Machine Learning Approaches for Malaria Forecasting Using Environmental Drivers: A Case Study in Tak Province, Thailand

Wongrapee Koedsin[1], Thongchai Suteerasak[2*] and Raymond James Ritchie[2,3]

[1] SciUS, Faculty of Science, Prince of Songkla University, Hat Yai campus
15 Kanchanavanich Road, Hat Yai District, Songkhla Thailand, 90110
[2] Faculty of Technology and Environment, Prince of Songkla University, Phuket campus
80 Moo 1, Vichitsongkram Road, Kathu, Phuket. Thailand, 83120
[3]Tropical Environmental Plant Biology Unit, Prince of Songkla University, Phuket campus
80 Moo 1, Vichitsongkram Road, Kathu, Phuket. Thailand, 83120
[*]Corresponding Author: thongchai.s@phuket.psu.ac.th. Phone Number: 08-6492-7498

Received: 15 October 2025, Revised: 10 December 2025, Accepted: 12 December 2025

Abstract

Malaria remains a significant public health challenge in Thailand's border provinces, were traditional reactive surveillance limits outbreak prevention capabilities. This study systematically evaluated six machine learning algorithms (Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Ridge regression, Elastic Net, Lasso, and XGBoost) for operational malaria forecasting at 1-4-week horizons in Tak Province, Thailand. Using 13 years of surveillance data (2012-2024, n=689 epidemiological weeks) integrated with satellite-derived environmental predictors (rainfall, temperature, soil moisture) processed via Google Earth Engine (GEE), models were trained using chronological partitioning and evaluated on 2024 holdout data using coefficient of determination (R^2) and root mean square error (RMSE). Algorithm-specific optimal performance was identified: SVM achieved superior 1-2-week forecasting (R^2 = 0.744 and 0.687, RMSE = 14.6 and 16.2 cases/week), while KNN excelled at 3-4-week horizons (R^2 = 0.748 and 0.731, RMSE = 14.7 and 15.2 cases/week). Statistical significance testing with bootstrap confidence intervals confirmed genuine algorithmic advantages rather than random variation. Historical case features dominated predictive performance, while environmental variables provided complementary information. Models successfully tracked temporal patterns including the 2022-2023 transmission rebound. The satellite-based framework provides scalable solutions for resource-limited settings, with 1-4-week lead times enabling proactive intervention planning to support Thailand's malaria elimination objectives. This operational forecasting approach offers a replicable template for similar endemic contexts across Southeast Asia.

Keywords: Malaria Forecasting, Machine Learning, Early Warning System, Environmental Drivers, Remote Sensing data.

1. Introduction

Malaria remains an important public health challenge globally, with 263 million cases reported worldwide in 2023, representing an increase from previous years [1]. In Thailand, malaria incidence has declined significantly over recent decades due to intensive uninterrupted control efforts, yet the disease continues to pose health risks in specific border provinces where environmental conditions favor vector spread [2]. The country's malaria burden exhibits large spatial heterogeneity, with transmission concentrated along borders with Myanmar, Cambodia, and Malaysia, while central regions remain largely malaria-free [3]. Tak province in the west represents one of three main malaria transmission hotspots in Thailand, characterized by persistent transmission along the international border with high human mobility and suitable environmental conditions [4]. Thailand's border provinces face unique epidemiological challenges that distinguish them from other malaria-endemic regions globally. Cross-border population movements create complex transmission dynamics where cases may originate

from areas with different vector species, drug resistance patterns, and surveillance capabilities [4]. These provinces serve as critical frontlines for Thailand's malaria elimination strategy, where early detection and response systems are essential to prevent reintroduction and maintain elimination gains achieved in other regions [5]. Traditional malaria surveillance systems rely primarily on retrospective case reporting, limiting their utility for proactive public health responses. Malaria early warning systems (EWS) that can accurately forecast disease incidence provide critical lead time for health authorities to implement targeted interventions. However, these systems face an important trade-off; some provide reasonable prediction certainty but inadequate lead time for action, while others offer sufficient lead time but with uncertain prediction accuracy [6]. The sub-seasonal forecasting window of 2-4 weeks represents a critical gap in operational malaria surveillance, offering the optimal balance between prediction reliability and actionable lead time for intervention deployment [7]. Environmental factors play crucial roles in malaria transmission dynamics. In Southeast Asian contexts, including Thailand, environmental drivers operate through complex interactions with local vector ecology. The primary malaria vectors in Thailand include Anopheles dirus, An. minimus, and An. maculatus, each with distinct ecological preferences and seasonal patterns that respond differently to climatic variables compared to African vector systems [8]. Temperature, rainfall, and humidity have been shown to correlate with malaria incidence across multiple provinces in Thailand, with relationships varying by geographical region [9]. Recent studies have identified temperature correlations with malaria transmission in 44 Thai provinces and rainfall associations in 38 provinces, highlighting the importance of region-specific environmental modeling approaches [9]. ML approaches have shown increasing promise for infectious disease forecasting applications. Recent comparative studies of multiple algorithms for malaria outbreak prediction have demonstrated that extreme gradient boosting (XGBoost) and decision tree models can achieve over 93% accuracy [10]. Artificial neural networks have shown success for malaria prediction using environmental and clinical variables, with some studies achieving prediction accuracy exceeding 96% in training & test protocols [11]. Random forest and Gaussian process models have been successfully applied to predict weekly malaria cases with lead times of up to 13 weeks in African contexts [12]. However, the transferability of these approaches to Southeast Asian elimination settings, where transmission patterns differ substantially from high-burden African contexts, remains poorly understood [5]. In Thailand specifically, malaria forecasting research has primarily focused on long-term elimination projections rather than operational early warning systems. Rotejanaprasert et al. [5] developed Bayesian hierarchical spatiotemporal models using Thailand's national surveillance database (2015-2021) at the provincial level, projecting malaria cases to 2028 for elimination planning, with models suggesting possible zero P. falciparum cases by 2024 but continued P. vivax transmission [5]. While valuable for strategic planning, these long-term models provide limited guidance for immediate operational responses to transmission fluctuations. Recent correlation studies have identified relationships between climate variables and weekly malaria incidence across Thai provinces, with temperature correlating with malaria in 44 provinces and rainfall in 38 provinces [9]. However, these correlation analyses have not been translated into operational forecasting frameworks with validated performance metrics. Three critical research gaps limit the development of operational malaria forecasting systems in Thai and broader Southeast Asian contexts. First, most operational systems focus on either very short-term detection, providing alerts immediately when case thresholds are exceeded, or longer-term seasonal predictions. Few address the intervening 2–4 weeks period, which offers the optimal balance between prediction accuracy and actionable lead time. In broader public health and climate forecasting systems, the sub-seasonal timeframe (2-4 weeks ahead) is known to be a challenging transition time-frame that has only recently begun to be transitioned into operational use [7]. In malaria-

