IOP Conference Series: Earth and Environmental Science

PAPER • **OPEN ACCESS**

Comparison of different variable selection methods for predicting the occurrence of Metisa Plana in oil palm plantation using machine learning

To cite this article: Y P Wang *et al* 2023 *IOP Conf. Ser.: Earth Environ. Sci.* **1274** 012008

Comparison of different variable selection methods for predicting the occurrence of Metisa Plana in oil palm plantation using machine learning

Y P Wang[1,5], N H Idris[1,2,*], F M Muharam[3], N Asib[4], and Alvin M S Lau[1,2]

[1]TropicalMap Research Group, Faculty of Built Environmental and Surveying, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia
[2]Department of Geoinformation, Faculty of Built Environment and Surveying, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia
[3]Department of Agriculture Technology, Faculty of Agriculture, Universiti Putra Malaysia, Selangor, 43400, Malaysia
[4]Department of Plant Protection, Faculty of Agriculture, Universiti Putra Malaysia, Selangor, 43400, Malaysia
[5]Department of Hydraulic Engineering, Hebei University of Water Resources and Electric Engineering, CangZhou, Hebei, 061001, China

E-mail: hawani@utm.my

Abstract. Monitoring and predicting the spatio-temporal distribution of crop pests and assessing related risks are crucial for effective pest management strategies. Machine learning techniques have shown potential in analysing agricultural data and providing accurate predictions. Variable selection plays a critical role in crop pest analysis by identifying the most informative and influential features that contribute to pest distribution and risk prediction. The current practice of choosing variable selection methods is mostly based on previous experience and may involve a certain degree of subjectivity. This paper aims to provide empirical comparisons of different variable selection methods for machine learning applications in crop pest spatio-temporal distribution and risk prediction. This study conducted various variable selection methods, including filter methods (information gain, chi-square test, mutual information), wrapper methods (RFE), and embedded methods (Random Forest), using worms pest (*Metisa plana*) in oil palm trees as the experimental subject. The initial set of variables included bioclimatic, vegetation indices, and terrain variables. The experimental results indicated that there was some overlap in the selected variables across different methods, bioclimatic variables (rainfall (RF), relative humidity (RH)) were selected as important variables by different methods; non-important variables like NDVI and elevation when added to the ANN modelling can clearly contribute to the improvement in prediction accuracy. These empirical findings can provide guidance for relevant data monitoring in the prediction of crop pest and disease outbreaks. Additionally, the results can serve as a reference for variable selection in spatiotemporal prediction of pests and diseases in other agricultural and forestry crops.

1. Introduction

Crop pest infestation poses significant challenges to global agricultural productivity and food security[1-3]. Monitoring and predicting the spatio-temporal distribution of crop pests and assessing related risks are crucial for effective pest management strategies[4]. For example, if left untreated, a bagworm (*Metisa plana*) infestation can cause significant damage to the oil palm industry in Malaysia, as it has the potential to completely skeletonize and eventually kill oil palm fronds, leading to devastating losses[5]. Machine learning techniques have shown potential in analyzing large-scale agricultural data and providing accurate predictions[6]. However, the success of these models heavily relies on the selection of relevant variables or features[7].

Variable selection, also known as feature selection, plays a critical role in machine learning applications by identifying the most informative and influential features that contribute to pest distribution and risk prediction. Various variable selection methods have been proposed and applied in the context of crop pest analysis. [8] compared Filter feature selection methods (correlation coefficient, variable inflation factor (VIF)) and Feature Dimension Reduction method (principal component analysis (PCA)) in their study of *M. pruinosa* and *S. litura*, and used MaxEnt modelling to compare the final experimental results. [9] used a Random Forest (RF) model to predict the spatial distribution of soil arthropods, where the RF model belongs to the embedded method, and the variable selection process is performed during the modelling process. [10, 11] used the Least Absolute Shrinkage Selection Operator (LASSO) for feature selection, mainly by shrinking the coefficients of unnecessary or highly correlated predictor factors to zero, to address the issue of multicollinearity. This method also belongs to the embedded method. [12] built a model for the spatial distribution and variation patterns of stingless bees based on machine learning. They used the Recursive Feature Elimination (RFE) in the wrapper method to obtain the smallest subset of predictive variables that could produce comparable results. Pearson correlation coefficient and VIF methods were also used to select non-collinear feature variables in this study.

Among various variable selection methods, different types of methods have different scopes of application and conditions for use[13, 14]. Filter methods are independent of specific machine learning algorithms and are applied during the pre-processing stage. Wrapper methods rely on the final model performance to select feature sets and are applied during the model training phase, which incurs significant computational costs and is suitable for small datasets. Embedded methods automatically select important features during model training and are suitable for cases where the relationship between features and the target variable is complex[15].

The selection of feature variables can significantly impact the final model accuracy[8]. Currently, in the context of agricultural pest distribution and risk prediction (PDRP), the current practice of choosing variable selection methods is mostly based on previous experience and may involve a certain degree of subjectivity[16]. It is not yet studied whether different variable selection methods yield consistent results in terms of the selected features and how they influence the final modelling accuracy in the context of PDRP. Understanding the consistency and influence of various variable selection methods on model accuracy can provide valuable insights for improving prediction models in agriculture.

Therefore, in this article, we delve into the empirical comparison of different variable selection methods for machine learning applications in crop pest spatio-temporal distribution and risk prediction. We aim to evaluate and analyse the performance of these methods in terms of their ability to identify relevant features and improve prediction accuracy. By understanding the strengths and limitations of each method, we can provide insights into selecting an optimal variable selection approach for crop pest analysis.

The remainder of this article is organized as follows: Section 2 firstly provides an overview of the existing variable selection methods commonly used in machine learning and *Metisa plana* and its living habitats. Section 3 presents the dataset and experimental setup used for our comparative analysis. Section 4 presents the results and discussion, highlighting the performance of different variable selection

methods. Finally, Section 5 concludes the article and discusses future directions in variable selection for crop pest analysis.

2. Literature Review

2.1 Variable Selection Methods in Machine Learning

The generally recognized fundamental process of machine learning includes five main steps: data collection, data cleaning, feature engineering, model building, and model monitoring[17], the detail has outlined in figure 1. This paper mainly focuses on feature selection within the feature engineering step. As has mentioned in section 1, variable selection is a crucial step in the process of building predictive models and extracting meaningful insights from data. It involves choosing a subset of relevant features or variables from a larger set of available options.

