Pertanika J. Sci. & Technol. 34 (2): 1025 - 1052 (2026)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

# Integrating Time Lag Effects in Predictive Modelling Using Random Forest for Early Detection of Bagworm Outbreaks

**Yi Peng Wang[1,5,6], Nurul Hawani Idris[1,2]*, Norhayu Asib[3], Mohamad Hafis Izran Ishak[4], and Alvin Lau Meng Shin[1,2]**

[1]*TropicalMap Research Group, Faculty of Built Environment and Surveying, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia*
[2]*Geoinformation Department, Faculty of Built Environment and Surveying, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia*
[3]*Faculty of Agriculture, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia*
[4]*Control and Mechatronics Department, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia*
[5]*Department of Hydraulic Engineering, Hebei University of Water Resources and Electric Engineering, 061001 Cangzhou, Hebei, China*
[6]*Hebei Technology Innovation Centre for Coastal Wetland Water Resources Allocation and Ecological Protection, 061001 Cangzhou, Hebei, China*

## ABSTRACT

The palm oil bagworm inflicts substantial economic losses on the palm oil industry in Malaysia annually. At present, integrated pest management remains the primary method for managing pests and diseases. However, early prediction of pest occurrence can help to assist in managing sustainable palm oils more effectively. Understanding the time window as well as the lag associated with environmental changes at the plot level is fundamental to such pest management strategies. However, few studies consider historical environmental data and recognise the biological and ecological time lags in pest population responses in their predictive models. This study aims to develop a predictive model based on spatiotemporal environmental data to support early warning systems for managing palm oil bagworm infestations. Therefore, this study examines the best time lags from one to six cycles before the census lagged variables, ranging from period 1 to 6 and trains the model using a random forest algorithm under 12 different time window configurations. Key driving factors influencing bagworm occurrences, including rainfall, land surface temperature, humidity, and road density, are investigated. Experimental results indicate that the combination of the first three cycles

ARTICLE INFO

*Article history:*
Received: 01 August 2025
Accepted: 16 March 2026
Published: 30 April 2026

DOI: https://doi.org/10.47836/pjst.34.2.18

*E-mail addresses*:
yipeng1992@graduate.utm.my (Yi Peng Wang)
hawani@utm.my (Nurual Hawani Idris)
norhayuasib@upm.edu.my (Norhayu Asib)
hafis@utm.my (Mohamad Hafis Izran)
alvinlau@utm.my (Alvin Lau)
* Corresponding author

e-ISSN: 2231-8526

(Lag 1, 2, and 3) of temporal data before the census date yields the best model performance with a recall rate of 0.76 and an AUC value of 0.68. Among these, relative humidity (RH_lag2; two cycles before the census) and surface temperature (LST_lag1; one cycle before the census) emerges as the most influential predictors of bagworm outbreaks. Finally, the visualisation of model predictions enables plot-based targeted preventive measures in high-risk areas, thereby supporting the Sustainable Development Goal (SDG) in palm oil management.

*Keywords:* Bagworm, Geographic Information System (GIS), prediction model, random forest, spatio-temporal, time lag

## INTRODUCTION

Palm oil is currently the most consumed and traded vegetable oil in the world. The global palm oil market in 2019 is estimated at 74.6 million tonnes. This figure is expected to further increase to 111.3 million tonnes by 2025 (Sundaraja et al., 2021). Malaysia is the world's second-largest producer of palm oil. In 2020, the country's palm oil production and exports accounted for 25.8% and 34.3% of the world's palm oil production and exports, respectively (Zakaria et al., 2024). Malaysia's palm oil industry has a long history of bagworm infestations, and the substantial yield reduction due to bagworm attacks is becoming a main threat in the palm oil industry (Sulaiman & Talip, 2021; Thaer et al., 2021).

The bagworm (Lepidoptera: Psychidae) is a major foliar pest capable of causing severe leaf damage and yield losses of up to 43% if left untreated (Sulaiman & Talip, 2021), as reported by the Malaysian Palm Oil Board (MPOB) in 2016. The rapid expansion and widespread monoculture of palm oil plantations provide abundant food resources, creating ideal conditions for the pest’s growth and spread. Several factors influence bagworm population dynamics, including environmental conditions (both biotic and abiotic), the balance of predator-prey interactions, the availability of alternate host plants, and human activities. When environmental conditions are favourable, these factors often trigger localised infestations at high densities, which can escalate into widespread outbreaks with serious consequences for crop productivity.

Precision agriculture, utilising technologies such as remote sensing, IoT, and LiDAR visualised through Geographic Information Systems (GIS), has become increasingly common globally (Sabtu et al., 2018). The integration of these digital agriculture tools with Artificial Intelligence (AI) enhances a country's ability to modernise farming, significantly accelerating agricultural processes, improving productivity, and streamlining the food supply chain. Furthermore, digital agriculture actively supports efforts toward achieving carbon neutrality by optimising resource use, reducing greenhouse gas emissions and chemical pesticide use, and improving agricultural resilience to climate change (Balasundram et al., 2023).

Bagworms (Lepidoptera: Psychidae) are primary leaf-eating pests of palm oil in Malaysia, which can cause about 33%- 40% yield losses (Sulaiman & Talip, 2021; Thaer et al., 2021). They feed on leaves at a rapid pace. There are some kinds of species of bagworms like *Metisa plana*, *Pteroma pendula* and *Mahasena corbetti* (Hoong & Hoh, 1992; Lever, 2008; Wahid et al., 1988). Studies have shown that *Metisa plana* is the most widely distributed palm oil insect in Peninsular Malaysia, followed by *Pteroma pendula*. The most severely affected areas are in Perak, Sabah and Johor (Kamarudin & Mohd, 2007; Napiah et al., 2023).

Chemical control is still the major control mechanism in managing bagworm outbreaks in most plantations, as compared to smallholding plantations (Kamarudin et al., 2017; Salim & Hamid, 2012). However, over-reliance and excessive use of chemical insecticides to control bagworm has often led to the development of other, more persistent problems. For example, the resistance of pests to treatment, the abundance of harmful chemical residues in the environment, and the disruption of beneficial insect populations (Kamarudin et al., 2017) are not conducive to sustainable development.

One way to improve the efficiency of pesticide use and reduce the use of insecticides is to make predictions in advance about where and when those pests and diseases are likely to occur, which requires analysing and modelling pest and disease outbreak patterns from both spatial and temporal perspectives, and accordingly using pesticides within a reasonable spatiotemporal range. Studying the problem from only a temporal or only a spatial perspective is not feasible. The primary rationale is that the essence of pest and disease outbreak dynamics is spatio-temporal in nature (Li et al., 2021), and considering time and space separately during modelling fails to capture the key ecological mechanisms involved: temporal-only modelling overlooks spatial relationships between outbreak locations (e.g., areas surrounding an outbreak site are more likely to be affected due to dispersal), while spatial-only modelling is likely to ignore the temporal cumulative effects inherent in outbreak mechanisms (Guo et al., 2020). From a practical application perspective, spatio-temporal models can simultaneously answer the question of "when" and "where" risks are high.

The occurrence of observed diseases or pests is not directly influenced by current meteorological conditions, but rather by those of a previous period, indicating a time lag between the emergence of diseases or pests and environmental fluctuations (Holloway et al., 2018). At the same time, the occurrence of diseases and pests is the result of the cumulative influence of environmental variables over a period of time windows (Hamer et al., 2020; Pol et al., 2016). This characteristic of disease and pest research requires that the impact of time lag and time windows be considered when modelling. The time window refers to the period of past days or weeks to understand the environment before a pest outbreak occurs. Whereas the time lag indicates how long it takes for environmental changes to show an effect on pests.

Although abundant environmental data have been utilised in establishing the relationship between ecological characteristic variables and pests and diseases, most studies have not explicitly considered the influence of time lag and cumulative environmental variables. In their research, immediate or period-averaged environmental variables are usually used to explain the spatial distribution or risk prediction of pests and diseases (Gachoki et al., 2024; Lee et al., 2023; Mahmoodi et al., 2024; Makori et al., 2024; Rouabah et al., 2022), without incorporating the lagged impact of historical variables on the dynamics of pests and diseases into the model.

As the biological characteristics (growth cycles and patterns) of different pests vary, the optimal time lag may also vary for different pests, and the time windows over which to gather environmental variables in a prediction model are likely to be highly species-specific (Holloway et al., 2018; Pol et al., 2016). In terms of the palm oil pest *Metisa plana*, there is currently a lack of systematic research determining the optimal time lag and time windows of environmental variables to help figure out its growth and development.

In fields of crop disease and pest prediction, the decisions made by prediction models can affect the efficiency of field operations. At such times, a good model can ensure the establishment of trust, meet regulatory requirements, and reduce decision-making errors (Dwivedi et al., 2023; Rojat et al., 2021). Currently, the interpretability of most model results related to agricultural pests and diseases is not detailed enough. Most model results only demonstrate the significance of environmental variables and whether their impacts are positive or negative (Hamer et al., 2020; Mahmoodi et al., 2024). However, there is a lack of models that consider when and where to act to prevent an outbreak before it occurs. Therefore, this study develops a prediction model for bagworm occurrence using Random Forest, a common machine learning algorithm (Guo et al., 2020; Rouabah et al., 2022), which considers the time lag before the effect and time windows of spatio-temporal environmental data. This study includes 3 objectives: (i) to develop an interpretable prediction model, (ii) to quantify time lag effects of environmental variables, and (iii) to identify key drivers. The time lag effects were examined by dividing environmental conditions into 12-time window schemes. The final model was quantitatively interpreted, and the key drivers and the most informative time lags for bagworm occurrence were identified through Shapley Additive exPlanations (SHAP) analysis (Lundberg et al., 2018).

## MATERIALS AND METHODS

This section describes the methodology applied in this study. The following Figure 1 is the research framework that starts with problem identification. The modelling stage is for research objectives (i) and (ii), the interpretability stage and the performance stage are for objective (iii).

