Big Data and Information Analytics

https://www.aimspress.com/journal/BDIA

BDIA, 9: 29–47.
DOI:10.3934/bdia.2025002
Received: 25 March 2025
Revised: 19 May 2025
Accepted: 01 June 2025
Published: 18 June 2025

Research article

Application of XGBoost model and multi-source data for winter wheat yield prediction in Henan Province of China

Hua Li[1,2,*], Jingyi Gao[1], Yadan Guo[2] and Xianzhi George Yuan[3]

[1] School of Mathematics and Statistics, Zhengzhou University, Zhengzhou 450001, China
[2] Henan Key Lab of Digital Finance, Zhengzhou University, Zhengzhou 450001, China
[3] Business School, Chengdu University, Chengdu 610106, China

* **Correspondence:** Email: huali08@zzu.edu.cn.

Abstract: Accurate prediction of winter wheat yield is essential for precision agricultural management. Traditional methods, which primarily rely on vegetation indices, often fall short in achieving high prediction accuracy. In this study, we proposed an innovative approach by integrating multi-source data, including vegetation, geographical, and soil features, with 61 features extracted across growth stages. Using these features, the XGBoost regression model was applied to predict winter wheat yield in Henan Province, China. The model was trained on data from 2016 to 2020 and validated using the 2021 dataset. The XGBoost model achieved a mean absolute error (MAE) of 371.36 kg/hm², a root mean square error (RMSE) of 516.97 kg/hm², and a coefficient of determination (R^2) of 0.85. These results significantly outperformed other models, such as random forest (RF), gradient boosting decision tree (GBDT), and Lasso. Notably, the XGBoost model provided high accuracy 2–3 growth stages before harvest, enabling earlier yield prediction than traditional methods. To enhance interpretability, SHAP (shapley additive explanations) was employed to quantify the influence of each input variable on yield and determine the direction of influence. This study underscores the potential of multi-source data and advanced machine learning techniques for accurate and interpretable winter wheat yield prediction, providing a valuable solution for precision agriculture.

Keywords: yield prediction; winter wheat; multi-source data; XGBoost model; SHAP explainable framework

1. Introduction

Ensuring grain security is a critical strategic issue for economic development and social stability, and it serves as a cornerstone of national security. For China, a country with a large population, the level of agricultural development and grain production capacity directly impact the national economy and

people's livelihoods [1]. Consequently, accurate monitoring and prediction of grain yields are essential for formulating national and regional socio-economic development plans. They not only aid in the development of agricultural import and export strategies but also ensure national grain security and provide a vital basis for guiding and regulating the macro-level structure of the planting industry [2].

Traditional grain yield prediction methods typically rely on field surveys conducted during the growing season or empirical knowledge of grain growth conditions. However, these approaches are often limited by small sample sizes and constraints in human resources, leading to sampling and non-sampling errors that can compromise the reliability of the results. In contrast, satellite remote sensing, with its extensive coverage and short revisit cycles, has emerged as a critical technological tool for large-scale grain yield prediction. Vegetation indices (VIs) derived from satellite data are the most commonly used means for predicting grain yields, with hundreds of VIs currently available. These VIs can be categorized into three types based on their calculation formulas and functional requirements: Simple VIs, modified VIs, and functional VIs. Here, we list different VIs and their related literature in Table 1, showing that the normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI), based on visible and near-infrared (NIR) bands, are the most widely used. However, studies have shown that other indices may outperform these traditional VIs in specific contexts. For example, Zhang et al. [3] found that the green chlorophyll vegetation index (GCVI) outperformed other VIs, although this result was not validated in this study. Bolton et al. [4] demonstrated that the enhanced vegetation index (EVI2) is the best indicator for predicting maize yield in non-semi-arid regions. Qu et al. [5] pointed out that the green normalized difference vegetation index (GNDVI) is highly correlated with nitrogen and is suitable for characterizing canopy biomass during the heading and tasseling stages of crops. Additionally, Qiao et al. [6] showed that combined indices such as the modified chlorophyll absorption ratio index (MCARI) and the optimized soil-adjusted vegetation index (OSAVI) have strong capabilities for estimating chlorophyll content.

VIs offer valuable insights into vegetation conditions. However, they alone cannot fully capture the diverse environmental stresses crops experience during growth and development. To address this limitation, recent studies have increasingly incorporated environmental variables that consider abiotic factors to enhance the accuracy of crop yield predictions. These variables, including temperature, precipitation, and soil data, have shown significant effectiveness in some cases [4]. Specifically, Zhou et al. [7] used the following environmental variables: maximum temperature (tmx), minimum temperature (tmn), daily precipitation (pre), reference evapotranspiration (pet), vapor pressure (vap), and vapor pressure deficit (vpd). Lang et al. [8] also employed the following environmental variables: the palmer drought severity index (pdsi), precipitation accumulation (pr), and soil moisture (soil). However, traditional air temperature (Tair) primarily reflects external controls, and rainfall alone is insufficient to represent soil moisture accurately [9]. To overcome these challenges, we introduce a more comprehensive set of environmental factors, detailed in Table 2. The data are sourced from the ERA5* global reanalysis climate dataset released by the European Centre for Medium-Range Weather Forecasts (ECMWF). Notably, the feasibility of daytime land surface temperature (LST_Day_1km) and nighttime land surface temperature (LST_Night_1km) has been validated in previous studies [9, 10].

*ERA5 is the fifth-generation ECMWF global climate atmospheric reanalysis data system, produced by C3S, covering data from January 1940 onwards. ERA5 provides hourly estimates of a large number of atmospheric, land, and oceanic climate variables, covering the Earth on a 30 km grid, with 137 levels from the surface up to 80 km in altitude.

Table 1. Different VIs and corresponding references.

Feature	Formula	References
Enhanced vegetation index (EVI)	$2.5\times(NIR-R)/(NIR+6\times R-7.5\times B+1)$	[3,5,6,11]
Normalized difference vegetation index (NDVI)	$(NIR-R)/(NIR+R)$	[4,7,8,12,13]
Green chlorophyll vegetation index (GCVI)	$NIR/G-1$	[3,14]
Green normalized difference vegetation index (GNDVI)	$(NIR-G)/(NIR+G)$	[3,10]
Wide dynamic range vegetation index (WDRVI)	$(0.2\times NIR-R)/(0.2\times NIR+R)$	[3,10]
Simple ratio vegetation index (SR)	NIR/R	[3,6]
Leaf area index (Lai_500m)		[13,15,16]
Greenness index (GI)	R/G	[8]
Difference vegetation index (DVI)	$NIR-R$	[5,8]
Enhanced vegetation index (EVI2)	$2.5\times(NIR-R)/(NIR+2.4\times R+1)$	[4,12]
Normalized difference greenness index (NDGI)	$(G-R)/(G+R)$	[5]
Modified chlorophyll absorption ratio index (MCARI)	$\left((NIR-R)-\sqrt{0.2\times(NIR-B)}\right)\times(NIR/R)$	[6,17]
Soil-adjusted vegetation index (SAVI)	$1.5\times(NIR-R)/(NIR+R+0.5)$	[6,18]
Optimized soil-adjusted vegetation index (OSAVI)	$1.6\times(NIR-R)/(NIR+R+0.16)$	[10,19]
Moisture index (wet)		[7]
Modified simple ratio vegetation index (MSR)	$(NIR/R-1)\,/\,\sqrt{(NIR/R+1)}$	[19]
Excess green index (EXG)	$2\times G-R-B$	[20]
Weighted difference vegetation index (WDVI)	$NIR-0.5\times R$	[21]
Visible-band difference vegetation index (VDVI)	$(2\times G-R-B)/(2\times G+R+B)$	[22]
Normalized green-blue difference index (NGBDI)	$(G-B)/(G+B)$	[22]

Table 2. Different environmental indices and corresponding references.