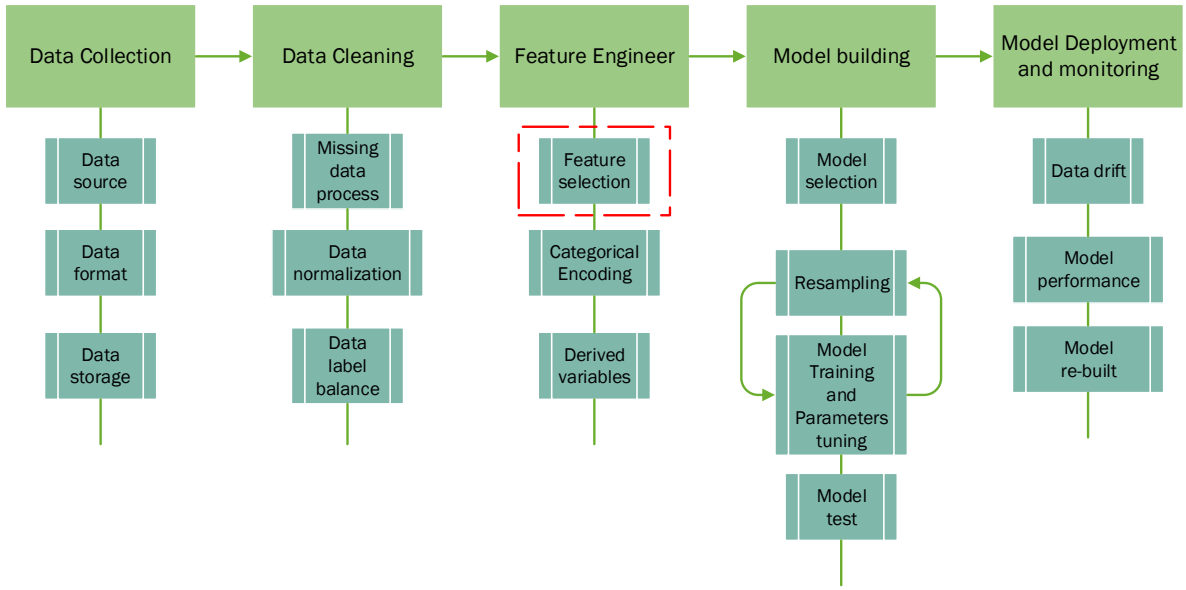


Figure 1. General process of Machine Learning.

This selection process helps to improve model performance, reduce overfitting, and enhance interpretability. The first method known as filter which typically rely on some filter index like statistical measures or scoring techniques to evaluate the relevance of variables, independent of any specific predictive model. Common filter techniques include correlation analysis, mutual information, chi-square test, and information gain[13]. The variables are ranked or assigned a score based on these measures, and a subset of top-ranking variables is selected.

The second method is wrapper approach which treats variable selection as a search problem, in which it repeatedly trains the predictive model with different subsets of variables and uses specific evaluation measures (for example, accuracy, AUC, or cross-validation performance) to guide the search. It can be seen that specific predictive models are incorporated into the selection process. The search can be exhaustive (considering all possible combinations) or employ heuristic techniques like forward selection, backward elimination, or recursive feature elimination[18]. The goal is to find the optimal subset of variables that maximizes the model's performance.

The third is embedded methods which combine feature selection with the model training process. These methods incorporate variable selection within the algorithm itself. Random Forest[19], regularization techniques like LASSO (Least Absolute Shrinkage and Selection Operator) and Ridge regression[7] are popular embedded methods. They calculate the importance of features in the training process of model directly and avoid the extra computational cost of independent feature selection. Each variable selection method has its own strengths and limitations. Filter methods are computationally

efficient but may overlook complex interactions among variables. Wrapper methods are more computationally intensive but can capture variable dependencies. Embedded methods strike a balance between the two and are often used when the number of variables is large. Here in figure 2 is the process framework of these three kinds of variable selection methods.

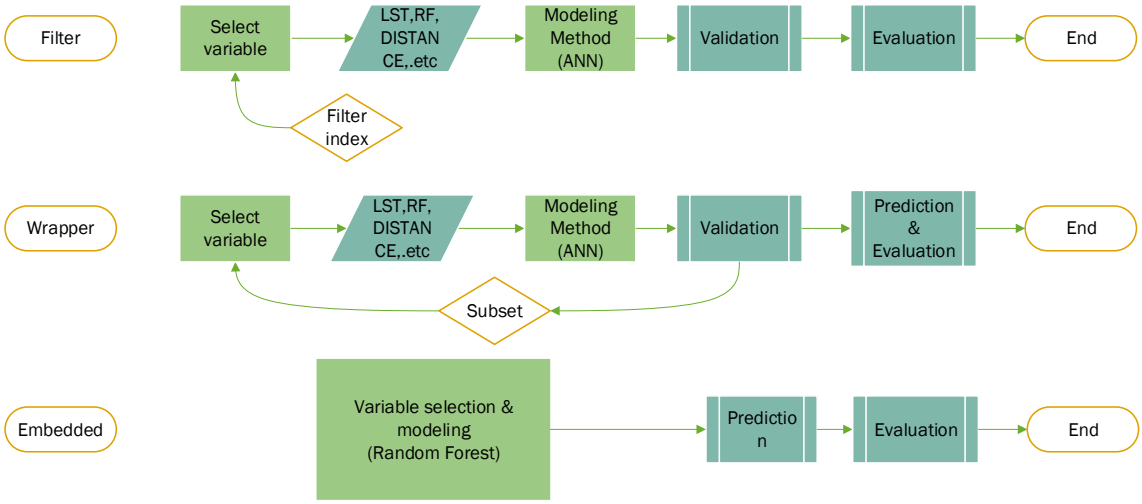


Figure 2. The process framework of different variable selection methods.

