Journal of Zhejiang University-SCIENCE B 2026 27(5):537-560
www.jzus.zju.edu.cn; www.springer.com/journal/11585
E-mail: jzus_b@zju.edu.cn

Research Article
https://doi.org/10.1631/jzus.B2500403

Deep learning-based phenology extraction and crop classification in arid oasis using Sentinel-2 time series

Chunli WANG[1,2*], Jianan CHI[1,2*], Xiao ZHANG[1,2], Nannan ZHANG[1,2✉]

[1]*College of Information Engineering, Tarim University, Alar 843300, China*
[2]*Key Laboratory of Oasis Agriculture in Tarim, Tarim University, Alar 843300, China*

Abstract: Multi-temporal remote sensing data in large-scale crop phenology identification and classification have become increasingly utilized, particularly for precision management in arid oasis agricultural regions with complex cropping systems. In this study, we developed a deep learning framework integrating Sentinel-2 multi-temporal imagery and normalized difference vegetation index (NDVI) time series for mapping cotton, winter jujube, and tiger nut crops in Tumushuke City, Xinjiang Uygur Autonomous Region, China. We employed the minimum redundancy maximum relevance (mRMR) algorithm for spectral and vegetation index feature selection, followed by Savitzky-Golay (S-G) filtering and double logistic function fitting, to automatically extract the key phenological parameters (start of season (SOS), peak of season (POS), and end of season (EOS)), significantly improving phenological feature extraction accuracy. By incorporating multi-temporal Sentinel-2 data and a multi-scale feature fusion approach, we could systematically compare five classification models (multi-layer perceptron (MLP), residual network-18 (ResNet-18), convolutional long short-term memory (ConvLSTM), Transformer, and random forest classifier (RFC)), demonstrating that high-resolution spatial details substantially enhance crop boundary delineation and classification consistency in complex environments. Further optimization of Transformer's spatial representation through multi-scale window analysis revealed that the use of 1×1+3×3+5×5 convolutional windows achieves an optimal balance between accuracy and computational efficiency. Independent validation on unseen areas confirmed robust model transferability, with F1 scores of 94.37%, 87.75%, and 86.35% for the three crops (winter jujube, cotton, and tiger nut), respectively. This study validates the high-precision identification potential of Sentinel-2 temporal data and deep neural networks for multi-crop environments, enabling the precise spatial mapping of crop distributions and providing methodological support for smart agricultural decision-making in arid oasis regions.

Key words: Remote sensing classification; Deep learning; Sentinel-2 imagery; Multi-scale feature fusion; Crop phenology

1 Introduction

Data on the phenological patterns of crops during growth, which directly reflect their key developmental stages, provide critical guidance for rational agricultural scheduling, pest control, yield prediction, and climate change adaptation (Kordi and Yousefi, 2022). Against the backdrop of escalating climate variability and demand for precision agriculture, the accurate acquisition of crop growth stage information has become particularly crucial for modern agricultural development (Ismaili et al., 2024). Taking Tumushuke in Xinjiang Uygur Autonomous Region as an example, located at the ecotone between arid zones and oasis belts, this region exhibits unique climatic characteristics. Its agricultural production relies entirely on irrigation, with diverse cropping systems showing distinct growth cycles among crops (He et al., 2025). As local agriculture is influenced by both natural conditions and human interventions, the timely and accurate monitoring of crop growth stages has become a key factor in maintaining stable agricultural output (Zhai et al., 2025). Compared to traditional manual field observations, satellite remote sensing—with advantages such as large-scale coverage, temporal continuity, and cost-effectiveness—is particularly suited for Xinjiang's vast and fragmented farmland distribution,

✉ Nannan ZHANG, zhangnannan@taru.edu.cn
* The two authors contributed equally to this work
Nannan ZHANG, https://orcid.org/0000-0003-2956-8815
Chunli WANG, https://orcid.org/0009-0009-1865-5404

Received July 11, 2025; Revision accepted Dec. 3, 2025;
Crosschecked Mar. 20, 2026; Published online Apr. 15, 2026

offering an effective solution for modern agricultural monitoring and smart management (Newete et al., 2024).

In recent years, the rapid development of remote sensing technology has generated innovative approaches for crop phenology monitoring. With continuous improvements in both temporal and spatial resolution, remote sensing imagery now offers enhanced capabilities for tracking crop growth processes dynamically (Qiu et al., 2024). Among the available platforms, Sentinel-2 stands out as a multispectral high-resolution satellite, providing 13 spectral bands that are particularly advantageous for constructing vegetation indices and identifying crop types (Nivedita et al., 2024). Wang C et al. (2024) demonstrated the effectiveness of Sentinel-1/2 data in early crop classification, highlighting its significant potential in improving overall classification accuracy. Du et al. (2024) integrated Sentinel-2 data with meteorological datasets to achieve the daily-scale precision monitoring of soil moisture in croplands, thereby enhancing the timeliness and reliability of regional water resource management. Chen et al. (2024) made notable progress in non-photosynthetic vegetation monitoring by fusing microwave remote sensing data with novel spectral indices. Moreover, Zhou JP et al. (2024) developed a new straw index based on Sentinel-2 data, which significantly improved the efficiency of maize residue identification. Meanwhile, Han et al. (2025) achieved an accurate classification of tobacco planting areas using unmanned aerial vehicle (UAV)-based remote sensing techniques. A growing body of research highlights the value of normalized difference vegetation index (NDVI) time series as an effective indicator of vegetation dynamics across ecological and agricultural applications. Ghilardi et al. (2025) systematically analyzed forest habitat degradation in Madagascar using Landsat-derived NDVI time series. Furthermore, de la Guardia et al. (2024) investigated water use patterns in Brazilian dry bean cultivation areas through Sentinel-2 NDVI time series, offering a novel approach for assessing irrigation efficiency. Collectively, these studies confirm the practical utility of NDVI time series in monitoring both natural ecosystem dynamics and agricultural production processes. Notably, Farbo et al. (2024) innovatively integrated artificial intelligence into NDVI time-series analysis, developing a dynamic prediction model for maize NDVI that enables the precise simulation of crop growth, thus providing robust data support for growth modeling.

Among the various modern agricultural monitoring technologies, remote sensing has emerged as a research hotspot due to its unique advantages. In fact, the integration of traditional machine learning with deep learning techniques is driving a paradigm shift in this field from simple observation to intelligent analysis. Khankeshizadeh et al. (2024) proposed an FBA-DPAttResU-Net model that effectively improved the identification accuracy of fire-damaged areas in Sentinel-1/2 imagery by incorporating dual-path residual structures and attention mechanisms, demonstrating the technical superiority of deep learning in complex agricultural scenarios. Mendes et al. (2025) further revealed that hyperparameter optimization plays a decisive role in the crop classification performance of convolutional neural networks (CNNs), providing valuable insights for the parameter tuning of agricultural remote sensing classification models. Notably, Martínez-Movilla et al. (2024) successfully achieved the efficient monitoring of intertidal macroalgae by combining large-scale UAV imagery with image recognition algorithms, showcasing the applicability of this technology in specialized agricultural ecosystems. In the field of time-series data analysis, Li CL et al. (2025) employed multiple input single output (MISO) and long short-term memory (LSTM) for flow prediction in the Yangtze-Dongting Lake system, offering a case study for research on the coupling of hydrology and remote sensing. Reviewing the evolution of remote sensing data processing techniques from early machine learning methods, such as random forests, to recently widely applied deep neural networks, including CNNs, LSTMs, and Transformers, these technological innovations have significantly enhanced the intelligence level of agricultural monitoring.

The integration of remote sensing technology and deep learning methods has significantly advanced crop phenology monitoring research. However, practical applications still face several key challenges. Current research primarily focuses on the phenology monitoring of major staple crops (such as wheat, rice, and corn) based on remote sensing recognition, while specialty economic crops have received much less attention in this regard, such as regions like Xinjiang (e.g., cotton, winter jujube, and tiger nut). Due to the overlapping phenological periods of different crops (e.g.,

the flowering period of cotton overlaps with the tasseling period of corn), accurately extracting the key phenological features of a target crop under mixed cropping conditions is challenging (Marino, 2023). For example, in regions like Tumushuke in Xinjiang where crops are interspersed, spectral mixing phenomena affect the quality of remote sensing data. Besides, the significant differences in the planting cycles of different crops further complicate the extraction of temporal features, posing a severe test to the generalization ability of models. Additionally, deep learning models rely on large-scale labeled samples, yet in the field of agricultural remote sensing, there are challenges such as high sample acquisition costs and a lack of field data during the critical phenological stages. Thus, effectively integrating multi-source remote sensing feature variables, temporal variation patterns, and crop physiological characteristics to achieve high-precision phenological identification in complex regions remains a hot topic and a key field in current remote sensing agricultural research.

To directly address these challenges, this study developed a dedicated framework focusing on the understudied specialty crops in Tumushuke, Xinjiang, China. Focusing on phenological feature extraction for typical crops (cotton, winter jujube, and tiger nut) in Tumushuke, we plan to utilize multi-temporal Sentinel-2 remote sensing imagery combined with field survey data and deep learning models. By applying Savitzky-Golay (S-G) filtering and double logistic function fitting to NDVI time-series data, we aim to accurately extract key phenological parameters, including the start of season (SOS), peak of season (POS), and end of season (EOS). Furthermore, we intend to use time-series Sentinel-2 imagery to investigate the role of deep learning and multi-scale window analysis in crop spatial distribution identification. Through this approach, we seek to develop a deep learning-based method for multi-temporal remote sensing image analysis to support crop spatial distribution mapping in the Tumushuke region and beyond.

2 Materials and methods

2.1 Study area

The study area is situated in Tumushuke City, Xinjiang, China (79°12′54.38″E, 39°53′42.16″N). According to the Köppen climate classification, this region has a severe, continental arid desert climate, classified as BWk (cold desert climate), where environmental conditions are harsh and unforgiving. In the summer, temperatures can soar to a blistering 40 ℃, while winter ushers in a sharp contrast, with temperatures plunging to as low as −25 ℃. The annual total solar radiation ranges from approximately 545.8 to 597.1 MJ/m^2, with scarce precipitation and intense evaporation (Zheng et al., 2022). The dominant soil types are sandy loam and loam, which provide specific conditions for crop growth. Benefiting from the region's unique combination of solar radiation, heat, and soil conditions, crops such as cotton, winter jujube, and tiger nut have successfully been selected for this environment and hold significant roles in the local agricultural system (Cao et al., 2022).

To ensure accurate correspondence between remote sensing classification samples and ground features, we used a handheld real-time kinematic (RTK) high-precision positioning instrument to record the geographic coordinates of each plot, enabling spatial point sampling. Our sampling strategy was designed to be representative by covering a range of soil types, crop categories, and land use types across the region. A total of 1317 ground truth plots were surveyed, including 270 RTK points for cotton, 150 RTK points for winter jujube, and 85 RTK points for tiger nut. To rigorously assess model generalization, these samples were divided into a training set and an independent test set based on spatial location. To address the class imbalance caused by differences in planting area, the experiment adjusted the class weights in the loss function of the model, assigning a higher weight to tiger nut, the crop with a smaller cultivation area. Sentinel-2 multispectral imagery was captured during three key phenological stages of cotton during the 2024 growing season, supporting both sample construction and temporal window definition. During the sowing stage, the land surface is mostly bare. Some early-sown fields exhibit initial canopy emergence, but the vegetation indices remain low and the spectral differences between plots are limited; imagery at this stage is useful for identifying tillage conditions and estimating planting density. The vigorous growth stage, with maximum vegetation coverage and enhanced greenness, makes it the optimal period for crop type and structure identification using remote sensing. At this

stage, the canopy reflectance characteristics are highly differentiated due to variations in leaf area index (LAI) and chlorophyll concentration (SPAD), resulting in distinct spectral boundaries—particularly useful for distinguishing among winter jujube, tiger nut, and cotton. In the senescence or harvest stage, where some crops have withered or been harvested, substantial changes in surface reflectance occur. On the one hand, increased stubble and bare soil make field textures more complex; on the other hand, certain late-maturing crops retain some greenness. Hence, imagery from this stage provides important reference information for determining the crop growth cycle and validating the classification boundaries, as well as serving as a basis for subsequent dynamic monitoring and model accuracy adjustment.

2.2 Integrated framework for crop phenology and spatial classification

Fig. 1 illustrates the complete workflow of this study from data acquisition to classification evaluation, which consists of five main stages: (1) acquisition and preprocessing of the multi-temporal Sentinel-2 imagery, including radiometric calibration, atmospheric correction, geometric correction, and cloud masking; (2) construction of the crop sample dataset based on Google Earth imagery interpretation, RTK field surveys, and local crop yearbook data; (3) generation of NDVI time-series curves using the TIMESAT tool with S-G smoothing and double logistic fitting, followed by feature selection using the minimum redundancy maximum relevance (mRMR) algorithm; (4) model construction and comparison, including multi-layer perceptron (MLP), residual network-18 (ResNet-18), convolutional long short-term memory (ConvLSTM), Transformer, and random forest classifier (RFC), with classification accuracy further optimized through multi-scale window analysis; and (5) accuracy evaluation and regional validation, ultimately achieving visualized outputs of both crop spatial distribution and phenological timelines.

