RESEARCH ARTICLE

Crop pest occurrence prediction and prevention strategies based on big data analysis

Weina Li[1, *], Liang Qin[2], Dongjie Ge[1]

[1]College of Chemical and Environmental Engineering, [2]College of Electronic and Computer Engineering, Jiaozuo University, Jiaozuo, Henan, China.

Received: March 4, 2025; **accepted:** August 3, 2025.

Precision agriculture has emerged as a critical response to increasing challenges in global food security and crop management, particularly under the pressure of climate change and intensified pest outbreaks. Crop pests such as rice thrips significantly reduce yield and quality, especially in humid and tropical regions. Pest prediction is an important part of precision agriculture, which can help agricultural managers take timely prevention and control measures to improve the yield and quality of crops. Accurate forecasting of pest occurrence remains a major challenge due to the complex interactions among meteorological, agronomic, and soil variables. This study built a prediction model for rice thrips pests through a deep learning method based on meteorological, crop growth, and soil environmental data. The research data covered meteorological data including temperature, humidity, precipitation, wind speed, crop growth data including rice growth stage and crop density, and soil data including soil pH, nitrogen, phosphorus, and potassium content from January 2018 to December 2020. Pearson correlation analysis and variance inflation factor (VIF) analysis were used to identify the features related to the number of rice thrips. Feature selection was performed through Lasso regression, and temperature, humidity, and crop density were selected as the optimal features. The research used a support vector regression (SVR) model for pest prediction and evaluated it on training and test sets. The results showed that the SVR model exhibited high prediction accuracy on both training and test sets with a small mean square error (MSE) and a high coefficient of determination (R^2), which proved the effectiveness of the model in pest prediction. The proposed model achieved high accuracy with a MSE of 0.42 and a R^2 of 0.88 on training set, and a MSE of 0.38 and a R^2 of 0.85 on test set. When applying to 2021 data, the prediction error ranged from 1.2% to 3.1%. Based on predicted pest thresholds, customized prevention and control strategies were implemented, improving early warning efficiency and reducing false alarms. The results confirmed that the proposed pest prediction model could provide scientific decision-making support for agricultural managers and promote the development of precision agriculture.

Keywords: rice thrips; pest prediction; precision agriculture; support vector regression (SVR); feature selection.

[*]**Corresponding author:** Weina Li, College of Chemical and Environmental Engineering, Jiaozuo University, Jiaozuo 454000, Henan, China. Email: LiWeina-lili@outlook.com.

Introduction

Agricultural production is the foundation for human society to maintain food security and economic development. However, with the intensification of global climate change and differences in farmland management levels, the occurrence of agricultural pests has posed a huge threat to the production and quality of crops, which affects not only the growth and yield of crops but also the health of crops by transmitting diseases and even causing serious food shortages in some cases [1]. In particular, the harm of pests in rice production has become increasingly prominent, especially the outbreak of certain pests in rice fields, which can lead to serious yield reductions or even devastating effects [2]. The rice thrip (*Stenotus rubrovittatus*) is a common sucking pest of rice, which is about 2 - 3 mm long, a brightly colored body usually red or orange in adult, and three stages of life cycle including egg, nymph, adult with the adult and nymph being the main stages of damage. The adult sucks the young leaves and panicles of rice plants, while the nymph mainly parasitizes on the back of rice leaves [3]. These sucking behaviors lead to the loss of plant cell fluid, which in turn causes the leaves to turn yellow and wither, ultimately affecting the growth and development of rice. The sucking of rice thrips can also reduce the disease resistance of plants and provide a transmission pathway for other diseases such as rice false smut and rice blast. In addition to direct physical damage, the saliva of rice thrip contains toxins, which destroy the normal function of plant cells and make plants more vulnerable [4]. The aggregation and reproduction of nymphs further aggravate the damage caused by the pest. Studies have shown that rice thrips have strong adaptability to environmental conditions, especially under conditions of high temperature and high humidity, where they reproduce quickly and the insect population density increases rapidly, causing serious damage [5]. In recent years, with the intensification of climate change, the distribution area of rice thrip has continued to expand, posing greater challenges to global rice production [6]. The insecticide resistance of rice thrip has gradually increased, and traditional chemical control methods are no longer effective, forcing agricultural producers to find new control methods. At present, control strategies for rice thrips include chemical control, biological control, and physical control [7]. However, due to its high reproductive capacity, strong adaptability, and concealment, existing control methods often face problems such as unstable effects and high costs. Therefore, studying the ecological behavior and environmental adaptability of rice thrip and exploring accurate prediction and control strategies based on big data have become research hotspots in agricultural pest management [8].

With the rapid development of big data technology, data-driven pest prediction has become an important direction of modern agricultural research [9]. Big data analysis can reveal the relationship between the occurrence patterns of pests and environmental factors by integrating and deeply mining meteorological data, pest monitoring data, crop growth data, *etc*. [10]. Comparing to the traditional experience-based pest early warning system, big data analysis provides a more accurate and dynamic prediction model that can identify the occurrence timing, spatial distribution, and occurrence intensity of pests in advance, thereby providing farmers with timely prevention and control suggestions [11]. Studies have shown that pest occurrence is often affected by multiple environmental factors such as climatic conditions including temperature, humidity, and precipitation, as well as agricultural management factors including crop growth conditions and planting density. Changes in these factors may lead to the outbreak or spread of pests. Therefore, how to construct an accurate pest prediction model through big data analysis of multi-source data has become an important challenge facing agricultural scientists [12]. Traditional pest prediction methods such as statistical models based on meteorological data and phenological models can reflect the trend of pest occurrence to a certain extent, but their accuracy and adaptability are limited. In recent years, pest prediction research based on machine learning, deep learning, and other algorithms has made significant progress. In the prediction of rice field pests, researchers have successfully predicted the outbreak period of rice field pests by establishing a regression model based on meteorological conditions and historical pest data [13]. However, the application of big data analysis in agricultural pest prediction still faces many challenges such as data quality, diversity of data sources, and generalization ability of the model. Therefore, how to build a pest prediction model that is suitable for different regions and environments and improve the prediction accuracy of the model is still a hot topic and difficulty in current research [14].