2.2 Metisa Plana – Oil Palm Bagworm

Bagworm can cause detrimental effect to the oil palm plantation in Malaysia every year. Studies have shown that *Metisa plana* is the most widely distributed oil palm insect in Peninsular Malaysia[20]. *Metisa plana*, commonly known as the bagworm or coconut caterpillar, is a serious pest that poses a significant threat to the oil palm industry[21]. The larvae of *Metisa plana* are the most damaging stage of the insect's life cycle (figure 3(a)), they construct bag-like structure made of silk and plant material, which they use as protective shelters. Their larvae feed voraciously on the foliage of oil palm trees, primarily targeting the fronds (figure 3(b)). As they consume the leaf tissue, they can cause complete skeletonization of the fronds, leaving behind only the leaf veins. Severe infestations can lead to the death of fronds and, in extreme cases, even the death of the entire oil palm tree. The loss of foliage reduces the photosynthetic capacity of the tree, impacting oil palm yield and overall productivity[22].

(a) (b)

Figure 3. (a) *Metisa plana* in different instars, from left to right is instar 1 to 4 and 7 respectively (b) oil palm tree leaves that has been infested.

3. Material and Method

All experiments in this study are based on the '*mlr*' package of R language, which constructs a unified machine learning process.

3.1 Study site and primary dataset

The study site is a 2000-ha oil palm estate locates in the Tabung Haji Plantation Berhad in Sungai Mengah, Muadzam Shah, with the coordination of 2° 57' 30" N, 102° 52' 30" E to 3° 1' 30" N, 102° 55' 0" E in the state of Pahang, Malaysia. There is a total of 24 planting blocks in the study site (figure 4).

The initial set of variables included *M. plana* census data, which was collected biweekly; bioclimatic (Land Surface Temperature (℃) (LST), Relative Humidity (%) (RH), Rainfall (mm/h) (RF)) and is collected every week; vegetation indices (Normalized Difference Vegetation Index (NDVI)) with biweekly collected also and constant variables - terrain (elevation). LST, RH, and NDVI data were obtained from Moderate Resolution Imaging Spectroradiometer (MODIS) sensor of TERRA satellite, RF was provided by the Tropical Rainfall Measuring Mission (TRMM) satellite. The time frame of these dataset is 2014-2015. These data were supplied from previous study by [23].

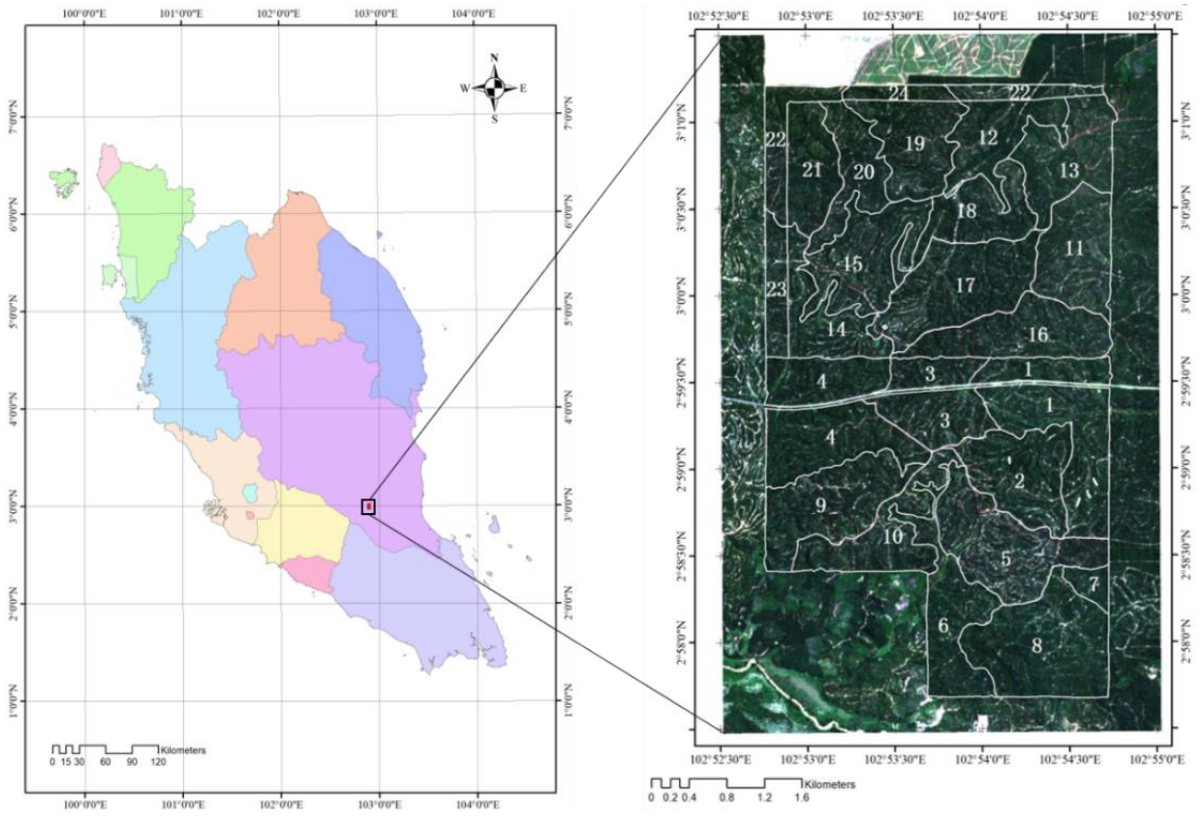


Figure 4. Location of the study site.

3.2 Data Pre-processing:

Firstly, the missing values and outliers are removed and spatiotemporal interpolation is carried out. Secondly, all of the independent data were scaled to values between 0 and 1 by the Min-Max normalization method. Thirdly, there is a time lag alignment between pest statistics and corresponding predictive variables, by considering that the observed outbreak of pest incidence of one time does not result from the weather variables of the same time but, instead, results much more from the previous weather conditions[24]. Here we refer to the study of [25] and chose to the values of the predictor variables in the past three weeks. The bioclimatic variables three weeks ahead were averaged and

corresponded to those pest census data while the vegetation indices (NDVI) two weeks ahead were used. Fourth, all the data (287 samples) were divided into 2 parts firstly; 1) 70% (200 samples) for network training and validation 2) 30% (87 samples) dataset for testing in the end.

3.3 Feature Engineering

Due to the different characteristics of different variable selection methods, different processes were adopted in the modelling process of this paper, as shown in figure 5: First, data cleaning was performed on the original data, and then variable selection methods were adopted in the experiments respectively: Filter method (information gain, chi-square test, mutual information), wrapper method (recursive feature elimination) and embedded method (random forest) to select essential features. Combined with the machine learning algorithm ANN in the next step, classification analysis is conducted. Since random forest allows the importance of variables to be determined automatically during model training, variable selection process is embedded in model training without the need to combine ANN. So totally, this research includes 5 experimental cases: 1) information gain + ANN, 2) chi-square test + ANN, 3) mutual information + ANN, 4) RFE + ANN, 5) Random Forest, the following is an introduction of these specific variable selection method (see figure 5).