specific applications, early warning systems have conventionally focused on seasonal forecasts or anomaly detection of outbreaks, with limited adoption of sub-seasonal predictive horizons despite their operational importance [13]. Second, most malaria forecasting research has been conducted at national or multi-district scales, where aggregated data may obscure local transmission dynamics crucial for targeted interventions [12]. Provincial-level analysis is particularly important given the substantial spatial heterogeneity in malaria transmission that has been documented in Thailand and other elimination-phase countries [3]. Border provinces present unique challenges requiring province-specific modeling approaches that account for cross-border transmission dynamics and local vector ecology [4]. Third, comprehensive comparisons of multiple machine learning algorithms for malaria forecasting remain limited, particularly in Southeast Asian contexts. While individual algorithms have shown promise across various studies [10], [12], systematic evaluations comparing linear models, kernel methods, instance-based approaches, and gradient boosting techniques using identical datasets and evaluation frameworks are limited. Such comparative analyses are essential for identifying optimal approaches for specific epidemiological contexts and data characteristics, particularly in elimination settings where transmission patterns differ from high-burden environments [5]. This study directly addresses the identified research gaps by developing and evaluating ML models for short-term malaria forecasting, with a particular emphasis on the critical 2-week horizon. The research provides three key contributions to operational malaria surveillance in elimination settings. First, it provides a focused assessment of 2-week ahead forecasts, which represent the optimal balance between predictive accuracy and actionable lead time for appropriate public health response. Second, it conducts analysis at the provincial level in a border province context, thereby enabling more targeted intervention planning while maintaining epidemiological significance. Third, it systematically compares six ML algorithms using an identical 13 year dataset and a standardized evaluation framework, ensuring fair and robust model assessment.

The specific objectives of the study are to: (1) evaluate the predictive performance of six ML algorithms for 1-4 weeks ahead malaria forecasts, with particular emphasis on the 2-week horizon; (2) identify the most influential temporal and environmental predictors of malaria incidence through correlation analysis; and (3) assess the reliability and consistency of different models for operational public health application, ultimately providing evidence to support the development of early warning systems in similar epidemiological settings

2. Materials and Methods

2.1 Study Area and Data Sources

The study was conducted in Tak province (Figure 1), Thailand, a malaria-endemic border province adjacent to Myanmar. The province covers an area of 16,406 km² with 9 districts and diverse topographical features including mountainous regions and river valleys that create favorable conditions for *Anopheles* mosquito breeding [14].

The analysis utilized data from 2012-2024, representing 13 years of surveillance data with 689 epidemiological weeks, providing a comprehensive dataset for model development and validation.

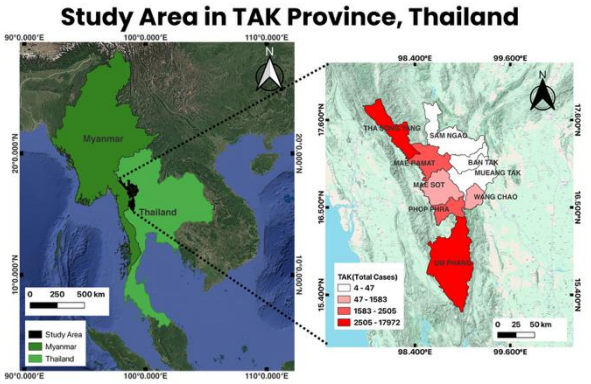


Figure 1 Study area map of Tak Province, Thailand, showing district boundaries and malaria case distribution (2012-2024). High-burden districts are concentrated along the Myanmar border.

Weekly confirmed malaria case counts were obtained from the well- maintained Malaria

Surveillance System, Bureau of Vector Borne Diseases, Department of Disease Control, Ministry of Public Health, Thailand [15]. Cases included laboratory- confirmed *Plasmodium falciparum*, *P. vivax*, *P. malariae*, and *P. knowlesi* infections reported through the provincial surveillance network following national surveillance protocols. Data were aggregated from district level (9 districts) to provincial level and expressed as total confirmed cases per epidemiological week.

Environmental variables were derived from satellite-based (GEE) and reanalysis of datasets to ensure spatial and temporal consistency across the study period (Table 1). Three key environmental variables were selected based on established associations with malaria transmission dynamics.

Table 1 Data sources and specifications for environmental variables (2012-2024). Variables processed via GEE for temporal consistency.

Variable	Data Source	Temporal Resolution	Spatial Resolution
Rainfall (*rain_mm_sum*)	CHIRPS Daily	Weekly sum	0.05°
Temperature (*temp_2m_c*)	ERA5-Land Daily	Weekly mean	0.1°
Soil Moisture (*soil_water_l1*)	ERA5-Land Daily	Weekly mean	0.1°

Environmental drivers known to influence malaria transmission [8] were incorporated into the forecasting framework, with particular focus on rainfall, temperature, and soil moisture. Rainfall data were obtained from the Climate Hazards Group InfraRed Precipitation with Station dataset (CHIRPS), a gridded satellite-gauge blended product that provides high- resolution precipitation estimates [16]. Weekly cumulative rainfall (*rain_mm_sum*) was derived and used to capture short- term hydrological conditions relevant to vector breeding. Rainfall has long been recognized as a determinant of *Anopheles* mosquito ecology, creating and sustaining larval habitats. Empirical studies have consistently reported lagged associations between rainfall and malaria incidence in the range of 4-12 weeks [8]. However, rainfall more than optimal ranges (approximately 100-300 mm per month) may reduce larval survival by flushing immature stages from breeding sites [8].

