



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}✉

¹College of Information Engineering, Tarim University, Alar 843300 City, Xinjiang Uygur Autonomous Region, China

²Key Laboratory of Oasis Agriculture in Tarim, Tarim University, Alar 843300, Xinjiang Uygur Autonomous Region, China

Abstract: Multi-temporal remote sensing data in large-scale crop phenology identification and classification has 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, 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 (MLP, ResNet-18, ConvLSTM, Transformer, and 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 $1 \times 1 + 3 \times 3 + 5 \times 5$ convolutional windows achieves 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, 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 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

✉ Nannan ZHANG, zhangnannan@taru.edu.cn

* The two authors contributed equally to this work

 Chunli WANG, <https://orcid.org/0009-0009-1865-5404>

Received July 11, 2025; Revision accepted Dec. 3, 2025;

Crosschecked xxx. xx, 20xx; Published online xxx. xx, 20xx

conditions and human interventions, the timely and accurate monitoring of crop growth stages has become key factors 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, 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 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 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-DPAAttResU-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 et al. (2025) employed MISO and LSTM for flow prediction in the Yangtze-Dongting Lake system, offering a case study for coupled hydrological-remote sensing research. 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. Focusing on phenological feature extraction for typical crops (cotton, winter jujube, and tiger nut) in Tumushuke, we utilized 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 accurately extracted key phenological parameters including the start of season (SOS), peak of season (POS), and end of season (EOS), significantly improving the accuracy and reliability of crop phenology extraction. Furthermore, we used time-series Sentinel-2 imagery to investigate the role of deep learning and multi-scale window analysis in crop spatial distribution identification. Compared to single-scale modeling, multi-scale feature fusion demonstrated superior capability in characterizing both spatial patterns and temporal variations of crops, leading to enhanced classification accuracy and model robustness. This approach showed particular adaptability and application potential in complex cultivation environments. The proposed deep learning method for multi-temporal remote sensing image analysis effectively supports crop spatial distribution mapping in the Tumushuke region and beyond. Therefore, we argue that this study not only provides agricultural management departments with precise data support but also offers a practical solution to the key challenges of crop identification in complex arid oasis environments, thereby facilitating the advancement of regional crop management toward scientific and intelligent practices.

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 °C, while winter ushers in a sharp contrast, with temperatures plunging to as low as -25 °C. The annual total solar radiation ranges from approximately 130.5 to 142.8 kCal/cm² (equivalent to 545.8–597.1 MJ/m²), with scarce precipitation and intense evaporation (Zheng et al., 2022). The dominant soil types are sandy loam and loam, which provide specific conditions for

✉ Nannan ZHANG, zhangnannan@taru.edu.cn

ORCID Nannan ZHANG, <https://orcid.org/0000-0003-2956-8815>

Received xxx. xx, 20xx; Revision accepted xxx. xx, 20xx
Crosschecked xxx. xx, 20xx; Published online xxx. xx, 20xx

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 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, making 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, causing substantial changes in surface reflectance. 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 Savitzky–Golay smoothing and double logistic fitting, followed by feature selection using the mRMR algorithm; (4) Model construction and comparison, including MLP, ResNet-18, ConvLSTM, Transformer, and RFC classifiers, with classification accuracy further optimized through multi-scale window analysis; (5) Accuracy evaluation and regional validation, ultimately achieving visualized outputs of both crop spatial distribution and phenological timelines.

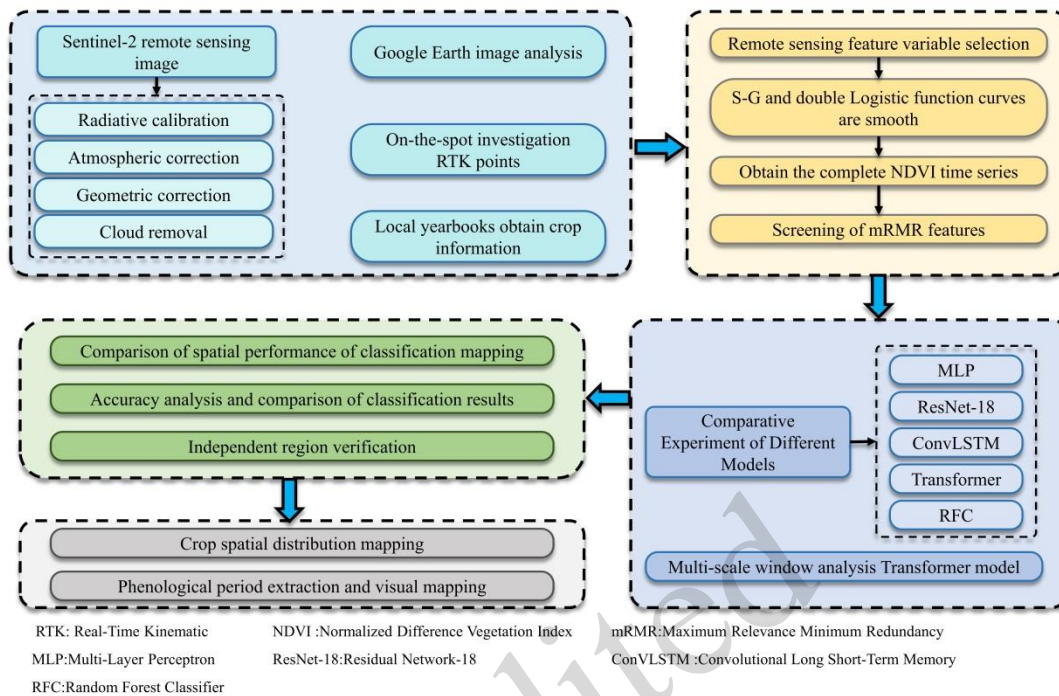


Fig. 1 Overall technical workflow adopted in this study

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 meters (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 City, Xinjiang Uygur Autonomous Region, 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 near-infrared (NIR) and red bands (Bands 4 and 8) form the basis for calculating fundamental vegetation indices such as NDVI. The red-edge bands (Bands 5–7) are highly sensitive to changes in chlorophyll content and subtle variations in plant physiology, which improves species differentiation. Meanwhile, the SWIR bands (Bands 11 and 12) 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).

