enabled high sensitivity (>90%), making the model suitable for early warning and risk prioritization rather than precise outbreak confirmation.

This work demonstrates that weather variables provide complementary predictive value beyond autoregressive dengue signals, while also highlighting trade-offs between sensitivity and precision in outbreak forecasting. Limitations include coarse spatial weather resolution, absence of entomological indices, and the need for external validation.