The prevention and control of rice thrips has always been one of the difficulties in rice pest management. Traditional prevention and control methods mainly rely on chemical pesticides, but as pests become more resistant to pesticides, the effectiveness of pesticides is declining. Scholars have begun to explore more diverse prevention and control strategies mainly including new methods such as biological control, physical control, and ecological control. Biological control is an important method of using natural enemies to control the number of the rice thrip. Studies found that the natural enemies of rice thrip included predatory insects such as ladybugs and spiders and parasitic insects such as parasitic wasps. These natural enemies inhibited the expansion of rice thrip populations by preying on or parasitizing the eggs, nymphs or adults of rice thrip [15]. The advantage of biological control is its sustainability and ecological safety, but its implementation effect is usually restricted by environmental factors and the number of natural enemy populations. Physical control mainly includes the use of trapping devices, light trapping, and sonic insect repellent. Studies showed that the use of sex attractants and light sources to trap adult rice thrip could effectively reduce insect population density, while avoiding the use of pesticides [16]. In recent years, the development of Internet of Things technology and remote sensing technology has provided new opportunities for pest monitoring and control. By setting up sensors and monitoring equipment, the dynamics of rice thrip can be tracked in real time, providing data support for pest control. Ecological control methods include changing rice cultivation patterns using insect-resistant varieties and rational crop rotation. By optimizing the agricultural ecological environment, the insect resistance of rice can be enhanced, and the possibility of pest occurrence can be reduced. These methods not only effectively reduce the risk of pest occurrence but also help achieve sustainable agricultural development [17]. Paddy fields are the main ecological environment for rice cultivation. The humid climate and special cultivation methods make rice an ideal habitat for many pests. As one of the common pests in rice cultivation, the rice thrip poses a great threat to rice with its strong reproductive ability and wide adaptability. Especially in a hot and humid environment, the occurrence of rice thrip not only directly affects the growth of rice, but may also cause secondary diseases, seriously reducing the yield and quality of rice. Therefore, studying the biological characteristics and damage mechanism of rice thrip and exploring how to effectively prevent and control its spread have become important topics in the field of agricultural research [18]. Currently, the research on rice thrip mainly focuses on its biological characteristics, ecological habits, and control strategies. The studies on biological characteristics revealed the reproduction cycle, ecological habits, and adaptability of rice thrip to different environmental conditions through several field surveys and laboratory studies. The control strategies research found that traditional control methods mainly relied on pesticide spraying, but due to the increasing resistance of rice thrip to pesticides, the effectiveness of this method had been weakened [19]. In recent years, biological control and physical control methods have gradually been proposed and achieved certain results including reducing the number of thrip by releasing natural enemies and installing physical barriers.

There are shortcomings in existing research, which include that there is no unified research conclusion on the climate adaptability and biological characteristics of rice thrip, especially the ecological differences in different regions and different cultivation modes. Although traditional control measures can control pests to a certain extent, due to the strong concealment and rapid reproduction of rice thrips, the control effect is usually not ideal [20]. This study proposed a pest prediction model based on big data analysis and corresponding prevention and control strategies by collecting data of meteorological, pest monitoring, rice growth and analyzing the relationship between the occurrence law of rice thrip and environmental factors. Machine learning algorithms were used to predict the occurrence time and intensity of pests including precision application, automated monitoring, and real-time feedback mechanism to reduce the use of pesticides and improve the efficiency of prevention and control. The actual effects of the proposed prediction model and prevention and control strategy were verified through field experiments to provide scientific data support and decision-making basis for agricultural production and promote the development of precision agriculture. This study provided a new perspective for pest ecology, provided farmers with scientific prevention and control decisions, and promoted the combination of big data and intelligent agricultural technology for sustainable agricultural development.

Materials and methods

Environmental and agronomic factors

Factors that affect the occurrence of rice thrips pests include temperature, humidity, precipitation, soil conditions, rice growth stage, crop density, and agricultural management measures. Each factor has different degrees of influence on the occurrence of pests. Temperature is a key factor affecting the growth and reproduction of rice thrips. Under high temperature conditions, the activity range of rice thrips expands, and the reproduction rate accelerates. In agricultural research, the degree-day (DD) model is often used to quantify the cumulative effect of temperature on insect growth and is calculated as below [21].

$$DD = \sum_{i=1}^{n} \left(\frac{T_{\max,i} + T_{\min,i}}{2} - T_{\text{base}} \right) \quad (1)$$

where $T_{\max,i}$ and $T_{\min,i}$ are the highest and lowest temperatures of the day. T_{base} is the reference temperature usually 10°C. n is the number of days for calculating the accumulated temperature. Through the accumulation of temperature, the effect of temperature on the reproduction of rice thrips can be quantitatively evaluated. Humidity is also critical to the growth and reproduction of rice thrips. Higher humidity can provide a more suitable habitat, promoting its reproduction and survival. Humidity is usually measured by relative humidity (RH) and can be calculated as follows [22].