2.3 Satellite data acquisition and preprocessing

The satellite imagery employed in this research was sourced from the Sentinel-2A and Sentinel-2B platforms operated by the European Space Agency (ESA). Both satellites are equipped with a multispectral instrument (MSI), offering 13 spectral bands with a maximum spatial resolution of 10 m (Jia et al., 2025), which was also the final spatial resolution of the Sentinel-2 images used in this study. The study area is located in typical farmland regions of Tumushuke,

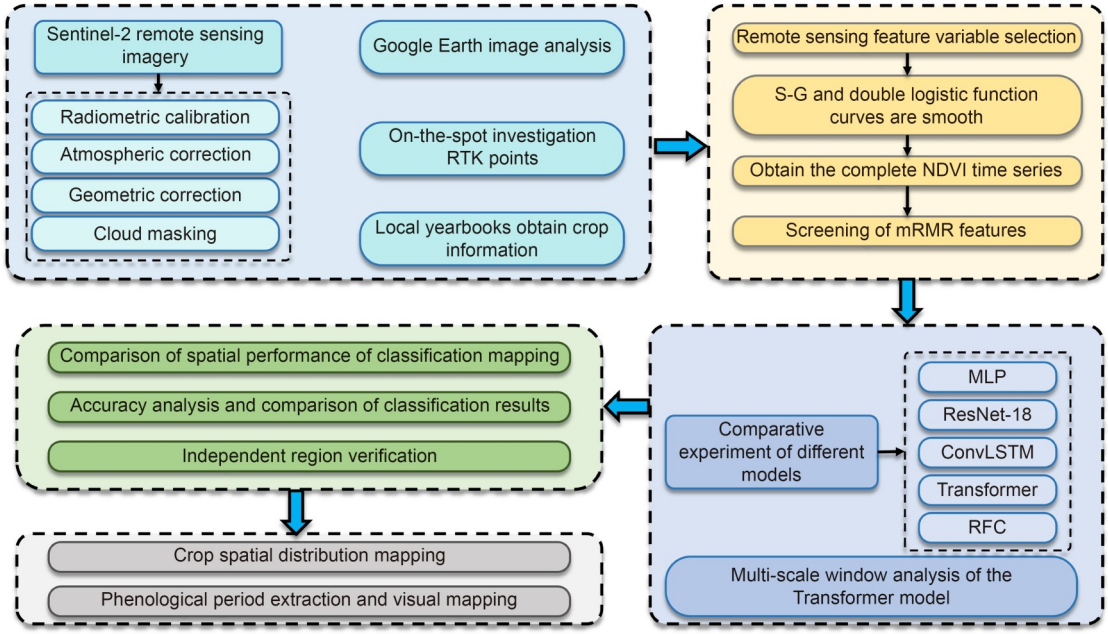

Fig. 1 Overall technical workflow adopted in this study. RTK: real-time kinematic; S-G: Savitzky-Golay; NDVI: normalized difference vegetation index; mRMR: minimum redundancy maximum relevance; MLP: multi-layer perceptron; ResNet-18: residual network-18; ConvLSTM: convolutional long short-term memory; RFC: random forest classifier.

Xinjiang, China. Sentinel-2 Level-2A images covering the entire crop growth cycle—from January 2 to December 29, 2024—were collected, totaling 75 scenes. Ten key spectral bands from the Sentinel-2 MSI, selected for their relevance to agricultural monitoring, were used in the analysis. The details of these bands are provided in Table 1. These bands span the visible, near-infrared (NIR), and shortwave infrared (SWIR) regions, effectively capturing the spectral responses of crops at various phenological stages. They are widely applied in vegetation index construction, phenological detection, and feature extraction for classification models. Specifically, the NIR and red bands (Band 4 (B4) and B8) form the basis for calculating fundamental vegetation indices such as NDVI. The red-edge bands (B5–B7) are highly sensitive to changes in chlorophyll content and subtle variations in plant physiology, which improves species differentiation. Meanwhile, the SWIR bands (B11 and B12) are primarily influenced by crop moisture content and stress levels but also respond to biochemical constituents such as proteins and cellulose. This makes them critical not only for monitoring drought conditions and identifying crop maturity stages in arid regions but also for discriminating crop types based on their biochemical properties (Faqeerzada et al., 2025).

The imagery was obtained through the Copernicus Open Access Hub and processed using ENVI software (version 5.6.3, L3Harris Geospatial, Boulder, CO, USA), where radiometric calibration, atmospheric correction, and geometric correction were performed to generate Level-2A products containing surface reflectance data. Additional data acquisition and preprocessing were carried out using the Google Earth Engine (GEE) platform. The workflow involved several key steps, including cloud and shadow detection and masking, image clipping, NDVI calculation, and the construction of time-series datasets. The role of GEE in this process was mainly for data acquisition and preprocessing tasks, such as handling large datasets, performing cloud/shadow masking, and calculating indices like NDVI, which can be done more efficiently on the cloud. These processed images were then used to generate crop-specific NDVI time series, perform vegetation index analyses, and extract relevant remote sensing features. While ENVI was utilized for detailed calibration and atmospheric correction, GEE provided an efficient platform for the preprocessing steps, especially for cloud masking, using methods like scene classification layer (SCL) and quality assurance (QA) bands, along with advanced algorithms such as S2cloudless (Gao et al., 2024). We employed algorithms such as S2cloudless to detect clouds and cloud shadows, followed by the application of masking to eliminate these regions. After cloud shadow masking, pixels with minimal cloud shadow were retained to mitigate the impact of cloud cover on the NDVI time series. Missing NDVI values were then imputed using values from preceding and subsequent time periods to preserve the continuity and integrity of the time series. Together, these steps ensured the provision of high-quality input data for the development of the classification model.

2.4 Spectral and vegetation index selection

Crop recognition primarily relies on the differences in spectral features, phenological characteristics,

Table 1 Selected Sentinel-2 bands

Band	Resolution (m)	Name	Wavelength (μm)	Bandwidth (nm)
2	10	Blue	0.458–0.523	65
3	10	Green	0.543–0.578	35
4	10	Red	0.650–0.680	30
5	20	VegetationRedEdge	0.698–0.712	14
6	20	VegetationRedEdge	0.733–0.747	14
7	20	VegetationRedEdge	0.773–0.793	20
8	10	NIR	0.785–0.899	114
8A	20	NarrowNIR	0.855–0.875	20
11	20	SWIR	1.565–1.655	90
12	20	SWIR	2.100–2.280	180

NIR: near-infrared; SWIR: shortwave infrared.

and temporal features. In terms of spectral characteristics, crop recognition is mainly based on the reflection and absorption properties of crops within the electromagnetic spectrum, particularly the response differences in the visible to NIR bands (Tufail et al., 2025). The reflectance differences between crops in the red-edge and NIR bands are particularly significant and are often used to construct vegetation indices to enhance crop differentiation capabilities. As shown in Table 2, this study selected commonly used spectral features, including NDVI and its derived indices (e.g., ratio vegetation index (RVI) and modified simple ratio (MSR)). Temporal feature analysis was based on multi-temporal remote sensing data and quantified the dynamic changes in vegetation parameters throughout the crop growth cycle, providing valuable insights into crop development and improving crop type identification. To address the inherent irregularity of time in the image data, we normalized the time series by interpolating the data to regular time intervals before applying the S-G filter. This ensured that the filter accurately accounted for the temporal irregularities. Given the differences in the growth cycles of different crops, their time-series curves exhibit unique morphological features, which can effectively reveal subtle differences between crops. These features are key parameters for constructing high-precision classification models, and in large-scale crop classification, temporal features can significantly improve classification accuracy.

Table 2 Vegetation indices used in this study

Full form	Definition	Calculation formula	Reference
Normalized difference vegetation index	NDVI	$\frac{\mathrm{NIR}-R}{\mathrm{NIR}+R}$	Peng et al., 2025
Ratio vegetation index	RVI	$\frac{\mathrm{NIR}}{R}$	Zhou ZJ et al., 2017
Modified simple ratio	MSR	$\frac{\frac{\mathrm{NIR}}{R}-1}{\sqrt{\frac{\mathrm{NIR}}{R}+1}}$	Liu Y et al., 2024
Green ratio vegetation index	GRVI	$\frac{\mathrm{NIR}}{G}$	Li WC et al., 2023
Renormalized difference vegetation index	RDVI	$\frac{\mathrm{NIR}-R}{\sqrt{\mathrm{NIR}+R}}$	Gámez et al., 2025
Modified chlorophyll absorption ratio index	MCARI	$\frac{(\mathrm{NIR}-R)-0.2(\mathrm{NIR}-G)}{\mathrm{NIR}/R}$	Marcone et al., 2024
Green-normalized difference vegetation index	GNDVI	$\frac{\mathrm{NIR}-G}{\sqrt{\mathrm{NIR}+G}}$	Ji et al., 2024
Wide dynamic range vegetation index	WDRVI	$\frac{0.12\mathrm{NIR}-R}{0.12\mathrm{NIR}+R}$	Testa et al., 2018
Soil-adjusted vegetation index	SAVI	$\frac{1.5(\mathrm{NIR}-R)}{\mathrm{NIR}+R+0.5}$	Haseeb et al., 2025
Enhanced vegetation index	EVI	$\frac{2.5(\mathrm{NIR}-R)}{\mathrm{NIR}+6R-7.5B+1}$	Tesfaye et al., 2025
Difference vegetation index	DVI	$\mathrm{NIR}-R$	Xu et al., 2025
Optimized soil-adjusted vegetation index	OSAVI	$\frac{1.16(\mathrm{NIR}-R)}{\mathrm{NIR}+R+0.16}$	Fern et al., 2018
Modified triangular vegetation index 2	MTVI2	$\frac{1.5\left[1.2(\mathrm{NIR}-G)-2.5(R-G)\right]}{\sqrt{(2\mathrm{NIR}+1)^2-(6\mathrm{NIR}-5R)-0.5}}$	Naqvi et al., 2021
Modified simple ratio based on green band	MSR_G	$\frac{\frac{\mathrm{NIR}}{G}-1}{\sqrt{\frac{\mathrm{NIR}}{G}+1}}$	Marin et al., 2021

R, *G*, *B*, and NIR represent the spectral reflectance values in the red, green, blue, and near-infrared bands, respectively.

2.5 mRMR feature selection algorithm

To address the issue of high dimensionality and suboptimal classification performance in time-series remote sensing data, this study employs the mRMR algorithm for feature selection, which uses a dual optimization criterion to perform feature selection. The first criterion is to maximize relevance, selecting features that are strongly correlated with the classification target. We employ the Pearson correlation coefficient and other statistical measures to quantify the degree of association between features and the target variable, ensuring that the selected features have significant discriminative power (Jiang et al., 2024). The second criterion is to minimize redundancy, which reduces the information redundancy between features. By calculating the mutual information and similarity between features, redundant features are removed, while highly complementary feature combinations are retained (Ren et al., 2021). This dual optimization mechanism allows the mRMR algorithm to effectively reduce the data dimensionality while preserving feature discriminability. In the experiment, we used the Pearson correlation and mutual information as quantifying metrics, and applied the feature subset selected by mRMR to improve the running efficiency of the classification algorithm and enhance the classification accuracy of the model, making it suitable for processing multi-temporal, high-dimensional remote sensing data.

$$\mathrm{mRMR}(S)=\frac{1}{|S|}\sum_{X_i\in S}I(X_i,Y)-\frac{1}{|S|^2}\sum_{X_i\in S,X_j\in S}I(X_i,X_j), \tag{1}$$

where $I(X_i, Y)$ refers to the mutual information shared between feature X_i and the target variable Y, $I(X_i, X_j)$ corresponds to the mutual information shared between features X_i and X_j, and S refers to the selected feature subset. The target variable Y represents labels that are highly related to the classification objective. The first term of the objective function indicates feature–target relevance and the second term represents redundancy among features. By optimizing this objective function, mRMR seeks to maximize the relevance between features and the target variable while minimizing redundancy among the selected features.

2.6 NDVI time-series feature construction and phenological extraction method

On the basis of official statistical data and field phenological observations, this study delineated the typical growth stages of cotton, winter jujube, and tiger nut from March to October (Fig. 2) and constructed a time-series feature framework covering the entire growth cycle of major crops in the study area. For each crop, the key phenological stages were identified according to their specific developmental processes. On this basis, we defined four key temporal windows with advantages for remote sensing identification: early sowing stage, rapid growth stage, peak growth stage, and maturity–senescence stage. The definition of these windows was grounded in the distinctive biological and ecological characteristics of each phase, which in turn manifested as unique temporal signatures in remote sensing-derived parameters such as NDVI. During the sowing stage, vegetation cover is minimal and NDVI values are relatively low, which facilitates the distinction between cultivated land, non-cultivated land, and other vegetation types. The rapid growth stage is characterized by a swift increase in biomass and leaf area, resulting in a steep upward slope in the NDVI profile. In contrast, during the peak growth stage, NDVI values reach their maximum and the differences in time-series NDVI curves among crop types become most pronounced, making this stage critical for achieving high-accuracy classification using remote sensing. Finally, the maturity–senescence stage is marked by a decline in photosynthetic activity and chlorophyll content, leading to a characteristic decrease in NDVI. The identification of these temporal windows provides a clear and scientifically grounded time reference for subsequent NDVI feature extraction and classification model construction.