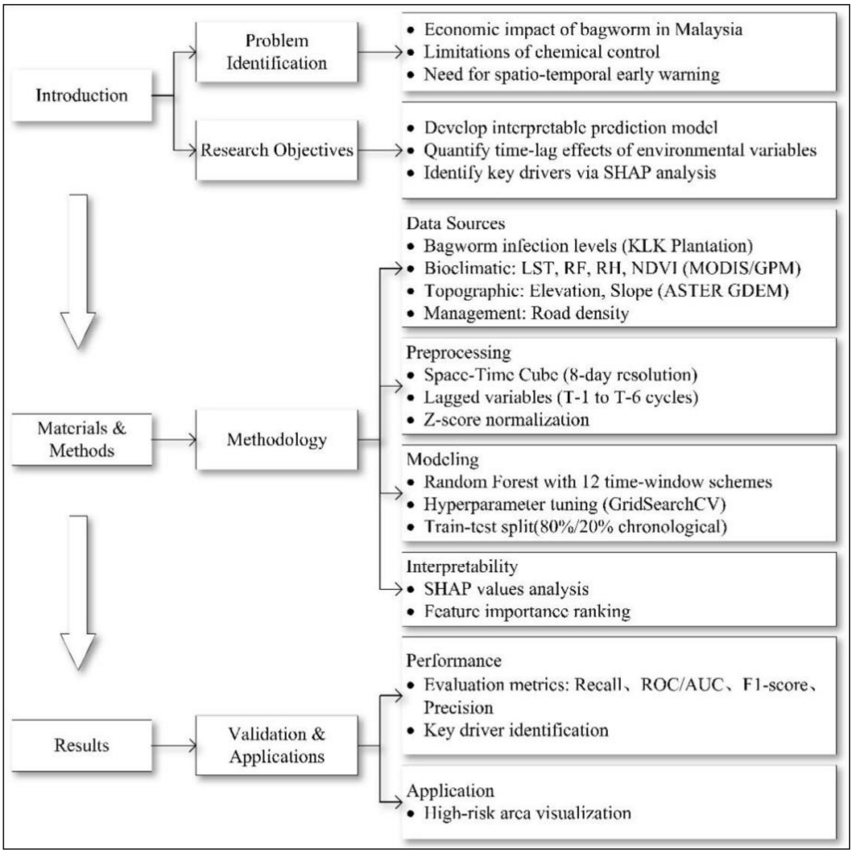



*Figure 1.* Research framework

## Study Area and Dataset

The research area is in Lekir, situated in the Perak State, with geographic coordinates ranging from 4° 4' 48" N to 4° 10' 48" N and from 100° 45' 36" E to 100° 50' 24" E. The site covers an area of approximately 33.29 km², as shown in Figure 2.

The study area was divided into different plots and numbered. The plantation area is well-structured, divided into 42 plots, each assigned a unique identification number. The northern part of the palm plantation has higher terrain compared to the southern part. The elevation of this study area ranges between 4m and 37m above sea level. The main palm oil pest in the study area is *Metisa plana*. The statistical data indicating the severity of the pest and disease was obtained from the KLK palm oil plantation. The original census location data was collected using a handheld GPS. This data marked the geographical coordinates (latitude and longitude) of each census tree and the infection status of *Metisa plana* at specific dates. The overall period of the data is from April to December 2020, January to July 2021, and the entire year of 2022.

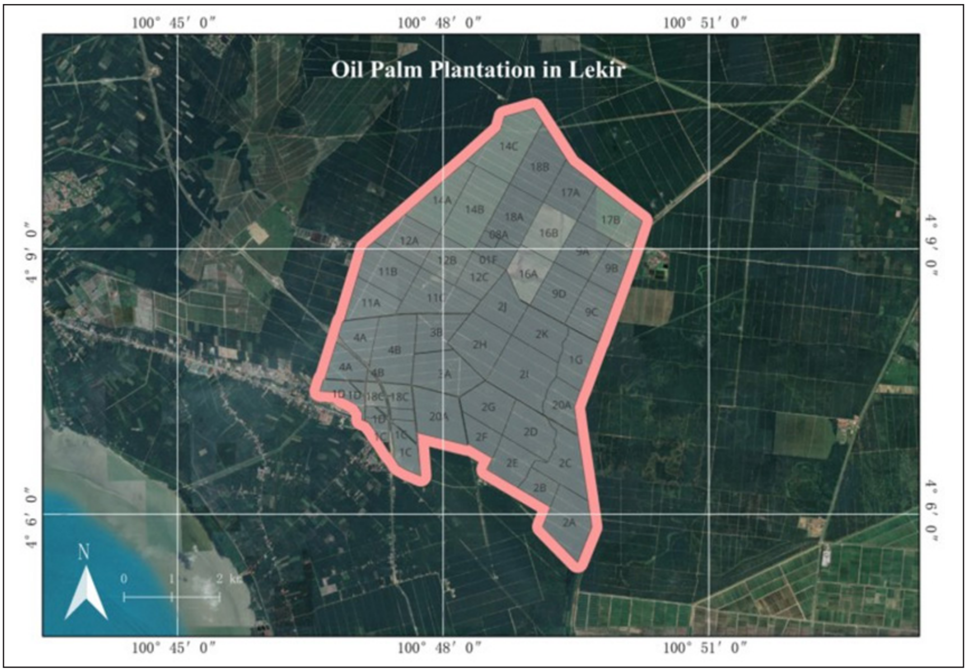



*Figure 2.* Location of the study site in Lekir, Perak

The census data were collected by manually counting the number of pests and assessing the severity of infestation in the plot. Some of the data collection cycles did not fully cover all plots in the study area. The infection levels were determined by the above regulation in Table 1.

The bioclimatic data used in this study include Land Surface Temperature (LST) from the MOD11A2 (Wan et al., 2021), Normalised Difference Vegetation Index (NDVI) from MOD13Q1 (Didan, 2021), and Relative Humidity (RH) derived from MOD05/06/07 (MODIS, A. S. T., 2017a, 2017b, 2017c), all of which were obtained from NASA's Earth science data at https://search.earthdata.nasa.gov/search. Rainfall (RF) data were obtained from the NASA Global Precipitation Measurement (GPM) dataset (Huffman, 2023), available at GES DISC Earth data (https://disc.gsfc.nasa.gov). The datasets covered the period from 2020 to 2022, aligning with the bagworm field census data. In addition, the predictor variables included spatial data such as road density, elevation, and slope. These topographic variables represent static features and were obtained from the official KLK plantation records and the ASTER GDEM V003 (Advanced Spaceborne Thermal Emission and Reflection Radiometer Global Digital Elevation Model, Version 003) dataset (NASA/METI/AIST/Japan, 2019), available at https://search.earthdata.nasa.gov/search/. Detailed information is provided in Table 2.

Table 1
*Infection level regulation*

| Explanation | Symbol | Indicator |
|---|---|---|
| Larvae (<1cm, 0 pcs) | Green flag | OK |
| Larvae (<1cm, 1-4 pcs) | Yellow flag | Monitor carefully |
| Larvae (<1cm, >5 pcs) | Red flag | Immediate treatment |
| Larvae (>1cm, >5 pcs) | Red block | Monitor carefully |
| Larvae (>1cm, <5 pcs) | Green block | No need to treat |

Table 2
*The data used in this study*

| Data layers | Data sources | Year |
|---|---|---|
| Bagworm Infectious Level | Official data-KLK Plantation | April 2020 to Dec 2020.<br>Jan 2021 to July 2021; 2022 |
| LST / (°C) | MODIS (MOD11A2) | 2020/2021/2022 |
| RF (mm/8days) | GPM | 2020/2021/2022 |
| RH | MODIS (MOD05/06/07) | 2020/2021/2022 |
| NDVI | MODIS (MOD13A1) | 2020/2021/2022 |
| Road density (km/km²) | Official data-KLK Plantation | 2022 |
| Elevation / m | ASTER GDEM V3 | 2022 |
| Slope / （ ） | ASTER GDEM V3 | 2022 |

## Data Preprocessing

The bagworm infestation data were applied to a Space Time Cube (STC) model to statistically analyse the spatio-temporal pattern of each plot. The STC was constructed using the ArcGIS Pro tool. The constructed STC is based on an 8-day unit to align with the satellite-based data cycle of bioclimatic data, with the data representing *Metisa plana* in each plot. Therefore, the severity of infected plots, from April 2020 to July 2021 and the entire year of 2022, within 8 days, can be assessed. The final dataset comprises 5,208 records of pest infection at 124 time points across 42 plots, with 842 records containing valid data and the remaining 4,366 records being missing or invalid. The main causes of this data missingness are that the bagworm census data were collected manually (Tailliez & Ballo Koffi, 1992). Ideally, bagworm counts should be collected and recorded regularly in both time and space. However, as the spatial and temporal coverage of the survey expands, the workload and difficulty of data collection increase exponentially, making it inevitable that some data will be missing. As this research concentrates on the relationship between the presence of bagworms and habitat factors, rather than on the relationship between the absence of bagworm data and habitat factors, here we exclude the missing parts.

The spatio-temporal distribution of bagworm infestation severity is presented in the form of a heat map as shown in Figure 3. The vertical axis on the left represents different plot numbers, the horizontal axis at the bottom represents time, and the colour band on the right, ranging from yellow to green, indicates the increasing severity of bagworm infestation.

In this study, the dependent variable, "bagworm infectious level" data, was standardised and converted into binary categories, represented by 0 and 1 to indicate the absence or presence of bagworms. Another reason for this approach is that data collection on bagworm infection levels in the lekir study area has not been uniformly conducted over time and space. Consequently, the spatio-temporal cube (data points aggregated within a time step interval) cannot distinguish the degree of infestation but can only distinguish the presence or absence of bagworm infestation.