$$H_{\text{rel}} = \frac{E_{\text{actual}}}{E_{\text{saturation}}} \times 100 \quad (2)$$

where E_{actual} and $E_{\text{saturation}}$ are the actual partial pressure of water vapor in the air and the saturated water vapor pressure. Too low or too high humidity will affect the survival and reproduction of rice thrips, so relative humidity (H_{rel}) is needed for quantitative analysis. Precipitation is also an important factor affecting rice thrips pests. Appropriate precipitation helps maintain the water supply required for rice growth and can affect the humidity of rice fields, thereby affecting the occurrence of pests. Precipitation can be quantified by daily precipitation as shown below.

$$P_{\text{daily}} = \sum_{i=1}^{n} P_i \quad (3)$$

where P_i is the precipitation, which not only directly affects the growth conditions of rice, but also affects the habitat of rice thrips by regulating soil moisture. Soil conditions are also one of the key factors affecting the occurrence of pests. The pH value, fertility, and moisture content of the soil directly affect the growth of rice, which indirectly affects the habitat of rice thrips. Soil conditions affect the growth stage and development speed of rice, which are closely related to the occurrence of pests. The quality, nutrient content, and moisture level of the soil will affect the number and distribution of rice thrips. There are significant differences in the suitability of different growth stages to rice thrips. Rice is relatively fragile during the tillering and jointing stages and is more susceptible to rice thrips. When rice enters the heading stage, the vegetation structure and physiological characteristics of the rice field change, and the survival and reproduction of rice thrips are inhibited. Therefore, the rice growth stage should be considered as an important feature in the pest prediction model. In addition to environmental and crop factors, agricultural management measures also play a vital role in the occurrence of pests. Factors such as crop density, irrigation methods, fertilization, and pesticide use directly affect the ecological environment of rice fields, which in turn affects the occurrence of rice thrips. Through reasonable agricultural management measures, the occurrence of pests can be effectively controlled.

Multicollinearity and feature independence analysis

In the process of predicting pest occurrence, the independent analysis between factors is crucial. If there is a strong correlation between features, it may cause multicollinearity problems in the model, thus affecting the reliability of the prediction results. The Pearson correlation coefficient and variance inflation factor was used in this study to evaluate the independence between features. The Pearson correlation coefficient is a common method to measure the linear correlation between two variables and is expressed as follows.

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{4}$$

By calculating the correlation coefficients between different environmental and crop factors, the factors that have significant linear relationships can be determined. If the correlation coefficients of some factors are high, those factors are redundant to a certain extent and can be removed or merged. The variance inflation factor (VIF) is a statistic used to assess multicollinearity. The larger the VIF value, the higher the correlation between the feature and other features. The VIF can be calculated below.

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2} \tag{5}$$

where R_j^2 is the coefficient of determination obtained by regressing a feature with other features. If the VIF value is greater than 10, it is generally considered that the feature has serious multicollinearity and may need to be adjusted in the model. Through independence analysis, the redundant relationships between features can be identified, and the appropriate features can be selected for modeling to improve the prediction accuracy of the model.

Feature selection

The goal of feature selection was to select the most important features for the occurrence of pests from many factors. To improve the accuracy and generalization ability of the model, chi-square test, Lasso regression, and random forest were applied to screen features. The Chi-square test was used to evaluate the relationship between agricultural management measures and pest occurrence and was calculated as follows.

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{6}$$

If the chi-square test result showed some characteristics that were significantly correlated with the occurrence of pests, these characteristics would be selected into the final model. The objective function of Lasso regression was shown below.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^{n} (y_i - X_i \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \tag{7}$$

By adjusting the regularization parameter λ , Lasso regression could effectively select the features that had the most significant impact on the occurrence of insect pests. Random forest is an ensemble learning method based on decision trees, which makes predictions by building multiple decision trees and combining their predictions. The importance of features was evaluated by calculating the contribution of each feature to the accuracy of the decision tree.

Predictive modeling approach

After feature selection, the support vector machine (SVM) model was used to predict the occurrence probability of rice thrips pests. The core of SVM regression is to find an optimal regression hyperplane that minimizes the error between the predicted value and the actual value. The objective function of the SVM regression model was shown below.

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \quad (8)$$

where $K(x_i, x)$ was the kernel function. α_i was the Lagrange multiplier. b was the bias term. In practical applications, the Gaussian radial basis kernel (RBF) function was usually used as the kernel function and was calculated as follows.

$$K(x_i, x) = \exp\left(-\gamma \| x_i - x \|^2\right) \quad (9)$$

By training the SVM regression model, the prediction value of each feature for the occurrence of pests could be obtained and provide a scientific basis for agricultural management.