To obtain a complete NDVI time series, we constructed a multi-temporal NDVI dataset covering the full crop growth cycle using 75 scenes of Sentinel-2 Level-2A imagery acquired in 2024. After atmospheric correction, geometric registration, and cloud masking, we calculated NDVI values using the red band (B4) and NIR band (B8). We then smoothed the NDVI time series using the S-G filter, which effectively suppressed high-frequency noise and reconstructed unimodal NDVI curves representing crop development from emergence through growth to senescence (Fig. 3). To further quantify key phenological stages, this study

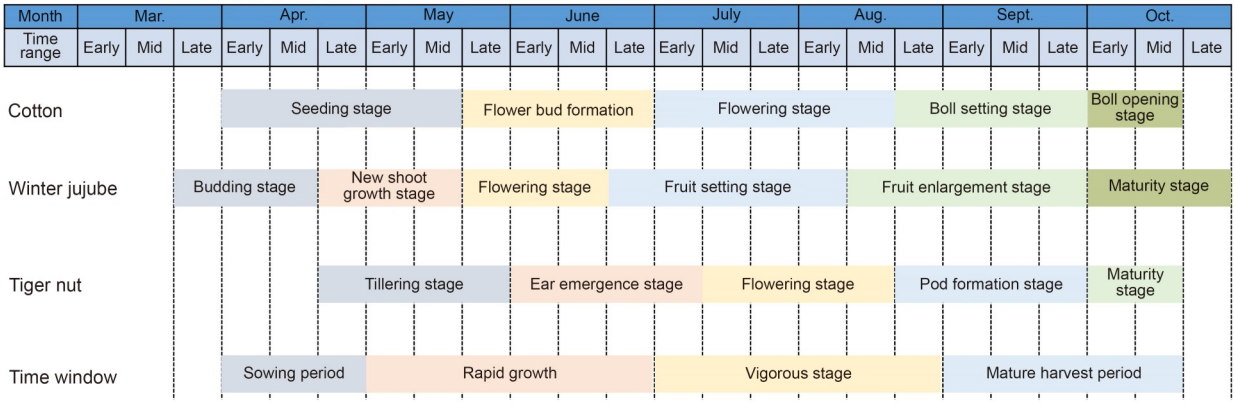

Fig. 2 Phenological calendar and remote sensing time windows for the three major crops—cotton, winter jujube, and tiger nut—in the study area.

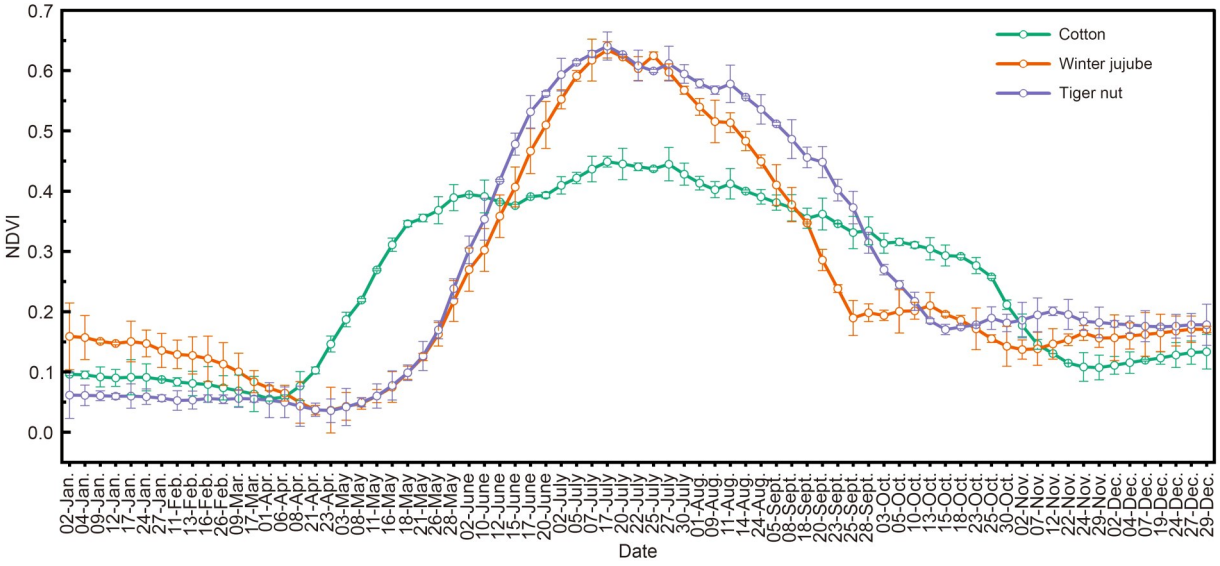

Fig. 3 Normalized difference vegetation index (NDVI) time-series curves of cotton, winter jujube, and tiger nut in 2024, smoothed using the Savitzky-Golay (S-G) filter.

employed the TIMESAT tool for curve fitting and analysis of the NDVI time series. By applying S-G smoothing and a double logistic fitting function, remote sensing-derived phenological parameters—such as SOS, POS, and EOS—could be automatically extracted for each pixel. To validate the accuracy of these extracted phenological parameters (SOS, POS, and EOS), we compared them with agricultural field survey data, including field observations of key phenological events such as crop emergence, flowering, and senescence dates, allowing us to assess the error margins of the TIMESAT-extracted values and evaluate their alignment with actual field observations. Although these parameters differ in definition from agronomic terms such as sowing period, rapid growth, peak growth, and maturity–senescence stages, they show good temporal and developmental correspondence. This enables effective alignment between remote sensing monitoring outputs and practical agricultural applications, providing support for time-window identification and enhancement of input features in classification models.

Tumushuke is located in an arid region of the mid-temperate zone where agricultural production typically follows a single-cropping system each year. As a result, the NDVI curves of all three crop types exhibit a typical unimodal pattern. However, the NDVI trajectories show clear inter-crop differences: tiger nut exhibits an earlier onset and peak, winter jujube presents a stable curve with a prolonged high-value period,

and cotton reaches its NDVI peak in July, reflecting a longer growth cycle. The phenological metrics extracted using TIMESAT—through S-G smoothing and double logistic fitting—enhanced the identification of key NDVI transition points, providing high-quality, structured input features for subsequent deep learning-based classification modeling. These phenological parameters, combined with spectral bands and vegetation indices selected by the mRMR algorithm, form a multidimensional feature set that is directly fed into all subsequent classification models. By integrating phenological temporal features with spectral characteristics, this strategy enables deep learning models to simultaneously leverage both spectral responses and phenological rhythm differences among crops, thereby achieving a more accurate discrimination of crop types in complex agricultural landscapes.

2.7 Model construction

2.7.1 MLP model

The MLP is one of the most fundamental feedforward neural network models in deep learning (Cheng et al., 2022). It consists of multiple layers—typically at least three—including an input layer, one or more hidden layers, and an output layer. Each layer comprises a number of neurons, and information is transmitted and processed through weighted connections between layers. MLPs are widely applied in a variety of tasks, including classification, regression, and pattern recognition (Gao Y et al., 2025). The MLP serves as a fundamental baseline model in this study, with its input being a high-dimensional feature vector derived from the same processed feature set as other models. Specifically, this vector integrates multi-scale spatial statistics computed from the key spectral bands and vegetation indices, alongside the three phenological parameters (SOS, POS, and EOS). We preselected these features from a larger candidate pool using the mRMR algorithm to ensure discriminative power and reduce redundancy. The MLP architecture comprises two to three hidden layers, each containing 64 to 128 neurons with rectified linear unit (ReLU) activation, and an output layer of three neurons with softmax for classification. We trained the model using the Adam optimizer (learning rate 0.001–0.01) for 50–100 epochs with a batch size of 32 or 64. To mitigate overfitting, we applied dropout (rate 0.2–0.5) and early stopping (patience 10–15). This configuration resulted in approximately 1.09 million trainable parameters, a scale that corresponds to the high-dimensional input designed to encapsulate both spatial-contextual and phenological information.

$$\boldsymbol{Z}^{(l+1)} = \boldsymbol{W}^{(l)}\boldsymbol{h}^{(l)} + \boldsymbol{b}^{(l)}, \tag{2}$$

$$\boldsymbol{h}^{(l+1)} = f\left(\boldsymbol{Z}^{(l+1)}\right), \tag{3}$$

$$\mathrm{ReLU}\left(\boldsymbol{x}\right) = \max\left(0, \boldsymbol{x}\right), \tag{4}$$

$$L = -\sum_{i=1}^{n} y_i \ln\left(\hat{y}_i\right), \tag{5}$$

where $\boldsymbol{W}^{(l)}$ denotes the weight matrix of size $n_{l+1}\times n_l$ encoding the connection strength between layer l+1 and layer l, $\boldsymbol{h}^{(l)}$ is the output vector of length n_l from layer l, $\boldsymbol{b}^{(l)}$ represents the bias vector of length n_{l+1} that shifts the neuron's activation, $\boldsymbol{Z}^{(l+1)}$ is the weighted input vector of length n_{l+1} to layer l+1, and f denotes the activation function that introduces non-linear transformations. ReLU is the ReLU activation function, $\boldsymbol{x}$ is the scalar input to the activation function, L is the cross-entropy loss, n is the total number of samples, y_i refers to the ground-truth label, and $\hat{y}_i$ indicates the predicted probability produced by the model.

2.7.2 ResNet-18 model

ResNet-18, as a significant milestone in deep CNNs, was first introduced by the Kaiming HE team at the 2015 Computer Vision and Pattern Recognition (CVPR) conference on computer vision (Pandey et al., 2025). As the number of layers in a network increases, model performance typically deteriorates rather than improves. Through the use of skip connections, the concept of residual learning, introduced by ResNet, allows the network to directly learn the residual mapping between input and output. This clever design effectively solves the vanishing gradient problem, making the training of deep networks feasible. Specifically, ResNet-18 consists of 16 convolutional layers and 2 fully connected layers (excluding pooling and activation layers). Its elegant residual block design and moderate computational complexity make it an ideal choice for many practical applications. In tasks such as remote sensing image classification, ResNet-18 offers sufficient feature extraction capability without imposing an excessive computational burden, demonstrating excellent cost-effectiveness (Ren ZQ et al., 2025).

The network input is derived from the same processed feature set as other models, integrating

multi-scale spatial statistics from key spectral bands and vegetation indices, along with the three phenological parameters (SOS, POS, and EOS). These selected features are organized into a multi-channel tensor suitable for convolutional processing. The first convolutional layer employs a 7×7 kernel with a stride of 2, generating 64 feature maps. Feature extraction is performed through residual blocks, each consisting of two 3×3 convolutional layers, followed by batch normalization and ReLU activation. Some convolutional layers use a stride of 2 for downsampling, while skip connections enable residual learning. The fully connected layer outputs three neurons corresponding to the cotton, jujube, and chufa crop categories, with a softmax function used to estimate class probabilities. The model is typically trained using the Adam optimizer with a learning rate ranging from 0.001 to 0.01, for 50 to 100 epochs, and augmented with early stopping to prevent overfitting. To enhance generalization, dropout rates between 0.2 and 0.5 are applied in the fully connected layers. The model contains approximately 11.7 million trainable parameters.

$$H(\boldsymbol{x})=F(\boldsymbol{x})+\boldsymbol{x}, \tag{6}$$

$$H(\boldsymbol{x})=F(\boldsymbol{x})+\boldsymbol{W}_s\boldsymbol{x}, \tag{7}$$

$$\boldsymbol{Y}_{i,j,k}=\sum_{m=1}^{C_{\text{in}}}\sum_{p=1}^{K}\sum_{q=1}^{K}\boldsymbol{W}_{k,m,p,q}\boldsymbol{X}^{m}_{i+p,j+q}, \tag{8}$$

$$\mathrm{BN}(x)=\gamma\frac{x-\mu}{\sqrt{\sigma^2+\epsilon}}+\beta, \tag{9}$$

where $\boldsymbol{x}$ denotes the input to the residual module and $F(\boldsymbol{x})$ denotes the intermediate output obtained after two convolutional layers, each followed by batch normalization (BN) and ReLU activation, with the exception that the final ReLU is applied after the residual addition. The final output of the residual block is denoted as $H(\boldsymbol{x})$, which is computed via a residual connection. Two types of connections are used: an identity mapping $\boldsymbol{x}$ when the dimensions match; and a projection mapping $\boldsymbol{W}_s\boldsymbol{x}$ when dimensions change, where $\boldsymbol{W}_s$ is a 1×1 convolution applied to $\boldsymbol{x}$ for dimensionality alignment and s denotes the stride. Let $\boldsymbol{X}$ be the input feature map with dimensionality $H\times W\times C_{\text{in}}$ and $\boldsymbol{W}$ be the convolutional kernel of size $K\times K\times C_{\text{in}}\times C_{\text{out}}$ (here, K indicates the kernel size, H is height, W is weight, C_{in} is input channels, and C_{out} is out channels). The output feature map is denoted as $\boldsymbol{Y}$, i and j are the spatial indices of the output feature map $\boldsymbol{Y}$, and k represents the output channel index. In Eq. (9), γ and β are learnable scale and shift parameters, respectively, μ and σ^2 are the mini-batch mean and variance, respectively, ϵ is a small constant for numerical stability, and x is the scalar input to the BN layer.