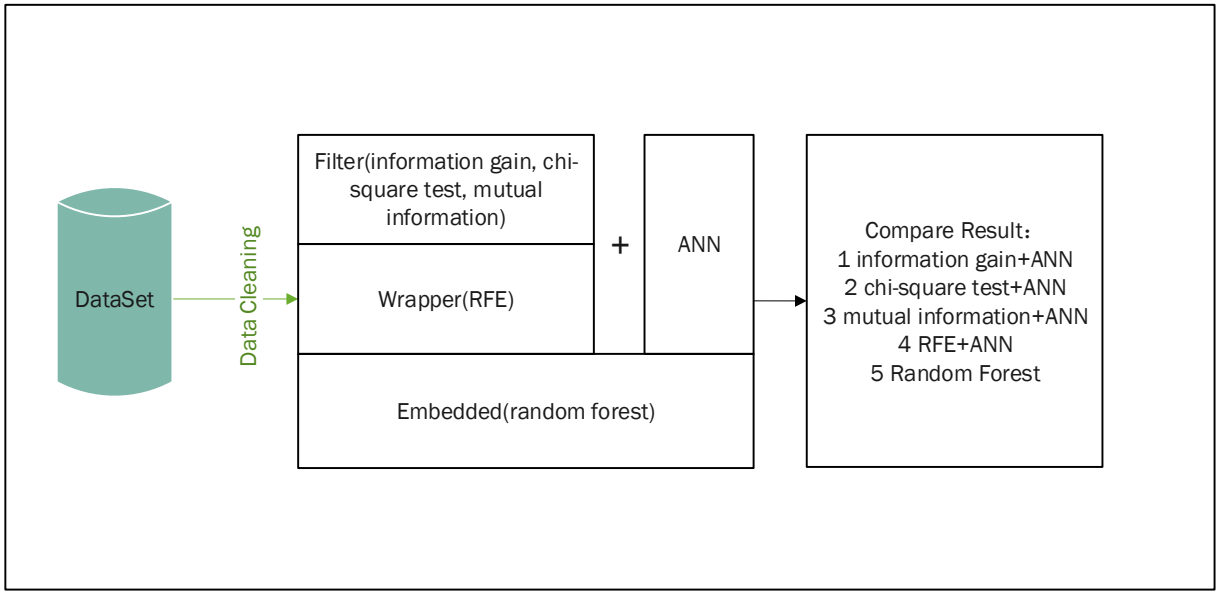


Figure 5. The overall processes of experimental design.

Information gain is based on the concepts of information theory, in particular Entropy and Conditional Entropy. In classification problems, the higher the Entropy, the more chaotic the data set, which means the more uncertain the classification of the sample. The formula for Entropy is[26]:

$$\text{Entropy}(D) = -\sum_{i=1}^{k} p_i * \log_2(p_i) \quad (1)$$

Where D represents the sample set, p_i represents the proportion of category i in the sample, and k is the total number of categories.

Conditional Entropy is the entropy of the dataset D given the conditions of feature A $\{a_1, a_2, \dots, a_v\}$, its formula can be expressed as[26]:

$$H(D|A) = \sum_{i=1}^{v} \frac{|D_i|}{|D|} * \text{Entropy}(D_i) \quad (2)$$

Where $|D|$ represents the total number of samples, $|D_i|$ represents the number of samples whose features A is a_i in the sample set.

Information gain is used to measure the contribution of feature A to the classification task, and the formula can be expressed as[26]:

$$\mathrm{IG}(D, A) = \mathrm{Entropy}(D) - \mathrm{H}(D|A) \tag{3}$$

The greater the information gain, the greater the contribution of feature A to the classification task.

Chi-Square Test is used to measure the correlation between a feature and a target variable. Its principle is based on the chi-square test in statistics to calculate the chi-square statistics between each feature and the target variable, which is used to determine the importance of the feature, and then select the important feature subset for model building and prediction. The steps are to establish the observed frequency table, calculate the expected frequency table, calculate the chi-square statistics (χ^2), and finally sort according to the χ^2 to determine the importance of features (see equation (4)).

$$\chi^2 = \sum \frac{(\text{observed frequency} - \text{expected frequency})}{\text{expected frequency}} \tag{4}$$

Mutual information is based on the concept of information theory and is used to measure the degree of information sharing between two random variables, and its formula is also based on the concept of entropy, including Entropy H(X), Joint Entropy H (X, Y) and Conditional Entropy H ($Y|X$) as follows[27]:

$$\mathrm{H}(X) = -\sum_{x \in X} \mathrm{P}(x) * \log_2(\mathrm{P}(x)) \tag{5}$$

$$\mathrm{H}(X, Y) = -\sum_{x \in X} \sum_{y \in Y} \mathrm{P}(x, y) * \log_2(\mathrm{P}(x, y)) \tag{6}$$

$$\mathrm{H}(Y|X) = -\sum_{x \in X} \sum_{y \in Y} \mathrm{P}(x, y) * \log_2\left(\frac{\mathrm{P}(x,y)}{\mathrm{P}(x)}\right) \tag{7}$$

Where P(x) represents the probability that the variables X takes the value x, and P (x, y) represents the joint probability that X takes the value x and Y takes the value y. The formula for mutual information I (X; Y) is[27]:

$$\mathrm{I}(X; Y) = \sum_{x \in X} \sum_{y \in Y} \mathrm{P}(x, y) * \log_2\left(\frac{\mathrm{P}(x,y)}{\mathrm{P}(x) * \mathrm{P}(y)}\right) \tag{8}$$

The greater the mutual information, the higher the degree of information sharing between the feature and the target variable, indicating that the feature has a greater contribution to the prediction of the target variable.

RFE (Recursive Feature Elimination) obtains the optimal combination variables that can maximize the performance of the model by adding or removing specific characteristic variables. The basic RFE algorithm can be summarized as follows[28]:

1) train the model using all features variables
2) calculate the importance of each feature variable and rank them
3) For each variable subset $S\_\{i\}$, i=1...S:
   - Extract the top $S\_\{i\}$ most important feature variables.
   - Train the model based on the new dataset.
   - Optionally, re-calculate the importance of each feature variable and rank them again.
   - Evaluate and compare the performance of the model obtained from each subset.
4) Decide on the optimal feature variable set
5) Select the model with the optimal feature variable set as the final model.