Feature	References
Daytime Land Surface Temperature (LST_Day_1km)	[9, 10]
Nighttime Land Surface Temperature (LST_Night_1km)	[9, 10]
Soil Temperature Level 1 (soil_temperature_level_1)	ERA5
Soil Temperature Level 2 (soil_temperature_level_2)	ERA5
Soil Temperature Level 3 (soil_temperature_level_3)	ERA5
Soil Temperature Level 4 (soil_temperature_level_4)	ERA5
Average Surface Skin Temperature (AvgSurfT_inst)	ERA5
Plant Canopy Surface Water (CanopInt_inst)	ERA5
Canopy Water Evaporation (ECanop_tavg)	ERA5
Root Zone Soil Moisture (RootMoist_inst)	ERA5
10cm Soil Moisture (SoilMoi0_10cm_inst)	ERA5

Table 3. Different ML algorithms and corresponding references.

Model	References
Long Short-Term Memory (LSTM)	[3, 11, 13, 15, 23, 24]
Random Forest (RF)	[3, 7, 8, 10, 12, 16]
Support Vector Regression (SVR)	[7, 8, 12]
Light Gradient Boosting Machine (LightGBM)	[3, 14]
Extreme Gradient Boosting (XGBoost)	[25–27]

Geographical location is a critical factor influencing crop yield, necessitating the incorporation of geographical features into yield prediction models. In this study, we introduce administrative division codes of counties and cities as one-dimensional geographical features, with "NAME" used to identify the geographical locations. Integrating such data enables a more in-depth analysis and understanding of the dynamic variations in crop yield.

The formation of crop yield is a complex and prolonged process, influenced by numerous factors that may vary in their impact across growth stages. For example, Jiang et al. [23] demonstrated that the transition from the vegetative to the reproductive development stage is most critical for maize yield prediction. Similarly, Ren et al. [16] found that the key features during the vegetative and reproductive growth stages of maize are growth status features and water-related features, respectively. These studies highlight the importance of analyzing various growth stages and their corresponding features for accurate yield prediction. However, in-depth research on the features of different growth stages remains relatively limited, and our comprehensive understanding of the mechanisms underlying crop yield formation requires further enhancement. To better comprehend the dynamic variations in crop yield, it is essential to conduct thorough investigations into the significance of various features during different growth stages for yield prediction.

Traditional crop yield prediction methods primarily rely on crop growth models and statistical regression techniques. However, crop growth models demand extensive input data, including climate,

variety, management, and soil conditions, which can be challenging to obtain and process. Moreover, empirical regression models are often limited by their locality and insufficient spatial generalization capabilities. To overcome these limitations, machine learning (ML) algorithms provide a more flexible and efficient alternative. Studies (see Table 3) have highlighted the significant potential of ML in crop yield prediction. For instance, Shahhosseini et al. [25] evaluated several ML models, including LASSO regression, RF, and XGBoost, and demonstrated that XGBoost is the most accurate model for yield prediction. The XGBoost model in Chen et al. [26] further showed that the XGBoost model exhibits strong generalization capabilities, allowing it to adapt to various types of data and complex relationships. Additionally, Li et al. [27] highlighted that XGBoost has clear advantages in feature selection and classification performance compared to logistic regression and other tree-based models.

While researchers [28] have achieved significant success in winter wheat yield prediction using deep learning, we focus on constructing an XGBoost regression model to explore the precise prediction of winter wheat yield in Henan Province. Notably, although the researchers in [29] also employed the XGBoost model for winter wheat yield prediction, we enhance feature selection by integrating multiple feature sources , including VIs, geographical factors, and soil features. In this study, we construct a combination model with multiple feature variable sets to achieve regression prediction of winter wheat yield in Henan Province and compare it with other ML models. Furthermore, we adopt the shapley additive explanations (SHAP) explainable framework [30] to conduct an in-depth analysis of the selected features. Our core objectives of this research are to address the following four objectives:

(1) Evaluate the predictive performance of the model by combining wheat growth stages, VIs, and newly introduced feature variables;

(2) Analyze the prediction results of winter wheat across growth stages to identify the stage with the highest prediction accuracy;

(3) Compare the performance of the XGBoost model with other ML models in winter wheat yield prediction;

(4) Assess the importance of selected features across growth stages to gain a deeper understanding of their contribution to yield prediction.

The remainder of this paper is organized as follows: In section 2, we introduce the research area and data, providing a detailed description of the specific region and data types used, along with preliminary analysis and organization of the collected data to ensure accuracy and reliability. In section 3, we present the framework of the prediction model, summarizing its core components. In section 4, we discuss the results and analysis, comprehensively validating the model's effectiveness through feature verification, prediction result presentation, multi-model comparison, and model interpretation. Finally, In section 5, we provide the major conclusions, summarizing the primary outcomes of the study and offer recommendations based on the research results.

2. Research area and research data

2.1. Research area

The study area is in Henan Province, which encompasses 102 counties (Figure 1). Located in the middle and lower reaches of the Yellow River, Henan Province spans geographical coordinates from 31°23′ to 36°22′ north latitude and 110°21′ to 116°39′ east longitude. Predominantly within the warm temperate zone, the province extends into the subtropical zone in the south, characterized by

a continental monsoon climate that transitions from the northern subtropical to the warm temperate zone. The region also features a climatic gradient from plains in the east to hilly and mountainous areas in the west, with distinct seasons and concurrent rainfall and heat. Over the past decade, Henan Province has experienced an average annual temperature ranging from 15.1°C to 15.9°C, with average annual precipitation of 512.6 mm to 1129.1 mm and average annual sunshine duration of 1774.5 to 2024.1 hours. These climatic conditions, combined with fertile soil, make Henan one of China's most important grain production areas. Within this region, wheat accounts for approximately one-quarter of the national planting area and yield, significantly contributing to China's overall grain production.