2.7.3 ConvLSTM model

In the field of deep learning, researchers have been exploring improved methods to handle spatio-temporal sequence data. In 2015, the Xingjian SHI team proposed an innovative solution—the ConvLSTM model (Wang LY et al., 2025). The brilliance of this model lies in its ability to combine the strengths of CNNs in extracting spatial features with the temporal modeling capability of LSTM. Traditional LSTM models, when dealing with image sequences or spatially structured data, often require flattening the two-dimensional features, which inevitably leads to the loss of important spatial information. ConvLSTM addresses this issue by replacing fully connected operations with convolutional operations (Li GG et al., 2024). Specifically, it retains the ability of LSTM to model temporal dynamics while simultaneously capturing local spatial features.

The input to the ConvLSTM model follows the same feature-construction pipeline as other models, forming a chronologically ordered sequence where each time step comprises a feature vector that encapsulates multi-scale spatial statistics and spectral-phenological attributes. This sequence structure enables the network to inherently model temporal dynamics. The architecture consists of a two-layer ConvLSTM stack, with each layer containing 64 to 128 units, utilizing the hidden states from all time steps, while the output layer consists of three neurons with softmax activation for multiclass classification. The model is optimized using Adam (learning rate 0.001–0.01), trained for 50–100 epochs with a batch size of 32. Dropout (rate 0.2–0.5) and early stopping are applied for regularization. This design enables the ConvLSTM to jointly capture localized spatial patterns and long-range temporal dependencies, effectively enhancing classification performance.

$$\boldsymbol{i}_s=\sigma\left(\boldsymbol{W}_{\text{xi}}\boldsymbol{X}_s+\boldsymbol{W}_{\text{hi}}\boldsymbol{H}_{s-1}+\boldsymbol{W}_{\text{ci}}\odot\boldsymbol{C}_{s-1}+\boldsymbol{b}_{\text{i}}\right), \tag{10}$$

$$\boldsymbol{f}_s=\sigma\left(\boldsymbol{W}_{\text{xf}}\boldsymbol{X}_s+\boldsymbol{W}_{\text{hf}}\boldsymbol{H}_{s-1}+\boldsymbol{W}_{\text{cf}}\odot\boldsymbol{C}_{s-1}+\boldsymbol{b}_{\text{f}}\right), \tag{11}$$

$$\boldsymbol{C}_s=\boldsymbol{f}_s\odot\boldsymbol{C}_{s-1}+\boldsymbol{i}_s\odot\tanh\left(\boldsymbol{W}_{\text{xc}}\boldsymbol{X}_s+\boldsymbol{W}_{\text{hc}}\boldsymbol{H}_{s-1}+\boldsymbol{b}_{\text{c}}\right), \tag{12}$$

$$\boldsymbol{o}_s=\sigma\left(\boldsymbol{W}_{\text{xo}}\boldsymbol{X}_s+\boldsymbol{W}_{\text{ho}}\boldsymbol{H}_{s-1}+\boldsymbol{W}_{\text{co}}\odot\boldsymbol{C}_s+\boldsymbol{b}_{\text{o}}\right), \tag{13}$$

$$\boldsymbol{H}_s=\boldsymbol{o}_s\odot\tanh\left(\boldsymbol{C}_s\right), \tag{14}$$

where $\boldsymbol{X}_s$ denotes the current input vector, $\boldsymbol{H}_s$ and $\boldsymbol{H}_{s-1}$ represent the hidden states (i.e., the output of the ConvLSTM unit) at the current and previous time steps, respectively, and $\boldsymbol{C}_s$ and $\boldsymbol{C}_{s-1}$ are the cell states at the current and previous time steps, respectively. $\boldsymbol{i}_s$, $\boldsymbol{f}_s$, and $\boldsymbol{o}_s$ refer to the input, forget, and output gate activation vectors, respectively. σ denotes the sigmoid activation function, $\odot$ denotes the Hadamard (element-wise) product, and tanh refers to the hyperbolic tangent activation function. The weight matrices $\boldsymbol{W}_{\mathrm{xi}}$, $\boldsymbol{W}_{\mathrm{xf}}$, $\boldsymbol{W}_{\mathrm{xc}}$, and $\boldsymbol{W}_{\mathrm{xo}}$ connect the input $\boldsymbol{X}_s$ to the input, forget, cell, and output gates, respectively; $\boldsymbol{W}_{\mathrm{hi}}$, $\boldsymbol{W}_{\mathrm{hf}}$, $\boldsymbol{W}_{\mathrm{hc}}$, and $\boldsymbol{W}_{\mathrm{ho}}$ connect the previous hidden state $\boldsymbol{H}_{s-1}$ to the corresponding gates; $\boldsymbol{W}_{\mathrm{ci}}$, $\boldsymbol{W}_{\mathrm{cf}}$, and $\boldsymbol{W}_{\mathrm{co}}$ are the peephole connections from the cell state to the input, forget, and output gates, respectively; and $\boldsymbol{b}_{\mathrm{i}}$, $\boldsymbol{b}_{\mathrm{f}}$, $\boldsymbol{b}_{\mathrm{c}}$, and $\boldsymbol{b}_{\mathrm{o}}$ denote the bias vectors for the corresponding gates.

2.7.4 Transformer model

The Transformer is a neural network architecture based primarily on the self-attention mechanism, first proposed by Vaswani and colleagues in 2017 (Lu et al., 2025). However, since its introduction, the Transformer architecture has transcended the boundaries of natural language processing (NLP), achieving remarkable progress in fields like computer vision, speech recognition, and even remote sensing classification (Han et al., 2025). In fact, its widespread success has spurred the creation of several specialized models—Bidirectional Encoder Representations from Transformers (BERT), Vision Transformer (ViT), and Generative Pre-trained Transformer (GPT)—each crafted to optimize performance for specific data types and tasks (Tran et al., 2025).

Importantly, the Transformer captures both spatial and temporal dependencies by employing a self-attention mechanism over multi-temporal feature sequences. These sequences are constructed by first extracting a rich set of candidate features—including multi-scale statistics from key spectral bands and vegetation indices, as well as phenological parameters (SOS, POS, and EOS)—and then selecting the most informative features via the mRMR algorithm. The selected features are organized chronologically into a sequence, where each time step contains a feature vector representing both the spectral response and spatial context at that moment. The model is optimized using the Adam optimizer with a learning rate set between 0.001 and 0.01, and trained for 50 to 100 epochs with a batch size of 64. Dropout rates ranging from 0.2 to 0.5 are applied to prevent overfitting. The classification head consists of three output neurons corresponding to the crop classes (cotton, winter jujube, and tiger nut), with a softmax activation function to predict class probabilities. Regularization techniques, including dropout and early stopping, are employed to further reduce the risk of overfitting. The early stopping mechanism monitors the validation set and halts training when the performance plateaus, preventing overfitting to the training data. Furthermore, the multi-scale fusion strategy enhances the generalization ability of the model, making it more robust to spatial variations across different crop types.

$$\mathrm{Attention}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{V}, \quad (15)$$

$$\mathrm{MultiHead}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\mathrm{Concat}\left(\mathrm{head}_1,\mathrm{head}_2,\cdots,\mathrm{head}_h\right)\boldsymbol{M}^{\mathrm{O}}, \quad (16)$$

$$\mathrm{head}_i=\mathrm{Attention}\left(\boldsymbol{Q}\boldsymbol{M}_i^{\mathrm{Q}},\boldsymbol{K}\boldsymbol{M}_i^{\mathrm{K}},\boldsymbol{V}\boldsymbol{M}_i^{\mathrm{V}}\right), \quad (17)$$

$$\mathrm{FFN}\left(\boldsymbol{x}\right)=\mathrm{ReLU}\left(\boldsymbol{x}\boldsymbol{M}_1+\boldsymbol{b}_1\right)\boldsymbol{M}_2+\boldsymbol{b}_2, \quad (18)$$

$$\mathrm{LayerOutput}\left(\boldsymbol{x}\right)=\mathrm{LayerNorm}\left(\boldsymbol{x}+\mathrm{SubLayer}\left(\boldsymbol{x}\right)\right), \quad (19)$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ denote the query, key, and value matrices, respectively, d_{k} denotes the dimensionality of the key vectors, and softmax is used for scaling the attention scores during normalization. $\boldsymbol{M}_i^{\mathrm{Q}}$, $\boldsymbol{M}_i^{\mathrm{K}}$, and $\boldsymbol{M}_i^{\mathrm{V}}$ represent the query, key, and value linear projection matrices for each attention head (head_i), respectively, $\boldsymbol{M}^{\mathrm{O}}$ is the output linear transformation matrix, h indicates the number of attention heads, and Concat is a concatenation operation. $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ denote the weight matrices of the two linear layers in the feed-forward network, respectively, and $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ represent the bias vectors of the two linear layers, respectively. $\boldsymbol{x}$ denotes the input vector to the feed-forward network (FFN) in Eq. (18), and the input vector to the sub-layer (SubLayer) before applying the residual connection and layer normalization (LayerNorm) in Eq. (19).

2.7.5 Random forest classifier

The RFC is an ensemble learning algorithm that enhances predictive accuracy and model stability by constructing a collection of decision trees and aggregating their outputs. Unlike neural networks that rely on gradient-based backpropagation to update parameters, RFC builds each tree independently using bootstrap

sampling and random feature selection. For classification tasks, the final predictions are determined by majority voting across all trees, while in regression scenarios, predictions are typically averaged (Chaudhary et al., 2016). A random forest is composed of an ensemble of decision trees, each trained using two independent sources of randomness to promote diversity within the model (Mishra et al., 2025). The first source is sample-level randomness, introduced through bootstrap sampling: each tree is built on a training set generated by sampling with replacement from the original dataset. The second source is feature-level randomness: at each node, only a randomly selected subset of features is evaluated when determining the optimal split. This fusion of sampling strategies significantly boosts the ability of the model to generalize, while it simultaneously mitigates the risk of overfitting. This translates into more robust, reliable models, making RFCs increasingly favored across a wide range of practical domains.

2.7.6 Multi-scale feature fusion

To enhance the capacity of the model to capture spatial heterogeneity in remote sensing imagery, this study introduces a multi-scale feature fusion strategy. By leveraging NDVI time-series raster data generated through the GEE platform, the model gains the ability to process intricate spatial variations with greater precision and depth. We utilized a local data processing workflow based on Python 3.10, incorporating specialized tools such as rasterio, NumPy, and Geospatial Data Abstraction Library (GDAL) to construct multi-scale spatiotemporal features. Specifically, rasterio was used to read Sentinel-2 NDVI time-series data (75 scenes throughout the year) and perform sliding window operations, while GDAL was employed for coordinate system alignment and spatial resampling to ensure geometric consistency. It is crucial to note that this is a spatial feature extraction process, not a temporal smoothing operation like the S-G filter. In the feature extraction phase, multi-scale analysis windows (ranging from 1×1 to 9×9) are applied and the vectorized operations of NumPy are used to compute the weighted average of the NDVI time series within each scale neighborhood, extracting local statistical features such as mean and variance. Finally, by aligning along the temporal dimension and concatenating along the channel dimension, a structured input tensor with spatiotemporal contextual awareness is constructed. The experimental design follows a controlled variable method (Huang et al., 2024). Initially, a comparative analysis of several deep-learning models—including MLP, ResNet-18, ConvLSTM, Transformer, and RFC—was conducted at a single spatial scale to identify the optimal model, and the Transformer model was found to perform the best. Based on this result, we further examined the performance of the Transformer model across various spatial scale configurations. With consistent training data, network architecture, and hyperparameters, we then conducted a comparative experiment using different window combinations (e.g., 1×1, 1×1+3×3, and 1×1+3×3+5×5+7×7+9×9), based on the model that demonstrated optimal performance. This enabled the quantitative assessment of the role of spatially aware scales in crop classification.