Besides, the bioclimatic independent variables (i.e. bioclimatic data) also need to be pre-processed and normalised, so that they can keep and transform their temporal resolution into the exact resolution, then the model can more easily interpret the data and identify potential features within it.

In terms of spatial scale, all bioclimatic data (raster format) were resampled using bilinear interpolation to increase their spatial resolution, ensuring that it is finer than the management units of the study area. Zonal statistical analysis was then applied to ensure spatial alignment for subsequent analyses. Thus, the responses of environmental variables identified in this model should be interpreted as indicative of plot-level averages pertinent to plantation management.

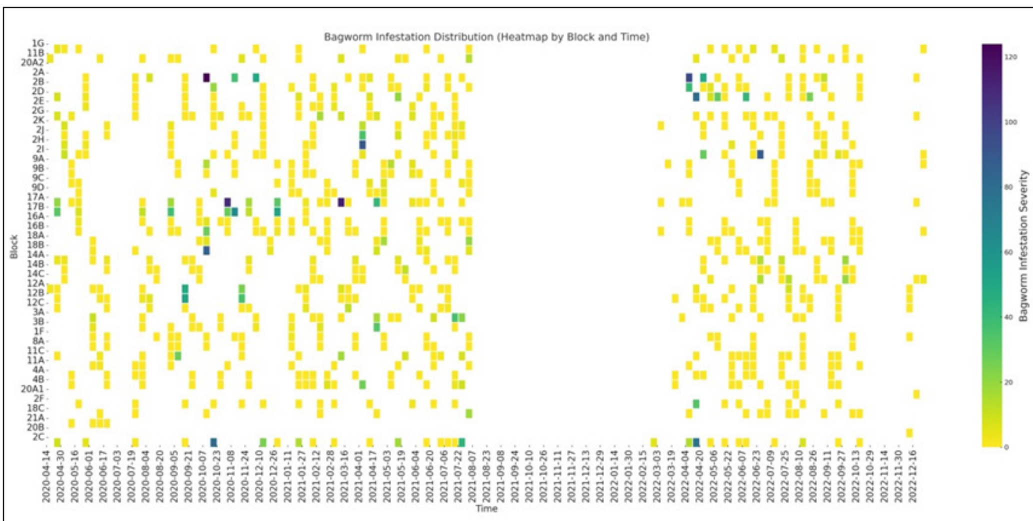
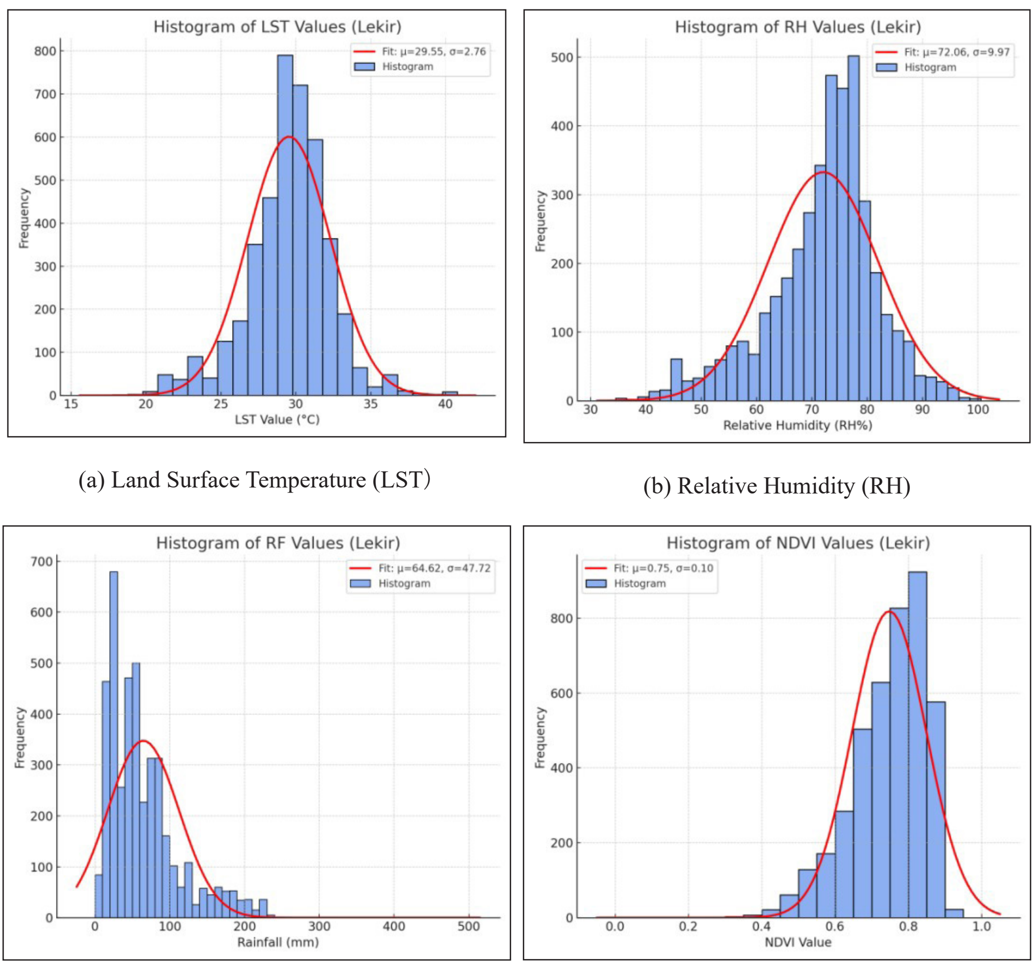



*Figure 3.* Heat map of bagworm infestation severity

After outliers are removed, the weather data are summarised per plot per 8-day intervals, and the frequency of the data range for each plot every 8 days is calculated. This yields the frequency distribution histogram (Figure 4). In addition, the data for each plot every 8 days are summarised in terms of mean, standard deviation, maximum value, and minimum value, resulting in the weather data statistics for the Lekir study area (Table 3).

In the pest and disease spatial-temporal prediction work, these two contexts - the time lag and time windows- should also be considered, as the effects of certain environmental factors on the bagworm at the current moment do not immediately manifest but occur after a period.




(a) Land Surface Temperature (LST)

(b) Relative Humidity (RH)

(c) Rainfall (RF)

(d) Normalised Difference Vegetation Index (NDVI)

*Figure 4.* Frequency distribution histograms of (a) LST; (b) RH; (c) RF; and (d) NDVI in Lekir

Table 3
*The statistics of Lekir weather data*

| Weather data | Average | Standard deviation | Max | Min |
|---|---|---|---|---|
| LST | 29.55°C | 2.76°C | 40.73°C | 16.78°C |
| NDVI | 0.75 | 0.10 | 0.93 | 0.22 |
| RH | 72.06% | 9.97% | 99.82% | 34.52% |
| RF | 64.62 mm | 47.72 mm | 232.91 mm | 3.08 mm |

*Note.* LST = land surface temperature; NDVI = normalised difference vegetation index; RH = relative humidity; RF = rainfall

In this study, a time lag (Lag T) feature represents the value of an environmental variable at a specific time period in the past (Ravljen et al., 2018), relative to the observation period of the bagworm census. This present study defined the time lag as Equation 1:

$$X_{lag,k}(T) = X(T - k * \Delta t) \quad [1]$$

Where X is an environmental variable (e.g., RH, LST), T is the time of the bagworm census, k is the lag number (1,2,3…), and $\Delta t$ is the time step (in our case, 8 days interval). While the time window (N cycle window) refers to including multiple consecutive lagged features to form a time window, thereby capturing recent historical conditions (Ravljen et al., 2018). In this study, it is defined as the set of features with Equation 2:

$$Window_N(T) = \{X(T - \Delta t), X(T - 2\Delta t), \cdots, X(T - N\Delta t)\} \quad [2]$$

Where N is the window size. In this research, a "3-cycle window" for variable X means including 3 individual lagged environmental variables (X_Lag1, X_Lag2, X_Lag3) as stacked, independent predictors in the model.

With an 8-day interval, here we pair the feature values of LST/NDVI/RH/RF lagged by 1 to 6 cycles with bagworm statistics based on the time relationship. After data processing, the data used for modelling are presented in the form of a table, including the bagworm data and other bioclimatic data components, all of which contain variable values with 1 to 6 cycle lags. The bioclimatic variables are shown in Table 4. After removing the data of invalid bagworm statistical periods, we finally get a total of 784 samples.

Table 4
*Independent variables (bioclimatic data) in the experiment*

| **Independent variable** | **Current** [a] | **Lag T** [b] | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| LST | √ | √ | √ | √ | √ | √ | √ |
| RF | √ | √ | √ | √ | √ | √ | √ |
| RH | √ | √ | √ | √ | √ | √ | √ |
| NDVI | √ | √ | √ | √ | √ | √ | √ |
| Road density | √ | - | - | - | - | - | - |
| Elevation | √ | - | - | - | - | - | - |
| Slope | √ | - | - | - | - | - | - |

*Note.* [a] Current = the period of the bagworm census [b]Lag T = the period that T (=1,2,3,4,5,6) cycle before the bagworm census period. For example, Lag1 indicates the period that is 1 cycle before the bagworm census period

## Model Construction

The model construction employed the Random Forest algorithm (Cutler et al., 2012), a supervised learning approach that integrates multiple decision trees to generate predictions based on the collective output of all individual models. This ensemble method utilises a resampling technique known as "bagging," in which each decision tree is independently constructed using bootstrap samples drawn randomly from both the dataset and the feature space (Cutler et al., 2012; Fox et al., 2017). The introduction of randomness significantly reduces the correlation among constituent trees, thereby minimising the risk of overfitting and enhancing generalisation performance (Guo et al., 2020). As the number of diverse and independent trees increases, the predictive accuracy of the ensemble improves accordingly. Furthermore, the Random Forest algorithm demonstrates robustness and effectiveness in scenarios involving high-dimensional data with relatively few informative features, owing to its reliance on stochastic sampling with replacement (Salman et al., 2024). Figure 5 shows the workflow of data processing and random forest model training.