3.4 Model development

Artificial Neural Network (ANN) and Random Forest (RF) were chosen as the modelling methods in this study because they have been proven to possess good flexibility and have demonstrated favourable accuracy in numerous comparative studies among various models[29].

ANN is a fundamental deep learning algorithm that mimics the functionality of the brain through interconnected layers and neurons, it has been used to predict the potential presence of insect species[30]. This study employed Multi-Layer Perceptron (MLP) models and three main layers characterize the architecture of neural network: the input, output, and hidden layers (see figure 6). The input layer receives the feature vector of an input sample, denoted as x, the feature vector contains m elements representing the m features of the input. One or more hidden layers may exist in the model, each consisting of multiple neuron. These neurons process the input data using weighted connections and activation functions. Output layers produce the classification result. In a classification task with C classes, the output layer typically consists of C neurons, each representing the probability of the input sample belonging to a particular class.

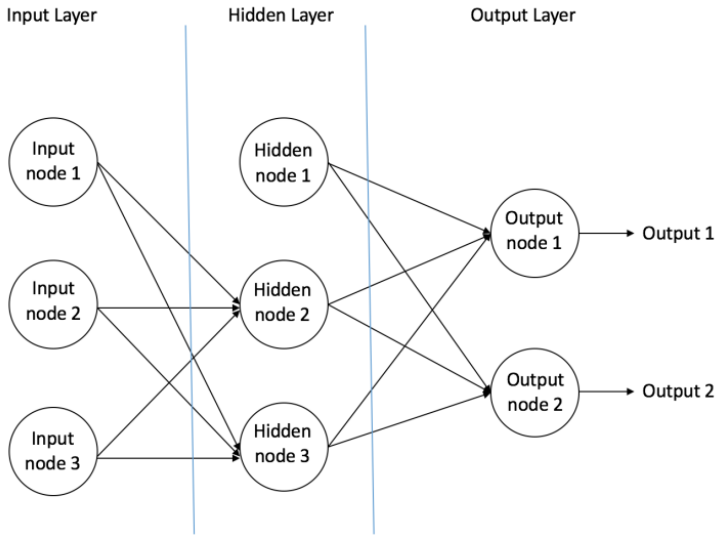


Figure 6. Structure of MLP.

RF is a machine learning techniques that combines the results of many classification trees developed from bootstrap samples randomly selected from the original data and is often praised for its robust performance[31]. The prediction of Random Forest can be represented as follows:

There are k decision trees forming the Random Forest, for a new input sample *x*, each decision tree will output a predicted class label. The Random Forest combines the predictions using a voting mechanism, selecting the class with the highest number of votes as the final predicted result. The specific formula is[32]:

$$\text{Predicted Result} = \text{argmax}(\sum \text{Vote}(t, x)) \quad (9)$$

Where Vote (t, *x*) represents the predicted class label by decision tree t for the input sample *x*, $\sum$ denotes the summation symbol, summing over all decision trees t, argmax is a function that selects the class label that maximizes the expression inside the parentheses, which corresponds to the class with the most vote.

3.5 Model Performance evaluation

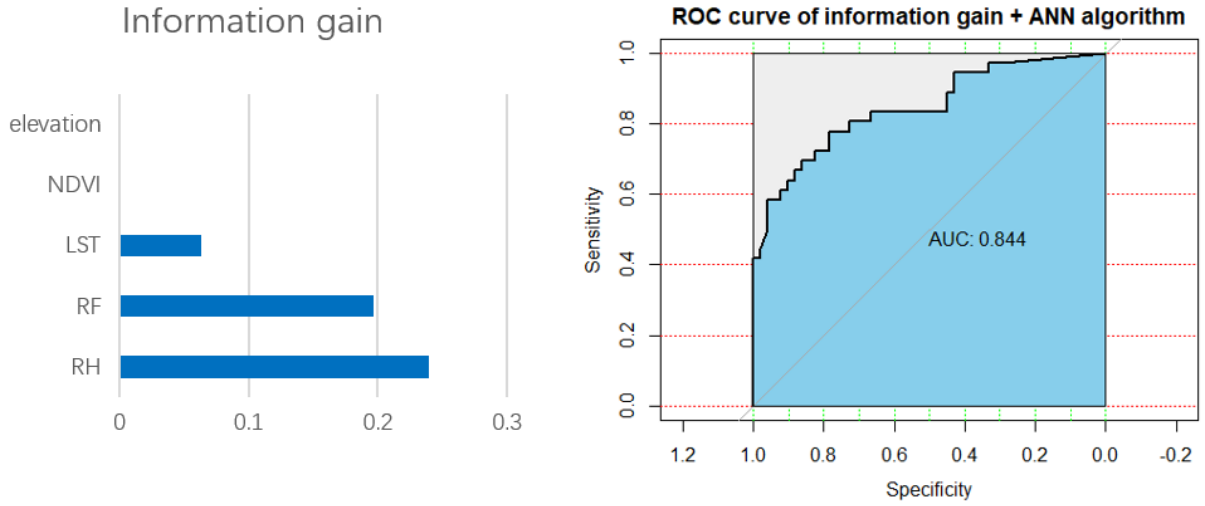
Since the focus of this study is not on the modelling method, only Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) were used to evaluate the goodness of the classifiers' prediction[33, 34].

4. Results and Discussion

4.1 Results

The experimental results are shown respectively in figure 7 with the relative importance of variables plot in the left and modelling prediction results' ROC/AUC in the right part. The five experimental cases were figured out with part (a) to (e) respectively.

