

A query-aware CNN–transformer framework for multi-temporal crop classification and phenological stage estimation using Sentinel-2 time series

Yusuf Öztıp¹, Berrin Bilgin², Sudenaz Şenbay³, Ozan Çelik⁴, Cavit Karakuş⁵

¹Department of Translation and Interpreting Studies, Boğaziçi University, South Campus, John Freely Hall, 34342 Bebek/İstanbul, Türkiye, ²Artificial Intelligence and Technology Academy, İstanbul/Türkiye, ³Department of Computer Engineering, Burdur Mehmet Akif Ersoy University, İstiklal Campus, 15030 Merkez/Burdur, Türkiye, ⁴Department of Computer Engineering, Atılım University, Kızılcaşar Mahallesi, 1184. Cad No:13, 06830 Incek, Gölbaşı Ankara/Türkiye, ⁵Department of Advanced Technologies, Institute of Natural and Applied Sciences, Kırşehir Ahi Evran University, 40100, Kırşehir, Türkiye

Corresponding author: Yusuf Öztıp, yusufoztop33@gmail.com

This study presents a high performance deep learning framework for multi-temporal agricultural monitoring using satellite image time series. A hybrid query-aware CNN–transformer architecture (C03) is proposed to jointly perform crop type classification and rice phenology stage prediction within a unified multitask learning setting. The model leverages spatial–temporal feature extraction by combining convolutional neural networks for local spatial context and transformer-based attention mechanisms for long-range temporal dependencies.

Initial baseline experiments with a CNN–TCN–BiLSTM–attention architecture revealed significant performance limitations, including early metric stagnation and severe mode collapse, where the model overfits to dominant classes while failing to generalize across minority crop types. To address these challenges, the proposed C03 model introduces explicit query conditioning based on day-of-year (DOY), enabling precise temporal alignment between input sequences and phenological stages.

The final C03 + P6 system achieved a Macro-F1_{Crop} of 0.9941, a Macro-F1_{RicePheno} of 0.9382, and a weighted total score of 0.9605. Ablation studies confirm that query-awareness is a critical component, as removing the queried day-of-year (DOY) information results in a dramatic performance degradation. A complementary input ablation shows that crop classification strongly depends on spectral inputs, whereas rice phenology performance is largely preserved when imagery is removed. Furthermore, an optimized inference pipeline (P6) significantly reduces computational overhead on the ground-based evaluation platform, indicating suitability for resource-constrained deployment settings; on-orbit use would additionally require validation on flight-representative hardware.

The findings demonstrate that integrating spatial context, temporal attention, and query-aware conditioning yields a robust and scalable solution for agricultural monitoring. The proposed framework not only advances state-of-the-art performance in multi-temporal remote sensing tasks but also provides a step toward deployment-oriented satellite intelligence systems. Future work will explore the integration of multi-modal data sources, including SAR imagery, and evaluate model generalization across diverse agro-ecological regions.

Keywords – Agricultural monitoring, CNN–transformer framework, crop classification, phenological stage estimation, Sentinel-2 time series

1. INTRODUCTION

Operational agricultural monitoring increasingly depends on the ability to transform multi-temporal satellite imagery into reliable crop type and phenology information. Agriculture constitutes the fundamental infrastructure of global stability, yet food security remains a critical challenge. According to the IPCC Sixth Assessment Report (AR6), climate change is already putting increased pressure on food production and accessibility, with global warming projected to further destabilize agricultural systems in vulnerable regions. As highlighted in the 2023 Synthesis Report, integrated adaptation measures supported by advanced monitoring technologies are essential to mitigate these risks and ensure sustainable development [1].

Reliable crop maps and phenological information are essential for monitoring agricultural land use and enabling decision makers to implement effective policies for sustainable development [2]. Sentinel-2 imagery has become a particularly important data source for this purpose because it provides multispectral and multi-temporal observations suitable for large-scale agricultural analysis. However, transforming raw Earth observation data into deployable agricultural intelligence remains challenging due to irregular acquisition intervals, cloud contamination, spatial heterogeneity, and the computational cost of processing multi-date image sequences. Historically, various approaches have been developed to transform raw Earth observation data into actionable agricultural intelligence. The work conducted between 2018 and 2020 relied heavily on classical machine learning methods such as Random Forest (RF), Support Vector Machine (SVM), CART, and ensemble classifiers for crop classification, marking a significant turning point in the use of Sentinel-2 data for agricultural monitoring [3–7]. These approaches commonly use manually engineered vegetation indices and carefully selected temporal windows to improve classification accuracy. For example, Sonobe et al. (2018) and Kobayashi et al. (2020) examined Sentinel-2-derived vegetation indices in ensemble learning frameworks, while Saini and Ghosh (2018) and Neetu and Ray (2019) demonstrated the viability of Sentinel-2 imagery at the agricultural parcel scale [3–6]. These studies established the value of Sentinel-2 time-series data for crop mapping, but they also revealed the dependence of traditional methods on feature engineering, temporal-window selection, preprocessing, and manual handling of cloud-affected observations [7].