Table 1. Selected Sentinel-2 bands

Band	Resolution(m)	Name	Wavelength (μm)	Bandwidth (nm)
Band 2	10	Blue	0.458–0.523	65
Band 3	10	Green	0.543–0.578	35
Band 4	10	Red	0.650–0.680	30
Band 5	20	VegetationRedEdge	0.698–0.712	14
Band 6	20	VegetationRedEdge	0.733–0.747	14
Band 7	20	VegetationRedEdge	0.773–0.793	20
Band 8	10	NIR	0.785–0.899	114
Band8 A	20	NarrowNIR	0.855–0.875	20
Band 11	20	SWIR	1.565–1.655	90
Band 12	20	SWIR	2.100–2.280	180

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 SCL (Scene Classification Layer) and QA (Quality Assurance) 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 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 near-infrared bands (Tufail, Tassinari and Torreggiani, 2025). The reflectance differences between crops in the red-edge and near-infrared 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., RVI, MSR, etc.). Temporal feature analysis was based on multi-temporal remote sensing data and quantifies 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 Savitzky-Golay 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

Name	Calculation Formula	References
NDVI	$\frac{NIR - R}{NIR + R}$	Peng et al., 2025
RVI	$\frac{NIR}{R}$	Zhou et al., 2017
MSR	$\sqrt{\left(\frac{NIR}{R}\right) - 1}$	Liu et al., 2024
GRVI	$\frac{NIR}{G}$	Li et al., 2023
RDVI	$\frac{NIR - R}{\sqrt{NIR + R}}$	Gómez et al., 2025
MCARI	$\frac{(NIR - R) - 0.2(NIR - G)}{NIR/R}$	Marcone et al., 2024
GRDVI	$\frac{NIR - G}{\sqrt{NIR + G}}$	Ji et al., 2024
WDRVI	$\frac{0.12NIR - R}{0.12NIR + R}$	Testa et al., 2018
SAVI	$(1 + 0.5) \frac{NIR - R}{(NIR + R + 0.5)}$	Haseeb et al., 2025
EVI	$\frac{2.5(NIR - R)}{NIR + 6R - 7.5B + 1}$	Tesfaye, Osgood and Aweke, 2025
DVI	$NIR - R$	Xu et al., 2025
OSAVI	$\frac{1.16(NIR - R)}{NIR + R + 0.16}$	Fern et al., 2018

MTVI2	$\frac{1.5[1.2(NIR - G) - 2.5(R - G)]}{\sqrt{(2NIR + 1)^2 - (6NIR - 5R) - 0.5}}$	Naqvi et al., 2021
MSR_G	$\frac{\left(\frac{NIR}{G}\right) - 1}{\sqrt{\left(\frac{NIR}{G}\right) + 1}}$	Marin et al., 2021

where 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 (Minimum Redundancy Maximum Relevance) 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.

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

where $I(X_i, Y)$ refers to the mutual information between feature X_i and the target variable Y , and $I(X_i, X_j)$ corresponds to the mutual information shared between features X_i and X_j ; 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 (see 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 manifest 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.

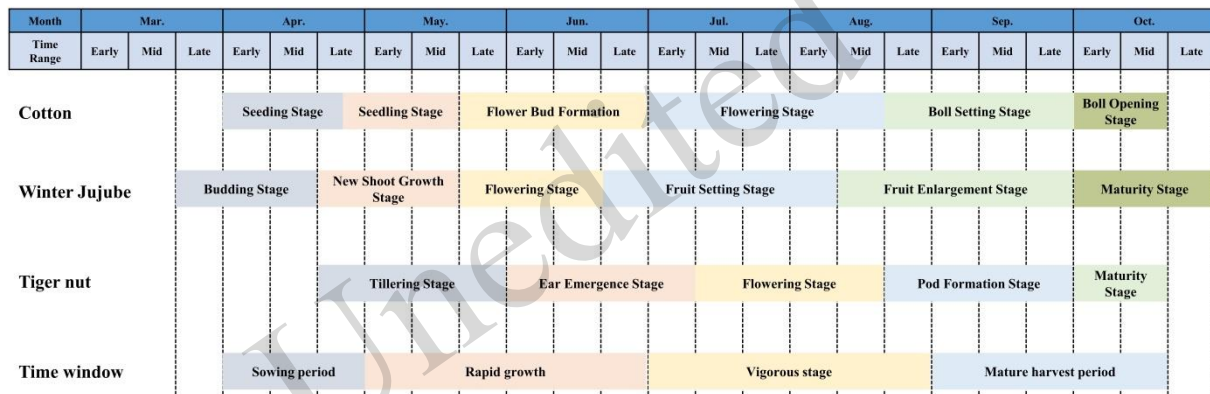


Fig. 2 Phenological calendar and remote sensing time windows for the three major crops—cotton, winter jujube, and tiger nut—in the study area

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 (Band 4) and near-infrared band (Band 8). We then smoothed the NDVI time series using the Savitzky–Golay (S-G) filter, which effectively suppressed high-frequency noise and reconstructed unimodal NDVI curves representing crop development from emergence through growth to senescence (see Fig 3). To further quantify key phenological stages, this study 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 start of season (SOS), peak of season (POS), and end of season (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.