Integration of early warning mechanism

In modern agricultural production, timely warning of pests is crucial to ensure crop yield and quality. The pest prediction model based on deep learning, spatial analysis, and meteorological data can provide farmers with accurate warning information. To achieve this goal, an early warning system was embedded in the existing pest prediction model. By combining spatial analysis, meteorological data, pest history records, and other information, the potential risks of pests could be timely identified and provide corresponding prevention and control measures. Pest prediction usually relies on multiple meteorological variables and historical data on pest occurrence. By establishing a pest occurrence probability model based on meteorological and environmental factors, the probability of pest occurrence in a certain area can be estimated.

$$P(t) = f(T(t), H(t), R(t), L(t)) \quad (10)$$

where $P(t)$ was the probability of pest occurrence at time t . $T(t)$ was the temperature. $H(t)$ was the humidity. $R(t)$ was the amount of precipitation. $L(t)$ was the influencing factor of historical pest records. Function f was a mapping relationship obtained by training a machine learning model. Once the probability of pest occurrence was predicted, the risk threshold $P_{\text{threshold}}$ could be set up to determine whether an early warning needed to be initiated. The specific early warning judgment conditions could be expressed as feed to colleagues below.

$$\text{Alert} = \begin{cases} \text{True} & \text{if } P(t) > P_{\text{threshold}} \\ \text{False} & \text{if } P(t) \leq P_{\text{threshold}} \end{cases} \quad (11)$$

If $\text{Alert} = \text{True}$, the system would notify farmers of high-risk areas for pests. The early warning system also provided farmers with corresponding prevention and control measures, which were intelligently recommended based on the severity of the pests, the time of occurrence, and environmental conditions. A prevention and control strategy recommendation model based on risk level was then designed. When the probability of pest occurrence and the pest intensity were high, the system could recommend the use of pesticides for prevention and control.

$$I(t) = w_1 \cdot T(t) + w_2 \cdot H(t) + w_3 \cdot R(t) + w_4 \cdot L(t) \quad (12)$$

where $I(t)$ was the pest intensity. w_1, w_2, w_3, w_4 were the weight coefficient obtained by model learning. When the pest intensity value was high, the system would recommend spraying pesticides to suppress pests.

$$\text{Pesticide\_use} = \begin{cases} \text{True} & \text{if } I(t) > I_{\text{threshold}} \\ \text{False} & \text{if } I(t) \leq I_{\text{threshold}} \end{cases} \quad (13)$$

When pest intensity was low and climatic conditions were suitable, biological control might be an effective alternative. The system could recommend appropriate biological control measures based on pest occurrence patterns, crop types, and climatic conditions as follows.

$$\text{Biological\_control} = \begin{cases} \text{True} & \text{if condition=favorable for biological control} \\ \text{False} & \text{if not favorable} \end{cases} \quad (14)$$

Some environmentally friendly non-chemical control measures could also be recommended such as physical control, reasonable crop rotation, *etc*. When the system found that pests occurred in the early stages of crop growth and climatic conditions were suitable, physical control through crop rotation or the use of barrier nets could be recommended. To improve the reliability and accuracy of the system, a circular feedback mechanism was adopted to continuously optimize the model parameters according to the actual pest occurrence situation to reduce false alarms and missed alarms. By continuously collecting information such as farmland monitoring data, meteorological changes, and pest control effects, the thresholds of the prediction model and prevention and control recommendations could be adjusted in real time.

Study area, data collection and processing

The data used for this study were collected from Zhaoqing, Guangdong, China. Zhaoqing is situated in the subtropical monsoon climate zone, characterized by high humidity and abundant rainfall with the average annual temperature approximately 22.5°C and the peak temperatures and pest activity typically occurring between June and September. All meteorological data, crop growth data, pest monitoring records, and soil parameters were sourced from local agricultural and meteorological monitoring systems with data spanning from January 2018 to December 2020. Meteorological variables including daily temperature, humidity, precipitation, and wind speed were retrieved from Guangdong Meteorological Data Service Platform (Guangdong Meteorological Bureau, Guangzhou, Guangdong, China). Crop growth data including rice development stages and plant density, as well as pest population data for rice thrips were sourced from the Guangdong Agricultural Technology Extension Center (Department of Agriculture and Rural Affairs of Guangdong Province, Guangzhou, Guangdong, China). Soil data including pH, nitrogen, phosphorus, potassium, and organic matter content were collected through the Guangdong Soil and Fertilizer Station (Guangdong Agricultural Resources Monitoring and Management Center, Guangzhou, Guangdong, China). Four types of data were collected including meteorological conditions, crop growth status, pest monitoring, and soil characteristics. In addition, pest monitoring data were collected daily, specifically recording the population of rice thrips, which served as the core target variable in model training and validation. Soil data were gathered monthly and included soil pH, nutrient concentrations of nitrogen, phosphorus, potassium, and organic matter content. The daily frequency of meteorological, crop, and pest data ensured high temporal resolution, allowing for precise correlation analysis between environmental conditions and pest outbreaks. Meanwhile, the monthly soil data added valuable contextual depth. The integration of these diverse but interrelated datasets formed the basis for a robust and dynamic pest prediction framework, supporting accurate modeling of rice thrips behavior under varying agroecological conditions. To better understand the overall distribution and variability of the datasets, a box plot was constructed for the four major categories of data. During data preprocessing, missing values were removed, and all continuous variables were standardized to bring them onto the same scale, which ensured comparability