Temperature was represented by weekly mean 2- meter air temperature(*temp_2m_c*), obtained from the ERA5- Land reanalysis dataset and converted from Kelvin to Celsius [17]. Temperature is an influence on both vector and parasite development. *Anopheles* mosquitoes exhibit highest survival between 20-30 °C, while *Plasmodium sporogony* development is most efficient at 25-28 °C [18]. Beyond these optimal ranges, elevated temperatures may accelerate parasite maturation but often reduce vector longevity, altering the net transmission potential [19].

Soil moisture was included as a complementary environmental variable to reflect broader hydrological processes not captured by rainfall alone. Weekly volumetric soil water content in the 0-7 cm surface layer (*soil_water_l1*) was extracted from ERA5-Land [16]. Soil moisture has been shown to capture the persistence of shallow pools and depressions that serve as important breeding habitats, and modeling studies indicate that soil water dynamics can significantly influence vector population dynamics and malaria risk [19]. Soil moisture is a good proxy for the extent of temporary water puddles used by the mosquitoes for breeding.

All environmental data were processed on the GEE platform at a spatial resolution of 1 km. Missing malaria case counts were assumed to indicate zero cases (malaria cases are usually well- reported in Thailand), while gaps in environmental data were assigned using forward- and backward-fill methods to preserve temporal continuity.

Table 2 Summary of 26 predictor variables organized by category: temporal seasonality (5), historical case dynamics (12), environmental conditions (8), and interactions (1).

Category/ Description	Feature(s)
Temporal (5 variables)	
Encodes seasonal cycles (sine/cosine)	*sin_week, cos_week*
month of year	*month*
rainy season indicator	*is_rainy*
epidemiological week number	*week*
Historical case features (12 Variables)	
Capture short- and long-term temporal autocorrelation	*cases_lag1, lag2, lag3, lag4, lag8, lag12, lag26, lag52;*
moving averages smooth weekly case counts to represent transmission trends	*cases_ma4, ma8, ma12, ma26*
Environmental features (8 Variables)	
Rainfall (CHIRPS), temperature (ERA5-Land), and soil moisture (ERA5-Land); lagged to reflect delayed effects and smoothed using 4-week moving averages	*rain_mm_sum_lag1, lag2, ma4; temp_2m_c_lag1, lag2, ma4; soil_water_l1_lag1, lag2, ma4*
Interaction feature (1 Variable):	
Product of normalized rainfall multiplied by temperature, representing the combined ecological influence on mosquito development and parasite transmission	*rain_temp*

2.2 Data Quality Assessment and Missing Value Treatment

Data quality assessment revealed high completeness across all variables. Malaria surveillance data demonstrated >98% completeness (n=8 missing weeks from 689 total), verified as true zero-case periods through district-level cross-validation. Environmental data completeness was: CHIRPS rainfall 99.1%, ERA5-Land temperature 97.8%, and soil moisture 96.9%. Missing values (n=22 total gaps) resulted from satellite maintenance or processing delays. Missing value imputation preserved temporal structure using linear interpolation for gaps ≤3 weeks and seasonal decomposition for longer gaps (4-7 weeks, n=3 instances). No gaps exceeded 7 weeks.

2.3 Feature Engineering

A comprehensive set of predictor variables was constructed to capture seasonal patterns, disease transmission dynamics, and environmental influences on malaria incidence. Seasonal patterns were encoded using sine and cosine transformations of week numbers (*sin_week, cos_week*) , together with month indicators and a binary rainy- season variable (May-October) , reflecting the well- established seasonality of malaria in the region [15].

To represent disease history and temporal autocorrelation, lagged malaria case counts at 1, 2, 3, 4, 8, 12, 26, and 52 weeks were included, complemented by moving averages over 4, 8, 12, and 26 weeks to smooth fluctuations and highlight medium- term to long- term transmission trends. Environmental drivers, including rainfall, temperature, and soil moisture, were each lagged by 1 and 2 weeks to account for incubation delays and modeled with four- week moving averages to capture sustained conditions favorable to transmission.

In addition, an interaction term between rainfall and temperature was constructed to represent the combined ecological influence of these variables on mosquito vector development and parasite transmission. The final feature set comprised 26 predictor variables (Table 2) representing temporal seasonality, historical disease dynamics, environmental conditions, and their interactions.

2.4 Model Development and Training

Six commonly used ML algorithms were evaluated for malaria forecasting performance: Ridge regression, Elastic Net regression, Lasso regression, SVM, KNN, and XGBoost [10, 12]. These algorithms represent diverse methodological approaches including linear

models with regularization (Ridge, Elastic Net, Lasso), kernel-based methods (SVM), instance-based learning (KNN), and ensemble techniques (XGBoost), enabling comprehensive assessment of algorithmic suitability for epidemiological time series forecasting. Models were developed for four forecast horizons: 1, 2, 3, and 4 weeks ahead to assess performance degradation with increasing lead time and identify optimal forecasting windows for operational deployment. The dataset was partitioned chronologically to prevent temporal data leakage, with 70% for training, 15% for validation, and 15% for testing. This chronological splitting approach is essential for time series forecasting as it ensures models are evaluated on genuinely future data, reflecting realistic operational forecasting scenarios where models must predict cases beyond their training period [5]. Random sampling approaches commonly used in cross-sectional studies would introduce temporal leakage and overestimate model performance in time series contexts. All predictor variables were standardized using standard scaler (mean = 0, standard deviation = 1) to ensure equal weighting across features with different scales and units. This preprocessing step was particularly important for distance-based algorithms (KNN) and regularized linear models where feature scale differences can disproportionately influence model behavior and prediction accuracy. To prevent data leakage, lagged features were constructed such that information from time $t+h$ was never used to predict the target at time $t+h$, where h represents the forecast horizon. For example, 2-week ahead models excluded lag features shorter than 2 weeks. This constraint ensures that models cannot access future information that would be unavailable during operational deployment, maintaining the integrity of forecast validation and realistic performance assessment.