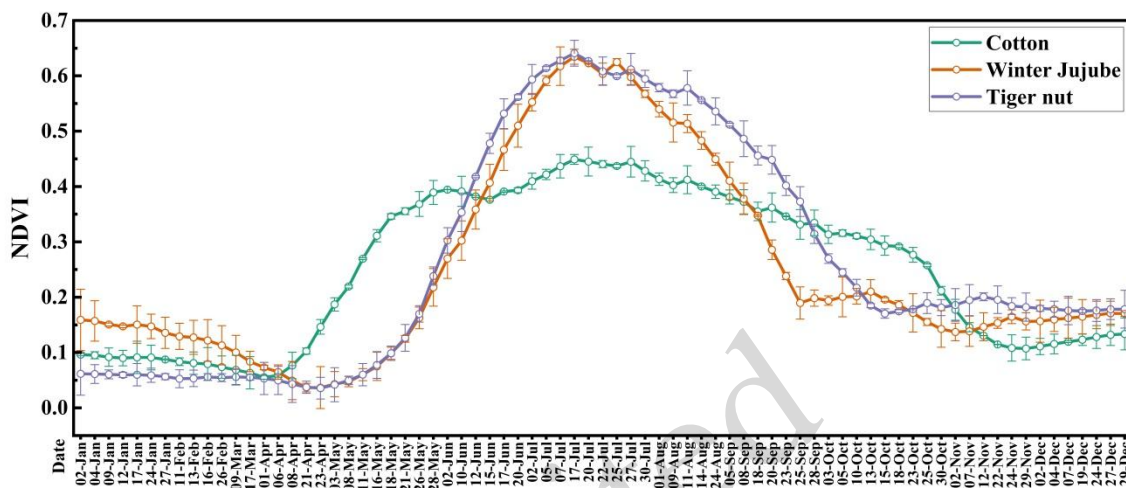


Fig. 3 NDVI time series curves of cotton, winter jujube and tiger nut in 2024, smoothed using the Savitzky–Golay (S-G) filter

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 early August, reflecting a longer growth cycle. The phenological metrics extracted using TIMESAT—through Savitzky–Golay 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 (minimum Redundancy Maximum Relevance) 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 Multi-Layer Perceptron (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 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 2 to 3 hidden layers, each containing 64 to 128 neurons with ReLU activation, and an output layer of 3 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.

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

$$h^{(l+1)} = f(z^{(l+1)}) \quad (3)$$

$$\text{ReLU}(x) = \max(0, x) \quad (4)$$

$$L = -\sum_{i=1}^n y_i \log(\hat{y}_i) \quad (5)$$

where $W^{(l)}$ denotes the weight matrix encoding the connection strength between layer l and layer $l + 1$; $b^{(l)}$ represents the corresponding bias vector that shifts the neuron's activation; f denotes the activation function that introduces non-linear transformations; 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 convolutional neural networks, was first introduced by the He Kaiming team at the 2015 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 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 3 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.

$$H_{(x)} = F_{(x)} + x \quad (6)$$

$$H_{(x)} = F_{(x)} + W_s x \quad (7)$$

$$Y_{i,j,k} = \sum_{m=1}^{C_{in}} \sum_{p=1}^K \sum_{q=1}^K W_{k,m,p,q} \cdot X_{i+p,j+q}^m \quad (8)$$

$$BN_{(x)} = \gamma \cdot \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \quad (9)$$

where x denotes the input to the residual module; $F(x)$ 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 $H(x)$, which is computed via a residual connection. Two types of connections are used: an identity mapping $H_{(x)} = F_{(x)} + x$ when the dimensions match, and a projection mapping $H_{(x)} = F_{(x)} + W_s x$ when dimensions change, where W_s is a 1×1 convolution applied to x for dimensionality alignment. Let X be the input feature map with dimensionality $H \times W \times C_{in}$ and W be the convolutional kernel. The output feature map is denoted as Y . Here, K indicates the kernel size, i and j are the spatial indices of the output feature map Y , s denotes the stride, and k represents the number of output channels.

2.7.3 ConvLSTM Model

In the field of deep learning, researchers have been exploring improved methods to handle spatiotemporal sequence data. In 2015, the Xingjian Shi team proposed an innovative solution—the Convolutional Long Short-Term Memory (ConvLSTM) model (Wang, Dai and Sun, 2025). The brilliance of this model lies in its ability to combine the strengths of Convolutional Neural Networks (CNNs) in extracting spatial features with the temporal modeling capability of Long Short-Term Memory (LSTM). Traditional LSTM models, when dealing with image sequences or data with spatial structure, 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 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.

$$i_s = \sigma(W_{xi}X_s + W_{hi}H_{s-1} + W_{ci} \odot C_{s-1} + b_i) \quad (10)$$

$$f_s = \sigma(W_{xf}X_s + W_{hf}H_{s-1} + W_{cf} \odot C_{s-1} + b_f) \quad (11)$$

$$C_s = f_s \odot C_{s-1} + i_s \odot \tanh(W_{xc}X_s + W_{hc}H_{s-1} + b_c) \quad (12)$$

$$o_s = \sigma(W_{xo}X_s + W_{ho}H_{s-1} + W_{co} \odot C_s + b_o) \quad (13)$$

$$H_s = o_s \odot \tanh(C_s) \quad (14)$$

where, X_s denotes the current input vector; H_{s-1} represents the hidden state from the previous time step; C_{s-1} is the cell state from the previous time step. i_s, f_s, o_s refers to the input gate, forget gate, and output gate; C_s is the current cell state; H_s denotes the hidden state (i.e., the output of the ConvLSTM unit); σ (as used previously) represents the sigmoid activation function; \odot also denotes the Hadamard (element-wise) product; and \tanh refers to the hyperbolic tangent (tanh) activation function.

2.7.4 Transformer Model

The Transformer is a neural network architecture based primarily on the self-attention mechanism, first proposed by Vaswani et al. 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—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.

$$Attention_{(Q,K,V)} = V * \text{softmax}\left(\frac{QK}{\sqrt{d_k}}\right) \quad (15)$$

$$MultiHead(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)M^O \quad (16)$$

$$\text{head}_i = Attention(QM_i^Q, KM_i^K, VM_i^V) \quad (17)$$

$$FFN(x) = ReLU(0, xM_1 + b_1)M_2 + b_2 \quad (18)$$

$$\text{LayerOutput} = \text{LayerNorm}(x + \text{SubLayer}(x)) \quad (19)$$

where $Q, K, V \in R^{n \times d_k}$; d_k denotes the dimensionality of the key vectors; *softmax* is used for scaling the attention scores during normalization; M_i^Q, M_i^K, M_i^V represent the linear projection matrices for each attention head; M^O is the output linear transformation matrix; h indicates the number of attention heads; M_1, M_2 denotes the weight matrix; and b_1, b_2 represents the bias term.