Following the dominance of traditional machine learning approaches during the 2018–2020 period, crop classification approaches have experienced two major turning points since 2021. First, studies such as those by Chakhar et al. (2021) and Pan et al. (2021) have focused on radar (SAR) fusion and phenological rule-based models to overcome the limitations of optical sensor data (Sentinel-2) [8, 9]. However, a key limitation of these studies is that performance degrades when the defined threshold values or selected manual features (feature engineering) are applied to geographically distinct regions [10, 11].

In contrast to these traditional approaches, deep learning architectures (CNNs, RNNs, and attention mechanisms) have begun to extract spectral-temporal features more effectively by automating classification processes [12, 13]. Seydi et al. (2022) and Feng et al. (2023) showed that attention-based architectures can improve the extraction of spectral-temporal patterns for crop classification [12, 13].

Between 2024 and 2026, deep learning-based crop classification studies have shifted their focus beyond merely increasing accuracy; they now increasingly prioritize computational efficiency, irregular data management, cloud contamination, class imbalance, and reliable early season prediction. In this period, hybrid architectures that combine the advantages of different neural network structures became more prominent. For example, attention-based Bi-LSTM and Temporal Convolutional Neural Networks (TCNs) have

been combined on the BreizhCrops dataset to capture both local and global temporal patterns. By addressing class imbalance through undersampling, this approach achieved 82% accuracy and demonstrated the complementary role of Bi-LSTM in modeling sequential growth cycles and TCN in extracting broader temporal patterns [14].

In parallel, lightweight architectures have been developed to reduce the computational burden of large-scale satellite image processing. MCTNet, a multi-stage CNN-transformer network, adopts a pixel-based processing strategy instead of conventional patch-based methods and uses Attention-based Learnable Positional Encoding (ALPE) to compensate for missing observations caused by cloud cover in Sentinel-2 imagery [15]. More recently, the quadratic complexity of transformer models has motivated the exploration of Mamba or state-space-based architectures, which offer linear scalability. Hybrid CNN-Mamba structures combine the local texture extraction ability of CNNs with Mamba’s capacity to model long-range dependencies at lower memory cost, showing promise for scalable remote sensing analysis under memory constraints [16].

Recent studies have also emphasized the importance of phenological information for early season and crop-specific monitoring. In in-season early classification, research using the TimeSen2Crop dataset showed that multilayer stacking, combining LR, RF, and XGBoost with phenological features, provided more stable early season results than deep models such as transformers when temporal observations were limited [17]. Similarly, in arid agricultural ecosystems such as Xinjiang, China, phenological parameters, including Start of Season (SOS), Peak of Season (POS), and End of Season (EOS), extracted through Savitzky-Golay filtering and double logistic functions, have been integrated into deep learning models. By combining spectral-temporal features selected with the mRMR algorithm and a multi-scale transformer architecture, these approaches reduced salt-and-pepper noise in fragmented fields and achieved high crop-specific F1 scores, such as 94.37% for winter jujube [18].

However, these approaches still face several limitations. Pixel-based models may lose local spatial context, recurrent models can be difficult to parallelize over a long time series, and feature-engineering-based methods remain dependent on manually selected indices or phenological parameters. Similarly, high-precision phenological fitting requires dense and clean temporal observations, while many high-performing models are tuned to specific crop types or geographic regions, limiting their transferability.

Despite these advances, most existing approaches are primarily designed to assign a single crop label to a pixel, parcel, or time series. However, practical agricultural monitoring also requires date-specific phenological inference. In rice monitoring, the crop identity of a location may remain stable throughout the season, while its phenological stage changes according to the requested date. Therefore, models that only encode the seasonal image sequence, without explicitly conditioning on the query date, may be insufficient for reliable rice phenology prediction.

To address this gap, this study proposes a compact query-aware Sentinel-2 time-series model for joint crop type and rice phenology prediction. The model extracts point-centered multi-temporal Sentinel-2 patches, combines normalized multispectral reflectance with valid-pixel mask channels, encodes each acquisition date using a lightweight CNN, and models seasonal dependencies through a transformer encoder. A separate query day-of-year encoding is then fused with the seasonal image representation, allowing the model to produce different phenology predictions for the same spatial point depending on the requested date.

2. METHODOLOGY

2.1 Dataset and data source

The dataset utilized in this study is derived from the Sentinel-2 Multispectral Instrument (MSI) Level-2A (L2A) product, which provides Bottom-Of-Atmosphere (BOA) reflectance measurements [19]. To comprehensively capture the phenological cycle of crops, a multi-temporal dataset consisting of 29 distinct acquisition dates was constructed for each agricultural plot.