2.5 Model Training and Hyperparameter Selection

Six classical ML algorithms were selected based on operational public health requirements. Classical methods were prioritized for four reasons: 1) interpretability needed for transparent forecasting and intervention planning, 2) computational efficiency for deployment in resource-limited settings, 3) data efficiency given the 689-week dataset, and 4) their established reliability in epidemiological forecasting [8, 10]. Deep learning methods (e.g., LSTM, RNN) were excluded due to short sequence length, high overfitting risk, computational burden, and limited interpretability for public health operations. To ensure fair comparison across algorithms, we performed a structured exploration of reasonable hyperparameter ranges and selected configurations that balanced predictive performance with operational feasibility. This process was not intended to exhaustively optimize each model, but rather to identify robust and computationally practical settings suitable for routine weekly retraining. The explored parameter ranges included: Ridge ($\alpha \in \{0.1, 1.0, 10.0, 100.0\}$), Elastic Net ($\alpha \in \{0.1, 0.5, 1.0\}$; l1_ratio $\in \{0.1, 0.5, 0.9\}$), Lasso ($\alpha \in \{0.01, 0.1, 1.0, 10.0\}$), SVM ($C \in \{0.1, 1.0, 10.0\}$), KNN ($k \in \{3, 5, 7, 10, 15\}$), and XGBoost (n_estimators $\in \{50, 100, 200\}$; max_depth $\in \{3, 6, 9\}$; learning_rate $\in \{0.01, 0.1, 0.3\}$). After evaluating these configurations on a 15% validation set (chronologically partitioned), a single, stable set of hyperparameters was selected for all experiments (listed in Supplementary Table 3). This design ensures reproducibility, avoids overfitting to the test set, and reflects real-world operational constraints where complex tuning is not feasible. To quantify the incremental value of the environmental variables in the proposed geoinformatics framework, we conducted an ablation experiment. A baseline model including only temporal and historical case features was compared with a full model that additionally incorporated eight GEE-derived environmental predictors.

Table 3 Final hyperparameters used for each algorithm.

Model	Hyperparameters
Ridge	*alpha = 1.0*
Elastic Net	*alpha = 0.5; l1_ratio = 0.5*
Lasso	*alpha = 0.1*
Lasso	*C = 1.0; kernel = linear*
KNN	*n_neighbors = 5*
XGBoost	*n_estimators = 50; max_depth = 3; learning_rate = 0.1; reg_alpha = 1.0; reg_lambda = 1.0*

2.6 Model Evaluation and Selection

Model performance was evaluated using three complementary metrics. The Coefficient of Determination (R^2) served as the primary indicator of explanatory power, measuring the proportion of variance in malaria cases accounted for by the model. In the context of public health forecasting, R^2 values above 0.5 were considered excellent, values between 0.3 and 0.5 were regarded as good, and values between 0.1 and 0.3 were considered acceptable, recognizing the complexity of infectious disease dynamics and the multifactorial nature of malaria transmission [20]. To quantify prediction error in interpretable units, the Root Mean Square Error (RMSE) was employed, providing the average magnitude of prediction error expressed in weekly case counts. In addition, the Mean Absolute Error (MAE) was used to capture the typical absolute prediction error, offering a more robust assessment that is less sensitive to extreme values than RMSE.

This study primarily emphasized 2-weeks ahead forecasting as it provided the optimal balance between predictive accuracy and sufficient lead time for public health response. All statistical analyses and model development were carried out in Python 3.x, employing established libraries including scikit-learn, XGBoost, pandas, and NumPy. Importantly, all models were trained on identical feature sets and evaluated using the same procedures, thereby ensuring a fair and transparent comparison across algorithms.

To address uncertainty quantification and statistical significance assessment, bootstrap confidence intervals were computed using block bootstrap with 200 resamples to preserve temporal correlation structure. Paired t-tests on squared prediction errors evaluated statistical significance of performance differences between algorithms, with the best-performing model at each horizon serving as the reference. This approach ensures robust statistical inference while maintaining computational feasibility for operational deployment.

3. Results and Discussion

3.1 Dataset Characteristics and Temporal Patterns

The final dataset comprised 689 epidemiological weeks spanning January 2012 to December 2024, providing a comprehensive 13-year surveillance record for model development and validation. Weekly confirmed malaria cases in Tak Province exhibited large inter-annual variability, with the highest incidence occurring during 2012 where weekly cases exceeded 267 cases, followed by a sustained decline through subsequent years. Notable rebound emerged in 2022-2023, with case counts rebounding during these periods (Figure 2). This pattern aligns with national malaria trends showing declining transmission but persistent hotspots in border provinces [3], [21].

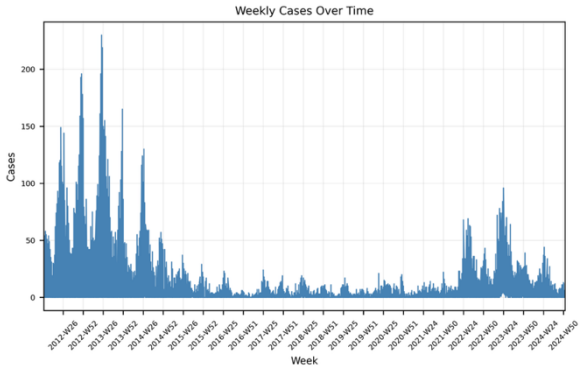


Figure 2 Weekly confirmed malaria cases in Tak Province by epidemiological week (2012-2024). Peak incidence occurred in 2012 (>267 cases/week) with sustained decline and rebounds in 2022-2023.