2.8 Model evaluation metrics

In remote sensing image classification tasks, the performance of a model needs to be systematically evaluated from multiple dimensions. This study comprehensively selects six evaluation metrics with clear physical meaning and complementarity, including intersection over union (IoU), mean accuracy (A), Kappa coefficient (K), precision (P), recall (R), and F1 score (F_1). The selection of these metrics provides a comprehensive reflection of model performance in terms of spatial localization, class differentiation, and misclassification control. IoU quantifies the ability of the model to identify object boundaries by calculating the ratio of the overlapping area between the predicted segmentation region and the true label (Sarkar et al., 2023). Meanwhile, the Kappa coefficient measures the consistency between the classification results and the true labels, and mean accuracy reflects the balance of recognition across different land cover classes from another perspective (Medelytė et al., 2025). Additionally, precision and recall, as paired metrics, reflect the model's error tendencies from different aspects. Finally, the F1 score, as the harmonic mean of precision and recall, provides a balanced evaluation of model performance.

$$\mathrm{IoU} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}} + N_{\mathrm{FN}}}, \tag{20}$$

$$A = \frac{1}{k}\sum_{i=1}^{k} \frac{N_{\mathrm{TP},i} + N_{\mathrm{TN},i}}{N_{\mathrm{TP},i} + N_{\mathrm{TN},i} + N_{\mathrm{FN},i} + N_{\mathrm{FP},i}}, \tag{21}$$

$$K = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r}(x_{i+}x_{+i})}{N^2 - \sum_{i=1}^{r}(x_{i+}x_{+i})}, \tag{22}$$

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}}, \tag{23}$$

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}}, \tag{24}$$

$$F_1 = \frac{2PR}{P + R}, \tag{25}$$

where N_{TP} (TP: true positive) denotes the number of pixels correctly identified as belonging to the target class, N_{TN} (NP: true negative) denotes the number of pixels correctly identified as not belonging to the target class, N_{FP} (FP: false positive) denotes the number of pixels erroneously classified as the target class, and N_{FN} (FN: false negative) denotes the number of target-class pixels incorrectly classified as other categories. k refers to the total number of classes, $N_{TP,i}$ represents the number of correctly classified pixels in class i, and $N_{FN,i}$ is the number of misclassified (omitted) pixels in class i. N is the total number of samples, r is the number of classes in the confusion matrix, x_{ii} denotes the number of samples correctly classified as class i (the diagonal element of the confusion matrix), $x_{i+}=\sum_{j=1}^{r} x_{ij}$ denotes the total number of samples belonging to class i (row sum), and $x_{+i}=\sum_{j=1}^{r} x_{ji}$ denotes the total number of samples predicted as class i (column sum).

3 Results

3.1 Spectral feature analysis and selection

Fig. 4a shows the Pearson correlation heatmap between the selected spectral bands and vegetation indices. The diagram includes multispectral bands from Sentinel-2 (e.g., B2–B12) and commonly used vegetation indices such as NDVI, enhanced vegetation index (EVI), soil-adjusted vegetation index (SAVI), and modified chlorophyll absorption ratio index (MCARI), totaling 24 variables. This analysis of linear relationships at the pixel level informed our feature selection for the subsequent non-linear models. Overall, most vegetation indices showed positive correlations with red-edge and NIR bands (e.g., B5, B6, B8, and B8A), underscoring the dominant role of these bands in vegetation index construction. Meanwhile, some shortwave infrared bands (e.g., B11 and B12) exhibited weak negative correlations with indices such as modified

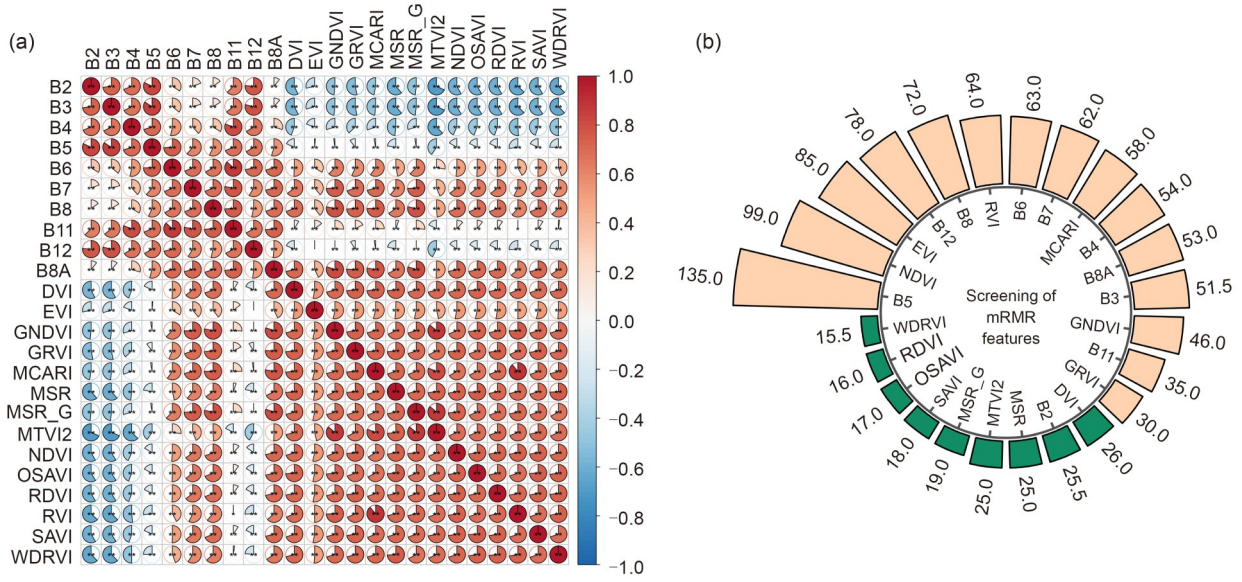

Fig. 4 Spectral correlation and feature selection results. (a) Pearson correlation heatmap among Sentinel-2 spectral bands and vegetation indices. The chromatic intensity in the spatial correlation map illustrates the degree of association: red hues represent positive associations, while blue tones signify inverse relationships, with intensified saturation corresponding to stronger correlations. Cells annotated with asterisks (*) and double asterisks (**) denote statistical significance at *P*=0.05 and *P*=0.01 thresholds, respectively. Furthermore, embedded circular quadrant visualizations within each grid encode both the directional orientation and scalar intensity of interdependencies. (b) Minimum redundancy maximum relevance (mRMR)-based ranking of feature importance for input variable selection. Details of Sentinel-2 bands (B2–B8A) and vegetation indices are shown in Tables 1 and 2.

triangular vegetation index 2 (MTVI2), suggesting their potential contribution to crop maturity detection. This correlation analysis provided a theoretical basis for subsequent feature dimensionality reduction and input variable selection.

To further optimize the model input variables, we applied the mRMR algorithm to assess the importance of 24 candidate features, including 10 Sentinel-2 spectral bands and 14 vegetation indices. To account for potential instability caused by the algorithm's inherent randomness, we determined the final importance score of each feature by averaging the results over 30 runs. Fig. 4b displays the features ranked by their mRMR scores in a descending order. Among them, B5, NDVI, and EVI showed the highest relevance and lowest redundancy, with scores of 135.0, 99.0, and 85.0, respectively. To determine the final input feature set, we adopted a threshold-based selection strategy using the "elbow point" of the mRMR score curve, combined with the trend of information gain and cumulative contribution. As a result, we selected the top 15 features with scores above GRVI (30 points), comprising 9 spectral bands (B5, B12, B8, B6, B7, B4, B8A, B3, and B11) and 6 vegetation indices (NDVI, EVI, RVI, MCARI, green-normalized difference vegetation index (GNDVI), and green ratio vegetation index (GRVI)). This strategy ensured both representativeness and complementarity of the selected inputs while effectively reducing feature dimensionality, thereby providing a solid data foundation for model development.

3.2 Comprehensive evaluation of classification models

3.2.1 Comparison of model performance on spatial classification

This study conducted modeling using the 15 selected features obtained from the mRMR algorithm. Fig. 5 presents the classification results of crop spatial distribution in three typical experimental zones (A, B, and C) within Tumushuke City, using five models—MLP, ResNet-18, ConvLSTM, Transformer, and RFC—all incorporating a multi-scale fusion mechanism (1×1+3×3+5×5). Significant differences in classification performance are observed across models. The Transformer model outperforms others in terms of boundary preservation, class discrimination, and noise suppression, particularly in transition zones between cotton and winter jujube. It produces classification maps that are spatially continuous and closely aligned with the actual field layout.

The ConvLSTM model, benefiting from temporal modeling, maintains structural coherence even in fragmented areas. However, some boundary blurring remains. ResNet-18, leveraging its spectral feature extraction capabilities, performs well in dominant crop zones but struggles with temporal adaptability, often leading to confusion among seasonally similar crops. The MLP model, due to its shallow architecture, exhibits isolated pixels and broken edges, failing to accurately capture the complexity of crop patterns. RFC shows the poorest performance in areas of intercropped fields, particularly between winter jujube and tiger nut, with frequent misclassifications—highlighting the limited generalization ability of traditional machine learning methods when applied to high-resolution remote sensing data.

Red rectangles in Fig. 5 indicate zoomed-in subregions used to validate the superiority of the Transformer model in terms of both accuracy and stability. As crop canopy structures stabilize, imagery from the late season provides stronger spectral contrast, enhancing model discrimination. Under late-season imagery, the Transformer model continues to deliver high boundary accuracy, especially in zone B with dense cropping where it effectively delineates small, irregular field parcels. ConvLSTM performs well in zones B and C, maintaining spatial coherence, but in zone A, it suffers from over-smoothing, leading to loss of local detail. While ResNet-18 maintains overall classification consistency, it performs less effectively at small field boundaries. MLP and RFC still exhibit severe misclassification and edge fragmentation, falling short of the precision required for high-resolution remote sensing-based crop classification.

3.2.2 Comparison of models' crop classification accuracy

Table 3 presents the overall classification accuracy of five models—MLP, ResNet-18, ConvLSTM, Transformer, and RFC—on three crop types (cotton, winter jujube, and tiger nut), all using the same multi-scale fusion configuration (e.g., 1×1+3×3+5×5). The results demonstrate clear differences in model performance. The Transformer model achieved the best results across all crops, benefiting from its global self-attention mechanism, which enhances adaptability to

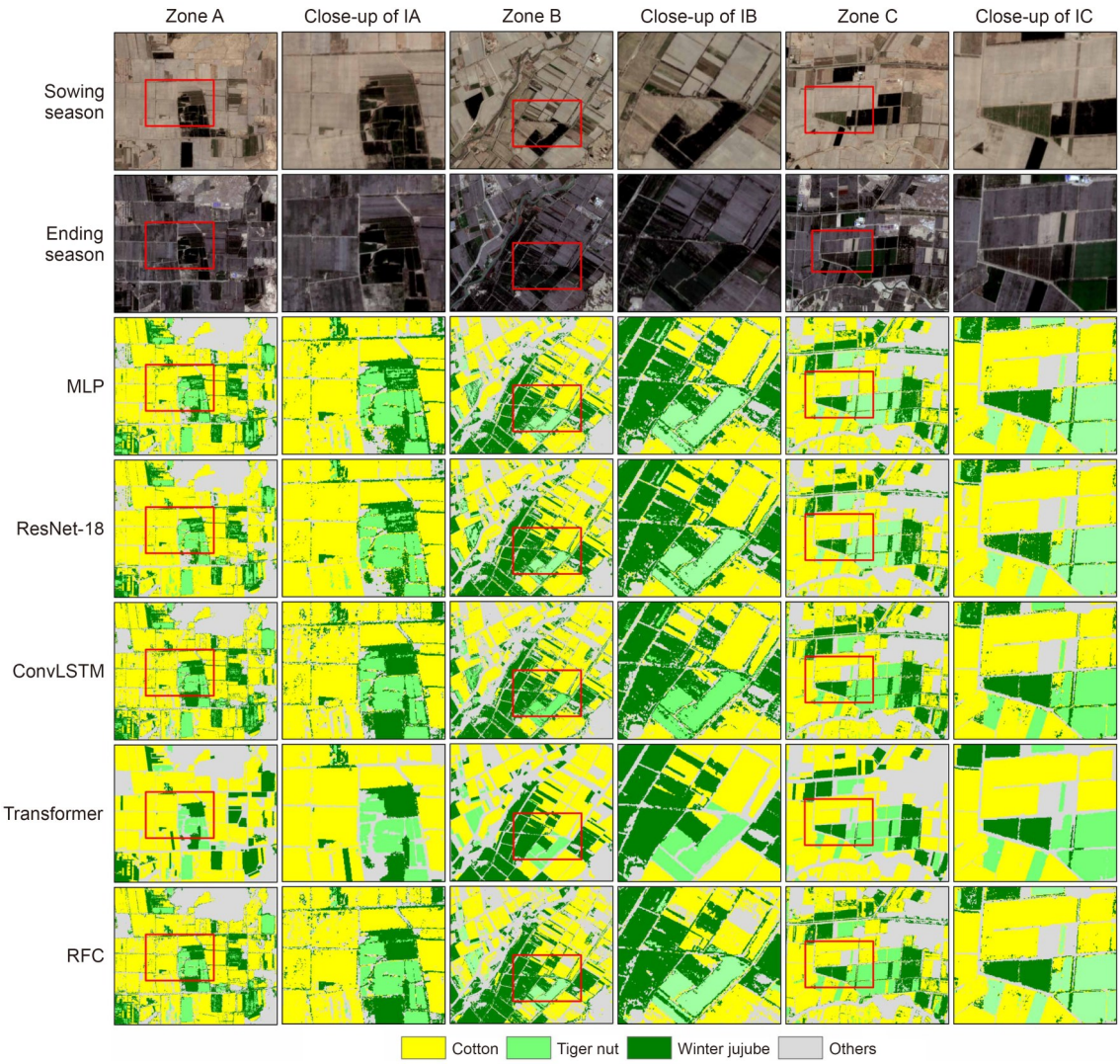

Fig. 5 Classification results of different models for the study area. MLP: multi-layer perceptron; ResNet-18: residual network-18; ConvLSTM: convolutional long short-term memory; RFC: random forest classifier.

field shapes and spectral variations. IoU scores of 92.74%, 91.33%, and 93.45% were recorded for cotton, winter jujube, and tiger nut, respectively, with F1 scores exceeding 90% for all three crops. This indicates its superior generalization ability and feature fusion capacity in capturing spatial information across different crop types. The ConvLSTM model ranked second, leveraging time-series information to capture phenological changes effectively, particularly excelling in tiger nut classification, with accuracy exceeding 87%. ResNet-18 performed well in identifying cotton due to its spectral feature extraction capabilities but struggled to distinguish between spectrally similar crops such as winter jujube and tiger nut. The MLP input layer size was 18; however, the final accuracy of this model was relatively low, likely due to the limitations of its shallow architecture in capturing and representing high-dimensional remote sensing features effectively. The RFC model showed moderate performance across all three crop types, underscoring the limited adaptability of traditional machine learning algorithms in handling complex, high-resolution remote sensing classification tasks.