2.7.5 Random Forest Classifier (RFC)

The Random Forest Classifier (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, Kolhe and Kamal, 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, Tiwari and Poonia, 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 Google Earth Engine (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 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 Savitzky-Golay 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 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, Wang and Li, 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 , $1 \times 1 + 3 \times 3$, and

$1 \times 1 + 3 \times 3 + 5 \times 5 + 7 \times 7 + 9 \times 9$), 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, Kappa coefficient, Precision, Recall, and F1 Score. 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.

$$IoU = \frac{TP}{TP + FP + FN} \quad (20)$$

$$Mean\ Accuracy = \frac{1}{k} \sum_{i=1}^k \frac{TP_i}{TP_i + FN_i} \quad (21)$$

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

$$Precision = \frac{TP}{TP + FP} \quad (23)$$

$$Recall = \frac{TP}{TP + FN} \quad (24)$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (25)$$

where TP (True Positive) denotes the number of pixels correctly identified as belonging to the target class; FP (False Positive) denotes the number of pixels erroneously classified as the target class; and FN (False Negative) denotes the number of target-class pixels incorrectly classified as other categories. K refers to the total number of classes; TP_i represents the number of correctly classified pixels in class i ; FN_i is the number of misclassified (omitted) pixels in class i ; p_o denotes the overall accuracy of the model; and p_e represents the expected accuracy under random classification.

3. Results

3.1 Spectral Feature Analysis and Selection

Fig 4(a) shows the Pearson correlation heatmap between the selected spectral bands and vegetation indices. The diagram includes multispectral bands from Sentinel-2 (e.g., B02–B12) and commonly used vegetation indices such as NDVI, EVI, SAVI, and MCARI, totaling 24 variables. 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. 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 near-infrared bands (e.g., B05, B06, B08, 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 MTVI2 and MSR_G, 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 Minimum Redundancy Maximum Relevance (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 4(b) 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, 99, and 85, 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, GRDVI, and 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.

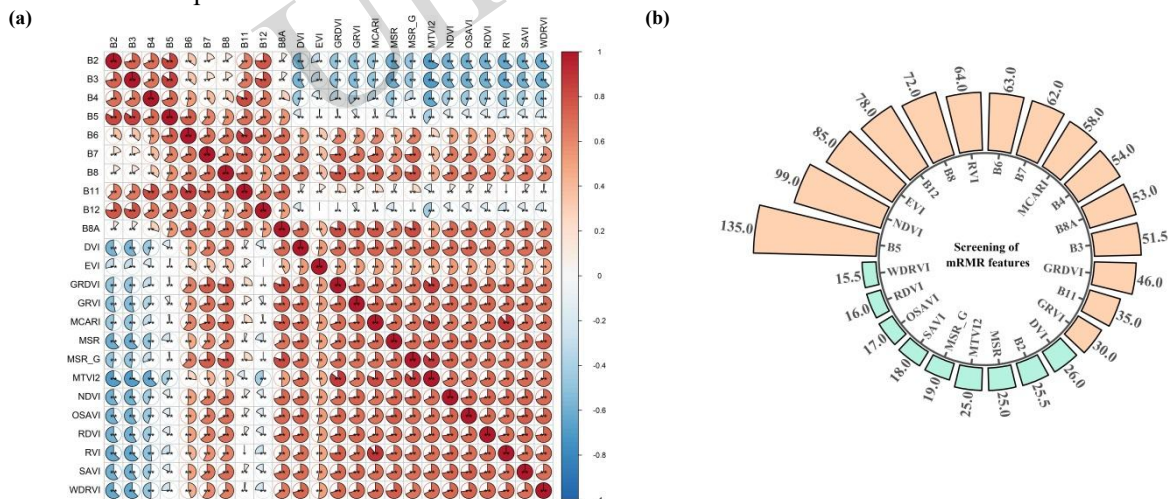


Fig 4. Spectral correlation and feature selection results. (a) Pearson correlation heatmap among Sentinel-2 spectral bands and vegetation indices; (b) mRMR-based ranking of feature importance for input variable selection.

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 ($1 \times 1 + 3 \times 3 + 5 \times 5$). 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 the Fig indicate zoomed-in sub-regions 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 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.