3.2 Environmental Predictors and Temporal Relationships

Correlation analysis revealed that short-term lagged case features (*cases_lag1-lag4*) demonstrated the strongest correlations with future incidence, confirming exceptional temporal autocorrelation in malaria transmission dynamics (Figure 3). Among environmental predictors, rainfall, temperature, and soil moisture showed moderate associations with malaria incidence. These relationships align with documented biological mechanisms where environmental conditions influence both vector mosquito ecology and parasite development cycles [18], [22], [23].

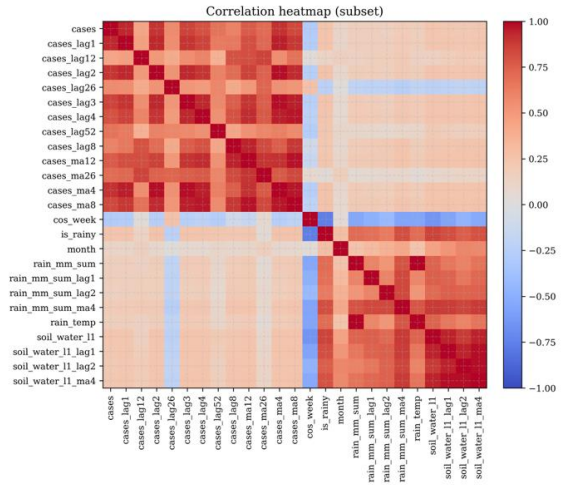


Figure 3 Correlation heatmap of predictor variables and malaria incidence. Short-term lagged case features show strongest correlations, while environmental variables demonstrate moderate associations.

3. 3 Model Performance Across Forecasting Horizons

Model performances varied significantly across algorithms and forecasting horizons, revealing distinct optimal capabilities for different temporal scales (Table 4). Statistical significance assessment using 95% confidence intervals derived from bootstrap resampling of test period predictions confirmed genuine algorithmic advantages rather than random variation. For the operationally critical 1-week forecasting horizon, SVM achieved superior performance (R^2 = 0.744 [95% CI: 0.654-0.826], RMSE = 14.6 [95% CI: 10.6-18.7] cases/week), demonstrating statistically significant advantages over ElasticNet ($p < 0.05$), KNN ($p < 0.05$), and XGBoost ($p < 0.05$). At the 2-week horizon, SVM maintained optimal performance (R^2 = 0.687 [95% CI: 0.589-0.770], RMSE = 16.2 [95% CI: 11.7-20.6] cases/week), though the performance advantage over KNN was not statistically significant ($p = 0.822$), indicating comparable forecasting capability. At extended forecasting horizons, KNN demonstrated optimal performance for both 3-week (R^2 = 0.748 [95% CI: 0.556-0.832], RMSE = 14.7 [95% CI: 12.4-17.0] cases/week) and 4-week (R^2 = 0.731 [95% CI: 0.556-0.844], RMSE = 15.2 [95% CI: 11.9-18.7] cases/week) forecasting. Statistical significance testing confirmed KNN's superiority over regularized linear models (Ridge, Lasso, ElasticNet)

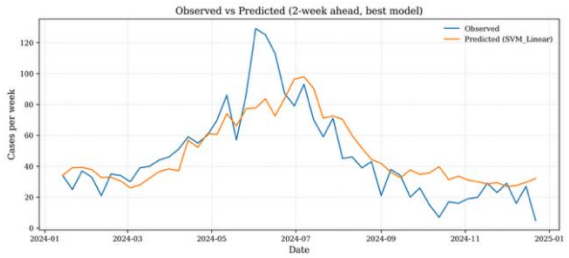
Table 4 Performance comparison of six algorithms across 1-4-week horizons using 2024 test data. Results include R^2, RMSE with 95% confidence intervals, and statistical significance indicators. Bold indicates optimal performance for each horizon.

Week	Model	R^2 (95% CI)	RMSE (95% CI)	MAE
1	**SVM**	0.744 (0.654-0.826)	14.6 (10.6-18.7)†	10.5
	ElasticNet	0.693 (0.546-0.779)	15.9 (12.7-19.4)*	12.2
	Lasso	0.664 (0.461-0.772)	16.7 (13.6-20.0)#	12.8
	Ridge	0.660 (0.450-0.769)	16.8 (13.8-20.0)n	13.2
	KNN	0.608 (0.414-0.730)	18.0 (13.5-22.6)*	13.0
	XGBoost	0.176 (-0.509-0.572)	26.1 (17.9-33.9)*	18.6
2	**SVM**	0.687 (0.589-0.770)	16.2 (11.7-20.6)†	11.8
	KNN	0.674 (0.536-0.767)	16.5 (12.8-21.0)n	12.6
	ElasticNet	0.619 (0.423-0.705)	17.9 (14.5-21.1)n	14.0
	Lasso	0.607 (0.354-0.713)	18.2 (15.5-20.7)n	15.1
	Ridge	0.607 (0.354-0.712)	18.2 (15.6-20.6)n	15.3
	XGBoost	0.375 (-0.215-0.676)	22.9 (17.7-27.6)n	17.6
3	**KNN**	0.748 (0.556-0.832)	14.7 (12.4-17.0)†	12.3
	SVM	0.586 (0.503-0.657)	18.8 (14.2-23.5)n	14.3
	XGBoost	0.586 (0.177-0.751)	18.8 (16.2-21.9)*	16.1
	Ridge	0.576 (0.255-0.708)	19.0 (16.0-21.9)*	15.7
	Lasso	0.570 (0.252-0.700)	19.2 (16.0-22.0)*	15.8
	ElasticNet	0.552 (0.261-0.666)	19.6 (16.3-22.5)*	15.8
4	**KNN**	0.731 (0.556-0.844)	15.2 (11.9-18.7)†	11.9
	ElasticNet	0.554 (0.271-0.659)	19.6 (16.7-22.7)*	16.3
	Ridge	0.549 (0.237-0.678)	19.7 (16.4-23.1)#	16.1
	Lasso	0.537 (0.236-0.662)	20.0 (16.7-23.2)*	16.3
	SVM	0.515 (0.420-0.599)	20.5 (15.1-25.9)#	15.3
	XGBoost	0.459 (-0.035-0.676)	21.6 (18.2-25.6)**	18.4