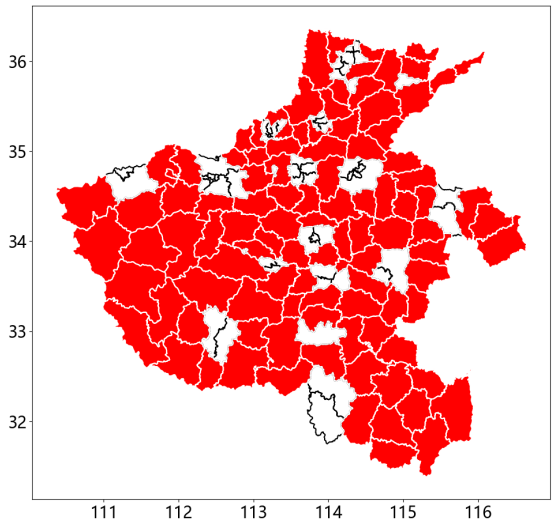

Figure 1. Map of the distribution of 102 counties in Henan Province (red areas).

2.2. Research data

2.2.1. Data source

Here, we leverage remote sensing data from Google Earth Engine (GEE), specifically two MODIS data products from the TERRA satellite: MYD11A2 and MYD15A2H. The MYD11A2 data provide detailed information on land surface temperature, including daytime and nighttime temperatures. The MYD15A2H dataset is utilized to obtain ecological parameters such as the leaf area index. Additionally, various spectral reflectance bands of the TERRA satellite data are combined linearly and nonlinearly to generate VIs for comprehensive vegetation growth monitoring. The ERA5 global reanalysis climate dataset from the ECMWF is also employed, encompassing major meteorological parameters like soil moisture and temperature at different depths. Furthermore, data from the Global Land Data Assimilation System (GLDAS) of the National Aeronautics and Space Administration (NASA) serve as supplementary data. The wheat yield data for Henan Province are sourced from the *Henan Statistical Yearbook*, which compiles the wheat planting area and yield across 102 counties and cities.

Table 4. Production cycles and corresponding features after selection.

Growth Stages	Selected Features
Tillering stage	LST_Night_1km
Elongation stage	EVI, NDVI, EVI2, DVI, GNDVI, Lai_500m, GCVI, MCARI, MSR, OSAVI, WDRVI, WDVI
Booting stage	NDVI, EVI2, DVI, EVI, GCVI, Lai_500m, MCARI, MSR, OSAVI, SVAI, WDVI
Heading stage	EVI, EVI2, DVI,WDRVI, WDVI, GCVI, MCARI, MSR, LST_Day_1km, soil_temperature_level_3, Lai_500m, SoilMoi0_10cm_inst, SoilMoi10_40cm_inst
Flowering stage	EVI, EVI2, DVI, soil_temperature_level_2, GCVI, Lai_500m, OSAVI, SVAI, WDVI, MCARI, SR, soil_temperature_level_3, SoilMoi0_10cm_inst, SoilMoi10_40cm_inst, wet
Grain filling stage	soil_temperature_level_2, soil_temperature_level_3, SoilMoi0_10cm_inst, soil_temperature_level_1, SoilMoi10_40cm_inst
Ripening stage	AvgSurfT_inst
Harvesting stage	soil_temperature_level_2, VDVI

Note: The same feature at different times, such as EVI during the heading stage and EVI during the flowering stage, are considered two distinct features. Therefore, the number of features is 61.

2.2.2. Data processing

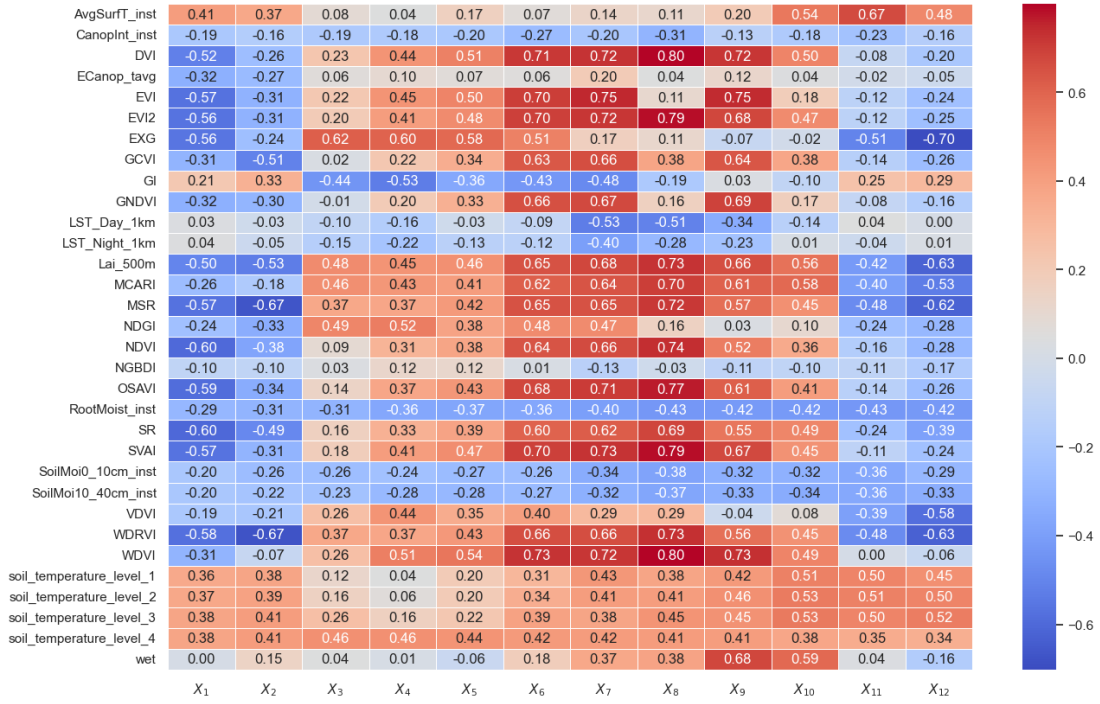

Figure 2. Heatmap of the correlation between yield and features.

Table 5. The growth stages of winter wheat.

Growth Stage	Date	
Sowing stage	10.05–10.20	X_1
Emergence stage	10.10–11.01	X_2
Tillering stage	11.01–12.20	X_3
Overwintering stage	12.20–02.10	X_4
Regreening stage	02.01–03.01	X_5
Elongation stage	03.01–04.05	X_6
Booting stage	04.05–04.14	X_7
Heading stage	04.14–04.20	X_8
Flowering stage	04.20–04.25	X_9
Grain filling stage	04.25–05.25	X_{10}
Ripening stage	05.25–06.05	X_{11}
Harvesting stage	06.01–06.10	X_{12}

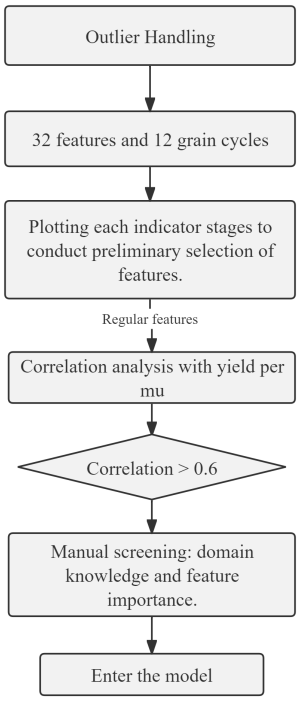

Figure 3. Flowchart feature selection process.