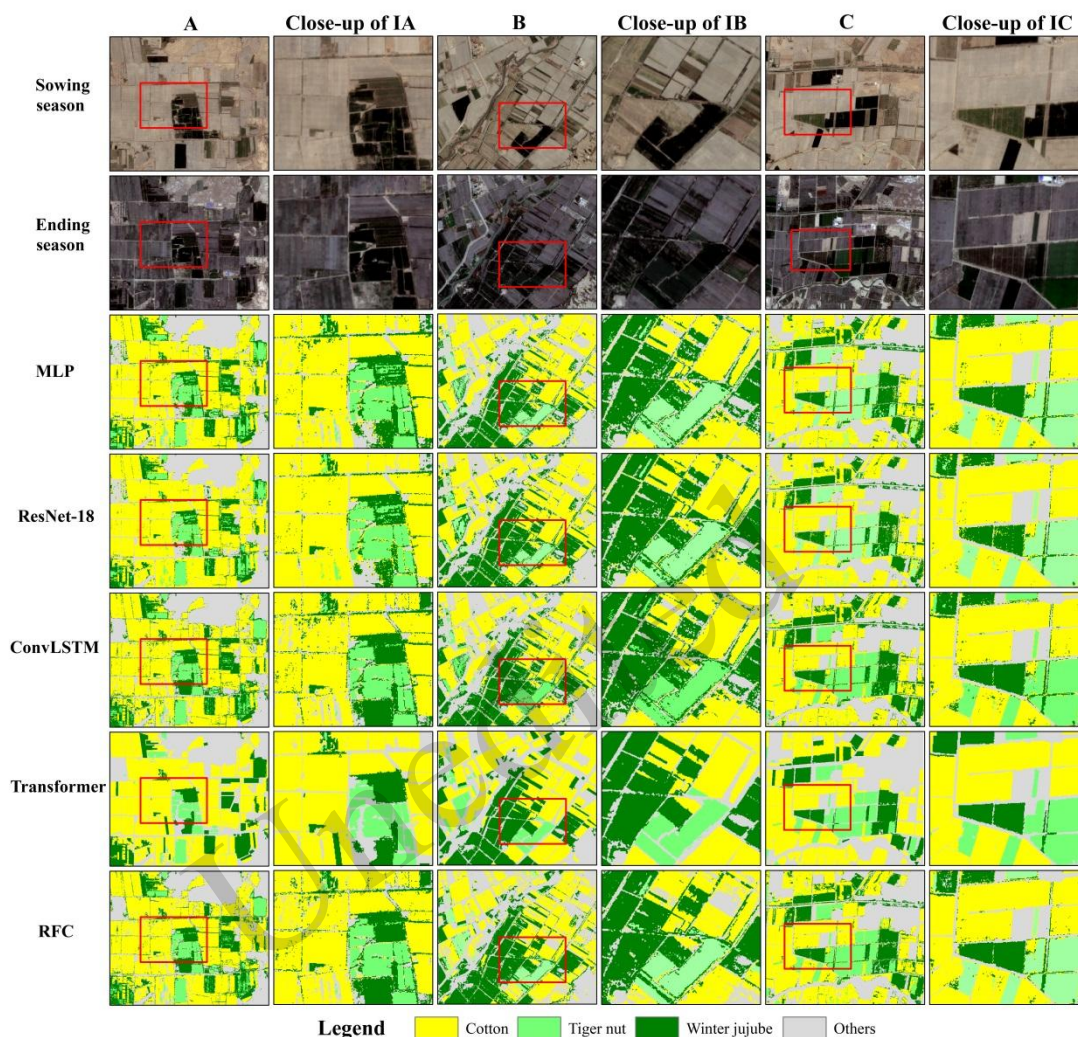


Fig. 5 Classification results of different models for the study area

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 ($1 \times 1 + 3 \times 3 + 5 \times 5$). 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 field shapes and spectral variations. It recorded IoU scores of 92.74%, 91.33%, and 93.45% 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 performed the worst across all three crop types, underscoring the limited adaptability of traditional machine learning algorithms in handling complex, high-resolution remote sensing classification tasks.

Table 3. Extraction Accuracy Comparison: 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	84.3	87.04	84.17	86.4	84.93	87.13
	Transformer	92.74	94.44	91.98	94.5	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

3.3 Evaluation of Multi-Scale Strategies in Transformer Models

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 (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 $1\times 1 + 3\times 3 + 5\times 5$ 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 preserve 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., $1\times 1 + 3\times 3 + 5\times 5$) 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.

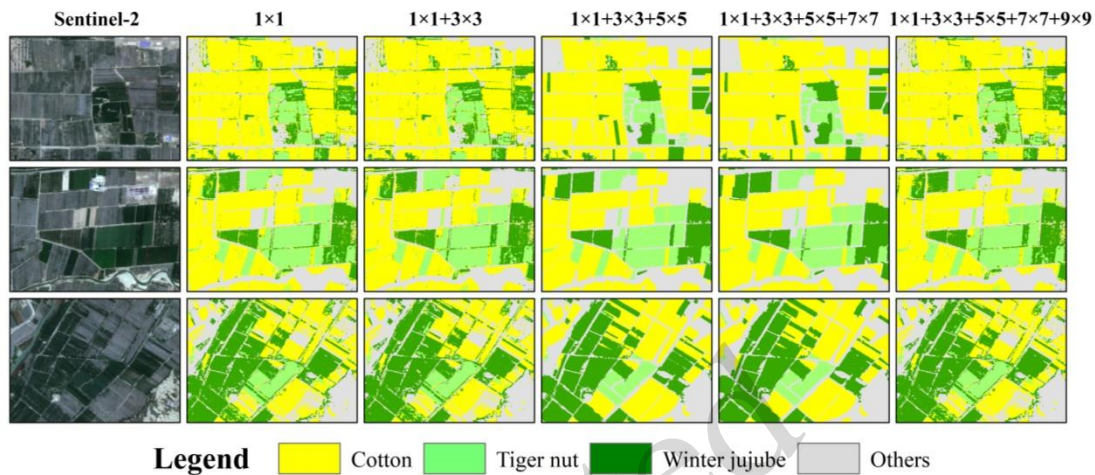


Fig.6 Effect of different scale spatial windows on the crop classification performance of the Transformer model

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 as the window size increases, the classification accuracy improves and noise interference is gradually reduced, 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: as more window scales are added, IoU increases from 88.67% to a peak of 92.81% and the F1 score remains consistently above 93%, 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, with IoU rising from 74.97% to 85.67% and the F1 score stabilizing around 84.8%. 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., $1 \times 1 + 3 \times 3$ or $1 \times 1 + 3 \times 3 + 5 \times 5$) strike the best balance between boundary integrity and noise suppression. In particular, the $1 \times 1 + 3 \times 3 + 5 \times 5$ 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 terms of both effectiveness and efficiency. In contrast, the 1×1 window, representing pixel-level classification, resulted in the poorest performance. This highlights 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.

Table 4. Comparison of Extraction Accuracy Across Different Windows

Different Scales	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+...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+...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

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 $1 \times 1 + 3 \times 3 + 5 \times 5$. 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. 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.