In the process of model development, four critical hyperparameters were optimized to enhance predictive performance: (i) the maximum depth of each decision tree (max_depth), (ii) the minimum number of samples required to perform an internal node split (min_samples_split), (iii) the minimum number of samples required at a leaf node (min_samples_leaf), and (iv) the total number of trees in the ensemble (n_estimators) (A Ilemobayo et al., 2024). To systematically identify the optimal parameter configuration, grid search was implemented in conjunction with k-fold cross-validation, ensuring a balance between model complexity and generalisation capability.

To prevent the influence of feature variable values and scales on model building and subsequent analysis, when using the tree-based models, there is a need to normalise the

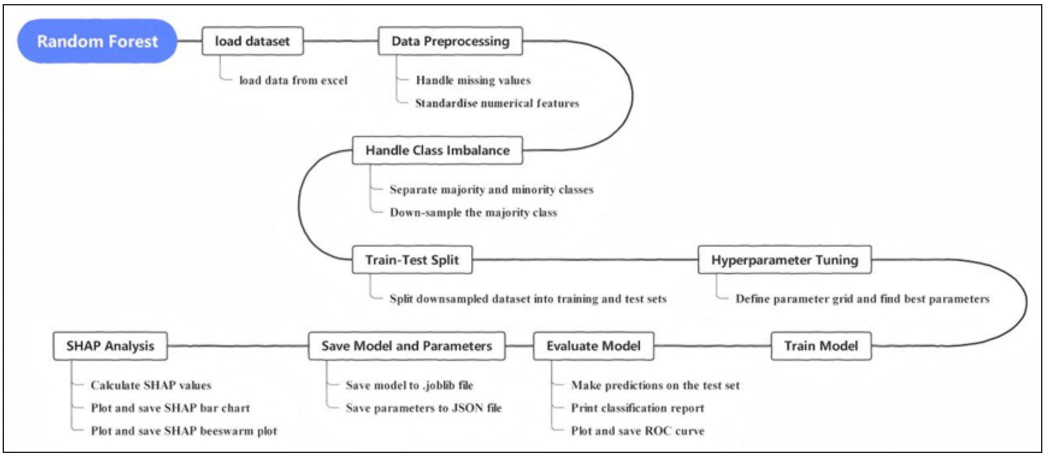



*Figure 5.* Workflow of the random forest modelling method

feature variables (Rajakumaran et al., 2024). Here we used z-score, which was implemented using the *sklearn.preprocessing.StandardScaler* function in Python. This method adjusts the distribution of each feature variable to a standard normal distribution with a mean of 0 and a standard deviation of 1.

At the same time, we matched the dependent variable and independent variables for different experimental schemes, as shown in the following Table 5. The data were divided into three different time window lengths, corresponding to 1 cycle, 2 cycles, and 3 cycles, to examine the temporal dynamics of environmental factors influencing bagworm infestation. Within each time window length, four different levels of time lag were incorporated to capture the delayed effects of environmental variables. The specific time window and time lag combination is shown in Figure 6.

Finally, the experimental design includes 12 experimental schemes. Considering that lagged environmental variables may exhibit correlations, we quantified these correlations to ensure the robustness of model evaluation. Specifically, for the environmental variables (including all lag periods) involved in the 12 experimental schemes, we calculated the Pearson correlation coefficient matrices and variance inflation factors (VIFs), as shown in Figure 7 and Figure 8, respectively. The results show that, although there is some correlation among different lag periods of the same variable, their VIF values all fall within acceptable thresholds (all less than 5), indicating a low risk of severe multicollinearity. Therefore, we ultimately retained all lagged variables to capture the complete temporal dynamics.

All variables were then matched with the bagworm census target variable based on the time and plot attributes to ensure accurate alignment between the feature variables and the target labels for model training and evaluation. Then the data were divided into training sets and test sets, with 80% and 20%, respectively (Raschka, 2018).

Table 5
*Experiment design schemes*

| Time window length [a] | Time lag [b] | Variable used |
|---|---|---|
| 1 cycle | 1 lag | Lag 1 |
| | 2 lag | Lag2 |
| | 3 lag | Lag3 |
| | 4 lag | Lag4 |
| 2 cycle | 1 lag | Lag 1-2 |
| | 2 lag | Lag 2-3 |
| | 3 lag | Lag 3-4 |
| | 4 lag | Lag 4-5 |
| 3 cycle | 1 lag | Lag1-3 |
| | 2 lag | Lag2-4 |
| | 3 lag | Lag3-5 |
| | 4 lag | Lag4-6 |

*Note.* [a] Time window length means the number of cycles of environmental conditions used in model building
[b] Time lag means the time delay between a statistic in environmental conditions and the resulting impact on pest populations

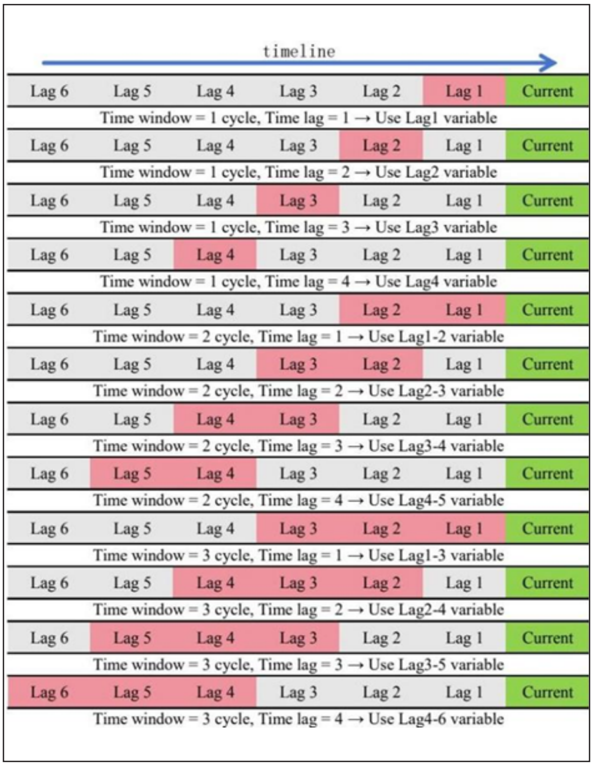



*Figure 6.* Time window & time lag for bagworm prediction

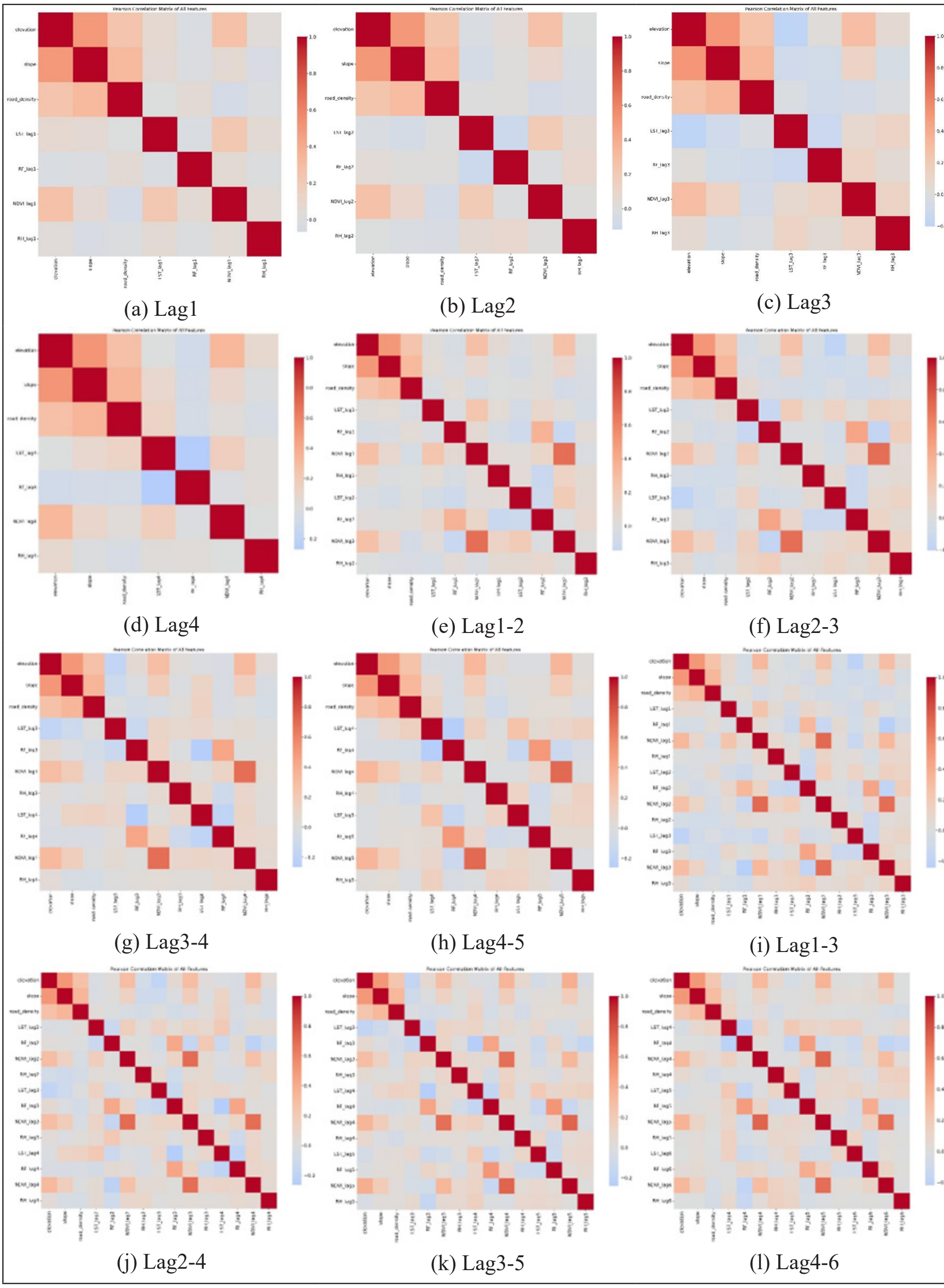



*Figure 7.* Pearson correlation coefficient matrices of 12 experiment schemes: (a) Lag1; (b) Lag2; (c) Lag3; (d) Lag4; (e) Lag1-2; (f) Lag2-3; (g) Lag3-4; (h) Lag4-5; (i) Lag1-3; (j) Lag2-4; (k) Lag3-5; (l) Lag4-6