To ensure data accuracy, we initially employ the boxplot method to identify and handle outliers in the feature data. Subsequently, the data are integrated according to the wheat growth stages, and growth stage images are plotted to facilitate preliminary feature selection through visualization. To further refine the feature selection, the correlation between 32 features and yield is calculated across different growth stages, and a correlation heatmap is generated (Figure 2). Features with an absolute

correlation value greater than 0.6 in each growth stage are selected to ensure their significant impact on wheat yield. Additionally, domain knowledge and feature importance evaluation are combined to further select features that have a substantial influence on wheat yield from the initially screened set. The detailed process of feature selection is illustrated in Figure 3. Ultimately, features corresponding to 8 growth stages are selected, along with the geographical feature "NAME", resulting in a total of 61 features across growth stages. Table 4 presents the selected growth stages and their corresponding features, while Table 5 provides the division information of the growth stages.

The data and code supporting this paper are available in https://gitee.com/zzufinlab/bookcode/blob/master/Prediction_of_Wheat_Yield.

3. Establishment of the prediction module framework.

3.1. Establishment of the XGBoost winter wheat yield prediction model

XGBoost, an efficient gradient boosting framework, is based on a C++ implementation of the Gradient Boosting Machine and can handle regression and classification problems. It leverages the multi-threading capabilities of the CPU to perform parallel computations, significantly enhancing the efficiency of model training. The core mechanism involves integrating multiple weak learners (decision trees) by selecting partial sample features to generate weak learners and continuously fitting the residuals of previous models to minimize the objective function. Through iterative processes, a robust predictive model is constructed.

In the process of building an XGBoost model for yield prediction, the selected feature dataset is divided into two parts based on time. Data from 2016 to 2020 are used as the training set, while data from 2021 serve as the test set. The training set comprises 510 records of 62-dimensional data extracted from 2016 to 2020, with the 61 selected features serving as independent variable x and yield data as the dependent variable y. The model training process is described as follows:

- Initialization: A simple decision tree is initialized as the base model to predict the approximate values of the initial target;
- Residual Calculation: The residuals between the predicted values of the current model and the actual target values are calculated;
- Weight Calculation: The weights of each sample are calculated based on these residuals, and these weights are used to train the next decision tree;
- Prediction Accumulation: In each iteration, the predictions of the newly trained decision tree are weighted and accumulated with the predictions of all previous models to obtain new predicted values;
- Iteration Termination: The iterative process terminated when the preset maximum number of trees is reached or when the model performance no longer improved significantly.

The formula for each tree can be expressed as follows:

$$\widehat{y}_i^{(t)} = \sum_{k=1}^{K} f_k(x_i) = \widehat{y}_i^{(t-1)} + f_t(x_i), \quad f_k \in F, \tag{3.1}$$

where F represents the function space containing all trees, K is the total number of trees (or boosting

iterations) up to the current step t, $f_k(x_i)$ denotes the weight of the i-th sample in the leaf node of the k-th tree, and $\widehat{y}_i^{(t)}$ represents the predicted value of the i-th sample at the t iteration.

To simplify the objective function of the model and effectively prevent overfitting, a regularization term is introduced, thereby improving the model's performance when handling large-scale and high-dimensional data.

After model training is complete, the trained model is applied to the test dataset for evaluation to verify its performance and generalization capability. To comprehensively evaluate the effectiveness and accuracy of the XGBoost prediction model, we adopte the root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2)as evaluation metrics. These metrics reflect the model's predictive performance from different perspectives, including the magnitude of errors, the proportion of deviation between predicted and actual values, and the model's ability to explain data variability.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}, \tag{3.2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{3.3}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|, \tag{3.4}$$

where y_i is the true value of the i sample, $\widehat{y_i}$ is the predicted value of the i sample, $\overline{y}$ is the sample mean, and n is the number of samples.

3.2. Overview of key points in model explanation

SHAP is a powerful explainable framework proposed by Lundberg et al. [30] for interpreting the predictions of ML models. It provides an intuitive and rigorous way to understand how a model makes predictions based on input features by calculating the contribution value of each feature to the prediction outcome. These contribution values can be positive or negative, indicating the direction of each feature's influence on the prediction. A positive value signifies that a feature positively impacts the prediction, while a negative value indicates a negative influence. Moreover, the greater the absolute value of a feature's contribution, the more important the feature is in the model.

In this study, leveraging the SHAP framework, we calculate the SHAP values for each feature to generate local explanations for each prediction. These SHAP values quantify the specific contribution of each feature to the model's predictions, thereby providing a measure of feature importance across different growth stages. By analyzing these values, we can gain insights into how the model utilizes input features to make predictions and identify the most influential features at each growth stage. This analysis not only enhances the interpretability of the XGBoost model but also addresses the "black box" problem often associated with machine learning models in yield prediction.

4. Results and analysis

In this study, data from 2016 to 2020 are selected as the training samples to construct the XGBoost model, while data from 2021 are used as the test samples. The prediction model is applied to obtain the predicted yields of winter wheat for 102 counties in Henan Province in 2021, and the prediction results for winter wheat across growth stages are analyzed. The results are then compared with those of RF, GBDT, and Lasso. Subsequently, based on the SHAP explainable framework, the SHAP values for each feature were calculated to generate a local explanation for each prediction.

4.1. Overview of feature validation

The yield data for winter wheat in each county and city are calculated by aggregating the wheat planting area and yield data. Missing values are imputed with the average yield of the respective county or city. Table 6 presents the winter wheat yields for selected regions in Henan Province. This approach ensures the completeness and accuracy of the yield data, providing a reliable basis for the subsequent analysis and modeling.

Table 6. Yield (kg/hm^2) in some areas of Henan Province from 2016 to 2021.

County	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
Linyi County	7600.07	7707.79	7408.60	7666.23	7667.55	7771.50
Xiayi County	7354.05	7499.40	7163.70	7432.05	7514.48	7591.20
Xinxiang County	7630.50	7849.43	7288.28	7437.00	7530.00	7537.88
Xin'an County	4439.93	4726.88	4624.28	4549.88	5214.75	5073.08
Xixia County	3399.98	3490.73	3213.83	3369.23	3352.50	3398.10
Xin County	3308.24	3286.50	3300.00	3315.00	3315.00	3315.45

The initially selected 32 features are validated through a comparative analysis of data from high-yield and low-yield areas in some regions of Henan Province in 2021 (as shown in Figure 4). Taking DVI, EVI, CanopInt_inst, LST_Night_1km, SoilMoi10_40cm_inst, and soil_temperature_level_3 as examples, the analysis results indicate that during the overwintering to grain-filling stages of wheat, the DVI and EVI values in high-yield areas are significantly higher than those in low-yield areas. Additionally, while the differences in LST_Night_1km, soil_temperature_level_3, and SoilMoi10_40cm_inst between high-yield and low-yield areas are not significant, these features exhibit clear periodic patterns. In contrast, CanopInt_inst shows no distinct differentiation between high-yield and low-yield areas, and its variation pattern is not evident. Therefore, CanopInt_inst is excluded from subsequent analyses.