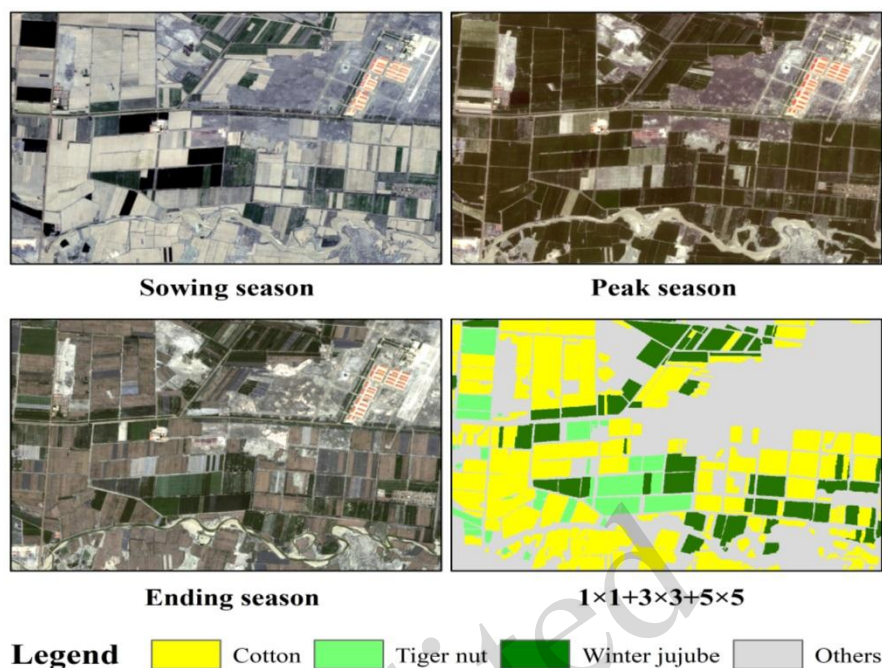


Fig. 7 Classification results for the validation region using the Transformer model with $1\times 1+3\times 3+5\times 5$ multi-scale spatial windows

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: 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.

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

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 generated by the Transformer model using a $1 \times 1 + 3 \times 3 + 5 \times 5$ 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 Start of Season (SOS), 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 peak of season (POS), with most fields reaching peak growth around DOY 190–198, while some late-sown plots being 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 End of Season (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 Savitzky–Golay filter and was subsequently fitted using a double logistic function. It should be clarified that the Savitzky–Golay 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 the start (SOS), peak (POS) and end (EOS) of the growing season. These features served as high-value temporal indicators for crop classification and phenological monitoring. This finding is consistent with [Shojaeezadeh et al. \(2025\)](#) results,

which showed low mean absolute errors (0.43 and 6 days) 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 (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 the Random Forest Classifier (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%, 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 [Guerra 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 (see Table 4 and Fig 6). The 5×5 window configuration emerged as the best performer, achieving an IoU of 94.1%, Mean Accuracy of 93.07% and an F1 score of 94.55%. This configuration effectively suppresses "salt-and-pepper noise" and excessive smoothing while maintaining 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 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 ([Yu et al., 2025](#)). 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 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 spatiotemporal 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 Savitzky–Golay filtering and fitted with a double logistic function. The key phenological metrics—Start of Season (SOS), Peak of Season (POS), and End of Season (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 an F1 score of 94.55%. The ConvLSTM model also showed 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 $1\times 1 + 3\times 3 + 5\times 5$, 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 demonstrate 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 multi-temporal 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 funded by the National Natural Science Foundation of China, grant numbers 32101621,62061041, and the Bingtuan Science and Technology Program, grant numbers 2022CB001-05, and Graduate Scientific Research Innovation project of Tarim University, grant number TDGRI2024092.

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.

Compliance with ethics guidelines

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

This study did not involve human participants or animal subjects. Therefore, ethical approval and informed consent were not required for this research. All procedures performed in studies were in accordance with the ethical standards of the institutional and/or national research committee.