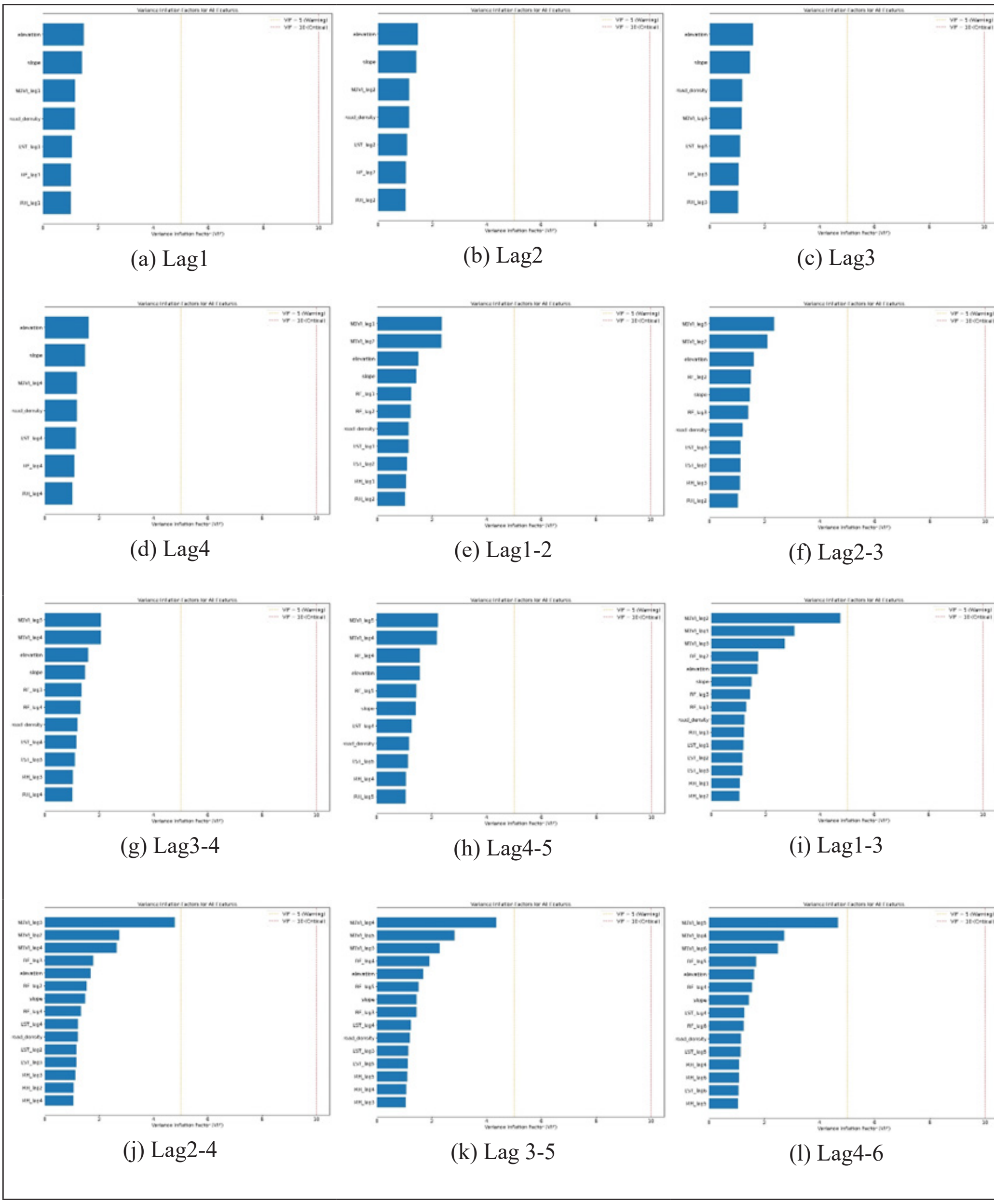



*Figure 8.* Variance inflation factors (VIFs) of 12 experiment schemes: (a) Lag1; (b) Lag2; (c) Lag3; (d) Lag4; (e) Lag1-2; (f) Lag2-3; (g) Lag3-4; (h) Lag4-5; (i) Lag1-3; (j) Lag2-4; (k) Lag3-5; (l) Lag4-6

Considering that the initial data has a time attribute, i.e., the data is time-sequenced, dividing the training set and test set according to the time sequence is a more reasonable choice. This can ensure that the model training and evaluation process conforms to the actual logic of time series analysis and avoids data leakage (Cerqueira et al., 2020).

We used 80% of the data with earlier time stamps as the training set and 20% of the data with later time stamps as the test set. Within the 80% training set, we adopted K-fold cross-validation (Wong & Yeh, 2020) to robustly tune the model parameters (K=5 in this study), and to get reliable estimates of the model performance. This method involved splitting the training set into K consecutive folds and iteratively training the model on K-1 folds, while using the remaining fold for validation, and finally averaging the results to guide model parameter selection. Then the remaining 20% data were used to test and make a prediction, thus obtaining the final evaluation. The final model represents the contribution of feature variables to the target variable by calculating SHAP values and thereby indicates the importance of features.

## Evaluation Metrics

In the model evaluation process, we adopt multiple methods, including precision, recall and F1-score, and also use the ROC-AUC value (Marković et al., 2021; Naidu et al., 2023).

Precision: Precision represents the number of samples with a true positive value among those predicted as positive. In this research, it refers to the number of samples that truly have bagworms among those predicted to have them. It's calculated with the Equation 3:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

Recall: Recall is employed to gauge the proportion of positive instances that the model successfully predicts as positive among all the positive samples. The higher the recall, the stronger the model's ability to recognise positive instances. It's calculated with Equation 4:

$$\text{Recall} = \frac{TP}{TP + F\text{N}} \tag{4}$$

In the field of agricultural pest infestation prediction, recall is particularly important. The primary goal is to accurately identify plots with pest infestations, minimising the risk of overlooking any actual occurrences of pests. Missing an infestation could lead to severe crop damage, economic losses, and the rapid spread of pests to surrounding areas.

F1-score: The F1-score is the harmonic mean of precision and recall, comprehensively considering the accuracy and identification ability of the model. The higher the F1-score, the better the comprehensive performance of the model. It's calculated with Equation 5:

$$\text{F1-score} = 2 \times \frac{\text{precision} \times \text{Recall}}{\text{precision} + \text{Recall}} \tag{5}$$

ROC-AUC: This metric stands for “Area Under the Receiver Operating Characteristic Curve”, which is a performance metric used to evaluate the quality of binary classification models(Fawcett, 2006; Hanley & McNeil, 1982). This quantitative indicator ranges from 0.5 to 1, summarising the overall ability of the classifier to rank positive examples ahead of negative ones across all possible decision thresholds. We compute it by using the standard implementation in *scikit-learn* (Pedregosa et al., 2011).

## RESULTS AND DISCUSSION

This study develop an interpretable prediction model and quantify the time lag effects of environmental variables, then identify the key drivers for bagworm occurrence. The following section provides an analysis of model performance and feature importance.

### Model Performance

Performance comparison across different time window configurations reveals distinct patterns in model effectiveness. As has been mentioned before, the time window refers to the length of segments of consecutive time steps used to structure environmental variables data for prediction, while time lag means the time between the response of pest emergence and environmental fluctuation. The experiment results are summarised in Table 6.

Among the 3-time window schemes, the Lag1-3 configuration demonstrates superior performance with a recall rate of 0.76 and an ROC-AUC value of 0.68, indicating its relatively high accuracy in predicting bagworm occurrences. For the 2-time window setting, the Lag1-2 configuration achieves the highest recall (0.71) and an ROC-AUC of 0.66. In contrast, under the single time window condition, the Lag2 configuration yields the best recall rate of 0.71 and an ROC-AUC of 0.74. Overall, the random forest model performs most effectively under the Lag1-3 configuration with three-time windows, followed by the Lag2 configuration under a single time window.

### Feature Importance Analysis

Given that the random forest model achieved the highest predictive performance under the Lag1-3-time condition, we selected this configuration for further feature importance analysis. SHAP values were calculated separately for each individual lagged feature variable within the Lag1-3 window, and corresponding visualisations were generated, including a standard bar chart (Figure 9) and a beeswarm plot (Figure 10). In Figure 9, each bar in the bar plot represents the mean absolute SHAP value for a specific feature, and in Figure 10, each point in the beeswarm plot represents a SHAP value for a specific feature for a single record in the test set.