The labeled data is provided at the point (parcel-sample) level. In total, 778 labeled spatial points are available, each annotated with a single crop type (corn, rice, or soybean). Every point is additionally paired with query dates spanning the growing season, and each point-date pair forms one query row, yielding 5 446 query rows. A query row consists of the full multi-temporal Sentinel-2 patch sequence of the point, the queried day-of-year, and the point-level crop label; for rice queries it additionally carries the phenological stage label valid at the queried date, drawn from seven stages (Greenup, MidGreenup, Peak, Maturity, MidSenescence, Senescence, and Dormancy). Stage supervision is applied to rice only, mirroring the evaluation task. The dataset was released by ITU (UN agency) and the Zhejiang Lab (ZJ Lab) within the AI and Space Computing Challenge, 2026 [20]. The labeled points described above constitute the public training partition; a separate test partition is retained by the provider and its labels are not released.

The dataset includes 12 spectral bands (B01–B12), covering the visible, Near-Infrared (NIR), and Short-Wave Infrared (SWIR) regions of the electromagnetic spectrum. In contrast to conventional single-date classification approaches, the adopted multi-temporal strategy enables the model to capture the temporal evolution of crops across key phenological stages, including green-up, peak growth, and maturity [21], [22].

2.2 Spatial-temporal data structuring

To overcome the limitations of single-pixel analysis and to better represent spatial context and textural characteristics of agricultural parcels, a spatial-temporal 5D tensor representation (N,T,C,H,W) was employed. Here, N denotes the number of agricultural samples, T represents the temporal dimension (29 timestamps), C corresponds to the number of input channels, and H×W defines the spatial patch size (15×15 pixels).

The model utilizes a total of 24 input channels, consisting of 12 normalized Sentinel-2 spectral bands and 12 corresponding binary valid-pixel mask channels. This mask-aware design explicitly enables the model to distinguish between valid observations and cloud-contaminated pixels, thereby improving data reliability and robustness [23].

2.3 Proposed architecture: Query-aware CNN-transformer

To effectively process the high-dimensional spatial-temporal data, a hybrid query-aware CNN-transformer architecture (denoted as C03) was developed.

Spatial CNN encoder: Each temporal acquisition is independently encoded using a localized convolutional neural network composed of successive 2D convolutional layers, batch normalization [24], and GELU activation functions [25]. This module compresses each 15×15 spatial patch into a 256-dimensional feature embedding.

Acquisition day encoding: To incorporate temporal information, a sinusoidal encoding scheme based on sine and cosine functions is applied to represent the day-of-year (DOY) of each acquisition. This encoding mechanism provides explicit temporal context and follows the positional encoding strategy introduced in transformer architectures [26].

Temporal transformer: The sequence of spatial embeddings is then processed by a transformer encoder consisting of four layers and eight attention heads. Through the self-attention mechanism, the model captures long-range temporal dependencies and adaptively focuses on the most informative timestamps within the seasonal sequence [26].

2.4 Query-aware prediction mechanism

The key innovation of this study lies in the query-aware prediction mechanism. Since phenological stages are inherently time-dependent, the queried Day-of-Year (DOY) is encoded separately and concatenated with the aggregated seasonal representation of the input time series.

This design enables the model to generate different phenological stage predictions for the same spatial location depending on the specified query date. In this way, the model effectively bridges the gap between seasonal temporal patterns and instantaneous crop conditions. This approach is conceptually aligned with task-conditioned representations in multitask learning frameworks [27].

2.5 Multitask learning and optimization

The proposed framework adopts a Multitask learning (MTL) paradigm, where the shared feature representation is branched into two separate output heads: one for crop type classification and another for phenological stage prediction [28], [29].

The total training loss is defined as:

$$L_{total} = L_{crop} + 0.6 \times L_{stage} \quad (1)$$

This weighted loss formulation emphasizes phenological stage prediction for rice crops, effectively implementing a "double-hit" objective tailored to food security monitoring applications.

To ensure stable training, z-score normalization was applied using statistics computed exclusively from valid pixels within the training dataset.

2.6 Handling class imbalance and optimization

Agricultural datasets inherently exhibit class imbalance, where dominant classes (e.g. widely cultivated crops) significantly outnumber minority classes. This imbalance often leads to biased model training, causing standard optimization procedures to converge toward majority-class predictions and resulting in mode collapse.

To address this issue, class-imbalance mitigation strategies, including dynamic class weighting and focal loss [30], [31], were explored during model development. In the final C03 configuration, however, both task losses are standard unweighted categorical cross-entropy: the crop head is supervised on all query rows, while the stage head is supervised only on rice rows. This rice-only stage supervision, combined with crop-stratified data partitioning and macro-averaged F1 monitoring, where each class contributes equally to the score regardless of its frequency, was sufficient to prevent majority-class collapse. As a result, the more complex weighting schemes were not retained in the final model.