References

- Kordi F, Yousefi H, 2022. Crop classification based on phenology information by using time series of optical and synthetic-aperture radar images. *Remote Sensing Applications: Society and Environment*, 27:100812. <https://doi.org/10.1016/j.rsase.2022.100812>
- 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>
- He J, Jia Y, Li Y, et al., 2025. Regional-scale precision mapping of cotton suitability using UAV and satellite data in arid environments. *Agricultural Water Management*, 307:109215. <https://doi.org/10.1016/j.agwat.2024.109215>
- Zhai L, Zan M, Ye M, et al., 2025. Time-series forest age estimation in Xinjiang based on forest disturbance and recovery detection. *Ecological Indicators*, 170:113043. <https://doi.org/10.1016/j.ecolind.2024.113043>
- 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. *The Egyptian Journal of Remote Sensing and Space Sciences*, 27(4):pp. 695-704. <https://doi.org/10.1016/j.ejrs.2024.10.001>
- Qiu Z, Liu D, Yan N, et al., 2024. Improving the observations of suspended sediment concentrations in rivers from Landsat to Sentinel-2 imagery. *International Journal of Applied Earth Observation and Geoinformation*, 134:104209. <https://doi.org/10.1016/j.jag.2024.104209>
- Nivedita V, Begum SS, Aldehim G, et al., 2024. Plastic debris detection along coastal waters using Sentinel-2 satellite data and machine learning techniques. *Marine Pollution Bulletin*, 209(Part A):117106. <https://doi.org/10.1016/j.marpolbul.2024.117106>
- Wang C, Zhang X, Wang W, et al., 2024. Understanding the potentials of early-season crop type mapping by using Landsat-8, Sentinel-1/2, and GF-1/6 data. *Computers and Electronics in Agriculture*, 224:109239. <https://doi.org/10.1016/j.compag.2024.109239>
- Du R, Xiang Y, Chen J, et al., 2024. The daily soil water content monitoring of cropland in irrigation area using Sentinel-2/3 spatio-temporal fusion and machine learning. *International Journal of Applied Earth Observation and Geoinformation*, 132:104081. <https://doi.org/10.1016/j.jag.2024.104081>
- Chen X, Ding Y, Zheng X, 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. *Ecological Informatics*, 84:102930. <https://doi.org/10.1016/j.ecoinf.2024.102930>
- Zhou J, Gu X, Liu C, et al., 2024. A new approach to extract the upright maize straw from Sentinel-2 satellite imagery using new straw indices. *Computers and Electronics in Agriculture*, 216:108506. <https://doi.org/10.1016/j.compag.2023.108506>
- Han W, Chen S, Xiao S, et al., 2025. Large-scale tobacco identification via a very-high-resolution unmanned aerial vehicle benchmark and a ConvFlow Transformer. *International Journal of Applied Earth Observation and Geoinformation*, 139:104549. <https://doi.org/10.1016/j.jag.2025.104549>
- 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. *Science of The Total Environment*, 974:179243.
<https://doi.org/10.1016/j.scitotenv.2025.179243>
- De La Guardia L, de Miranda JH, dos Santos Luciano AC, et al., 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. *Agricultural Water Management*, 305:109128.
<https://doi.org/10.1016/j.agwat.2024.109128>
- Farbo A, Sarvia F, De Petris S, et al., 2024. Forecasting corn NDVI through AI-based approaches using sentinel 2 image time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 211:pp. 244-261.
<https://doi.org/10.1016/j.isprsjprs.2024.04.011>
- Khankeshizadeh E, Tahermanesh S, Mohsenifar A, et al., 2024. FBA-DPAAttResU-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. *Ecological Indicators*, 167:112589.
<https://doi.org/10.1016/j.ecolind.2024.112589>
- Mendes J, Lima J, Costa L, et al., 2025. Impact of hyper-parameter tuning on CNN accuracy in agricultural image classification. *Smart Agricultural Technology*, 11:101016.
<https://doi.org/10.1016/j.atech.2025.101016>
- Martínez-Movilla A, Rodríguez-Somoza JL, Román M, et al., 2024. Rapid diagnosis of the geospatial distribution of intertidal macroalgae using large-scale UAVs. *Ecological Informatics*, 83:102845.
<https://doi.org/10.1016/j.ecoinf.2024.102845>
- Li C, Guo S, Cui Z, et al., 2025. Flow simulation based on MISO and LSTM models in the Yangtze River-Dongting Lake System. *Journal of Hydrology: Regional Studies*, 60:102469.
<https://doi.org/10.1016/j.ejrh.2025.102469>
- 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>
- 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. In: *Proceedings of the 3rd International Conference on Humanities, Arts, and Social Sciences (HASS 2022)*. pp. 109-119.
<https://doi.org/10.26914/c.cnkihy.2022.084080>
- Cao J, Ma S, Yuan W, et al., 2022. Characteristics of diurnal variations of warm-season precipitation over Xinjiang Province in China. *Atmospheric and Oceanic Science Letters*, 15(2):100113.
<https://doi.org/10.1016/j.aosl.2021.100113>
- Jia T, Shamseldin A Y, Liu T, et al., 2025. Soil moisture inversion method for semi-arid regions using multi-temporal Sentinel-1 and Sentinel-2 data. *Journal of Hydrology*, 661(Part A):133603.
<https://doi.org/10.1016/j.jhydrol.2025.133603>
- Faqeerzada M A, Kim H, Kim M S, et al., 2025. Hyperspectral imaging VIS-NIR and SWIR fusion for improved drought-stress identification of strawberry plants. *Computers and Electronics in Agriculture*, 237(Part C):110702.
<https://doi.org/10.1016/j.compag.2025.110702>
- Gao X, Chi H, Huang J, et al., 2024. Comparison of Cloud-Mask Algorithms and Machine-Learning Methods Using Sentinel-2 Imagery for Mapping Paddy Rice in Jiangnan Plain. *Remote Sensing*, 16(8):1305.
<https://doi.org/10.3390/rs16071305>
- 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 Sensing Applications: Society and Environment*, 38:101525.
<https://doi.org/10.1016/j.rsase.2025.101525>
- Peng K, Liu Y, Zhang K, 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>
- Zhou Z, Plauborg F, Thomsen A G, 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. *European Journal of Agronomy*, 87:1-7.
<https://doi.org/10.1016/j.eja.2017.04.002>
- Liu Y, Yang F, Yue J, et al., 2024. Crop canopy volume weighted by color parameters from UAV-based RGB imagery to estimate above-ground biomass of potatoes. *Computers and Electronics in Agriculture*, 227(Part 2):109678.
<https://doi.org/10.1016/j.compag.2024.109678>
- Li W, Chen R, Ma D, et al., 2023. Tracking autumn photosynthetic phenology on Tibetan plateau grassland with the green-red