Table 6
*Experiment result*

| Evaluation index | Precision | Recall | F1-score | ROC-AUC |
|---|---|---|---|---|
| Lag1-3 | 0.31 | 0.76 | 0.44 | 0.68 |
| Lag2-4 | 0.28 | 0.6 | 0.38 | 0.6 |
| Lag3-5 | 0.23 | 0.45 | 0.31 | 0.51 |
| Lag4-6 | 0.32 | 0.69 | 0.44 | 0.66 |
| Lag1-2 | 0.3 | 0.71 | 0.42 | 0.66 |
| Lag2-3 | 0.23 | 0.55 | 0.32 | 0.59 |
| Lag3-4 | 0.28 | 0.63 | 0.38 | 0.56 |
| Lag4-5 | 0.33 | 0.7 | 0.45 | 0.62 |
| Lag1 | 0.31 | 0.68 | 0.42 | 0.64 |
| Lag2 | 0.34 | 0.71 | 0.46 | 0.74 |
| Lag3 | 0.25 | 0.52 | 0.33 | 0.58 |
| Lag4 | 0.3 | 0.58 | 0.4 | 0.63 |
| Maximum | 0.34 | 0.76 | 0.46 | 0.74 |
| Average | 0.29 | 0.63 | 0.40 | 0.62 |

As illustrated in Figures 9 and 10, RH_lag2 and LST_lag1 emerged as the two most influential predictors of bagworm outbreaks, ranked first and second in terms of feature importance. The swarm plot reveals nonlinear relationships between habitat-related variables and outbreak likelihood. Specifically, RH_lag2 (relative humidity from two periods prior) exhibits a strong negative influence on bagworm occurrence. When RH_lag2 is below 60% (indicated by dense blue dots), the risk of an outbreak two periods later increases (positive SHAP values), which aligns with the biological observation that bagworm larvae hatch more readily under low-humidity conditions (Napiah et al., 2023). Conversely, when RH_lag2 exceeds 75%, SHAP values become negative, suggesting that high humidity may suppress adult activity. LST_lag1 (land surface temperature from one period prior) demonstrates optimal predictive power within the range of 28-32°C, where SHAP values peak, indicating favourable conditions for bagworm development. Outside this range (either too high or too low), the predictive strength diminishes, implying less favourable conditions for outbreaks one period later, which follows the conclusion of publication before (Enting & Latip, 2021). Road density ranks third in feature importance and shows notable influence. It is calculated as the total road length within a block divided by the block area and is used as a feature variable to represent the road dust level. The insecticidal effect of inert dusts like road dust and powdered clay has long been known; the mode of action is from its desiccant and adsorptive functions (Ebeling, 1971; Stejskal et al., 2021). However, according to Ho. (2002), the bagworm can survive on high fronds of tall palms.

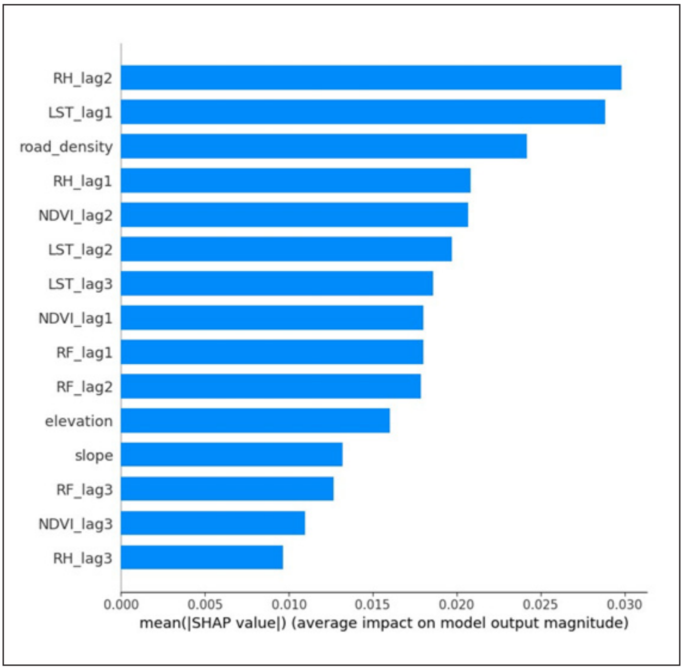



*Figure 9.* Bar plot of each variable SHAP values

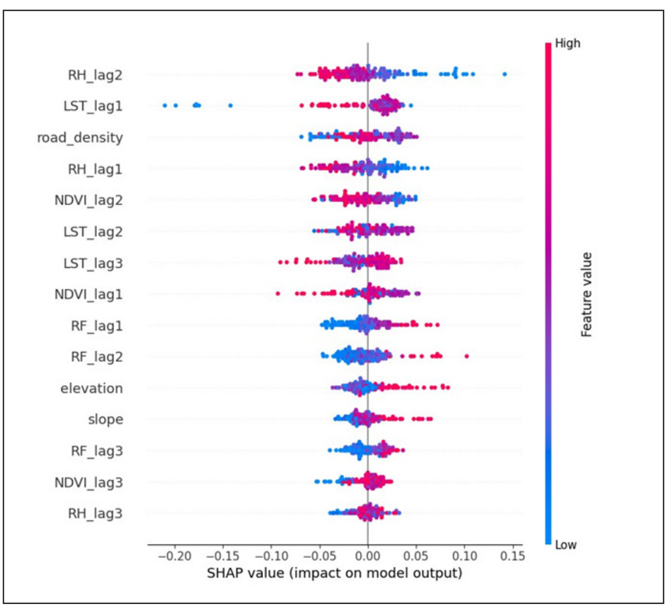



*Figure 10.* Beeswarm plot of each variable SHAP values (the colour ranging from blue to red corresponds to the magnitude of the independent feature value)

Its SHAP distribution lacks a consistent directional trend, which suggests that the relationship between road density and bagworm outbreaks varies spatially, potentially reflecting differences in land management intensity. RH_lag1 (relative humidity from one cycle prior) displays a nonlinear pattern like RH_lag2. RF_lag1, RF_lag2, and RF_lag3 exhibit relatively lower feature importance overall, indicating the impact of rainfall on the population of bagworm is very limited and is not the main driving factor (Napiah et al., 2023). They share a consistent positive influence mechanism whereby increased precipitation across lagged periods correlates with higher likelihoods of bagworm outbreaks 1-3 periods later.

## Temporal and Spatial Patterns

The data from the Lekir experimental zones were partitioned into two distinct subsets. To prevent data leakage, we implemented a chronological split method, wherein the initial 80% of the dataset was allocated for model training and validation, while the remaining 20% from a later time was reserved for prediction testing. Subsequently, we visualised the performance of the Random Forest model on the test set. The results are presented in Figure 11. In the visualisation, red-marked plots indicate areas predicted to experience bagworm outbreaks, green plots represent areas predicted to remain unaffected, and blank-coloured plots correspond to regions where no data were available in the test dataset for that specific period.

This study achieved an interpretable prediction of bagworm pest outbreaks in palm oil plantations by integrating multi-temporal remote sensing satellite data with the random forest model. Key findings revealed that lagged effects and nonlinear responses of environmental variables (such as relative humidity (RH), land surface temperature (LST), normalised difference vegetation index (NDVI), and rainfall (RF)) provided novel insights into the mechanisms underlying bagworm outbreaks. Compared to traditional models that rely solely on instantaneous meteorological data or averaged environmental variables over fixed periods (Gachoki et al., 2024; Lee et al., 2023; Mahmoodi et al., 2024) The incorporation of time-lagged environmental factors aligns more closely with the biological and temporal logic of pest development.

The SHAP analysis further uncovered significant nonlinear threshold effects of key predictors. Specifically, when RH_lag2 fell below 60%, the risk of outbreak increased markedly, which is consistent with the biological requirement of bagworm eggs for dry conditions during hatching (Napiah et al., 2023). Additionally, LST_lag1 exhibited an optimal temperature window of 28-32°C, within which the likelihood of outbreak peaked. Deviations from this range resulted in reduced SHAP values, indicating less favourable conditions for bagworm development, which corroborates laboratory observations of larval metabolic activity (Enting & Latip, 2021).

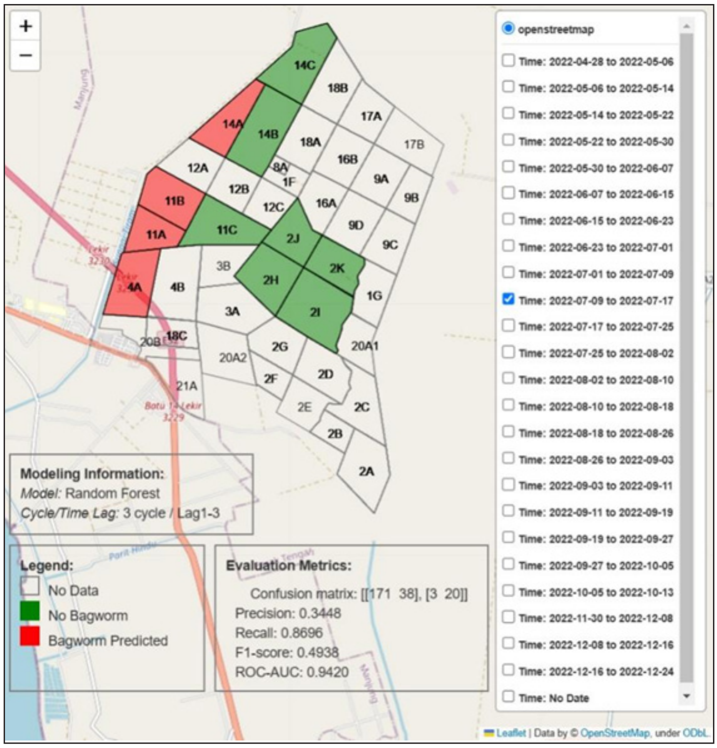



*Figure 11.* Visualisation of random forest prediction result on test data (randomly choose 2022-07-09 to 2022-07-17 as an example)

The application of SHAP analysis in this research played a crucial role in shifting from correlational insights to explanatory insights. Unlike traditional feature importance measures in Random Forests (such as Gini importance), which only reflect the overall contribution of a variable (Archer & Kimes, 2008; Loecher, 2022). SHAP values quantify both the direction and magnitude of each feature's influence on every individual prediction. This enables us to identify specific environmental thresholds that trigger high risk. Although traditional logistic regression models are interpretable (Carvalho et al., 2019), but they typically assume linear relationships and would struggle to reveal such complex, threshold-driven interactions without pre-specifying interaction terms.