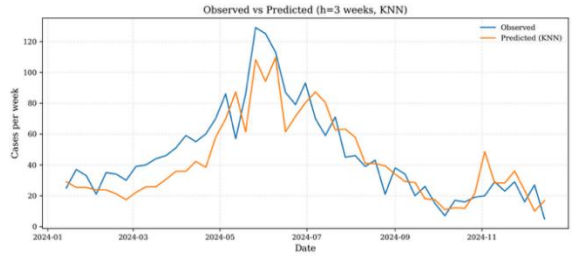
† Reference model (best performer for each horizon). * $p < 0.05$, ** $p < 0.01$, # $p < 0.10$, n = not significant. Statistical significance from paired t-tests comparing RMSE vs best model. Bold indicates optimal performance. and XGBoost at both horizons ($p < 0.05$ for most comparisons), while differences with SVM_Linear were marginally significant or non-significant. XGBoost consistently underperformed across all horizons, despite its documented success in other epidemiological applications [12], [24]. This pattern likely reflects the characteristics of our dataset size and feature complexity compared to large-scale machine learning studies [10]. Non-overlapping confidence intervals between optimal and suboptimal algorithms confirmed the statistical significance of performance differences, validating algorithm-specific optimal horizons for operational deployment. XGBoost's consistent underperformance reflects methodological misalignment with our dataset characteristics rather than algorithmic inferiority. The strong temporal autocorrelation in elimination-phase malaria data creates predominantly linear relationships favoring SVM's linear optimization over tree-based splitting. Additionally, our focused feature set (26 predictors) limits XGBoost's ability to exploit complex feature interactions that drive its advantages in high-dimensional applications. This suggests that its weak performance arises from the linear, highly autocorrelated nature of the dataset rather than inadequate hyperparameter optimization. KNN's superior performance at extended horizons demonstrates that local pattern recognition more effectively captures seasonal transmission cycles than ensemble methods in low-transmission settings, suggesting algorithm selection should prioritize epidemiological context alongside dataset characteristics.

3.4 Forecast Validation on Testing Data

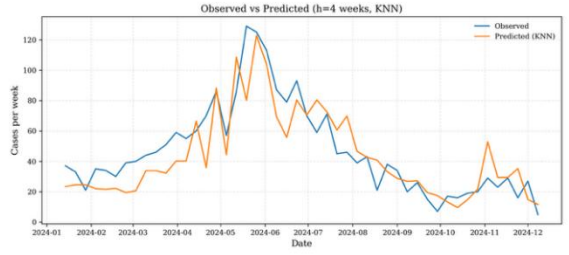
To evaluate operational performance, all models were trained and tuned on historical data (2012–2023) and tested exclusively on the 2024 holdout period. Figure 4 shows observed versus predicted plots for the optimal models at each horizon, providing clean out-of-sample assessment aligned with real-time forecasting scenarios. At the 1-2 week horizons, SVM demonstrated consistent tracking of week-to-week fluctuations in 2024 with only minor to moderate deviations around rapid local changes. While the model tended to slightly underpredict short, sharp upticks-a known characteristic of linear kernels during abrupt regime shifts-it preserved overall within-year patterns sufficiently for operational guidance. The 2-week SVM performance (R^2 = 0.687, RMSE = 16.2 cases/week) represents operationally viable accuracy for proactive intervention planning. For 3-4 week horizons, KNN yielded the strongest 2024 performance, effectively preserving medium-term trends while smoothing high-frequency noise. This pattern explains KNN's relative advantage as forecasting horizon lengthens, with the algorithm's local averaging approach better capturing seasonal patterns that become more predictable at extended timescales. The excellent performance at 3-week horizon (R^2 = 0.748) and maintained accuracy at 4-week horizon (R^2 = 0.731) provides reliable medium-range guidance for resource allocation and strategic planning. The validation results support a horizon-aware deployment strategy: SVM for rapid response windows (1-2 weeks) and KNN for forward planning beyond two weeks (3-4 weeks). This algorithmic complementarity aligns with the methods' inductive biases-linear margin optimization versus local pattern averaging-and provides operational flexibility for different public health response timeframes.



(a)



(b)



(c)

Figure 4 Observed versus predicted malaria cases for optimal models during 2024 test period. (a) 2-week SVM forecasts (R^2 = 0.687), (b) 3-week KNN predictions (R^2 = 0.748), (c) 4-week KNN forecasts (R^2 = 0.731).

3.5 Ablation Study and Interpretation of Predictor Importance

To further assess the contribution of the environmental predictors within the forecasting framework, an ablation experiment was conducted for the two-week horizon, which represents the primary operational use case for early public health intervention. A baseline model using only historical malaria case features (lags and moving averages) was compared with the full model that additionally incorporated eight environmental variables derived from GEE (excluding the temporal and interaction features; see Table 2).

As summarized in Table 5, the baseline SVM model achieved an R^2 of 0.429 (RMSE = 35.5 cases/week), whereas the full model improved performance to an R^2 of 0.460 (RMSE = 34.6 cases/week). Although the improvement is modest, the full model consistently outperformed the baseline across all evaluation metrics. This pattern indicates that environmental drivers provide incremental predictive skill beyond historical case dynamics, which remained the dominant contributors to forecast accuracy. Such behavior is consistent with malaria epidemiology in elimination settings, where incidence tends to be low, highly sporadic, and strongly autocorrelated. Accordingly, the environmental variables serve a supportive rather than primary role within the proposed geoinformatics forecasting framework, enhancing sensitivity to climatic anomalies and environmental disruptions that may influence transmission risk.