3.3 Evaluation of multi-scale strategies in the Transformer model

Fig. 6 compares the crop classification results produced by the Transformer model under different spatial perception window combinations (1×1, 3×3, 5×5, 7×7, and 9×9) in three typical experimental zones

Table 3 Comparison of extraction accuracy among multiple classification models

Crop	Model	IoU (%)	Mean accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)
Cotton	MLP	74.72	77.45	71.91	76.73	73.11	73.85
	ResNet-18	88.85	89.47	87.12	89.88	88.16	89.18
	ConvLSTM	84.30	87.04	84.17	86.40	84.93	87.13
	Transformer	92.74	94.44	91.98	94.50	93.17	92.85
	RFC	80.63	83.01	78.32	82.12	81.03	80.58
Winter jujube	MLP	68.67	70.74	66.02	66.46	68.65	67.32
	ResNet-18	82.77	87.11	80.58	85.12	81.68	84.01
	ConvLSTM	86.41	88.03	84.42	87.55	85.92	87.14
	Transformer	91.33	93.87	88.52	92.83	90.93	90.49
	RFC	77.85	80.22	76.02	79.94	77.75	76.94
Tiger nut	MLP	69.55	71.88	68.63	70.21	69.14	70.97
	ResNet-18	85.69	86.47	82.45	85.11	85.86	85.26
	ConvLSTM	87.35	87.99	85.86	90.16	88.02	87.24
	Transformer	93.45	97.06	94.16	94.76	93.45	95.11
	RFC	82.41	85.09	79.68	82.86	82.93	80.93

IoU: intersection over union; MLP: multi-layer perceptron; ResNet-18: residual network-18; ConvLSTM: convolutional long short-term memory; RFC: random forest classifier.

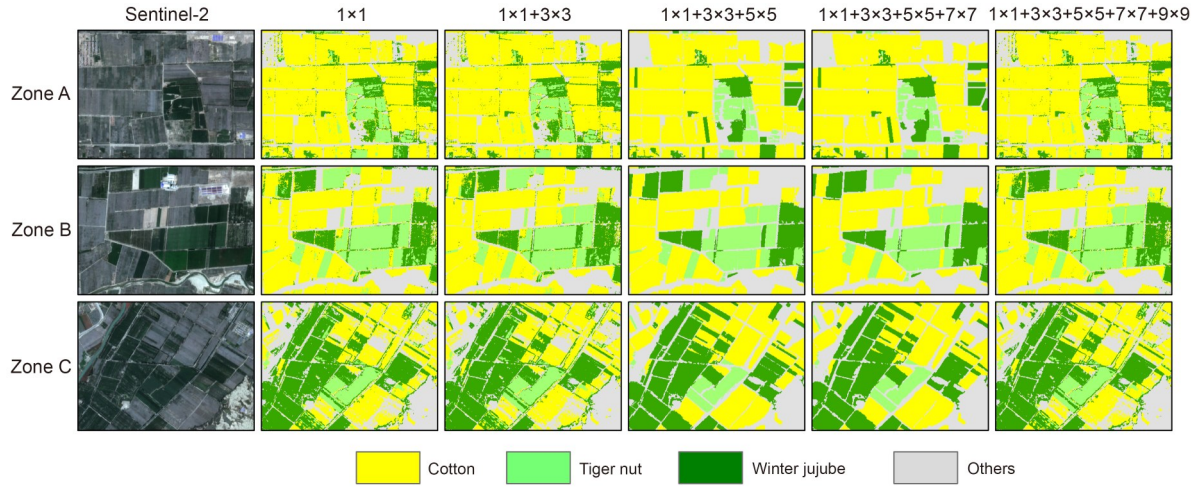

Fig. 6 Effects of different scale spatial windows on the crop classification performance of the Transformer model.

(A, B, and C) in Tumushuke City. The results indicate that using a single-scale window (e.g., 1×1) leads to unstable performance at field boundaries, with noticeable "salt-and-pepper" noise and the misclassification of small patches, resulting in poor spatial consistency. As additional convolutional windows (such as 3×3 and 5×5) are incorporated, classification noise is significantly reduced, field boundaries become clearer, and the spatial distribution of crop types becomes more stable and better aligned with the actual agricultural landscape. When fusing up to a 1×1+3×3+5×5 window combination, the model achieves a highly accurate delineation of crop boundaries, with greatly improved internal consistency and external continuity. Patch fragmentation is effectively suppressed, and elongated plots, especially in the central areas, maintain their structural integrity more clearly. Further expanding to 7×7 and 9×9 windows provides only marginal improvements, with diminishing returns on boundary enhancement while substantially increasing model complexity and computational cost. Large-scale and uniform farmlands typically exhibit more regular spatial structures. In such cases, a 5×5 convolutional kernel is effective in capturing broader spatial features, as a larger receptive field enables the capture of more comprehensive global information. In contrast, for smaller

and more fragmented farmlands, a 3×3 kernel is better suited to preserving local details, preventing the loss of important features due to excessive smoothing, thereby maintaining the accuracy of crop boundaries. However, the selection of the convolutional kernel should not solely depend on the size of the farmland but should also take into account specific factors such as crop type, soil characteristics, and vegetation cover. These findings suggest that a moderate multi-scale fusion strategy (e.g., 1×1+3×3+5×5) offers the best trade-off between accuracy and efficiency, as it effectively captures spatial contextual information while avoiding excessive computation and redundancy. Overall, in remote sensing-based crop classification, the design of an appropriate receptive field is crucial for accurate boundary detection, patch consistency, and discrimination between different crop types.

Table 4 investigates the classification performance of the Transformer model under different neighborhood window scales (from 1×1 to 9×9). The results show that incorporating multi-scale windows generally improves classification accuracy and reduces noise interference, indicating that an appropriate spatial context perception range has a significant positive impact on remote sensing-based crop recognition. In zone C, the improvement is most notable: by incorporating an optimal combination of window scales (1×1+3×3+5×5), IoU reaches a peak of 96.31% and the F1 score reaches 93.55%, suggesting that the spatial structure of this region is well-suited for deep model extraction. In zone B, a relatively challenging area for classification, multi-scale combinations substantially enhance the performance of the Transformer model. The IoU rises from 74.97% to a peak of 91.95% with the 1×1+3×3 combination, before stabilizing at 85.67% with larger window scales; the F1 score follows a similar trend, peaking at 93.17% and then leveling off around 84.80%. It is worth noting that zone A exhibits more variability across different window combinations. While the single-scale configuration (1×1) achieves an IoU of 88.87%, the inclusion of larger windows (e.g., 7×7 or 9×9) leads to a slight decline (down to 75.63%), possibly due to the compact distribution of fields in this area, where excessive receptive fields introduce redundant information.

Overall, moderate multi-scale combinations (e.g., 1×1+3×3, 1×1+3×3+5×5) strike the best balance between boundary integrity and noise suppression. In particular, the 1×1+3×3+5×5 setting in zone C achieved the highest performance, with IoU, mean accuracy, and F1 score reaching 96.31%, 96.85%, and 93.55%, respectively—outperforming all other configurations. Further expansion to 7×7 and 9×9 yielded diminishing returns in accuracy while significantly increasing network complexity and computational cost, which may hinder model deployment and generalization. Therefore, under the conditions of this study, the 5×5 window is considered the optimal configuration in

Table 4 Comparison of extraction accuracy across different windows

Different scale	Study area	IoU (%)	Mean accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)
1×1	Zone A	88.87	86.31	84.98	88.63	86.64	84.05
	Zone B	74.97	78.29	75.22	78.61	75.26	77.51
	Zone C	88.67	94.07	96.44	95.29	92.99	93.13
1×1+3×3	Zone A	84.86	86.38	88.26	86.22	88.98	87.89
	Zone B	91.95	93.25	91.08	96.32	92.94	93.17
	Zone C	86.13	83.65	83.34	86.32	88.63	85.32
1×1+3×3+5×5	Zone A	75.63	73.68	73.45	75.64	76.51	78.94
	Zone B	85.67	85.53	87.03	84.72	85.64	84.83
	Zone C	96.31	96.85	92.43	96.62	95.62	93.55
1×1+3×3+5×5+ 7×7	Zone A	75.63	73.68	73.45	75.64	76.51	78.94
	Zone B	85.67	85.53	87.03	84.72	85.64	84.83
	Zone C	91.95	93.25	91.08	96.32	92.94	93.17
1×1+3×3+5×5+ 7×7+9×9	Zone A	75.63	73.68	73.45	75.64	76.51	78.94
	Zone B	85.67	85.53	87.03	84.72	85.64	84.83
	Zone C	92.81	94.07	96.44	95.29	92.99	93.13

IoU: intersection over union.

terms of both effectiveness and efficiency. In contrast, the 1×1 window, representing pixel-level classification, often underperformed relative to optimal multi-scale combinations, highlighting the difficulty of accurately identifying crop types without spatial structural context. These experiments confirm the capability of Transformer in integrating spatial information and emphasize the need to balance generalization and redundancy when designing multi-scale structures.

3.4 Model generalization and classification accuracy in the independent validation region

To further validate the generalization capability of the proposed method under different regions and complex cropping structures, we selected an independent validation area located east of the main experimental zone in Tumushuke City. This area differs in field configurations and planting structures and was not used in model training or parameter tuning. The validation dataset included Sentinel-2 imagery at three key phenological stages: sowing, peak growth, and late growth. We compared model predictions with the actual remote sensing imagery and supported by high-resolution Google Earth images and a limited number of manually verified reference points to construct a "weakly labeled" ground truth layer.

Fig. 7 illustrates the classification results of the Transformer model for the validation area, using a multi-scale spatial perception window of 1×1+3×3+5×5. The model successfully delineates field boundaries for the major crops—cotton (yellow), winter jujube (dark green), and tiger nut (light green)—with high accuracy. The spatial patterns are closely aligned with actual field layouts. Even in areas with interleaved and mixed crop types, the model maintains edge continuity, and the classification results are clean, structured, and nearly noise-free. These findings demonstrate the generalization ability and robustness of the proposed method for cross-regional crop identification based on multi-temporal remote sensing data.

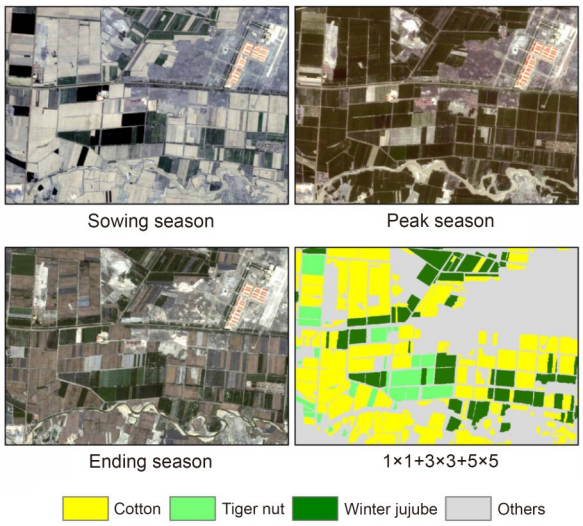

Fig. 7 Classification results for the validation region using the Transformer model with 1×1+3×3+5×5 multi-scale spatial window. The first three images depict Sentinel-2 scenes at different phenological stages, highlighting spectral and spatial variations in crop development. The lower right image shows the corresponding classification map.

To quantitatively assess classification performance for the validation set, Table 5 presents the evaluation metrics for the main crop categories. Among them, winter jujube achieved the highest classification accuracy, with an IoU of 92.03% and an F1 score of 94.37%. Cotton and tiger nut also exhibited consistent performance, with F1 scores of 87.75% and 86.35%, respectively. The training dataset in this study covers sufficient diversity and is large enough to effectively train the model in different environments. External validation provides a reliable assessment of the generalization ability of the Transformer model. This study used external validation to test its practical application performance, closely aligning with real-world scenarios. This is particularly important in remote sensing imagery and agricultural applications, where the model's ability to adapt to different regions is crucial. These results confirm the generalization ability and robustness of the proposed method for cross-regional crop classification:

Table 5 Accuracy metrics of major crop types in the independent validation region

Crop	IoU (%)	Mean accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)
Cotton	89.21	91.87	93.03	90.26	89.75	87.75
Tiger nut	85.51	87.83	89.12	86.74	85.96	86.35
Winter jujube	92.03	94.25	91.72	94.83	93.92	94.37

IoU: intersection over union.

it not only performed well within the main experimental area but also demonstrated transferability and practical applicability in regions with new field configurations and planting structures.