This final formulation keeps the optimization objective aligned with the macro-averaged evaluation metrics, under which misclassification of minority classes is penalized in the reported scores exactly as strongly as misclassification of dominant classes.

2.7 Spatial feature encoder (CNN module)

The primary task of the spatial feature encoder is to transform each 15x15 pixel patch obtained from Sentinel-2 data into a low-dimensional latent vector representing the textural and spectral signatures of the parcel. This process, whose general workflow is presented in Fig. 1, is defined as "spatial tokenization" and is performed independently for each time step t . In the model architecture, the input layer is configured as 24 channels by concatenating 12 spectral bands with 12 binary valid-pixel masks that provide spatial reliability and quality data to the model. At this stage of the model, instead of high-parameter standard architectures such as ResNet [32] or EfficientNet [33], an original lightweight CNN architecture named PatchCNNEncoder, optimized for small-scale input sizes and minimizing the risk of overfitting, has been preferred.

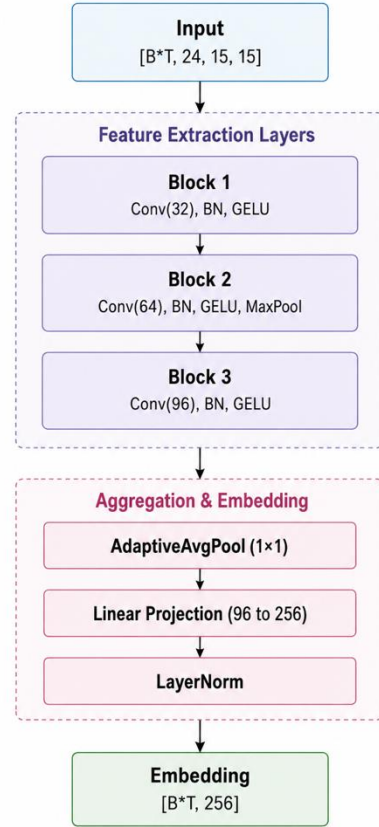


Figure 1 – PatchCNNEncoder feature extraction and embedding module

The proposed encoder architecture consists of three sequential convolutional blocks that provide hierarchical feature extraction. The first layer captures basic spectral relationships from the 24-channel raw data, and the subsequent layers code more complex textural patterns by increasing the number of channels to 32, 64, and 96, respectively. To optimize gradient flow and increase the nonlinear learning capacity of the network, Batch Normalization (BN) [24] and Gaussian Error Linear Unit (GELU) activation functions are applied after each convolutional layer [25].

$$x_l = GELU(BN(Conv_{3 \times 3}(x_{l-1}))) \quad (2)$$

Here, x_l represents the feature map of the respective layer, while $Conv_{3 \times 3}$ denotes the convolution operation performed with a stride of 1 and padding of 1.

To keep the spatial dimensions under control, a 2x2 Max-Pooling operation is applied following the second convolutional block. The architecture, whose hierarchical layer structure and tensor dimensions are visualized in Fig. 2, utilizes a Global Adaptive Average Pooling (GAP) mechanism in the final layer to summarize spatial information into a single vector [34]. This approach allows for the generation of a fixed-length feature representation regardless of the input size, while also maximizing parameter efficiency [35]. Following this process, for which layer-based operational details and parameter configurations are presented in Table 1, the resulting 96-channel feature map is

Table 1 – Detailed architectural specifications of the PatchCNNEncoder. The table outlines the operation types, input/output dimensions, and specific layer parameters (kernel size, normalization, and activation) utilized for lightweight spatial encoding

Operation / Layer	Input Size (C×H×W)	Parameters / Details	Output Size
Input (Patch)	24×15×15	12 Bands + 12 Valid Pixel Masks	24×15×15
Conv Block 1	24×15×15	Conv(3x3), BN, GELU	32×15×15
Conv Block 2	32×15×15	Conv(3x3), BN, GELU	64×15×15
Max Pooling	64×15×15	Kernel: 2, Stride: 2	64×7×7
Conv Block 3	64×7×7	Conv(3x3), BN, GELU	96×7×7
Global Avg Pool	96×7×7	AdaptiveAvgPool2d(1)	96×1×1
Flatten	96×1×1	-	96
Linear Proj.	96	Linear(96 → 256), LayerNorm	256

flattened and then projected into the 256-dimensional "token embedding" space required by the transformer module through a linear projection layer followed by a LayerNorm (Post-LN) step [36]. This projection and normalization step ensures that the raw features coming from the spatial encoder are transformed into a semantic and stabilized representation compatible with the transformer's self-attention mechanism.