4.2. Presentation and discussion of prediction results

Applying the XGBoost model for yield prediction reveals that models incorporating geographical and soil variables, as well as combining these variables with NAME, consistently outperform models using only VIs. Specifically, the model that integrates VIs, soil variables, and NAME demonstrates the best predictive performance. As shown in Table 7, compared to the model using only VIs, the model incorporating VIs and soil variables achieves a higher R^2, indicating enhanced explanatory power.

Additionally, the RMSE and MAE decrease, signifying improved prediction accuracy. Furthermore, the model combining VIs, soil variables, and NAME exhibits even better predictive performance than the model that includes only VIs and soil variables. These results highlight the significant role of soil variables and NAME in enhancing the predictive performance of the model.

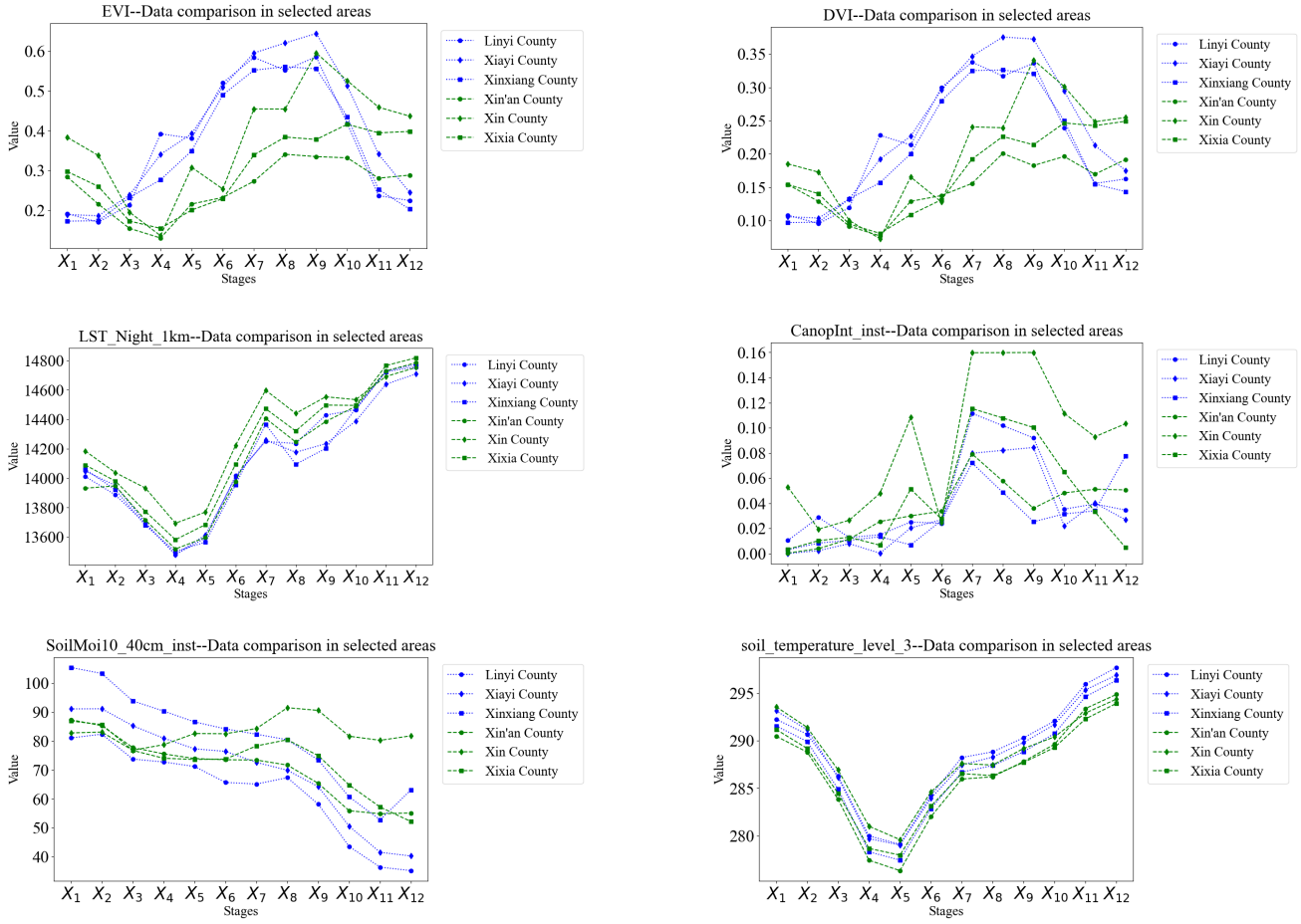

Figure 4. Data comparison chart for selected areas in Henan Province in 2021.

Table 7. Prediction results of the XGBoost model.

Model	R^2	RMSE (kg·hm^{-2})	MAE (kg·hm^{-2})
VIs + Soil Variables + NAME	0.85	516.97	371.36
VIs + Soil Variables	0.72	698.57	536.82
VIs Only	0.68	756.56	563.17

4.3. Prediction of winter wheat across growth stages

In this section, we conduct yield prediction for cumulative growth stages of the winter wheat (tillering stage, tillering to elongation stage, tillering to booting stage, tillering to heading stage, tillering to flowering stage, tillering to grain filling stage, tillering to ripening stage, and tillering to harvesting stage) to evaluate the optimal time window for the XGBoost model in predicting winter

wheat yield. The variable NAME is excluded during yield estimation, and the results are shown in Figure 5. However, even when including this variable, the trends remain similar.

The XGBoost yield prediction model performs poorly in the early stages of winter wheat growth, specifically exhibiting higher RMSE and MAE values and a lower R^2. This phenomenon may be attributed to the model's inability to capture sufficient crop growth, geographical, and soil information during the initial stages. As time progresses, the model gradually integrates more information, leading to improved performance. From the elongation stage to the booting stage, the RMSE and MAE decrease compared to the tillering stage, while R^2 increases, indicating a significant improvement in prediction accuracy. From the booting stage to the grain filling stage, the evaluation metrics exhibit minor fluctuations but remain relatively stable. After the grain filling stage, these metrics stabilize. Throughout the growth stages, the grain filling stage achieves the lowest MAE of 527.40 kg/hm², with an RMSE of 670.50 kg/hm². In contrast, the heading stage achieves the lowest RMSE of 662.29 kg/hm², a 1.2% reduction compared to the grain filling stage, while the MAE increases by 3% to 547.74 kg/hm². These results indicate that the XGBoost model provides high accuracy 2–3 growth stages before harvest, enabling earlier yield prediction than traditional methods. Therefore, the conclusions of this study align with the findings in the literature [28], and the prediction performance of this study has further improved compared to [28].