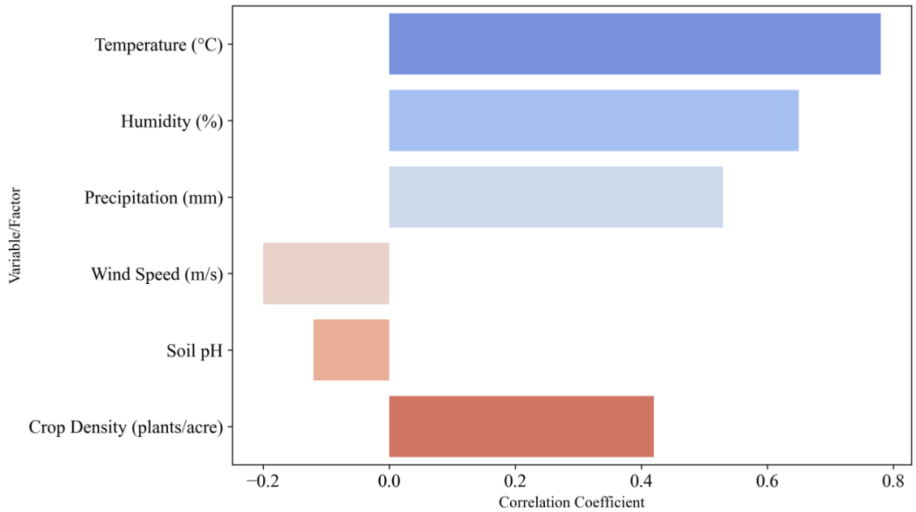

Figure 1. Correlation between environmental factors and pest count.

across different data types and improved model convergence during training. The visual distribution analysis confirmed the necessity of robust feature selection and normalization techniques to handle heterogeneous data characteristics before applying machine learning algorithms.

Statistical analysis

All statistical analyses and modeling tasks were conducted by using Python 3.9.13 (https://www.python.org/) and supported by open-source packages including scikit-learn 1.1.3 for Lasso regression, support vector regression (SVR), pandas, numpy, and matplotlib for data handling and visualization. All software tools were executed using Jupyter Notebook under the Anaconda distribution 2022.10 (Anaconda, Austin, Texas, USA).

Results and discussion

Correlation analysis

The linear relationship between each feature and the number of rice thrips was obtained using Pearson correlation analysis. The results demonstrated that the correlation coefficient between temperature and the number of rice thrips was 0.78, showing a strong positive correlation and indicating that rising temperature might promote the occurrence of pests. Humidity and precipitation were also positively correlated with the number of pests, but their correlations were slightly weaker than temperature at 0.65 and 0.53, respectively. The relationship between wind speed and the number of pests was more complicated with a correlation coefficient of -0.20, indicating that an increase in wind speed would have a certain inhibitory effect on the number of pests. The correlation between soil pH and the number of pests was low with only -0.12, indicating that the acidity and alkalinity of the soil had little effect on the occurrence of pests. The correlation coefficient between crop density and the number of pests was 0.42, indicating that, when the crop density was high, the possibility of pest occurrence also increased accordingly (Figure 1). Through this correlation matrix, the potential impact of various environmental and crop factors on pests were identified, which provided a reference for subsequent feature selection and model training.

Independence analysis

The variance inflation factor (VIF) values of different features showed that wind speed and soil pH had relatively low VIF values of 1.5 and 1.4, respectively, indicating that both factors

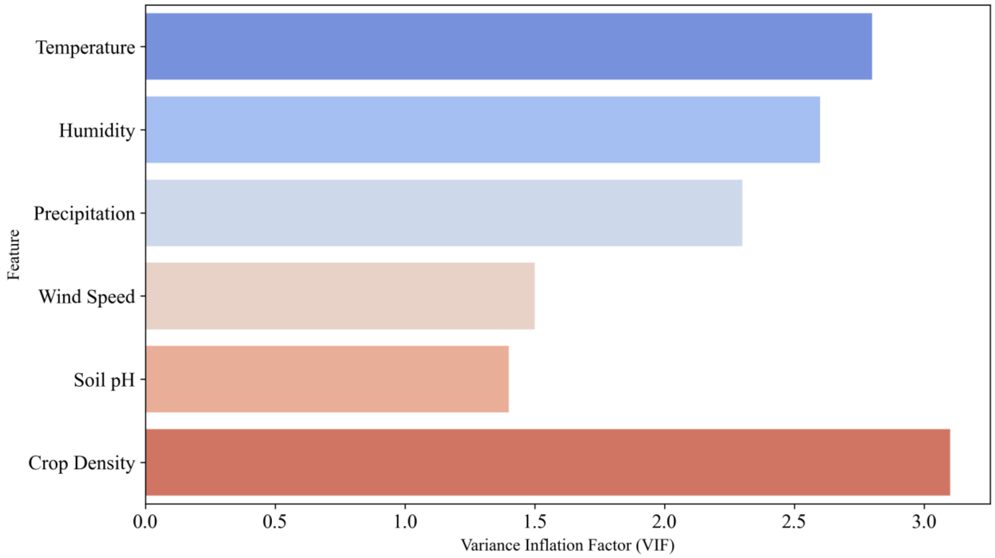

Figure 2. The variance inflation factor (VIF) values of different features.

were less correlated with other features and contributed more to the independence of the model (Figure 2). VIF is a common indicator for measuring multicollinearity between independent variables in a regression model, reflecting the degree of correlation between a feature and other features. The higher the VIF value, the stronger the correlation between the feature and other features, which may introduce multicollinearity problems and affect the stability and explanatory power of the regression model. Through VIF analysis, features that had high collinearity were identified and removed. The stability and prediction accuracy of the model were optimized, which improved the reliability and interpretability of the model.