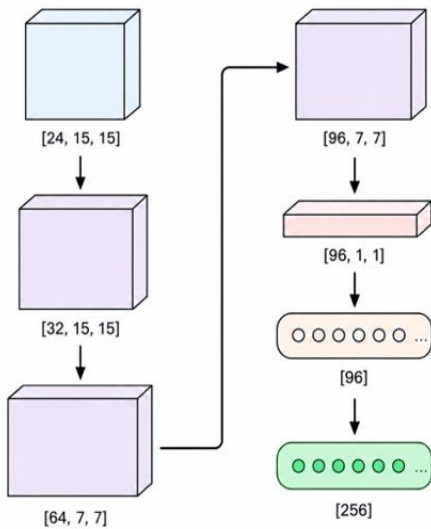


Figure 2 – CNN encoder architecture: Hierarchical feature extraction and transition from 24×15×15 input to 256-dimensional token embedding.

The PatchCNNEncoder successfully filters features representing the spectral and textural signatures of the parcel for each time step. However, agricultural classification depends on the change of these static images over time (phenological cycle) [37]. Therefore, the generated spatial features are transferred as input to the next stage, the temporal encoder module, to model the temporal dependencies between sequential observations.

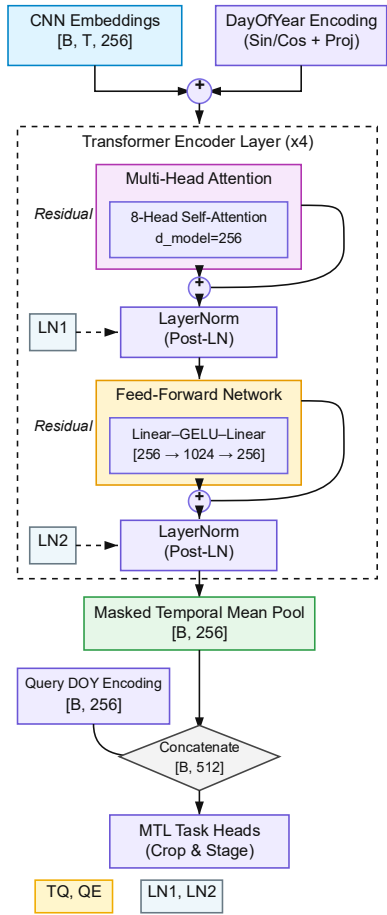


Figure 3 – Temporal feature encoder architecture: Integration of CNN embeddings with DOY encoding and four-layer multi-head self-attention transformer pipeline

2.8 Temporal learning and query-aware self-attention mechanism (transformer module)

The 256-dimensional feature sequence obtained from the PatchCNNEncoder module, which carries the spectral-spatial signature of the parcel for each time step, is

transferred to the temporal learning unit of the architecture. At this stage, a transformer-based structure capable of processing variable time intervals without being dependent on a fixed sequence length has been defined [26]. However, presenting satellite data to the transformer in its raw form may result in the loss of the temporal sequence (phenological cycle) within the series. To overcome this problem, the DayOfYearEncoding mechanism, whose sinusoidal projection process is detailed in Fig. 3, provides time-sensitive positional encoding by mapping raw day values (DOY) into the latent space via sinusoidal functions [26].

$$E_t = LN(GELU(W[\sin\theta, \cos\theta] + b))$$

These obtained temporal embeddings are added to the CNN vectors and presented to the multi-head self-attention layers [26]. The model architecture, whose architectural flow is presented in Fig. 3, utilizes four sequential transformer blocks, each containing eight parallel heads, to analyze the agricultural parcel's short-term growth bursts and long-term seasonal trends in different subspaces.

To maximize the representation capacity of the model, normalization processes were applied after the addition step (Post-LN), and the training process exhibited stable convergence [36]. GELU, which provides a smooth non-linear response, is preferred as the activation function [25].

One of the most critical advantages of the architecture is its ability to manage irregular time steps caused by cloud cover or lack of data through the time_mask [38]. During the self-attention calculation, invalid steps are masked and isolated from the gradient flow, and through the masked mean pooling applied at the end of the process, the summary of the entire season is reduced to a single vector. This method, preferred over the standard [CLS] token approach, provides a robust representation with lower variance because phenological evidence is distributed throughout the entire season [38].

At this stage, summarized with a 3D flow chart in Fig. 4, the obtained temporal summary vector reaches a critical matching point in the model's decision phase. The query_doy information, representing the moment of prediction, is integrated by being concatenated with the pooled vector at the transformer output, rather than being injected into the transformer's internal temporal sequence at the embedding level. This preference is based on fusion strategies that ensure features are preserved without distorting each other [39]. This strategic approach serves two primary purposes:

- **Preservation of temporal purity:** It allows the transformer layers to learn pure phenological changes and temporal dependencies within the sequence, independent of the query day.
- **Dominance filter:** It prevents the specific context of the query moment (query context) from dominating the multi-head attention weights within the transformer and overshadowing the seasonal summary.