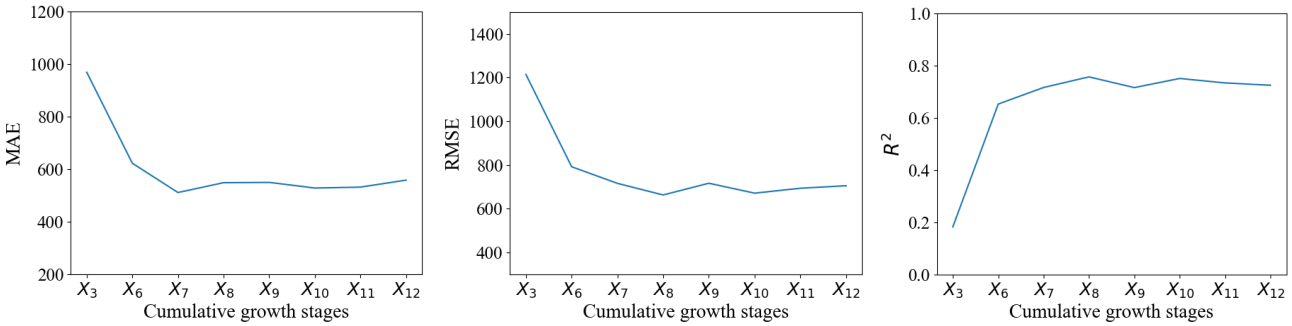

Figure 5. The accuracy of the XGBoost yield estimation model for various growth states.

4.4. Comparison and discussion of data results from different models

To further evaluate the performance of the XGBoost model in predicting winter wheat yield,we compare it with several other machine learning models, including RF, GBDT, and Lasso. Figure 6 shows scatter plots of the predicted versus actual yield values for counties in Henan Province using these four prediction models, while Table 8 provides the evaluation metrics for the different models. The XGBoost model, with an MAE of 371.36 kg/hm², RMSE of 516.97 kg/hm², and R^2 of 0.85, significantly outperforms the GBDT, RF, and Lasso models. This conclusion is further supported by Figure 6. Compared to the researchers in [29], we not only use VIs but also incorporate geographical and soil features, resulting in a higher R^2 than [29]. Although the RMSE is slightly higher than in [29], the difference is not significant, with only a 3% increase. This result indicates that, despite the slight increase in RMSE, the overall performance of the model has improved, particularly in terms of explaining the relationship between variables and yield.

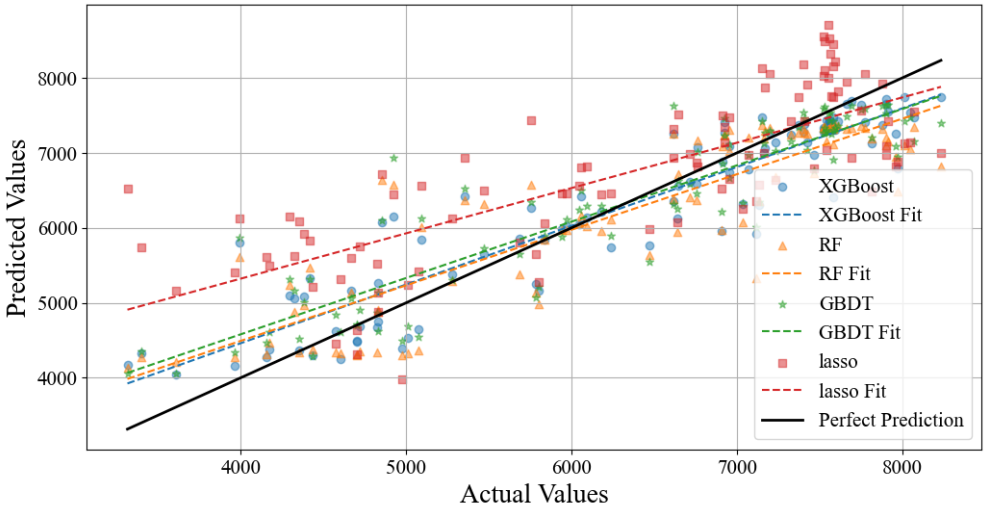

Figure 6. Performance of different yield prediction models.

Table 8. Evaluation metrics of different models.

Model	**R^2**	**RMSE (kg·hm^{-2})**	**MAE (kg·hm^{-2})**
XGBoost	0.85	516.97	371.36
GBDT	0.82	557.26	409.54
RF	0.77	632.72	483.89
Lasso	0.513	914.46	713.51

4.5. Model interpretation and findings

1) Analysis of individual samples.

Each sample in the test set has a corresponding explanation result, where the SHAP values represent the impact of each feature on the predicted yield for that sample. Here, we take the 10th sample in the test set as an example to analyze how the model generates the final prediction result. For the XGBoost model in this study, the SHAP explainable framework provides a baseline value of 6152.88, meaning that in the absence of any information, the model's predicted yield is 6152.88. The impact of each of the 61 features on the prediction result is then analyzed individually. The SHAP value of the feature variable NAME is 139.44, indicating that it increases the predicted yield from 6152.88 to 6292.32. This suggests that the feature has a positive effect on the yield prediction, increasing the predicted value. The SHAP value of the feature variable AvgSurfT_inst during the ripening stage is 282.70, indicating that it increases the predicted yield from 6292.32 to 6575.02. This suggests that the feature has a positive effect on the yield prediction, further increasing the predicted value. The SHAP value of the feature variable LST_Day_1km during the heading stage is −29.28, indicating that it decreases the predicted yield from 6575.02 to 6545.74. This suggests that the feature has a negative effect on the yield prediction, reducing the predicted value.

By analogy, the impact of the remaining 58 features on the yield prediction can be analyzed. Under the combined influence of all 61 features, the SHAP predicted value for this sample increases from the baseline value of 6152.88 to 7029.09. The SHAP explainable framework clearly demonstrates the role

each feature plays in this prediction process. The results are visualized using the SHAP explainable framework, as shown in Figure 7.

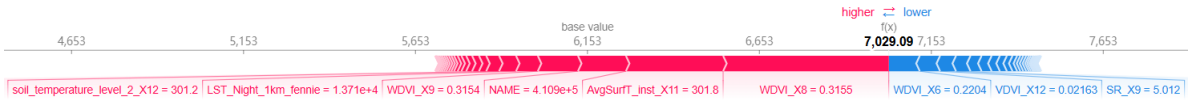

Figure 7. Visualization of SHAP Values for the 10th Sample.