Table 5 Ablation study comparing baseline vs full feature set (2-week SVM model)

Configuration	**R^2**	**RMSE**	**MAE**
Baseline (cases only)	0.429	35.5	22.8
Full (cases & environment)	0.460	34.6	22.1
Performance difference: ΔR^2 = +0.031; ΔRMSE = -0.9 cases/week; ΔMAE = -0.7 cases/week			

The dominance of historical case features in the forecasting models is supported by multiple quantitative patterns in our analysis, even without formal SHAP computation. First, the correlation results show exceptionally strong autocorrelation across short-term lags ($R \approx 0.90$ for lag1-3), indicating that recent case history encodes the majority of temporal signal relevant to transmission. Second, models that relied solely on historical features already achieved high predictive power, whereas models excluding these features performed substantially worse. Third, the ablation experiment demonstrated that adding environmental variables improved the two-week forecast only modestly (ΔR^2 = +0.031), suggesting that these predictors contribute supplementary information rather than serving as primary drivers. Together, these findings provide convergent evidence that historical case dynamics remain the strongest determinants of forecast accuracy, while environmental variables enhance performance incrementally under the proposed geoinformatics framework.

Regional evidence further supports our interpretation of the modest but meaningful contribution of environmental variables. In Central Vietnam [25], reported strong lagged effects of NDVI and temperature on malaria incidence, while in Lao PDR [26], found rainfall and humidity to be dominant predictors with spatially heterogeneous impacts across provinces. Compared with these settings, Thailand's elimination-phase transmission exhibits lower case counts and stronger short-term autocorrelation, which helps explain why environmental variables in our models provided incremental rather than dominant predictive value. Nonetheless, the consistent improvement observed in the ablation analysis indicates that environmental anomalies still supply useful supplementary information, particularly when climatic deviations may trigger localized increases in transmission risk.

3.6 Limitations and Future Directions

While this study emphasizes the importance of understanding local transmission dynamics, the modeling framework relied on provincially aggregated malaria data. Combining weekly cases from the nine districts in Tak Province introduces both strengths and limitations. District-level variability and localized hotspots may be obscured through aggregation, potentially masking spatial heterogeneity relevant for micro-targeted interventions. Conversely, aggregation enhances signal stability by reducing stochastic noise inherent in low-incidence, elimination-phase settings and yields a clearer temporal pattern for forecasting. This provincial-scale perspective aligns with Thailand's operational decision-making structure, where early warning systems and resource allocation are typically coordinated at the provincial level. Nonetheless, extending the framework to district-level or multi-scale modeling represents an important direction for future work as more granular surveillance data become available.

Several additional limitations warrant consideration. The dataset contains 689 weekly observations, which constrains the capacity of complex models to learn higher-order temporal relationships and increases the risk of overfitting-particularly for nonlinear or high-parameter algorithms. Deep learning architectures (e.g., LSTM, RNN) were not evaluated for this reason; their effectiveness generally depends on long temporal sequences, large training corpora, and substantial computational resources, and they offer limited interpretability for routine public health operations. Future research may revisit deep learning approaches once expanded datasets-incorporating longer historical records, additional spatial units, or cross-border epidemiological information-become available.

4. Conclusions

This study demonstrates a comprehensive geoinformatics framework for malaria forecasting that integrates satellite remote sensing and machine learning technologies. The systematic evaluation of six algorithms across 1-4 week forecasting horizons revealed algorithm-specific optimal capabilities, with SVM achieving superior short-term predictions (R^2 = 0.744 and 0.687 for 1-2 week horizons) and KNN excelling at extended forecasts (R^2 = 0.748 and 0.731 for 3-4 week horizons). Statistical significance testing confirmed these performance differences represent genuine algorithmic advantages rather than random variation, with bootstrap confidence intervals providing robust uncertainty quantification for operational deployment. The horizon-specific optimal performance advances understanding of how different computational approaches capture distinct temporal patterns in epidemiological time series. SVM's linear margin optimization effectively captures immediate transmission dynamics, while KNN's local averaging approach better identifies seasonal and environmental patterns that become more predictable at extended timescales. The cloud-based geospatial data processing pipeline through GEE provides automated access to satellite archives, overcoming infrastructure limitations of ground-based meteorological networks in border provinces. The framework requires minimal computational resources, supporting implementation in typical public health computing environments while providing actionable lead times (1-4 weeks) for outbreak detection and

resource allocation. Implementation within Thailand's malaria elimination program involves integration with existing surveillance infrastructure, providing automated weekly forecasts to provincial health offices. The 1-2 week SVM forecasts enable rapid response deployment, while 3-4 week KNN forecasts support strategic planning and cross-border coordination. The satellite-based pipeline addresses surveillance gaps in remote border areas, with computational efficiency supporting deployment on standard provincial infrastructure. Scalability across other border provinces requires minimal development, establishing a reusable template complementing Thailand's elimination strategy. The demonstrated forecasting accuracy ($R^2 > 0.68$ across all optimal horizons) supports transitions from reactive to predictive surveillance approaches. The statistical rigor provided through confidence intervals and significance testing ensures reliable performance assessment essential for operational public health decision-making. The geoinformatics framework has broad applicability beyond malaria, providing a template for integrating satellite environmental data with epidemiological surveillance systems for other vector-borne diseases across diverse endemic contexts. Future technical development priorities include implementation of ensemble forecasting approaches combining the horizon-specific optimal algorithms, integration of higher-resolution satellite imagery, and establishment of automated model retraining protocols for real-time operational deployment. The validated uncertainty quantification framework provides a foundation for incorporating forecast confidence into intervention threshold setting and resource allocation decisions.