(a)



(b)

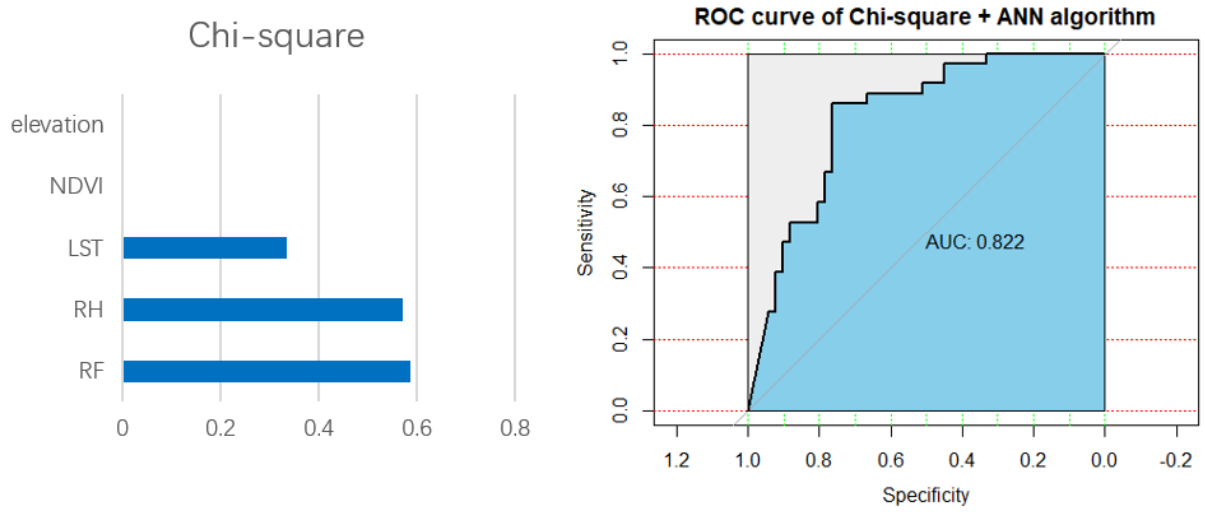
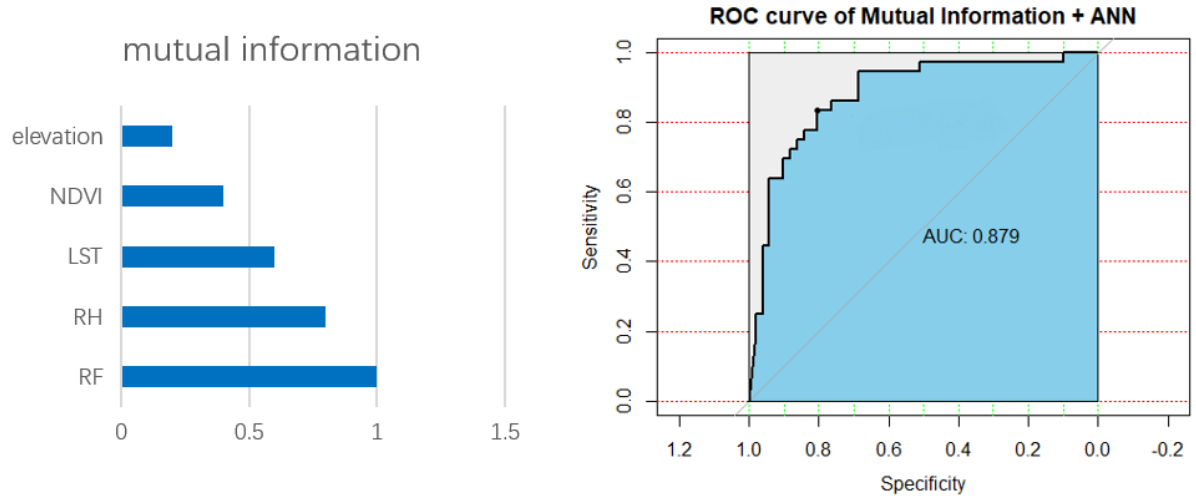
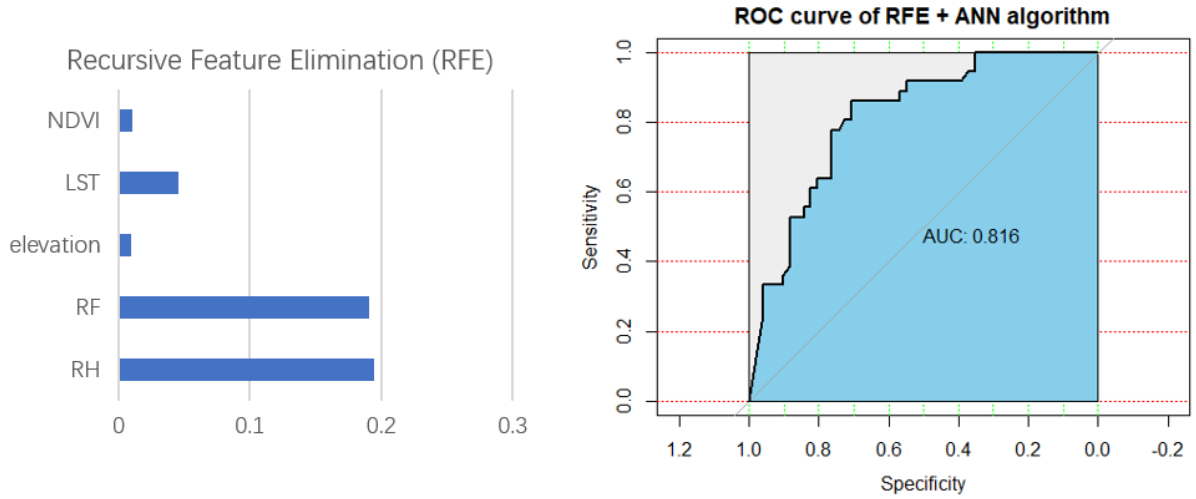


Figure 7. Variable relative importance & receiver operating characteristic curve (ROC/AUC) of these five research cases (a): information gain + ANN (b): chi-square + ANN (c): mutual information + ANN (d) RFE+ANN (e) random forest. RH: relative humidity, NDVI: normalized difference vegetation index, LST: land surface temperature, RF: rainfall.