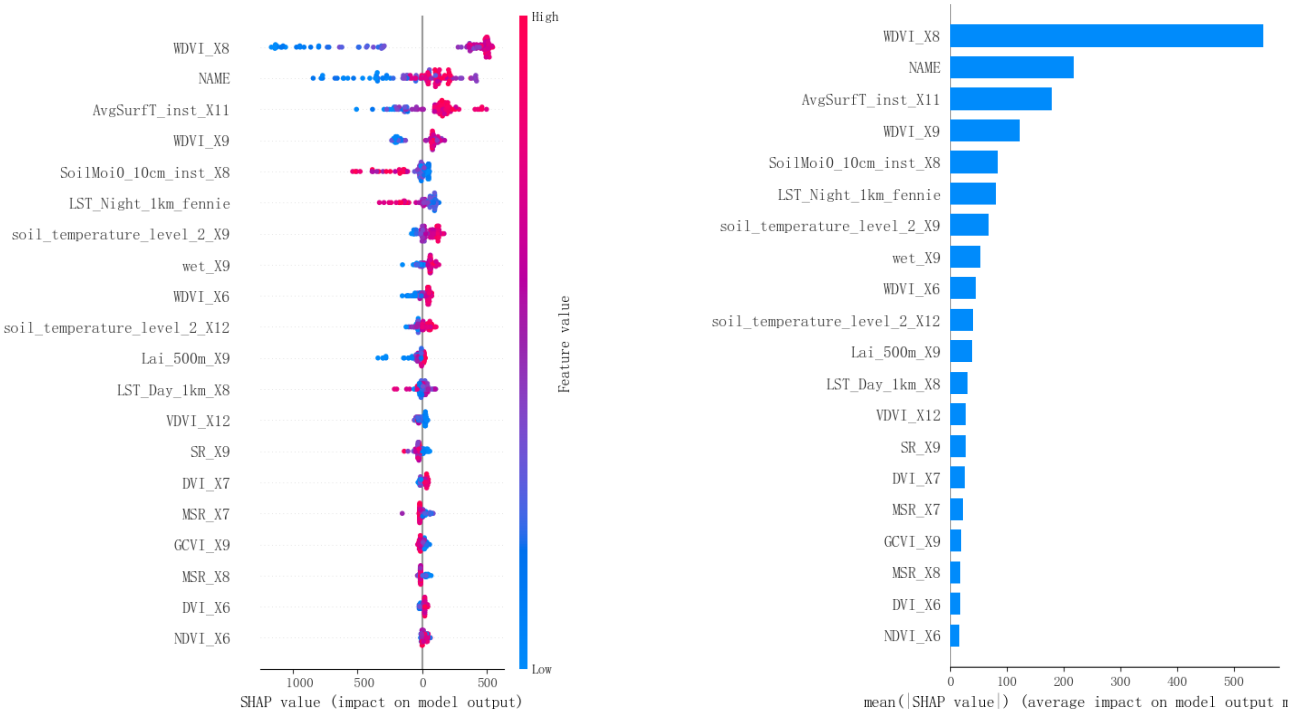

Figure 8. Visualization of SHAP values for overall features (left), and feature importance output from the SHAP explainability framework (right).

2) Analysis and explanation of the overall samples.

To gain a clearer understanding of the impact of features on the prediction results, we present the explanation results and feature importance for all samples, as shown in Figure 8. In the left plot, each point corresponds to the SHAP value of a sample, with the color gradient indicating the magnitude of the feature value: red represents higher feature values, while blue represents lower feature values. The SHAP value of 0 serves as a dividing line, where points to the left indicate a negative impact on the final prediction result and points to the right indicate a positive impact, with the distance from the line reflecting the magnitude of the effect. In the right plot, the x-axis represents the mean of the absolute SHAP values for each feature across all explained samples, with larger values indicating a greater average influence on the model's output.

By combining the insights from both figures, it can be observed that the following features have a significant impact on the model's prediction results: (1) WDVI during the heading stage, (2) the

geographical feature NAME, (3) AvgSurfT_inst during the ripening stage, (4) WDVI during the flowering stage, and (5) SoilMoi0_10cm_inst during the heading stage.

These findings underscore the importance of integrating multisource-data, including vegetation indices, geographical, and soil features, in enhancing the accuracy and interpretability of winter wheat yield prediction. The SHAP explainable framework not only quantifies the contribution of each feature to the model's predictions but also identifies the most influential features across different growth stages, thereby addressing the "black box" problem often associated with machine learning models. This comprehensive analysis provides valuable insights for precision agricultural management, enabling more informed decision-making and contributing to enhanced food security and sustainable agricultural development in Henan Province.

5. Conclusions

Based on data from 2016 to 2020, we establish a prediction model for winter wheat yield in Henan Province in 2021 using the XGBoost model and conducts explainability analysis and feature impact assessment using the SHAP method. The following key findings and conclusions are drawn:

(1) Feature selection and model performance: Through feature selection, we identify 8 growth stages and their corresponding features, along with the NAME, resulting in a total of 61 features included in the model. Compared to models using only VIs or VIs combined with soil variables, the model incorporating all newly introduced features performs slightly better. This highlights the significant role of soil variables and NAME in enhancing the model's predictive performance.

(2) Yield prediction across growth stages: By predicting yield across cumulative growth stages, we found that the proposed yield prediction model can accurately predict winter wheat yield during the mid-to-late growth stages, with the grain filling stage showing the best performance. Notably, the XGBoost model provides high accuracy 2–3 growth stages before harvest, enabling earlier yield prediction than traditional methods.

(3) Model evaluation metrics: The evaluation metrics of the XGBoost model indicate an MAE of 371.36 kg/hm², an RMSE of 516.97 kg/hm², and an R^2 of 0.85. When comparing the predictive performance of different models, the XGBoost model slightly outperforms other models in predicting winter wheat yield in Henan Province.

(4) SHAP analysis for model explainability: Using SHAP to explain the model, the analysis of individual samples reveals the impact of each feature on the prediction results. Additionally, the analysis of the overall samples demonstrates that the following features have a significant influence on the model's predictions: 1) WDVI during the heading stage, 2) the geographical feature NAME, 3) AvgSurfT_inst during the ripening stage, 4) WDVI during the flowering stage, and 5) SoilMoi0_10cm_inst during the heading stage.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (71971031), and General Program of National Social Science Foundation of China (18BJT021).