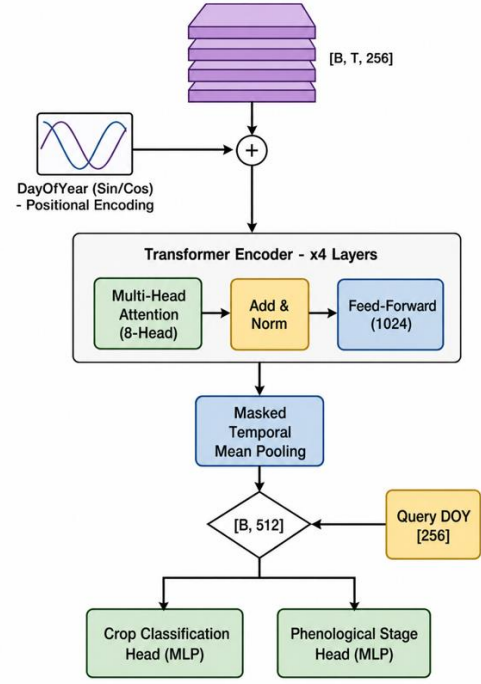


Figure 4 – Multitask Learning (MTL) framework: 3D schematic of sequence-to-vector transformation, featuring token pooling and date fusion for joint crop class and rice phenology prediction

This strategic fusion choice, whose configuration details, such as the number of layers, attention head depth, and total parameter load, are presented in Table 2, ensures the blending of the temporal sequence's internal dynamics with the specific context of the query moment without mutual distortion, and forms the final feature set for the Multitask Learning (MTL) heads [40].

Table 2 – Structural and training hyperparameters of the temporal transformer encoder

Parameter	Value
Model Architecture	nn.TransformerEncoder
Number of Layers (L)	4
Latent Dimension (dmodel)	256
Attention Heads (h)	8
FFN Hidden Dim (dff)	1024
Activation Function	GELU
Normalization Style	Post-LN
Dropout Rate	0.18
Positional Encoding	DayOfYear (SinCos)
Pooling Mechanism	Masked Mean Pool
Encoder Parameters	~3.16 Million

2.9 Multitask objective function

The proposed framework is trained using a Multitask Learning (MTL) objective, which jointly optimizes crop type classification and rice phenology estimation [40]. This approach mandates the model to learn both spatial features (crop type) and temporal-oriented features (stage) within the same shared feature space. This "joint learning" process

triggers information exchange between the tasks (inductive transfer) and enhances the model's generalization capacity. Consequently, it enables the generation of more robust features compared to models focused on singular tasks [41].

The interdependent task optimization strategy, which forms the foundation of the model design, aims to maximize the correlation between the two tasks. Since reliable agricultural monitoring requires both correct crop identification and date-specific phenological estimation, the loss function combines both objectives. This conditional optimization approach establishes a structural link between precision in the rice class and stage estimation accuracy [42]. To teach this constraint to the model as a mathematical objective, the total loss function is designed to consist of two components:

$$L_{total} = L_{crop} + \lambda_1 \cdot L_{stage} \quad (3)$$

Here, the hyperparameter λ_1 is set to 0.6 to regulate the impact of the rice phenology task on the total gradient flow. L_{crop} represents the categorical cross-entropy loss calculated for the corn, rice, and soybean classes [43].

This loss enables the model to distinguish crop types based on spectral patterns. L_{stage} , on the other hand, represents the cross-entropy loss calculated for seven different phenological stages (Greenup, MidGreenup, Peak, Maturity, MidSenescence, Senescence, and Dormancy). In the proposed model, stage estimation is modeled as a classification problem where stages are treated as distinct categories, rather than a regression (numerical prediction). This formulation allows the model to represent non-linear stage transitions and class-specific probability distributions [44]. Conflicting task gradients are a known failure mode in multitask optimization [45]. To avoid introducing noisy phenological gradients from non-rice samples, the stage loss was computed only for rice instances, while the crop classification loss was applied to all crop samples. This conditional loss design encourages the shared representation to learn general crop-discriminative features while preserving rice-specific temporal sensitivity for phenological estimation.

$$L(S) = -\frac{1}{|S|} \sum_{i \in S} \log p(y_i | x_i, d_{q,i}) \quad (4)$$

where S denotes the supervised set for the corresponding task, x_i is the input feature vector, $d_{q,i}$ is the associated query/domain information, and y_i is the ground-truth label. The crop classification loss is computed over all $N = 5\,446$ training query rows, whereas the stage classification loss is computed only over the subset of queries corresponding to rice samples. In the final configuration, $\lambda_1 = 0.6$.

2.10 Training strategy and implementation details

The training strategy of the proposed model was designed to support accurate crop classification and rice phenological stage estimation under resource-constrained inference scenarios [46]. The model was implemented using the

PyTorch deep learning library [47]. As illustrated in Fig. 5, the training pipeline updates the weights of the query-aware CNN–transformer architecture, which integrates spatial patch representations with temporal query information to generate joint crop and phenology predictions.