(c)



(d)



(e)

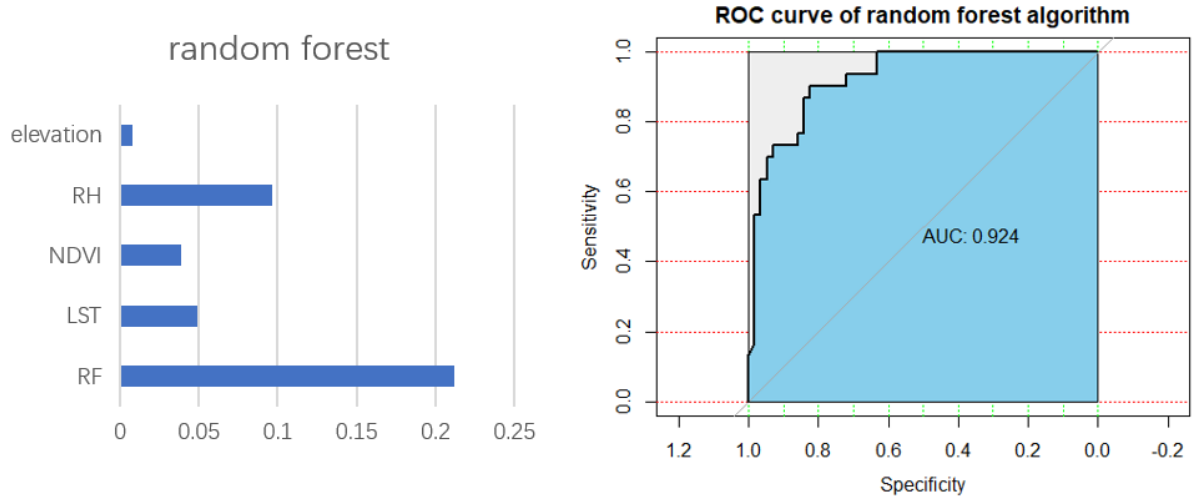


Figure 7. (continue).

From the results, it can be observed that in the information gain-based variable selection, relative humidity (RH) is the most important feature, followed by rainfall (RF) and land surface temperature (LST), while normalized difference vegetation index (NDVI) and elevation have zero importance, hence could be excluded. The chi-square test yields similar results, with 'RF' having slightly higher importance than 'RH', and 'NDVI' and 'elevation' still contributing little to the target variable and are excluded. In the mutual information-based variable selection, all variables are considered important, with 'RF,' 'RH', and 'LST' being the top three. 'NDVI' and 'elevation' remain less influential.

These above are the results obtained from the filter-based variable selection methods. As for the wrapper-based method, using RFE as an example, it shows that 'RH' and 'RF' remain the most important variables, followed by 'LST,' while 'NDVI' and 'elevation' have little impact. The last method, which is the embedded method using random forests, selects 'RF' as the most important variable, followed by 'RH,' but the importance of 'RF' is significantly greater than that of 'RH,' almost twice as much. Next are 'LST,' 'NDVI,' and 'elevation.'

Regarding the accuracy of the final models from the five experiment cases, the AUC metric was used for evaluation. It can be seen that the random forest model has the highest prediction accuracy with an AUC value of 0.924. The 'mutual information + ANN' model comes next with an AUC of 0.879, followed by the 'information gain + ANN' model with an AUC of 0.844. Finally, the 'chi-square + ANN' and 'RFE + ANN' models have similar prediction accuracy with 0.822 and 0.816 respectively.

4.2 Discussion

Experimental results have demonstrated that in the three types of variable selection methods (filter, wrapper, and embedded), relative humidity (RH) and rainfall (RF) consistently play a significant role in the selected variables. However, their relative importance may vary in different methods. On the other hand, land surface temperature (LST) has a relatively smaller effect, while normalized different vegetation index (NDVI) and elevation consistently rank last in the list of importance, even having no impact. This aligns with previous studies[21, 25] on the habitat preferences of the palm pest *M. plana*, where 'RH' has a strong influence, and 'RF' affects relative humidity variations, while excessive rainfall increases mortality and decrease the bagworm population. Temperature has a minor impact on *M. plana*[5].

Secondly, the consistency in variable selection results across different methods suggests that the selected variables are less influenced by the specific variable selection method used. Previous research has also pointed out that modelling and variable selection methods need to be tailored to different research objects and data, as the suitable methods may vary for different subjects and datasets[8]. Combined with the results of these experiments, it can be further analysed and indicated that the differences in modelling and variable selection for different research objects and data may not primarily due to the selected variables by different methods.

Regarding the modelling performance with different variable selection methods combined with ANN, there are some differences in prediction accuracy. Several points can be discussed:

- Comparing "information gain + ANN" with "chi-square + ANN," the differences are relatively small, possibly because their selected variables are similar, and the modelling approach is the same, resulting in similar model accuracy.
- Comparing "information gain + ANN" with "mutual information + ANN," it can be observed that including non-important variables like "NDVI" and "elevation" in the ANN model can still improve the final prediction accuracy.

According to ROC/AUC results, random forest has the highest classification accuracy in all five experiment results, which also explains the reason why random forest is a widely used machine learning algorithm[9].

5. Conclusion and future work

According to the results of five different variable selection experiment cases in this research, it can be concluded that the importance results of *M. Plana* variables obtained by different variable selection

methods have certain similarities, especially for important variables such as RH and RF, except that their importance order may be slightly different. However, this kind of difference does not significantly impact the final model prediction accuracy. On the other hand, non-important variables like "NDVI" and "elevation" clearly contribute to the improvement in prediction accuracy in ANN modelling.

The consistency of the variables importance that emerged in different variable selection methods also indicates that the differences in modelling and variable selection methods for different pest objects may not solely caused by the types of variable selection methods applied in one study.

This study focused on only some of the methods within the three types of variable selection (filter, wrapper, and embedded), while each category has other specific methods such as VIF and correlation for the filter, genetic algorithms for the wrapper, and LASSO for the embedded method. Including these methods in the experimental case could be a future research direction. However, the important variables have been identified, but the influence is positively correlated or inverse has not been demonstrated yet, which needs further research. Additionally, this study only considered ANN and random forest as modelling methods, which is also a limitation.

Nevertheless, the conclusions of this study can provide support for modelling the spatiotemporal distribution and risk prediction of agricultural pests. The findings can help optimize model construction strategies, improve model accuracy, and provide more accurate and efficient support for decision-making and pest control.