5. References

[1] World Health Organization. World malaria report 2024 [Internet]. Geneva: World Health Organization; 2024 [cited 2025 Aug 29]. Available from: https://www.who. int/teams /global- malaria- programme/ reports/world-malaria-report-2024 [Accessed 29th August 2025]

[2] Centers for Disease Control and Prevention. Thailand [Internet]. In: *CDC Yellow Book*. Atlanta (GA): CDC; [cited 2025 Aug 29]. Available from: https://www.cdc.gov/yellow-book/hcp/asia/thailand.html [Accessed 29th August 2025]

[3] Bisanzio D, Sudathip P, Kitchakarn S, Kanjanasuwan J, Gopinath D, Pinyajeerapat N, et al. Malaria stratification mapping in Thailand to support prevention of reestablishment. *Am J Trop Med Hyg*. 2023;110(1):79-82.

[4] Chang HH, Chang MC, Kiang M, Mahmud AS, Ekapirat N, Engø-Monsen K, et al. Low parasite connectivity among three malaria hotspots in Thailand. Sci Rep. 2021;11(1):23348.

[5] Rotejanaprasert C, Lawpoolsri S, Sa-angchai P, Khamsiriwatchara A, Padungtod C, Tipmontree R, et al. Projecting malaria elimination in Thailand using Bayesian hierarchical spatiotemporal models. Sci Rep. 2023;13(1):7799.

[6] Mategula D, Gichuki J, Barnes KI, Giorgi E, Terlouw DJ. Advancing early warning systems for malaria: progress, challenges, and future directions-a scoping review. *PLOS* Glob Public Health. 2025;5(5):e0003751.

[7] NOAA Science Advisory Board. Subseasonal and seasonal (S2S) forecasting innovation: plans for the twenty-first century [report]. Washington (DC): National Oceanic and Atmospheric Administration; 2022.

[8] Gunda R, Chimbari MJ, Shamu S, Sartorius B, Mukaratirwa S. Malaria incidence trends and their association with climatic variables in rural Gwanda, Zimbabwe, 2005-2015. *Malar J*. 2017;16(1):393.

[9] Kotepui M, Kotepui KU. Impact of weekly climatic variables on weekly malaria incidence throughout Thailand: a country-based six-year retrospective study. *J Environ Public Health*. 2018;2018:8397815.

[10] Khan O, Ajadi JO, Hossain MP. Predicting malaria outbreak in The Gambia using machine learning techniques. PLoS One. 2024;19(5):e0299386.

[11] Ozsahin DU, Duwa BB, Ozsahin I, Uzun B. Quantitative forecasting of malaria parasite using machine learning models: MLR, ANN, ANFIS and random forest. Diagnostics. 2024;14(4):385.

[12] Harvey D, Valkenburg W, Amara A. Predicting malaria epidemics in Burkina Faso with machine learning. PLoS One. 2021;16(6):e0252689.

[13] Roll Back Malaria Partnership. Malaria early warning systems: concepts, indicators and partners. Geneva: World Health Organization; 2001.

[14] Thomson MC, Ukawuba I, Hershey CL, Bennett A, Ceccato P, Lyon B, et al. Using rainfall and temperature data in the evaluation of national malaria control programs in Africa. *Am J Trop Med Hyg*. 2017;97(3 Suppl):32-45.

[15] Ministry of Public Health. Thailand malaria elimination [Internet]. Bangkok: Ministry of Public Health; [cited 2013 Jun 1]. Available from:https:// malaria. ddc. moph. go. th/ [Accessed 1st June 2013]

[16] Funk C, Peterson P, Landsfeld M, Pedreros D, Verdin J, Shukla S, et al. The climate hazards infrared precipitation with stations-a new environmental record for monitoring extremes. Sci Data. 2015;2:150066.

[17] Muñoz-Sabater J, Dutra E, Agustí-Panareda A, Albergel C, Arduini G, Balsamo G, et al. ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. Earth Syst Sci Data. 2021;13(9):4349-4383.

[18] Mordecai EA, Paaijmans KP, Johnson LR, Balzer C, Ben-Horin T, de Moor E, et al. Optimal temperature for malaria transmission is dramatically lower than previously predicted. Ecol Lett. 2013;16(1):22-30.

[19] Asare EO, Tompkins AM, Bomblies A. A regional model for malaria vector developmental habitats evaluated using explicit pond-resolving surface hydrology simulations. PLoS One. 2016;11 (3): e0150626.

[20] Gupta A, Stead TS, Ganti L. Determining a meaningful R-squared value in clinical medicine. Acad Med Surg. 2024:1-6.

[21] Parker DM, Matthews SA, Yan G, Zhou G, Lee MC, Sirichaisinthop J, et al. Microgeography and molecular epidemiology of malaria at the Thailand-Myanmar border in the malaria pre-elimination phase. *Malar J*. 2015;14:198.

[22] Shapiro LL, Whitehead SA, Thomas MB. Quantifying the effects of temperature on mosquito and parasite traits that determine the transmission potential of human malaria. PLoS Biol. 2017;15(10):e2003489.

[23] Tompkins AM, Ermert V. A regional-scale high-resolution dynamical malaria model that accounts for population density, climate and surface hydrology. *Malar J*. 2013;12:65.

[24] im Y, Ratnam JV, Doi T, Morioka Y, Behera S, Tsuzuki A, et al. Malaria predictions based on seasonal climate forecasts in South Africa: a time series distributed lag nonlinear model. Sci Rep. 2019;9:17882.

[25] Tam LT, Thinkhamrop K, Suttiprapa S, Clements ACA, Wangdi K, Suwannatrai AT. Bayesian spatiotemporal modelling of environmental, climatic and socio-economic influences on malaria in Central Vietnam. *Malar J*. 2024;23(1):258.

[26] Rotejanaprasert C, Malaphone V, Mayxay M, Chindavongsa K, Banouvong V, Khamlome B, et al. Spatiotemporal patterns and association with climate for malaria elimination in Lao PDR: a hierarchical modelling analysis with two-step Bayesian model selection. *Malar J*. 2024;23(1):231.