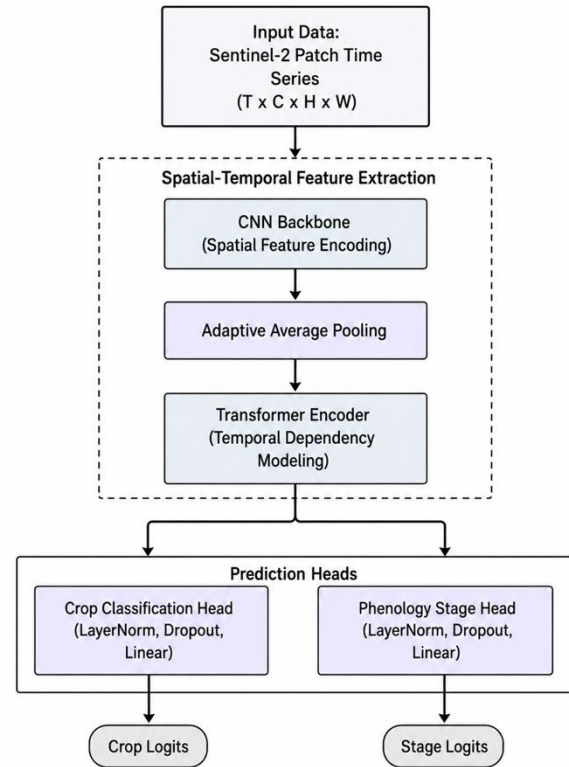


Figure 5 – Flowchart of Sentinel-2 time series transformation via CNN-Transformer backbone into joint crop and phenology logits

The training data consisted of 5 446 query rows derived from 778 unique spatial points. Each query row included the full multi-temporal Sentinel-2 patch sequence and a corresponding query day-of-year, allowing the model to learn both crop-level spatial patterns and date-specific rice phenology responses. The total loss function was implemented as a weighted combination of crop type classification loss and rice phenology stage loss, following the objective formulation described in Section 2.9.

Table 3 – Structural and operational configuration of the proposed model

Parameter	Value
Trainable parameters	3,275,146
Exported model size	12.55 MB
Input sequence length	29 time steps
Spectral bands	12
Valid mask channels	12

The structural and operational configuration of the proposed model is summarized in Table 3. The final C03 model contains 3 275 146 trainable parameters and has an exported weight size of 12.55 MB. The input sequence was represented using a fixed maximum temporal length of 29 acquisition slots, with an average of approximately 18 valid

Table 4 – Test performance, model size, and inference cost of evaluated systems

System	Model family	Model complexity	Checkpoint size	Macro-F1 _{Crop}	Macro-F1 _{RicePheno}	Total Score	Card Time (Hours)
C03 + P6	Query-aware CNN–Transformer	3.28M trainable parameters	12.55 MB	0.9941	0.9382	0.9605	0.0053
C03 baseline	Query-aware CNN–Transformer	3.28M trainable parameters	12.55 MB	0.9941	0.9382	0.9605	0.0119
C20 Fourier	CNN–Transformer with Fourier time encoding	3.28M trainable parameters	12.57 MB	1.0000	0.9074	0.9444	0.0106
C20 / XGBoost hybrid	Neural + tabular correction	3.28M trainable parameters + 21,587 tree nodes	18.77 MB	0.9941	0.8893	0.9312	0.0083
C29 TSViT	Temporal-spatial Transformer	23.17M trainable parameters	88.51 MB	0.9792	0.8960	0.9292	0.0117

Sentinel-2 observations per spatial point after masking unavailable dates. This representation allowed the model to capture seasonal crop development while preserving a consistent input structure for batch-based training. The model was optimized using the Adam optimizer, which was selected for its stable convergence behavior in deep neural network training [48]. The initial learning rate was set to 4×10^{-4} , with a batch size of 512 and a total training duration of 75 epochs. To reduce overfitting, the training duration was determined based on validation set behavior [43]. The compact parameter count and small exported model size make the proposed architecture appropriate for resource-constrained agricultural monitoring scenarios. However, runtime performance and deployment efficiency are evaluated separately in the experimental results, since model size alone does not fully determine end-to-end inference latency.

3. RESULTS

3.1 Evaluation protocol

The evaluation was based on two task-specific Macro-F1 metrics. The first metric, Macro-F1_{Crop}, measures crop-type classification performance for rice, corn, and soybean by averaging class-wise F1 scores. The second metric, Macro-F1_{RicePheno}, evaluates rice phenology recognition under a stricter joint criterion: a prediction contributes positively only when both the crop type and the phenological stage are correctly identified for the queried date. Therefore, the rice phenology metric is more restrictive than ordinary stage classification, since crop misclassification also affects the phenology score.