3.5 Time-series extraction and mapping of key crop growth periods

To scientifically reveal the temporal growth characteristics and spatial distribution of major crops in Tumushuke City, this study constructed a remote sensing recognition framework integrating multi-temporal Sentinel-2 imagery, NDVI time-series data, and deep learning with phenology models. This framework enables spatial classification and key phenological mapping for three major crops: cotton, winter jujube, and tiger nut. The spatial distribution map of major crops was generated by the Transformer model using a 1×1+3×3+5×5 multi-scale spatial perception window. The classification map displays clear boundaries, continuous field contours, and distinct spatial distribution patterns among the crop types. Cotton is primarily concentrated in the northern and eastern-central regions; winter jujube is mainly located in the southwest; and tiger nut appears scattered across the central and southern parts, reflecting typical spatial heterogeneity and intercropping features. The spatial distribution of the SOS was extracted by fitting and smoothing the NDVI time series using the TIMESAT tool. Most SOS values fall between day of year (DOY) 106–120, showing a north-to-south advancing pattern influenced by both planting schedules and regional climate gradients. The spatial distribution of the POS shows that most fields reach peak growth around DOY 190–198, while some late-sown plots are delayed beyond DOY 202. The POS map reveals a mosaic distribution of active growth zones, indicating asynchronous crop development caused by mixed cropping and varied sowing times—posing additional challenges for precision agricultural management. The spatial distribution of the EOS indicates the onset of crop maturity or harvest. Most fields reach EOS between DOY 284–301, with noticeable temporal variation across subregions, reflecting heterogeneity in the growing cycles and maturity rates within the study area. In summary, the developed classification and phenology extraction system enables the accurate spatiotemporal representation of crop distribution and growth dynamics, confirming the adaptability and scalability of the proposed approach for complex agricultural systems in arid regions through the integration of multi-source remote sensing data and deep learning techniques.

4 Discussion

The proposed method for extracting crop phenology, constructed based on multi-temporal Sentinel-2 imagery and NDVI time-series analysis, demonstrated robust performance and clear temporal differentiation. Tumushuke City, a representative arid oasis agricultural area in China, was selected as the study site to extract phenological features and classify three major crops: cotton, winter jujube, and tiger nut. Significant spectral differences were observed in the key phenological windows, particularly in April and September. Each crop exhibited distinct NDVI trajectories: cotton reached its peak around DOY 170, winter jujube sustained high NDVI values until approximately DOY 190–198, and tiger nut peaked earlier in the season. These temporal divergences reflect unique phenological cycles and provide a sound basis for remote sensing-based crop discrimination. To improve data accuracy and minimize noise, the NDVI time series underwent smoothing through the S-G filter and was subsequently fitted using a double logistic function. It should be clarified that the S-G filter was applied solely as a smoothing technique to the NDVI time series, and no derived index was input to the classification or phenology model. This preprocessing pipeline facilitated the automated identification of key phenological metrics, specifically SOS, POS, and EOS. These features served as high-value temporal indicators for crop classification and phenological monitoring. This finding is consistent with the results reported by Shojaeezadeh et al. (2025), which showed a low mean absolute error of 6 d and a reasonable precision of R^2>0.43 for Sentinel-2-based phenology prediction, thereby reinforcing the robustness of this method. The mapping results reveal clear crop boundary delineation and coherent spatial distribution, confirming the stability and adaptability of the method in diversified cropping systems. Overall, our findings validate the high application potential of time-series high-resolution remote sensing for accurate phenology extraction and crop mapping in complex agricultural environments.

In the multi-model comparison experiment, the Transformer model delivered the highest classification accuracy across the three selected regions (zones

A, B, and C) in Tumushuke City. Specifically, the IoU scores reached 92.74% for cotton, 91.33% for winter jujube, and 93.45% for tiger nut, with corresponding F1 scores of 92.85%, 90.49%, and 95.11%, respectively. These results considerably exceed the performance of traditional models such as ResNet-18, MLP, and RFC. The ConvLSTM model, which incorporates temporal sequence learning, also performed strongly—particularly in tiger nut and winter jujube classification—achieving F1 scores of 87.24% and 87.14%, respectively, slightly outperforming ResNet-18 and surpassing the shallow MLP and RFC models. The classification maps (Fig. 5) further confirm the effectiveness of the Transformer and ConvLSTM architectures in delineating crop boundaries, preserving field structures, and enhancing class separability. Notably, the superiority of Transformer can be attributed to its self-attention mechanism, which enables global contextual learning across spatial and temporal domains, consistent with findings by Guerri et al. (2025). In contrast, the RFC model exhibited significant boundary confusion and misclassification in intercropped zones, underscoring its limited capacity to handle complex, high-dimensional remote sensing inputs. These results highlight the need for deep architectures with temporal encoding capabilities in high-resolution agricultural classification tasks.

To further elevate the spatial representation capacity of the Transformer model, we assessed the impact of varying perception window sizes—ranging from 1×1 to 9×9—on its classification performance (Table 4 and Fig. 6). The 1×1+3×3+5×5 window configuration emerged as the best performer, achieving an IoU of 96.31%, mean accuracy of 96.85%, and an F1 score of 93.55%. This configuration effectively suppresses "salt-and-pepper noise" and excessive smoothing while maintaining a clear boundary definition, particularly achieving more natural boundary transitions in complex crop boundary regions. In comparison, the 1×1 window lacks spatial context information, leading to noticeable breaks in the classification results. While the 7×7 and 9×9 windows enhance contextual expression, they introduce excessive redundant information, causing boundary blurring. Yu et al. (2025) introduced an enhanced U-Net model, integrating multi-scale features to overcome the single receptive field limitation of U-Net, which also improved the model's ability to learn multi-scale features of typical land covers in complex coastal wetland areas. This experiment similarly validates that a well-designed multi-scale fusion mechanism is a key strategy for improving model discriminative ability and stability in high-resolution remote sensing crop identification, providing empirical support for the structural design of deep learning models.

Although this study achieved relatively high classification accuracy in the Tumushuke Oasis region, the proposed method exhibits some dependency on specific climatic zones, soil conditions, and cropping patterns; therefore, its regional adaptability remains to be further validated. Future work should incorporate multi-region transfer learning strategies to improve model robustness across different regions and under varying temporal conditions. Furthermore, while NDVI is a widely used metric, it is inherently sensitive to cloud contamination, soil background variations, and other sources of noise. This limitation suggests that future studies could benefit from integrating additional indicators or refining the use of NDVI to mitigate these issues. Incorporating multi-source data, such as Sentinel-1 Synthetic Aperture Radar (SAR) for structure, thermal infrared imagery for canopy temperature, and meteorological inputs, could improve the continuity and reliability of phenological monitoring. Additionally, although the Transformer and ConvLSTM models demonstrated superior accuracy, their computational demands are nontrivial, especially when applied with multi-scale input and long time series. Future work may investigate lightweight architectures, such as Swin-Transformer Lite or EfficientFormer, or adopt adaptive convolution mechanisms to reduce complexity while maintaining performance. Finally, while the proposed framework provides clear crop boundaries, it still struggles with fine-grained discrimination in intercropped areas. Enhancing ground truth quality through high-resolution sampling and spatio-temporal label augmentation would help refine training datasets and improve model generalization in complex planting systems.

5 Conclusions

This study investigated a typical oasis agricultural region in Tumushuke City, Xinjiang, China, conducting full-season remote sensing monitoring and crop classification for three representative crops—cotton, winter jujube, and tiger nut—by integrating multi-temporal

Sentinel-2 imagery with NDVI time-series data. A multidimensional feature system combining spectral bands and vegetation indices was constructed, followed by optimal feature selection using the mRMR algorithm. To improve time-series robustness, NDVI curves were smoothed using S-G filtering and fitted with a double logistic function. The key phenological metrics—SOS, POS, and EOS—were extracted automatically via the TIMESAT platform, supporting phenology-based crop classification. Among the tested models, the Transformer architecture, equipped with a global attention mechanism, delivered the highest classification performance across all crop types, with a mean IoU exceeding 91% and the highest F1 score reaching 95.11% for tiger nut. The ConvLSTM model also showed a certain temporal learning capacity, particularly for crops with distinct growth rhythms, and performed well in capturing spatial structure and boundary information under complex planting conditions. The proposed multi-scale perception module, integrating convolution windows from 1×1+3×3+5×5, demonstrated superior performance in maintaining field integrity and reducing classification noise. In particular, the 5×5 component contributed significantly to image consistency and boundary clarity. Independent validation confirmed the generalization ability of the selected models in heterogeneous cropping systems. Supported by the spatial resolution and revisit frequency of Sentinel-2, the results of this study demonstrated the potential of the proposed method as an effective toolset for fine-scale crop mapping and phenological monitoring at the regional level. Overall, the proposed framework highlights the practical value of combining multitemporal remote sensing with deep learning for scalable, high-precision agricultural management in arid environments. Future research can prioritize validating the model's temporal robustness across multiple years and pursuing model lightweighting for simpler operational deployment.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 32101621 and 62061041), the Bingtuan Science and Technology Program (No. 2022CB001-05), and the Graduate Scientific Research Innovation Project of Tarim University (No. TDGRI2024092), China.

Author contributions

Chunli WANG contributed to conceptualization, methodology, software development, formal analysis, data curation, writing–original draft, visualization, and writing–review & editing. Jianan CHI contributed to methodology, software development, and formal analysis. Xiao ZHANG contributed to conceptualization, validation, resources, supervision, and writing–review & editing. Nannan ZHANG contributed to methodology, validation, investigation, resources, supervision, project administration, funding acquisition, and writing–review & editing. All the authors have read and approved the final manuscript, and therefore, have full access to all the data in the study and take responsibility for the integrity and security of the data.

Compliance with ethics guidelines

Chunli WANG, Jianan CHI, Xiao ZHANG, and Nannan ZHANG declare that they have no conflicts of interest.

This article does not contain any studies with human or animal subjects performed by any of the authors.

Declaration on the use of generative AI tools

No generative AI tools were used in the preparation of this manuscript.

References

Cao J, Ma SP, Yuan WH, et al., 2022. Characteristics of diurnal variations of warm-season precipitation over Xinjiang Province in China. *Atmos Oceanic Sci Lett*, 15(2):100113. https://doi.org/10.1016/j.aosl.2021.100113

Chaudhary A, Kolhe S, Kamal R, 2016. An improved random forest classifier for multi-class classification. *Inf Process Agric*, 3(4):215-222. https://doi.org/10.1016/j.inpa.2016.08.002

Chen XM, Ding YL, Zheng XM, et al., 2024. Improved estimation of non-photosynthetic vegetation cover using a novel multispectral slope difference index with soil information, Sentinel-1 data, and machine learning. *Ecol Inform*, 84:102930. https://doi.org/10.1016/j.ecoinf.2024.102930

Cheng JR, Li H, Li DB, et al., 2022. A survey on image semantic segmentation using deep learning techniques. *Comput Mater Contin*, 74(1):1941-1957. https://doi.org/10.32604/cmc.2023.032757

de la Guardia L, de Miranda JH, dos Santos Luciano AC, 2024. Assessment of irrigation water use for dry beans in center pivots using ERA5 Land climate variables and Sentinel 2 NDVI time series in the Brazilian Cerrado. *Agric Water Manag*, 305:109128. https://doi.org/10.1016/j.agwat.2024.109128

Du RQ, Xiang YZ, Chen JY, et al., 2024. The daily soil water content monitoring of cropland in irrigation area using Sentinel-2/3 spatio-temporal fusion and machine learning. *Int J Appl Earth Obs Geoinf*, 132:104081. https://doi.org/10.1016/j.jag.2024.104081

Faqeerzada MA, Kim H, Kim MS, et al., 2025. Hyperspectral

imaging VIS-NIR and SWIR fusion for improved drought-stress identification of strawberry plants. *Comput Electron Agric*, 237:110702. https://doi.org/10.1016/j.compag.2025.110702

Farbo A, Sarvia F, de Petris S, et al., 2024. Forecasting corn NDVI through AI-based approaches using Sentinel 2 image time series. *ISPRS J Photogramm Remote Sens*, 211:244-261. https://doi.org/10.1016/j.isprsjprs.2024.04.011

Fern RR, Foxley EA, Bruno A, et al., 2018. Suitability of NDVI and OSAVI as estimators of green biomass and coverage in a semi-arid rangeland. *Ecol Indic*, 94(Pt 1): 16-21. https://doi.org/10.1016/j.ecolind.2018.06.029

Gámez AL, Segarra J, Vatter T, et al., 2025. Alfalfa yield estimation using the combination of Sentinel-2 and meteorological data. *Field Crops Res*, 326:109857. https://doi.org/10.1016/j.fcr.2025.109857