- vegetation index. *Agricultural and Forest Meteorology*, 339:109573.
<https://doi.org/10.1016/j.agrformet.2023.109573>
- Gómez A L, Segarra J, Vatter T, et al., 2025. Alfalfa yield estimation using the combination of Sentinel-2 and meteorological data. *Field Crops Research*, 326:109857.
<https://doi.org/10.1016/j.fcr.2025.109857>
- 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. *Scientia Horticulturae*, 325:112641.
<https://doi.org/10.1016/j.scienta.2023.112641>
- Ji Y, Liu Z, Liu R, et al., 2024. High-throughput phenotypic traits estimation of faba bean based on machine learning and drone-based multimodal data. *Computers and Electronics in Agriculture*, 227(Part 2):109584.
<https://doi.org/10.1016/j.compag.2024.109584>
- 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. *International Journal of Applied Earth Observation and Geoinformation*, 64:132-144.
<https://doi.org/10.1016/j.jag.2017.08.006>
- Haseeb M, Tahir Z, Mahmood S A, et al., 2025. Winter wheat yield prediction using linear and nonlinear machine learning algorithms based on climatological and remote sensing data. *Information Processing in Agriculture*, 14:107.
<https://doi.org/10.1016/j.inpa.2025.02.004>
- Tesfaye A A, Osgood D E, Aweke B G, 2025. Application of a novel vegetation condition index using MODIS EVI for structuring crop index insurance under a smallholder system. *Environmental and Sustainability Indicators*, 26:100696.
<https://doi.org/10.1016/j.indic.2025.100696>
- Xu L, Su X, Wang K, et al., 2025. Enhancing canopy nitrogen estimation in *Torrey grandis* based on advanced SLIC-EVI and HMT-seCNN methods using hyperspectral UAV data. *Computers and Electronics in Agriculture*, 231:109977.
<https://doi.org/10.1016/j.compag.2025.109977>
- Fern R R, Foxley E A, Bruno A, et al., 2018. Suitability of NDVI and OSAVI as estimators of green biomass and coverage in a semi-arid rangeland. *Ecological Indicators*, 94(Part 1):16-21.
<https://doi.org/10.1016/j.ecolind.2018.06.029>
- Naqvi S M Z A, Awais M, Khan F S, 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 Sensing Applications: Society and Environment*, 23:100596.
<https://doi.org/10.1016/j.rsase.2021.100596>
- Marin D B, Ferraz G A e S, Santana L S, et al., 2021. Detecting coffee leaf rust with UAV-based vegetation indices and decision tree machine learning models. *Computers and Electronics in Agriculture*, 190:106476.
<https://doi.org/10.1016/j.compag.2021.106476>
- Jiang D, Shi X, 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>
- Ren K, Zhang D, Wan M, et al., 2021. An infrared and visible image fusion method based on improved DenseNet and mRMR-ZCA. *Infrared Physics & Technology*, 115:103707.
<https://doi.org/10.1016/j.infrared.2021.103707>
- Cheng J, Li H, Li D, et al., 2022. A survey on image semantic segmentation using deep learning techniques. *Computers, Materials and Continua*, 74(1):1941-1957.
<https://doi.org/10.32604/cmc.2023.032757>
- Gao Y, Sun Z, 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 Chemistry*, 465(Part 1):141944.
<https://doi.org/10.1016/j.foodchem.2024.141944>
- Pandey G K, Mittal K, Bansal A, et al., 2025. Fire Detection with ResNet 18: Comparative Analysis Across Different Hyperparameters. *Procedia Computer Science*, 260:708-716.
<https://doi.org/10.1016/j.procs.2025.03.250>
- Ren Z, Tian F, Wang S, et al., 2025. Research on maize leaves surface action potential recognition method based on ResNet-18SE. *Smart Agricultural Technology*, 10:100819.
<https://doi.org/10.1016/j.atech.2025.100819>
- Wang L, Dai L, Sun L, 2025. ConvLSTM-based spatiotemporal and temporal processing models for chaotic vibration prediction of a microbeam. *Communications in Nonlinear Science and Numerical Simulation*, 140(Part 2):108411.
<https://doi.org/10.1016/j.cnsns.2024.108411>
- Li G, 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 Engineering*, 312(Part 1):119064.

- <https://doi.org/10.1016/j.oceaneng.2024.119064>
- Lu D, Xu L, Zhou J, et al., 2025. 3DLST: 3D Learnable Supertoken Transformer for LiDAR point cloud scene segmentation. *International Journal of Applied Earth Observation and Geoinformation*, 140:104572.
<https://doi.org/10.1016/j.jag.2025.104572>
- Han W, Chen S, Xiao S, et al., 2025. Large-scale tobacco identification via a very-high-resolution unmanned aerial vehicle benchmark and a ConvFlow Transformer. *International Journal of Applied Earth Observation and Geoinformation*, 139:104549.
<https://doi.org/10.1016/j.jag.2025.104549>
- Tran K H, Zhang X, Zhang H K, 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 Sensing of Environment*, 320:114656.
<https://doi.org/10.1016/j.rse.2025.114656>
- Chaudhary A, Kolhe S, Kamal R, 2016. An improved random forest classifier for multi-class classification. *Information Processing in Agriculture*, 3(4):215-222.
<https://doi.org/10.1016/j.inpa.2016.08.002>
- Mishra K, Tiwari H L, Poonia V, 2025. An integrated approach of machine learning methods coupled with cellular automaton for monitoring and forecasting of land use and land cover. *Journal of Arid Environments*, 226:105293.
<https://doi.org/10.1016/j.jaridenv.2024.105293>
- Huang X, Wang H, Li X, 2024. A multi-scale semantic feature fusion method for remote sensing crop classification. *Computers and Electronics in Agriculture*, 224:109185.
<https://doi.org/10.1016/j.compag.2024.109185>
- Sarkar A, Maity P P, Ray M, et al., 2023. Inclusion of fractal dimension in four machine learning algorithms improves the prediction accuracy of mean weight diameter of soil. *Ecological Informatics*, 74:101959.
<https://doi.org/10.1016/j.ecoinf.2022.101959>
- Medelytė S, Ržhanov Y, Šiaulyš A, et al., 2025. Evaluating textural descriptors for automated image classification of stony reefs in turbid temperate waters. *Ecological Informatics*, 90:103236.
<https://doi.org/10.1016/j.ecoinf.2025.103236>
- Shojaeezadeh A S, Elnashar A, Weber D K T, 2025. A novel fusion of Sentinel-1 and Sentinel-2 with climate data for crop phenology estimation using machine learning. *Science of Remote Sensing*, 9:100227.
<https://doi.org/10.1016/j.srs.2025.100227>
- Guerri M F, Distante C, Spagnolo P, et al., 2025. Boosting hyperspectral image classification with Gate-Shift-Fuse mechanisms in a novel CNN-Transformer approach. *Computers and Electronics in Agriculture*, 237(Part A):110489.
<https://doi.org/10.1016/j.compag.2025.110489>
- Yu D, Ren L, Chen C, et al., 2025. An AttSDNet model for multi-scale feature perception enhanced remote sensing classification of coastal salt-marsh wetlands. *Marine Environmental Research*, 204:106899.
<https://doi.org/10.1016/j.marenvres.2024.106899>