Feature selection and model building

A total of 13 environmental and agronomic variables were evaluated under different regularization strengths as λ being set at 0.01, 0.1, 1, 10 to determine their relative importance in predicting rice thrips occurrence. Temperature, humidity, precipitation, and crop density consistently maintained higher coefficient values across lower λ levels, indicating strong predictive power. Variables such as wind speed, soil pH, nitrogen, phosphorus, potassium content, leaf area index, rice growth stage, and field humidity were categorized as weaker features as their coefficients declined rapidly with increasing regularization and were eventually shrunk to zero. Wind direction was consistently excluded with its coefficient remaining zero across all λ levels, indicating negligible influence on pest occurrence. Although soil organic matter content showed minor influence at lower λ values, it was also classified as a weaker feature due to low coefficient magnitude. The selection process highlighted that temperature, humidity, precipitation, and crop density played a dominant role in pest prediction, while other environmental and soil variables had limited independent contribution when these key factors were already included. The feature selection results of Lasso regression under different regularization parameter values automatically selected the most important features through penalty terms. The results showed that the coefficients of temperature, humidity, precipitation, and crop density were large, indicating that these features had a strong effect in predicting pests. When the values of these factors increased, the coefficients of other features gradually decreased, and finally only temperature, humidity, precipitation, and crop density were retained. Therefore, these four variables were ultimately identified as optimal

Table 1. Prediction of pest population for 2021.

Date	Predicting pest population	Actual pest population	Error
June 1, 2021	1,200	1,185	1.3%
July 1, 2021	1,450	1,405	3.1%
August 1, 2021	1,800	1,750	2.8%
September 1, 2021	2,000	2,025	1.2%

features. Through Lasso regression, the model could effectively reduce redundant features and the complexity of the model, thereby improve prediction accuracy and avoid overfitting.

Model training and evaluation

After selecting optimal features, the model was trained and evaluated using both training and test datasets. The model demonstrated strong predictive accuracy across both sets. On the training dataset, the mean squared error (MSE) was 0.42, and the coefficient of determination (R^2) reached 0.88, indicating an excellent fit between predicted and actual pest counts. When tested on unseen data, the MSE was slightly reduced to 0.38, and the R^2 was 0.85, confirming the model's ability to generalize effectively to new inputs. The results suggested that the model not only avoided overfitting but also maintained stability when applied to real-world data. The combination of low MSE and high R^2 across both datasets highlighted the robustness of the proposed approach in modeling complex, nonlinear interactions between multiple environmental variables and pest behavior. By accurately predicting pest population dynamics, the SVR model provided a reliable decision-support tool for agricultural managers to implement timely and targeted prevention measures. This capability is essential for minimizing crop losses and optimizing pest control strategies under varying climate and cultivation conditions.

Model application and effect evaluation

The trained SVR model was applied to the 2021 pest prediction. The results showed that the model accurately predicted the number of pests in different months with a small error range from 1.2 to 3.1% (Table 1). These results suggested that the SVR model could effectively capture the changing trend of the pests number, and the prediction error was maintained within a reasonable range, which proved the practical application ability of the model. Through the prediction results, agricultural managers can take corresponding prevention and control measures in advance according to the predicted number of pests, thereby reducing the impact of pests on crops.

Prevention and control decision support

Based on the pest prediction results, scientific prevention and control decisions for agricultural management can be provided. Control strategies of different intensities were formulated according to the number of pests predicted by the model. When the predicted number of pests was between 0 - 500, it was recommended to conduct routine monitoring and moderate fertilization, and the control effect was relatively mild. When the number of pests was between 1,000 – 1,500, it was recommended to apply pesticides in advance and irrigate moderately, and the control effect was relatively severe. However, when the number of pests exceeded 1,500, the pest control measures should be strengthened, and enhanced monitoring and emergency measures should be conducted (Table 2). These control measures can help agricultural managers respond promptly according to pest prediction data, thereby achieving refined management and improving crop yield and quality. The predictive performance of the pest detection model was further evaluated using a confusion matrix, which provided a detailed view of its classification accuracy in distinguishing between pest occurrence and non-occurrence. Based on the results, the model correctly identified 250

Table 2. Protection measures based on prediction values.

Predicted pest population	Recommended prevention and control measures	Estimation of control effect
0 - 500	Regular monitoring and appropriate fertilization	Minor effects, no special intervention required
500 – 1,000	Increase pest monitoring frequency	Moderate impact, prevention and treatment recommended
1,000 - 1,500	Apply pesticides in advance and irrigate appropriately	Serious impact, necessary prevention and treatment
> 1,500	Strengthen pest control and monitoring	Extreme impact, urgent treatment required

instances where pests were present and 660 cases where pests did not occur using the confusion matrix. However, there were also 50 false negatives that were the cases with pests occurring but not detected by the model and 40 false positives that were the cases with the model incorrectly predicting pest presence but nothing happening. These results indicated that the model had a strong ability to differentiate positive and negative classes with a high number of true positives and true negatives. While the overall accuracy is commendable, the presence of false negatives suggested a risk of underestimating pest outbreaks, which might delay timely interventions. Similarly, false positives could lead to unnecessary prevention measures. The current confusion matrix highlighted the importance of further refining model parameters and thresholds to reduce misclassification. Enhancing precision and recall will ensure the model offers more reliable support for pest monitoring and control decisions in agricultural settings. By optimizing the model, the false positive and false negative rates can be further reduced, thereby improving the reliability of the prediction.