Gao XY, Chi H, Huang JL, et al., 2024. Comparison of cloud-mask algorithms and machine-learning methods using Sentinel-2 imagery for mapping paddy rice in Jianghan Plain. *Remote Sens*, 16(7):1305. https://doi.org/10.3390/rs16071305

Gao Y, Sun ZZ, Hu D, et al., 2025. GMOPNet: a GAN-MLP two-stage network for optical properties measurement of kiwifruit and peaches with spatial frequency domain imaging. *Food Chem*, 465(Pt 1):141944. https://doi.org/10.1016/j.foodchem.2024.141944

Ghilardi F, de Petris S, Torti V, et al., 2025. A possible role of NDVI time series from Landsat Mission to characterize lemurs habitats degradation in Madagascar. *Sci Total Environ*, 974:179243. https://doi.org/10.1016/j.scitotenv.2025.179243

Guerri MF, Distante C, Spagnolo P, et al., 2025. Boosting hyperspectral image classification with Gate-Shift-Fuse mechanisms in a novel CNN-Transformer approach. *Comput Electron Agric*, 237:110489. https://doi.org/10.1016/j.compag.2025.110489

Han W, Chen SH, Xiao SL, et al., 2025. Large-scale tobacco identification via a very-high-resolution unmanned aerial vehicle benchmark and a ConvFlow Transformer. *Int J Appl Earth Obs Geoinf*, 139:104549. https://doi.org/10.1016/j.jag.2025.104549

Haseeb M, Tahir Z, Mahmood SA, et al., 2025. Winter wheat yield prediction using linear and nonlinear machine learning algorithms based on climatological and remote sensing data. *Inf Process Agric*, 12(4):431-444. https://doi.org/10.1016/j.inpa.2025.02.004

He JQ, Jia YL, Li Y, et al., 2025. Regional-scale precision mapping of cotton suitability using UAV and satellite data in arid environments. *Agric Water Manag*, 307:109215. https://doi.org/10.1016/j.agwat.2024.109215

Huang XZ, Wang H, Li XB, 2024. A multi-scale semantic feature fusion method for remote sensing crop classification. *Comput Electron Agric*, 224:109185. https://doi.org/10.1016/j.compag.2024.109185

Ismaili M, Krimissa S, Namous M, et al., 2024. Mapping soil suitability using phenological information derived from MODIS time series data in a semi-arid region: a case study of Khouribga, Morocco. *Heliyon*, 10(2):e24101. https://doi.org/10.1016/j.heliyon.2024.e24101

Ji YS, Liu ZH, Liu R, et al., 2024. High-throughput phenotypic traits estimation of faba bean based on machine learning and drone-based multimodal data. *Comput Electron Agric*, 227(Pt 2):109584. https://doi.org/10.1016/j.compag.2024.109584

Jia TY, Shamseldin AY, Liu TX, et al., 2025. Soil moisture inversion method for semi-arid regions using multi-temporal Sentinel-1 and Sentinel-2 data. *J Hydrol Reg*, 661:133603. https://doi.org/10.1016/j.jhydrol.2025.133603

Jiang DG, Shi XY, Liang Y, et al., 2024. Feature extraction technique based on Shapley value method and improved mRMR algorithm. *Measurement*, 237:115190. https://doi.org/10.1016/j.measurement.2024.115190

Khankeshizadeh E, Tahermanesh S, Mohsenifar A, et al., 2024. FBA-DPAttResU-Net: forest burned area detection using a novel end-to-end dual-path attention residual-based U-Net from post-fire Sentinel-1 and Sentinel-2 images. *Ecol Indic*, 167:112589. https://doi.org/10.1016/j.ecolind.2024.112589

Kordi F, Yousefi H, 2022. Crop classification based on phenology information by using time series of optical and synthetic-aperture radar images. *Remote Sens Appl Soc Environ*, 27:100812. https://doi.org/10.1016/j.rsase.2022.100812

Li CL, Guo SL, Cui Z, et al., 2025. Flow simulation based on MISO and LSTM models in the Yangtze River-Dongting Lake System. *J Hydrol Reg Stud*, 60:102469. https://doi.org/10.1016/j.ejrh.2025.102469

Li GG, Zhang H, Lyu T, et al., 2024. Regional significant wave height forecast in the East China Sea based on the Self-Attention ConvLSTM with SWAN model. *Ocean Eng*, 312(Pt 1):119064. https://doi.org/10.1016/j.oceaneng.2024.119064

Li WC, Chen R, Ma D, et al., 2023. Tracking autumn photosynthetic phenology on Tibetan plateau grassland with the green–red vegetation index. *Agric For Meteorol*, 339: 109573. https://doi.org/10.1016/j.agrformet.2023.109573

Liu Y, Yang FQ, Yue JB, et al., 2024. Crop canopy volume weighted by color parameters from UAV-based RGB imagery to estimate above-ground biomass of potatoes. *Comput Electron Agric*, 227(Pt 2):109678. https://doi.org/10.1016/j.compag.2024.109678

Lu DN, Xu LL, Zhou J, et al., 2025. 3DLST: 3D Learnable Supertoken Transformer for LiDAR point cloud scene segmentation. *Int J Appl Earth Obs Geoinf*, 140:104572. https://doi.org/10.1016/j.jag.2025.104572

Marcone A, Impollonia G, Croci M, et al., 2024. Estimation of above ground biomass, biophysical and quality parameters of spinach (*Spinacia oleracea* L.) using Sentinel-2 to support the supply chain. *Sci Hortic*, 325:112641. https://doi.org/10.1016/j.scienta.2023.112641

Marin DB, Ferraz GAES, Santana LS, et al., 2021. Detecting coffee leaf rust with UAV-based vegetation indices and decision tree machine learning models. *Comput Electron*

Agric, 190:106476.
https://doi.org/10.1016/j.compag.2021.106476

Marino S, 2023. Understanding the spatio-temporal behaviour of the sunflower crop for subfield areas delineation using Sentinel-2 NDVI time-series images in an organic farming system. *Heliyon*, 9(9):e19507.
https://doi.org/10.1016/j.heliyon.2023.e19507

Martínez-Movilla A, Rodríguez-Somoza JL, Román M, et al., 2024. Rapid diagnosis of the geospatial distribution of intertidal macroalgae using large-scale UAVs. *Ecol Inform*, 83:102845.
https://doi.org/10.1016/j.ecoinf.2024.102845

Medelytė S, Rzhanov Y, Šiaulys A, et al., 2025. Evaluating textural descriptors for automated image classification of stony reefs in turbid temperate waters. *Ecol Inform*, 90: 103236.
https://doi.org/10.1016/j.ecoinf.2025.103236

Mendes J, Lima J, Costa L, et al., 2025. Impact of hyperparameter tuning on CNN accuracy in agricultural image classification. *Smart Agric Technol*, 11:101016.
https://doi.org/10.1016/j.atech.2025.101016

Mishra K, Tiwari HL, Poonia V, 2025. An integrated approach of machine learning methods coupled with cellular automaton for monitoring and forecasting of land use and land cover. *J Arid Environ*, 226:105293.
https://doi.org/10.1016/j.jaridenv.2024.105293

Naqvi SMZA, Awais M, Khan FS, et al., 2021. Unmanned air vehicle based high resolution imagery for chlorophyll estimation using spectrally modified vegetation indices in vertical hierarchy of citrus grove. *Remote Sens Appl Soc Environ*, 23:100596.
https://doi.org/10.1016/j.rsase.2021.100596

Newete SW, Abutaleb K, Chirima GJ, et al., 2024. Phenology-based winter wheat classification for crop growth monitoring using multi-temporal Sentinel-2 satellite data. *Egypt J Remote Sens Space Sci*, 27(4):695-704.
https://doi.org/10.1016/j.ejrs.2024.10.001

Nivedita V, Begum SS, Aldehim G, et al., 2024. Plastic debris detection along coastal waters using Sentinel-2 satellite data and machine learning techniques. *Mar Pollut Bull*, 209:117106.
https://doi.org/10.1016/j.marpolbul.2024.117106

Pandey GK, Mittal K, Bansal A, et al., 2025. Fire detection with ResNet 18: comparative analysis across different hyperparameters. *Procedia Comput Sci*, 260:708-716.
https://doi.org/10.1016/j.procs.2025.03.250

Peng KY, Liu YM, Zhang KY, et al., 2025. Regional NDVI reconstruction based on tree-ring width of *Pinus massoniana* Lamb. in the north-south transition zone of China. *Dendrochronologia*, 92:126373.
https://doi.org/10.1016/j.dendro.2025.126373

Qiu ZQ, Liu D, Yan NX, et al., 2024. Improving the observations of suspended sediment concentrations in rivers from Landsat to Sentinel-2 imagery. *Int J Appl Earth Obs Geoinf*, 134:104209.
https://doi.org/10.1016/j.jag.2024.104209

Ren K, Zhang DW, Wan MJ, et al., 2021. An infrared and visible image fusion method based on improved DenseNet and mRMR-ZCA. *Infrared Phys Technol*, 115:103707.
https://doi.org/10.1016/j.infrared.2021.103707

Ren ZQ, Tian FM, Wang SQ, et al., 2025. Research on maize leaves surface action potential recognition method based on ResNet-18SE. *Smart Agric Technol*, 10:100819.
https://doi.org/10.1016/j.atech.2025.100819

Sarkar A, Maity PP, Ray M, et al., 2023. Inclusion of fractal dimension in four machine learning algorithms improves the prediction accuracy of mean weight diameter of soil. *Ecol Inform*, 74:101959.
https://doi.org/10.1016/j.ecoinf.2022.101959

Shojaeezadeh SA, Elnashar A, David Weber TK, 2025. A novel fusion of Sentinel-1 and Sentinel-2 with climate data for crop phenology estimation using machine learning. *Sci Remote Sens*, 11:100227.
https://doi.org/10.1016/j.srs.2025.100227

Tesfaye AA, Osgood DE, Aweke BG, 2025. Application of a novel vegetation condition index using MODIS EVI for structuring crop index insurance under a smallholder system. *Environ Sustain Indic*, 26:100696.
https://doi.org/10.1016/j.indic.2025.100696

Testa S, Soudani K, Boschetti L, et al., 2018. MODIS-derived EVI, NDVI and WDRVI time series to estimate phenological metrics in French deciduous forests. *Int J Appl Earth Obs Geoinf*, 64:132-144.
https://doi.org/10.1016/j.jag.2017.08.006

Tran KH, Zhang XY, Zhang HK, et al., 2025. A Transformer-based model for detecting land surface phenology from the irregular harmonized Landsat and Sentinel-2 time series across the United States. *Remote Sens Environ*, 320: 114656.
https://doi.org/10.1016/j.rse.2025.114656

Tufail R, Tassinari P, Torreggiani D, 2025. Assessing feature extraction, selection, and classification combinations for crop mapping using Sentinel-2 time series: a case study in northern Italy. *Remote Sens Appl Soc Environ*, 38: 101525.
https://doi.org/10.1016/j.rsase.2025.101525

Wang C, Zhang XY, Wang WJ, et al., 2024. Understanding the potentials of early-season crop type mapping by using Landsat-8, Sentinel-1/2, and GF-1/6 data. *Comput Electron Agric*, 224:109239.
https://doi.org/10.1016/j.compag.2024.109239

Wang LY, Dai LM, Sun L, 2025. ConvLSTM-based spatio-temporal and temporal processing models for chaotic vibration prediction of a microbeam. *Commun Nonlinear Sci Numer Simul*, 140(Pt 2):108411.
https://doi.org/10.1016/j.cnsns.2024.108411

Xu LC, Su XY, Wang KT, et al., 2025. Enhancing canopy nitrogen estimation in Torreya grandis based on advanced SLIC-EVI and HMT-seCNN methods using hyperspectral UAV data. *Comput Electron Agric*, 231:109977.
https://doi.org/10.1016/j.compag.2025.109977

Yu DF, Ren LR, Chen C, et al., 2025. An AttSDNet model for multi-scale feature perception enhanced remote sensing classification of coastal salt-marsh wetlands. *Mar Environ Res*, 204:106899.
https://doi.org/10.1016/j.marenvres.2024.106899

Zhai LL, Zan M, Ye M, et al., 2025. Time-series forest age estimation in Xinjiang based on forest disturbance and recovery detection. *Ecol Indic*, 170:113043. https://doi.org/10.1016/j.ecolind.2024.113043

Zheng B, Liu X, Wu Y, 2022. Evaluation of urban human settlement environment in corps based on the entropy method: a case study of Tumushuke City. Proceedings of the 3rd International Conference on Humanities, Arts, and Social Sciences (HASS 2022), New York, USA, p.108-118. https://doi.org/10.54691/bcpssh.v18i.943

Zhou JP, Gu XH, Liu CL, et al., 2024. A new approach to extract the upright maize straw from Sentinel-2 satellite imagery using new straw indices. *Comput Electron Agric*, 216:108506. https://doi.org/10.1016/j.compag.2023.108506

Zhou ZJ, Plauborg F, Thomsen AG, et al., 2017. A RVI/LAI-reference curve to detect N stress and guide N fertigation using combined information from spectral reflectance and leaf area measurements in potato. *Eur J Agron*, 87:1-7. https://doi.org/10.1016/j.eja.2017.04.002