Reference

[1] Alfarisy AA, Chen Q, Guo M. Deep learning based classification for paddy pests & diseases recognition. Proceedings of 2018 International Conference on Mathematics and Artificial Intelligence; Chengdu, China. deep learning image classification: Association for Computing Machinery; 2018. p. 21–5.
[2] Wan H, Lu Z, Qi W, Chen Y. Plant Disease Classification Using Deep Learning Methods. Proceedings of the 4th International Conference on Machine Learning and Soft Computing; Haiphong City, Viet Nam. deep learning and image classification: Association for Computing Machinery; 2020. p. 5–9.
[3] Lelana NE, Utami S, Darmawan UW, Nuroniah HS, Darwo, Asmaliyah, et al. Bagworms in Indonesian Plantation Forests: Species Composition, Pest Status, and Factors That Contribute to Outbreaks. Diversity-Basel. 2022;14(6):20.
[4] Osborn D, Cutter A, Ullah F. Universal sustainable development goals. Understanding the transformational challenge for developed countries. 2015;2(1):1-25.
[5] Ruslan SA, Muharam FM, Omar D, Zulkafli ZD, Zambri MP. Development of geospatial model for predicting Metisa plana's prevalence in Malaysian oil palm plantation. IOP Conference Series: Earth and Environmental Science. 2019;230:012110.
[6] Charaya MU, Upadhyay A, Bhati HP, Kumar A. Plant disease forecasting: Past practices to emerging technologies. Plant Disease: Management Strategies; Nehra, S, Ed; Agrobios Research: Rajasthan, India. 2021:1-30.
[7] Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing. 2018;300:70-9.
[8] Yoon S, Lee WH. Methodological analysis of bioclimatic variable selection in species distribution modeling with application to agricultural pests (Metcalfa pruinosa and Spodoptera litura). Computers and Electronics in Agriculture. 2021;190:14.
[9] Guo X, Bian Z, Wang S, Wang Q, Zhang Y, Zhou J, et al. Prediction of the spatial distribution of soil arthropods using a random forest model: A case study in Changtu County, Northeast China. Agriculture, Ecosystems & Environment. 2020;292:106818.
[10] Munro HL, Montes CR, Gandhi KJK, Poisson MA. A comparison of presence-only analytical techniques and their application in forest pest modeling. Ecological Informatics. 2022;68:10.
[11] Mangeon S, Spessa A, Deveson E, Darnell R, Kriticos DJ. Daily mapping of Australian Plague Locust abundance. Scientific Reports (Nature Publisher Group). 2020;10(1).
[12] Makori DM, Abdel-Rahman EM, Ndungu N, Odindi J, Mutanga O, Landmann T, et al. The use of multisource spatial data for determining the proliferation of stingless bees in Kenya. Giscience & Remote Sensing. 2022;59(1):648-69.
[13] Kaur A, Guleria K, Trivedi NK, editors. Feature selection in machine learning: Methods and comparison. 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE); 2021: IEEE.
[14] Khalid S, Khalil T, Nasreen S, editors. A survey of feature selection and feature extraction techniques in machine learning. 2014 science and information conference; 2014: IEEE.
[15] Chandrashekar G, Sahin F. A survey on feature selection methods. Computers & Electrical Engineering. 2014;40(1):16-28.
[16] Pol M, Bailey LD, McLean N, Rijsdijk L, Lawson CR, Brouwer L, et al. Identifying the best climatic predictors in ecology and evolution. Methods in Ecology and Evolution. 2016;7(10):1246-57.
[17] Zhou Z-H. Machine learning: Springer Nature; 2021.
[18] El Aboudi N, Benhlima L, editors. Review on wrapper feature selection approaches. 2016 International Conference on Engineering & MIS (ICEMIS); 2016: IEEE.
[19] Chen R-C, Dewi C, Huang S-W, Caraka RE. Selecting critical features for data classification based on machine learning methods. Journal of Big Data. 2020;7(1):52.
[20] Kamarudin N, Mohd BW. Status of common oil palm insect pests in relation to technology

adoption. Planter. 2007;83(975):371-85.
[21] TUCK HC. Ecological studies on Pteroma pendula Joannis and Metisa plana Walker (Lepidoptera: Psychidae) towards improved integrated management of infestations in oil palm. 2002.
[22] plana Walker M. LIFE HISTORY AND FEEDING BEHAVIOUR OF THE OIL PALM BAGWORM.
[23] Ruslan SA. Development of geospatial model for Metisa plana (Walker) outbreak and outbreak prediction in oil palm plantations in Malaysia 2018.
[24] Hamer WB, Birr T, Verreet J-A, Duttmann R, Klink H. Spatio-Temporal Prediction of the Epidemic Spread of Dangerous Pathogens Using Machine Learning Methods. ISPRS International Journal of Geo-Information. 2020;9(1):44.
[25] Ruslan SA, Muharam FM, Zulkafli Z, Omar D, Zambri MP. Using satellite-measured relative humidity for prediction of Metisa plana's population in oil palm plantations: A comparative assessment of regression and artificial neural network models. PLoS One. 2019;14(10):e0223968.
[26] Lian W, Nie G, Jia B, Shi D, Fan Q, Liang Y. An intrusion detection method based on decision tree-recursive feature elimination in ensemble learning. Mathematical Problems in Engineering. 2020;2020:1-15.
[27] Latham PE, Roudi Y. Mutual information. Scholarpedia. 2009;4(1):1658.
[28] Kuhn M. Predictive modeling with R and the caret package. useR. 2013.
[29] de Oliveira Aparecido LE, de Souza Rolim G, da Silva Cabral De Moraes JR, Costa CTS, de Souza PS. Machine learning algorithms for forecasting the incidence of Coffea arabica pests and diseases. International Journal of Biometeorology. 2020;64(4):671-88.
[30] Lee WH, Song JW, Yoon SH, Jung JM. Spatial Evaluation of Machine Learning-Based Species Distribution Models for Prediction of Invasive Ant Species Distribution. Appl Sci-Basel. 2022;12(20):19.
[31] Liang L, Li X, Huang Y, Qin Y, Huang H. Integrating remote sensing, GIS and dynamic models for landscape-level simulation of forest insect disturbance. Ecological Modelling. 2017;354:1-10.
[32] Kalaiselvi B, Thangamani M. An efficient Pearson correlation based improved random forest classification for protein structure prediction techniques. Measurement. 2020;162:107885.
[33] Ma Q, Jin-Long G, Guo Y, Guo Z, Lu P, Xiang-Shun H, et al. Prediction of the Current and Future Distributions of the Hessian Fly, Mayetiola destructor (Say), under Climatic Change in China. Insects. 2022;13(11):1052.
[34] Xiao Q, Li W, Kai Y, Chen P, Zhang J, Wang B. Occurrence prediction of pests and diseases in cotton on the basis of weather factors by long short term memory network. BMC Bioinformatics. 2019;20(Suppl 25):688.