Early warning system evaluation

The evaluation results of the model under different risk thresholds including accuracy, recall, precision, F1 score, false positive rate, and false negative rate demonstrated that, as the threshold increased, the accuracy of the model increased, but the recall rate decreased. When the risk threshold was set to 0.3, the accuracy, recall rate, and precision rate were 0.85, 0.92, and 0.78, while, when the risk threshold was set to 0.9, the accuracy, recall rate, and precision rate were 0.87, 0.76, and 0.89, respectively (Figure 3). The evaluation results under different thresholds provided a basis for model tuning. The most appropriate threshold can be selected according to the specific application scenario, thereby balancing the accuracy and recall rate and maximizing the prediction efficiency of the model. Comparison of evaluation indicators on different datasets including 2022, 2023, tropical climate region, and the temperate climate region showed that the model on the 2022 dataset performed best with an accuracy of 0.91, a recall of 0.85, and an F1 score of 0.86, while the performance on the 2023 dataset declined slightly with an accuracy of 0.89, a recall of 0.82, and an F1 score of 0.83 (Figure 4). The evaluation results for the tropical climate region and the temperate climate region were also good, but there were certain regional differences, which might be affected by climate and environmental factors. These results showed that there were certain differences in the performance of the model under different datasets and environmental conditions. In the future, the model can be tuned for specific regions to improve its prediction ability in different scenarios.

Conclusion

This study proposed a prediction model for rice thrips pest based on multidimensional environmental data. Through in-depth analysis of

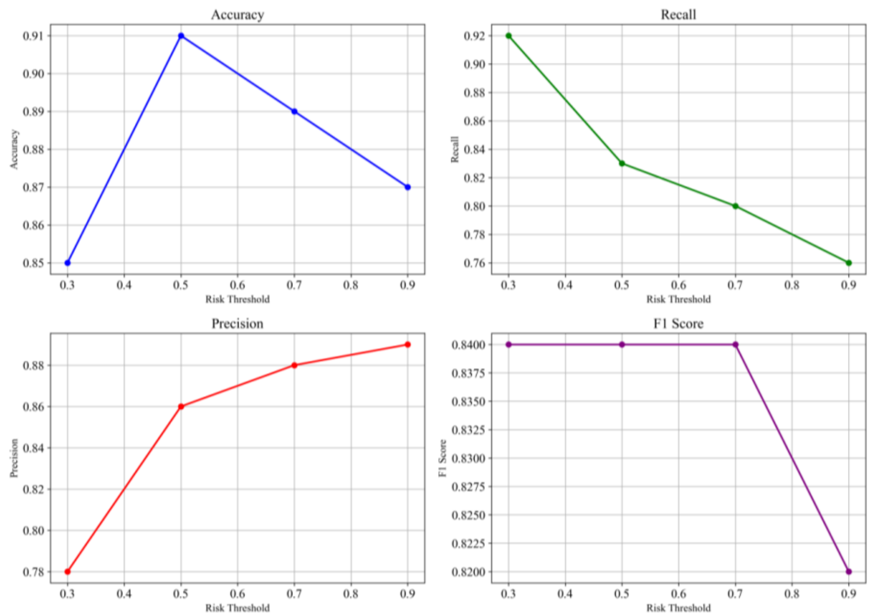

Figure 3. Evaluation under different risk thresholds.

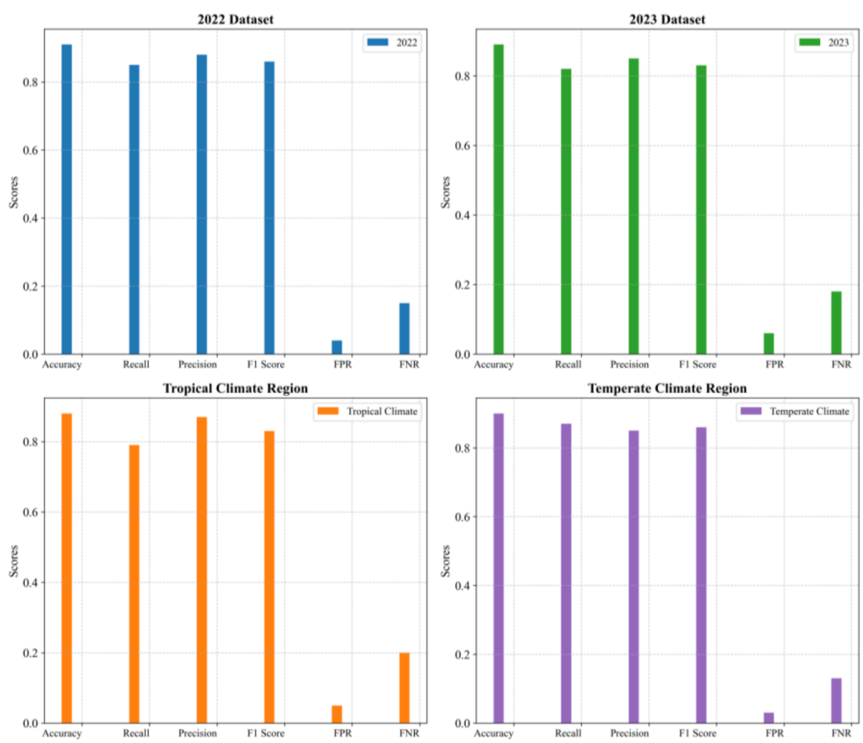

Figure 4. Comparison of indicators evaluation on different datasets.

meteorological, crop growth, and soil data, a model that could accurately predict the occurrence of pests was successfully constructed. The results showed that temperature, humidity, precipitation, and crop density were important factors affecting the number of rice thrips. The proposed model demonstrated high prediction ability on both the training and the test datasets, proving that the model could accurately capture the nonlinear relationship between environmental characteristics and pest occurrence. Through the prediction of pest data in 2021, the model showed a small error and provided timely decision-making support for agricultural management. According to different pest prediction results, scientific and reasonable prevention and control measures were formulated to ensure the healthy growth and high yield of crops. In addition, the evaluation under different risk thresholds showed the performance changes of the model under different circumstances. The high accuracy and low false positive rate showed that the model had strong adaptability and reliability under different environmental conditions. When faced with data from different climate regions, the model was able to effectively predict the probability of pest occurrence, proving its broad application prospects. Although the proposed pest prediction model has high accuracy and stability, there is still some room for improvement. The input features of the model can be further expanded to combine more agricultural management data and historical pest data to improve the prediction accuracy of the model. Meanwhile, the performance of the model can be further improved in the future through integrated learning methods or deep learning technology, especially when dealing with more complex climate change and crop growth conditions, deep learning may bring more significant advantages.