In contrast to conventional pest models that focus exclusively on climatic variables (Napiah et al., 2023), this study identified road density as a non-climatic factor with substantial predictive power, highlighting the ecological implications of agricultural infrastructure. Prior research has indicated that dust generated by roads can influence bagworm survival (Ebeling, 1971; Ho, 2002; Lever, 2008). Our findings quantify this relationship, revealing complex but meaningful patterns in SHAP value distributions. These suggest potential management-related influences; for instance, high road density may facilitate rapid pesticide application, whereas low road density may support parasitoid wasp populations. Future studies could explore the multilevel interactions between 'landscape structure, microclimate, and pest outbreaks' to deepen understanding of these ecological linkages.

From a practical perspective, the developed model can be operationalised into a practical application. For example, when habitat conditions such as RH_lag2 < 60% and LST_lag1 > 28°C are met, a high-risk alert is triggered. Subsequently, control strategies can be tailored based on road density, like drone-based spraying in low-density areas. High-risk spatial predictions can also guide targeted interventions, for instance, by integrating the spatio-temporal predictions shown in Figure 8 to reduce the use of pesticides, thereby directly supporting sustainable development goals in Malaysia's palm oil industry (Parveez et al., 2021).

Despite its contributions, this study has several limitations. The pest dataset used in this research contains a total of 5,208 records after preprocessing, of which 842 records contain valid data, while the remaining 4,366 records are missing or invalid, which means that nearly 84% of the records are unusable. Such a high proportion of missing data makes it likely that the valid subset cannot fully represent the conditions of the entire area. As a result, the trained model may inherently exhibit systematic selection bias (Phillips et al., 2009), with poor spatial generalisation: its predictions are most reliable for areas that resemble the well-monitored subset used for training, whereas greater uncertainty exists for areas with insufficient monitoring.

The time-series split approach adopted during validation to assess predictive performance on unseen time periods within the same spatial domain provides an evaluation of temporal generalisation to some extent (Cerqueira et al., 2020). However, its applicability to other regions or future climate scenarios requires additional scrutiny.

In terms of data, other factors not considered in this study include: pest occurrence was treated as a binary classification problem rather than a regression task to reflect infestation intensity; the bagworm pests are also not differentiated by species or sex (Loong & Chong, 2012); other potential independent variables such as wind direction, speed (Rhainds et al., 2002) and solar radiation is not included in this study; the influence of soil type and other static environmental variables was not controlled; finally, the landscape structure data used were static and did not account for the impact of temporary agricultural paths, which may affect local dispersal patterns of bagworms.

Future research will specifically concentrate on assessing the model's spatial scalability by implementing it in plantations with varying management practices and microclimates, as well as its temporal scalability by integrating multi-year data to evaluate its resilience under changing climatic conditions. Another direction for future work is to systematically compare the Random Forest model used here with other advanced algorithms (such as XGBoost, LSTM).

Such studies would require larger datasets to ensure reliable training and validation and would further refine the way managers select the most appropriate predictive tools in application.

## CONCLUSION

This study constructed a spatio-temporal prediction model for palm oil bagworm by integrating remote sensing imagery with the random forest algorithm. The model demonstrated effective performance in predicting the spatial and temporal patterns of bagworm outbreaks, achieving a maximum recall rate of 0.76. Visualisation of the prediction results enables targeted management of high-risk areas, thereby enhancing the practical applicability of the model. These outcomes contribute to improving pest control efficiency and support the sustainable development goals of Malaysia's palm oil industry. Future research could incorporate additional habitat-related variables, such as soil type, to further refine model accuracy. Moreover, integrating weather forecasting data into predictive modelling and exploring advanced ensemble techniques may enhance the robustness and generalisability of the prediction framework.

## ACKNOWLEDGEMENT

The authors would like to acknowledge the financial support from the Ministry of Education Malaysia under the Fundamental Research Grant Scheme (FRGS) (FRGS/1/2021/WAB04/UTM/02/2) and Universiti Teknologi Malaysia under R.J130000.7852.5F485. The authors also thanked Lekir Plantation, Kuala Lumpur Kepong Berhad, for the data used in this study. We also acknowledge the insight of this study from Mr Muhammad Idrus Bin Shukor.

## LIST OF ABBREVIATIONS

| | | |
|---|---|---|
| AI | : | Artificial Intelligence |
| FN | : | False negative |
| FP | : | False positive |
| GIS | : | Geographic information system |
| GPM | : | Global precipitation measurement |
| GPS | : | Global positioning system |
| IoT | : | Internet of Things |
| LiDAR | : | Light detection and ranging |
| LST | : | Land surface temperature |
| LSTM | : | Long short-term memory |
| MODIS | : | Moderate resolution imaging spectroradiometer |
| MPOB | : | Malaysian Palm Oil Board |
| NDVI | : | Normalised difference vegetation index |
| RF | : | Rainfall |
| RH | : | Relative humidity |

| | | |
|---|---|---|
| ROC-AUC | : | Area under the receiver operating characteristic curve |
| SDG | : | Sustainable development goal |
| SHAP | : | Shapley additive exPlanations |
| STC | : | Space-time cube |
| TP | : | True positive |
| VIF | : | Variance inflation factor |
| XGBoost | : | eXtreme gradient boosting |

## REFERENCES

A Ilemobayo, J., Durodola, O., Alade, O., J Awotunde, O., T Olanrewaju, A., Falana, O., Ogungbire, A., Osinuga, A., Ogunbiyi, D., & Ifeanyi, A. (2024). Hyperparameter tuning in machine learning: A comprehensive review. *Journal of Engineering Research and Reports, 26*(6), 388-395.https://doi.org/10.9734/jerr/2024/v26i61188

Archer, K. J., & Kimes, R. V. (2008). Empirical characterisation of random forest variable importance measures. *Computational Statistics & Data Analysis, 52*(4), 2249-2260. https://doi.org/10.1016/j.csda.2007.08.015

Balasundram, S. K., Shamshiri, R. R., Sridhara, S., & Rizan, N. (2023). The role of digital agriculture in mitigating climate change and ensuring food security: An overview. *Sustainability, 15*(6), 5325. https://doi.org/10.3390/su15065325

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics, 8*(8), 832. https://doi.org/10.3390/electronics8080832

Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning, 109*(11), 1997-2028. https://doi.org/10.1007/s10994-020-05910-7

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In C. Zhang & Y. Ma (Eds.), *Ensemble machine learning: Methods and applications* (pp. 157-175). Springer New York. https://doi.org/10.1007/978-1-4419-9326-7_5

Didan, K. (2021). MODIS/Terra vegetation indices 16 days L3 global 250 m SIN/Grid Version V061 [Data set]. https://doi.org/10.5067/MODIS/MOD13Q1.061

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., & Morgan, G. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys, 55*(9), 1-33. https://doi.org/10.1145/3561048

Ebeling, W. (1971). Sorptive dusts for pest control. *Annual Review of Entomology, 16*(1), 123-158. https://doi.org/10.1146/annurev.en.16.010171.001011

Enting, C. E., & Latip, S. N. H. M. (2021). Life cycle of palm oil bagworm, *Metisa plana* Walker (Lepidoptera: Psychidae) at different temperatures under controlled environment. *Serangga, 26*(2), 151-165.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Fox, E. W., Hill, R. A., Leibowitz, S. G., Olsen, A. R., Thornbrugh, D. J., & Weber, M. H. (2017). Assessing the accuracy and stability of variable selection methods for random forest modelling in ecology. *Environmental Monitoring and Assessment, 189*(7), 316. https://doi.org/10.1007/s10661-017-6025-0

Gachoki, S., Groen, T. A., Vrieling, A., Skidmore, A., & Masiga, D. (2024). Towards accurate spatial prediction of *Glossina pallidipes* relative densities at country-scale in Kenya. *Ecological Informatics, 81*, Article 102610. https://doi.org/10.1016/j.ecoinf.2024.102610

Guo, X., Bian, Z., Wang, S., Wang, Q., Zhang, Y., Zhou, J., & Lin, L. (2020). Prediction of the spatial distribution of soil arthropods using a random forest model: A case study in Changtu County, Northeast China. *Agriculture, Ecosystems & Environment, 292*, 106818. https://doi.org/10.1016/j.agee.2020.106818

Hamer, W. B., Birr, T., Verreet, J.-A., Duttmann, R., & Klink, H. (2020). Spatio-temporal prediction of the epidemic spread of dangerous pathogens using machine learning methods. *ISPRS International Journal of Geo-Information, 9*(1), 44. https://doi.org/10.3390/ijgi9010044

Ho, C. T. (2002). Ecological studies on Pteroma pendula Joannis and Metisa plana Walker (Lepidoptera: Psychidae) towards improved integrated management of infestations in palm oil [Doctoral dissertation, Universiti Putra Malaysia]. Universiti Putra Malaysia Publications. https://core.ac.uk/download/pdf/43000616.pdf

Holloway, P., Kudenko, D., & Bell, J. R. (2018). Dynamic selection of environmental variables to improve the prediction of aphid phenology: A machine learning approach. *Ecological Indicators, 88*, 512-521.https://doi.org/10.1016/j.ecolind.2017.10.032

Hoong, H. W., & Hoh, C. (1992). Major pests of palm oil and their occurrence in Sabah. *Planter, 68*(793), 193-210.