References

1. Mo Z, Liu ZQ, Wu YC, (2008) The present situation analysis of agricultural production monitoring and output forecasting system at home and abroad. *Chin Agric Sci Bull* 24: 434–437.
2. Yang BJ, Lu DH, Pei ZY, Zhao HJ, Wu YY, (1997) The structure design of a national crop condition monitoring system. *Trans Chin Soc Agricu Eng* 13: 16–19.
3. Zhang L, Zhang Z, Luo Y, Cao J, Xie R, Li S, (2021) Integrating satellite-derived climatic and vegetation indices to predict smallholder maize yield using deep learning. *Agric For Meteorol* 311: 108666. https://doi.org/10.1016/j.agrformet.2021.108666
4. Bolton DK, Friedl MA, (2013) Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agric For Meteorol* 173: 74–84. https://doi.org/10.1016/j.agrformet.2013.01.007
5. Qu G, Shuai Y, Shao C, Peng X, Huang J, (2023) County scale corn yield estimation based on multi-source data in Liaoning Province. *Agronomy* 13: 1428. https://doi.org/10.3390/agronomy13051428
6. Qiao L, Tang W, Gao D, Zhao R, An L, Li M, et al. (2022) UAV-based chlorophyll content estimation by evaluating vegetation index responses under different crop coverages. *Comput Electron Agric* 196: 106775. https://doi.org/10.1016/j.compag.2022.106775
7. Zhou W, Liu Y, Ata-Ul-Karim ST, Ge Q, Li X, Xiao J, (2022) Integrating climate and satellite remote sensing data for predicting county-level wheat yield in China using machine learning methods. *Int J Appl Earth Obs Geoinformation* 111: 102861. https://doi.org/10.1016/j.jag.2022.102861
8. Lang P, Zhang L, Huang C, Chen J, Kang X, Zhang Z, et al. (2023) Integrating environmental and satellite data to estimate county-level cotton yield in Xinjiang Province. *Front Plant Sci* 13: 1048479. https://doi.org/10.3389/fpls.2022.1048479
9. Pede T, Mountrakis G, Shaw SB, (2019) Improving corn yield prediction across the US Corn Belt by replacing air temperature with daily MODIS land surface temperature. *Agric For Meteorol* 276: 107615. https://doi.org/10.1016/j.agrformet.2019.107615
10. He Y, Qiu B, Cheng F, Chen C, Sun Y, Zhang D, et al. (2023) National scale maize yield Estimation by integrating multiple spectral indexes and temporal aggregation. *Remote Sens* 15: 414. https://doi.org/10.3390/rs15020414
11. Zhang L, Zhang Z, Luo Y, Cao J, Tao F, (2019) Combining optical, fluorescence, thermal satellite, and environmental data to predict county-level maize yield in China using machine learning approaches. *Remote Sens* 12: 21. https://doi.org/10.3390/rs12010021

12. Bian C, Shi H, Wu S, Zhang K, Wei M, Zhao Y, et al. (2022) Prediction of field-scale wheat yield using machine learning method and multispectral UAV data. *Remote Sens* 14: 1474. https://doi.org/10.3390/rs14061474

13. Di Y, Gao M, Feng F, Li Q, Zhang H, (2022) A new framework for winter wheat yield prediction integrating deep learning and Bayesian optimization. *Agronomy* 12: 3194. https://doi.org/10.3390/agronomy12123194

14. Cao J, Zhang Z, Tao F, Zhang L, Luo Y, Han J, et al. (2020) Identifying the contributions of multi-source data for winter wheat yield prediction in China. *Remote Sens* 12: 750. https://doi.org/10.3390/rs12050750

15. Wang J, Si H, Gao Z, Shi L, (2022) Winter wheat yield prediction using an LSTM model from MODIS LAI products. *Agriculture* 12: 1707. https://doi.org/10.3390/agriculture12101707

16. Ren Y, Li Q, Du X, Zhang Y, Wang H, Shi G, et al. (2023) Analysis of corn yield prediction potential at various growth phases using a process-based model and deep learning. *Plants* 12: 446. https://doi.org/10.3390/plants12030446

17. Maimaitijiang M, Sagan V, Sidike P, Hartling S, Esposito F, Fritschi FB, (2020) Soybean yield prediction from UAV using multimodal data fusion and deep learning. *Remote Sens Environ* 237: 111599. https://doi.org/10.1016/j.rse.2019.111599

18. Kumar C, Mubvumba P, Huang Y, Dhillon J, Reddy K, (2023) Multi-stage corn yield prediction using high-resolution UAV multispectral data and machine learning models. *Agronomy* 13: 1277. https://doi.org/10.3390/agronomy13051277

19. Zhang J, Zhao Y, Hu Z, Xiao W, (2023) Unmanned aerial system-based wheat biomass estimation using multispectral, structural and meteorological data. *Agriculture* 13: 1621. https://doi.org/10.3390/agriculture13081621

20. Zhou X, Zheng HB, Xu XQ, He JY, Ge XK, Yao X, et al. (2017) Predicting grain yield in rice using multi-temporal vegetation indices from UAV-based multispectral and digital imagery. *ISPRS J Photogramm Remote Sens* 130: 246–255. https://doi.org/10.1016/j.isprsjprs.2017.05.003

21. Clevers J (1989) Application of a weighted infrared-red vegetation index for estimating leaf area index by correcting for soil moisture. *Remote Sens Environ* 29: 25–37. https://doi.org/10.1016/0034-4257(89)90076-X

22. Wang XQ, Wang MM, Wang SQ, Wu YD, (2015) Extraction of vegetation information from visible unmanned aerial vehicle images. *Trans Chin Soc Agric Eng* 31: 152–159. https://doi.org/10.3969/j.issn.1002-6819.2015.05.022

23. Jiang H, Hu H, Zhong R, Xu J, Xu J, Huang J, et al. (2020) A deep learning approach to conflating heterogeneous geospatial data for corn yield estimation: A case study of the US Corn Belt at the county level. *Global Change Biol* 263: 1754–1766. https://doi.org/10.1111/gcb.14885

24. Cao J, Zhang Z, Tao F, Zhang L, Luo Y, Zhang J, et al. (2021) Integrating multi-source data for rice yield prediction across China using machine learning and deep learning approaches. *Agric For Meteorol* 297: 108275. https://doi.org/10.1016/j.agrformet.2020.108275

25. Shahhosseini M, Martinez-Feria RA, Hu G, Archontoulis SV, (2019) Maize yield and nitrate loss prediction with machine learning algorithms. *Environ Res Lett* 14: 124026. https://doi.org/10.1088/1748-9326/ab5268

26. Chen T, Guestrin C, (2016) Xgboost: A scalable tree boosting system. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2016: 785–794. https://doi.org/10.1145/2939672.2939785

27. Li H, Cao Y, Li S, Zhao J, Sun Y, (2020) XGBoost model and its application to personal credit evaluation. *IEEE Intell Syst* 35: 52–61. https://doi.org/10.1109/MIS.2020.2972533

28. He XH, Luo HT, Qiao MJ, ZH Tian, GS Zhou, (2021) Yield estimation of winter wheat in China based on CNN-RNN network. *Trans Chin Soc Agric Eng* 37: 124–132. https://doi.org/10.11975/j.issn.1002-6819.2021.17.014

29. Zhang YB, Li X, Man WD, Liu MY, Fan JH, HR Hu, et al. (2024) Research on yield estimation method of winter wheat based on Sentinel-1/2 data and machine learning algorithms. *Acta Agric Zhejiangensis* 36: 2812–2822. https://doi.org/10.3969/j.issn.1004-1524.20231368

30. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 30.

AIMS Press