Acknowledgements

The study was supported by the Key Research Project of Higher Education Institutions in Henan Province in 2024 (Grant No. 24A180016).

References

1. Tonnang HEZ, Salifu D, Mudereri BT, Tanui J, Espira A, Dubois T, *et al*. 2022. Advances in data-collection tools and analytics for crop pest and disease management. Curr Opin Insect Sci. 54:100964.
2. Chen M, Liu XD. 2023. Estimating insect pest density using the physiological index of crop leaf. Front Plant Sci. 14:1152698.
3. Aparecido LED, Rolim GD, De Moraes J, Costa CTS, de Souza PS. 2020. Machine learning algorithms for forecasting the incidence of *Coffea arabica* pests and diseases. Int J Biometeorol. 64(4):671-688.
4. Tsai MF, Lan CY, Wang NC, Chen LW. 2023. Time series feature extraction using transfer learning technology for crop pest prediction. Agronomy. 13(3):792.
5. Liu KG, Mu YM, Chen XH, Ding ZH, Song M, Xing D, *et al*. 2022. Towards developing an epidemic monitoring and warning system for diseases and pests of hot peppers in Guizhou, China. Agronomy. 12(5):1034.
6. Chen P, Xiao QX, Zhang J, Xie CJ, Wang B. 2020. Occurrence prediction of cotton pests and diseases by bidirectional long short-term memory networks with climate and atmosphere circulation. Comput Electron Agric. 176:105612.
7. Reynolds M, Kropff M, Crossa J, Koo J, Kruseman G, Milan AM, *et al*. 2018. Role of modeling in international crop research: Overview and some case studies. Agronomy. 8(12):291.
8. Mangala RR, Padmapriya A. 2019. Prediction based agro advisory system for crop protection. In: Proc Int Conf Intell Data Commun Technol Internet Things (ICICI). Aug 7-8. Coimbatore, India.
9. Emon MMH, Rabbani GMMDN, Nath A. 2023. Challenges and opportunities in the implementation of big data analytics in management information systems in Bangladesh. Acta Informatica Malaysia. 7(2):122-130.
10. Sharma RP, Ramesh D, Pal P, Tripathi S, Kumar C. 2022. IoT-enabled IEEE 802.15.4 WSN monitoring infrastructure-driven fuzzy-logic-based crop pest prediction. IEEE Internet Things J. 9(4):3037-3045.
11. Saleem RM, Bashir RN, Faheem M, Haq MA, Alhussen A, Alzamil ZS, *et al*. 2023. Internet of things based weekly crop pest prediction by using deep neural network. IEEE Access. 11:85900-85913.
12. Lee S, Yun CM. 2023. A deep learning model for predicting risks of crop pests and diseases from sequential environmental data. Plant Methods. 19(1):8.
13. Balduque-Gil J, Lacueva-Pérez FJ, Labata-Lezaun G, del-Hoyo-Alonso R, Ilarri S, Sánchez-Hernández E, *et al*. 2023. Big data and machine learning to improve European grapevine moth (*Lobesia botrana*) predictions. Plants. 12(3):633.
14. Alexandridis N, Marion G, Chaplin-Kramer R, Dainese M, Ekroos J, Grab H, *et al*. 2021. Models of natural pest control: Towards predictions across agricultural landscapes. Biol Control. 163:104761.
15. Nadeem RM, Jaffar A, Saleem RM. 2022. IoT and machine learning based stem borer pest prediction. Intell Autom Soft Comput. 31(3):1377-1392.
16. Talaat FM. 2023. Crop yield prediction algorithm (CYPA) in precision agriculture based on IoT techniques and climate changes. Neural Comput Appl. 35(23):17281-17292.
17. Su ZB, Luo JQ, Wang Y, Kong QM, Dai BS. 2023. Comparative study of ensemble models of deep convolutional neural

networks for crop pests classification. Multimed Tools Appl. 82(19):29567-29586.

18. Latif MS, Kazmi R, Khan N, Majeed R, Ikram S, Ali-Shahid MM. 2022. Pest prediction in rice using IoT and feed forward neural network. KSII Trans Internet Inf Syst. 16(1):133-152.
19. Sharma RP, Dharavath R, Edla DR. 2023. IoFT-FIS: Internet of farm things based prediction for crop pest infestation using optimized fuzzy inference system. Internet Things. 21:100658.
20. Mittal M, Gupta V, Aamash M, Upadhyay T. 2024. Machine learning for pest detection and infestation prediction: A comprehensive review. Wiley Interdiscip Rev Data Min Knowl Discov. 14(5):e1551.
21. Attri I, Awasthi LK, Sharma TP. 2024. Machine learning in agriculture: A review of crop management applications. Multimed Tools Appl. 83(5):12875-12915.
22. Wadhwa D, Malik K. 2024. A generalizable and interpretable model for early warning of pest-induced crop diseases using environmental data. Comput Electron Agric. 227:109472.