Huffman, G. J., Stocker, E. F., Bolvin, D. T., Nelkin, E. J., & Tan, J. (2023). *GPM IMERG final precipitation L3 1 day 0.1 degree × 0.1 degree Version V07* [Data set]. Goddard Earth Sciences Data and Information Services Centre (GES DISC). https://doi.org/10.5067/GPM/IMERGDF/DAY/07

Kamarudin, N., Ali, S. R. A., Masri, M. M. M., Ahmad, M. N., Manan, C. A. H. C., & Kamarudin, N. (2017). Controlling *Metisa plana* Walker (Lepidoptera: Psychidae) outbreak using *Bacillus thuringiensis* at an palm oil plantation in Slim River, Perak, Malaysia. *Journal of Palm oil Research, 29*(1), 47-54. https://doi.org/10.21894/jopr.2017.2901.05

Kamarudin, N., & Mohd, B. W. (2007). Status of common palm oil insect pests in relation to technology adoption. *Planter, 83*(975), 371-385.https://doi.org/10.56333/tp.2007.005

Lee, D. S., Lee, T. G., Bae, Y. S., & Park, Y. S. (2023). Occurrence prediction of western conifer seed bug (*Leptoglossus occidentalis*: Coreidae) and evaluation of the effects of climate change on its distribution in South Korea using machine learning methods. *Forests, 14*(1), 14. https://doi.org/10.3390/f14010117

Lever, R. J. A. W. (2008). Pests of palm oils in Malaysia and their control by Brian J. Wood, Kuala Lumpur: Incorporated Society of Planters (1968), pp. 204, $25 Malaysian. *Experimental Agriculture, 6*(1), 79-80. https://doi.org/10.1017/S0014479700005950

Li, J.-Y., Chen, Y.-T., Shi, M.-Z., Li, J.-W., Xu, R.-B., Pozsgai, G., & You, M.-S. (2021). Spatio-temporal distribution patterns of *Plutella xylostella* (Lepidoptera: Plutellidae) in a fine-scale agricultural landscape based on geostatistical analysis. *Scientific Reports, 11*(1), 13622. https://doi.org/10.1038/s41598-021-92562-9

Loecher, M. (2022). Unbiased variable importance for random forests. *Communications in Statistics-Theory and Methods, 51*(5), 1413-1425. https://doi.org/10.1080/03610926.2020.1764042

Loong, C., & Chong, T. (2012). Understanding pest biology and behaviour for effective control of palm oil bagworms. The Planter, 88(1039), 699-715.https://doi.org/10.56333/tp.2012.008

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent individualised feature attribution for tree ensembles. *arXiv* (arXiv:1802.03888). https://doi.org/10.48550/arXiv.1802.03888

Mahmoodi, S., Ganje, M. B., Ahmadi, K., Dalvand, Y., Naghibi, A., & Newlands, N. K. (2024). Modelling spatiotemporal distribution of yellow rust wheat pathogen using machine learning algorithms: Insights from environmental assessment. *Environmental Technology & Innovation, 36*, 103865.https://doi.org/10.1016/j.eti.2024.103865

Makori, D. M., Abdel-Rahman, E. M., Odindi, J., Mutanga, O., Landmann, T., & Tonnang, H. E. Z. (2024). Multi-pronged abundance prediction of bee pests' spatial proliferation in Kenya. *International Journal of Applied Earth Observation and Geoinformation, 128*, 103738. https://doi.org/10.1016/j.jag.2024.103738

Marković, D., Vujičić, D., Tanasković, S., Đorđević, B., Ranđić, S., & Stamenković, Z. (2021). Prediction of pest insect appearance using sensors and machine learning. *Sensors, 21*(14), 4846. https://doi.org/10.3390/s21144846

MODIS, A. S. T. (2017a). *MODIS/Terra clouds 5-min L2 swath 1 km and 5 km* [Data set]. NASA LANCE MODIS at the MODAPS. https://doi.org/10.5067/MODIS/MOD06_L2.NRT.061

MODIS, A. S. T. (2017b). *MODIS/Terra temperature and water vapour profiles 5-min L2 swath 5 km* [Data set]. NASA LANCE MODIS at the MODAPS. https://doi.org/10.5067/MODIS/MOD07_L2.NRT.061

MODIS, A. S. T. (2017c). *MODIS/Terra total precipitable water vapour 5-min L2 swath 1 km and 5 km* [Data set]. NASA LANCE MODIS at the MODAPS. https://doi.org/10.5067/MODIS/MOD05_L2.NRT.061

Naidu, G., Zuva, T., & Sibanda, E. M. (2023). A review of evaluation metrics in machine learning algorithms. In *Artificial Intelligence application in networks and systems*. Springer. https://doi.org/10.1007/978-3-031-35314-7_2

Napiah, N. R. A. M. A., Kamarudin, N., Bakeri, S. A., Zainuddin, N., Keni, M. F., & Masri, M. M. M. (2023). Impact of environmental factors on the larval population of bagworm, *Metisa plana* Walker (Lepidoptera: Psychidae) in palm oil smallholdings. *Serangga, 28*(2), 149-161.

NASA/METI/AIST/Japan, U.S./Japan ASTER Science Team. (2019). *ASTER global digital elevation model version V003* [Data set]. NASA Land Processes Distributed Active Archive Centre.https://doi.org/10.5067/ASTER/ASTGTM.003

Parveez, G. K. A., Tarmizi, A. H. A., Sundram, S., Loh, S. K., Ong-Abdullah, M., Palam, K. D. P., Salleh, K. M., Ishak, S. M., & Idris, Z. (2021). Palm oil economic performance in Malaysia and R&D progress in 2020. *Journal of Palm oil Research, 33*(2), 181-214. https://doi.org/10.21894/jopr.2021.0026

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.https://doi.org/10.48550/arXiv.1201.0490

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications, 19*(1), 181-197. https://doi.org/10.1890/07-2153.1

Pol, M., Bailey, L. D., McLean, N., Rijsdijk, L., Lawson, C. R., Brouwer, L., & Gimenez, O. (2016). Identifying the best climatic predictors in ecology and evolution. *Methods in Ecology and Evolution, 7*(10), 1246-1257. https://doi.org/10.1111/2041-210X.12590

Rajakumaran, M., Arulselvan, G., Subashree, S., & Sindhuja, R. (2024). Crop yield prediction using multi-attribute weighted tree-based support vector machine. *Measurement: Sensors, 31*, 101002. https://doi.org/10.1016/j.measen.2023.101002

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv* (arXiv:1811.12808). https://doi.org/10.48550/arXiv.1811.12808

Ravljen, M., Hovelja, T., & Vavpotič, D. (2018). Immediate, lag and time window effects of meteorological factors on ST-elevation myocardial infarction incidence. *Chronobiology International, 35*(1), 63-71. https://doi.org/10.1080/07420528.2017.1381847

Rhainds, M., Gries, G., Ho, C. T., & Chew, P. S. (2002). Dispersal by bagworm larvae, *Metisa plana*: Effects of population density, larval sex, and host plant attributes. *Ecological Entomology, 27*(2), 204-212.https://doi.org/10.1046/j.1365-2311.2002.00389.x

Rojat, T., Puget, R., Filliat, D., Del Ser, J., Gelin, R., & Díaz-Rodríguez, N. (2021). Explainable Artificial Intelligence (XAI) on time series data: A survey. *arXiv* (arXiv:2104.00950). https://doi.org/10.48550/arXiv.2104.00950

Rouabah, A., Meiss, H., Villerd, J., Lasserre-Joulin, F., Tosser, V., Chabert, A., & Therond, O. (2022). Predicting the abundances of aphids and their natural enemies in cereal crops: Machine learning versus linear models. *Biological Control, 169*, 104866. https://doi.org/10.1016/j.biocontrol.2022.104866

Sabtu, N., Idris, N., & Ishak, M. (2018). The role of geospatial in plant pests and diseases: An overview. *IOP Conference Series: Earth and Environmental Science*. https://doi.org/10.1088/1755-1315/169/1/012013

Salim, H., & Hamid, N. H. (2012). Evaluation of several chemical control approaches against bagworm, *Metisa plana* Walker (Lepidoptera: Psychidae) in Felda palm oil plantations. *The Planter, 88*, 785-799.https://doi.org/10.56333/tp.2012.009

Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random forest algorithm overview. *Babylonian Journal of Machine Learning, 2024*, 69-79. https://doi.org/10.58496/BJML/2024/007

Stejskal, V., Vendl, T., Aulicky, R., & Athanassiou, C. (2021). Synthetic and natural insecticides: Gas, liquid, gel and solid formulations for stored-product and food-industry pest control. *Insects, 12*(7), 590. https://doi.org/10.3390/insects12070590

Sulaiman, M. N., & Talip, M. S. A. (2021). Sustainable control of bagworm (Lepidoptera: Psychidae) in palm oil plantation: A review paper. *International Journal of Agriculture, Forestry and Plantation, 11*, 47-55.

Sundaraja, C. S., Hine, D. W., & Lykins, A. D. (2021). Palm oil: Understanding barriers to sustainable consumption. *PLoS ONE, 16*(8), e0254897. https://doi.org/10.1371/journal.pone.0254897

Tailliez, B., & Ballo Koffi, C. (1992). A method for measuring palm oil leaf area. *Oléagineux, 47*(8-9), 537-545.

Thaer, S., Kassim, F. A., Hasbullah, N. A., & Al-Obaidi, J. R. (2021). Evaluation of bagworm, *Metisa plana* (Lepidoptera: Psychidae) infestation and beneficial parasitoid in an palm oil plantation, Perak, Malaysia. *Journal of Science and Mathematics Letters, 9*(1), 19-35. https://doi.org/10.37134/jsml.vol9.1.3.2021

Wahid, M. B., Hassan, H. A. H., & Masijan, Z. (1988). Bagworms (Lepidoptera: Psychidae) of palm oils in Malaysia. *PORIM Occasional Paper, 23*, 1-38.

Wan, Z., Hook, S., & Hulley, G. (2021). *MODIS/Terra land surface temperature/emissivity 8-day L3 global 1 km SIN grid version V061* [Data set]. https://doi.org/10.5067/MODIS/MOD11A2.061

Wong, T.-T., & Yeh, P.-Y. (2020). Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering, 32*(8), 1586-1594. https://doi.org/10.1109/TKDE.2019.2912815

Zakaria, K., Hassan, N. A. M., Azam, A. H. M., Salleh, K. M., Kamil, N. N., & Abdullah, N. (2024). Factors affecting export demand for Malaysian palm-based finished products. *Asian Academy of Management Journal, 29*(2), 109-132. https://doi.org/10.21315/aamj2024.29.2.5