Formally, for a task with K classes,

$$\text{Macro-F1} = \frac{1}{K} \sum_{k=1}^K \frac{2P_k R_k}{P_k + R_k} \quad (5)$$

where P_k and R_k are the precision and recall of class k and classes with undefined precision or recall contribute zero. The ranking metric is the weighted total score $S = 0.4 \cdot \text{Macro-F1}_{\text{Crop}} + 0.6 \cdot \text{Macro-F1}_{\text{RicePheno}}$, where Macro-F1_{Crop} is computed over the $K = 3$ crop classes on all test queries and Macro-F1_{RicePheno} over the $K = 7$ stages on rice queries under the double-hit criterion described above.

Two evaluation regimes are used. For model development and the ablation studies, the 5 446 labeled query rows were partitioned with a region-grouped split where entire geographic regions were assigned to either training or validation, stratified by crop type. This produced 41 training regions (609 points; 4 263 query rows) and 8 spatially disjoint validation regions (169 points; 1 183 query rows, of which 259 stage-supervised rows belong to 37 rice points). Because grouping is performed at the region level, neither individual points nor the query rows derived from them can appear in both partitions, thereby preventing point or query-level leakage between the training and validation sets. For the final submitted model, the network was retrained on all 5 446 query rows with the 75-epoch fixed in advance from the validation runs, and this full data model was evaluated exclusively on the retained test set described below.

Test results are obtained through the provider’s evaluation service: a containerized inference pipeline is executed against the retained test partition of 942 point–date queries over 171 points, and only the aggregate metrics reported in Table 4 are returned. Per-sample predictions and per-class breakdowns are not accessible, so all per-class, confusion-matrix and significance analyses in this paper are computed

on the held-out validation regions. Because test labels are not accessible at any stage, the reported test scores cannot be affected by test set leakage or by model selection on test data.

3.2 Test performance and model comparison

Table 4 summarizes the test results and model footprint of the main evaluated systems. The final C03 + P6 system achieved a Macro-F1_{Crop} of 0.9941 and a Macro-F1_{RicePheno} of 0.9382, corresponding to a weighted total score of 0.9605. The same C03 model had previously produced the same component scores with a less optimized inference path, but with higher card consumption time. This indicates that the P6 pipeline improved inference efficiency without changing the predictive behavior of the trained model.

The comparison shows that the final system was not selected only because of its crop classification score. The C20 Fourier variant reached a perfect test set crop score, but its rice phenology score was lower than that of C03. Since the total metric gives greater weight to rice phenology, this reduced its overall score. The hybrid variant also did not improve the rice phenology component. Therefore, the final C03 + P6 system provided the strongest balance between crop recognition, rice stage prediction, model compactness, and inference cost.

The component scores also indicate that crop classification and rice phenology prediction have different difficulty profiles. Crop classification remained near-saturated across several candidates, whereas rice phenology varied more strongly and determined the final ranking.

3.3 Ablation study

To assess the contribution of the main architectural components, validation ablations were performed on query-date conditioning, acquisition-date encoding, and valid-pixel mask channels. These experiments are reported separately from the test set comparison because their purpose is diagnostic rather than final ranking. All ablations use the region-grouped validation split defined in Section 3.1.

Table 5 – Validation ablation study of temporal conditioning, mask-aware input representation, and spectral input attribution.

Variant	Ablated component	Macro-F1 _{Crop}	Macro-F1 _{RicePheno}	Total score
C03	None	1.0000	1.0000	1.0000
C07	Query DOY	0.9892	0.0673	0.4361
C08	Acquisition DOY	1.0000	0.9961	0.9976
C06	Valid mask channels	1.0000	0.9568	0.9740
C03N	Spectral bands → noise	0.1200	0.9884	0.6410
C03Z	Spectral bands → zero	0.2160	0.9613	0.6632

The most pronounced degradation occurred when query day-of-year (DOY) was removed. Crop classification remained high, with Macro-F1_{Crop} = 0.9892, but rice phenology decreased sharply to Macro-F1_{RicePheno} = 0.0673. This indicates that the Sentinel-2 time series alone is sufficient for crop recognition, but not for assigning a date-specific phenological stage. The label in the phenology task is therefore better understood as a point-date label rather than a point-only label.

A complementary input ablation isolates the contribution of the imagery. When the twelve spectral band channels of the trained C03 model are replaced by Gaussian noise at inference time, so that the input carries no image information, crop Macro-F1 on the validation regions falls to 0.1200 while rice stage Macro-F1 remains 0.9884. Setting the same channels to zero gives 0.2160 and 0.9613. Crop classification therefore depends on the spectral input as intended, whereas the rice stage score is largely preserved without it and reflects the queried date rather than the observed imagery.

Removing acquisition DOY produced a much smaller decrease, suggesting that acquisition timing helps the model represent the seasonal sequence but is less decisive than the queried date. Removing valid mask channels also reduced rice phenology performance, which supports the mask-aware input representation used in the proposed architecture.

The perfect validation entries in Table 5 (1.0000) warrant explicit interpretation. They are obtained on spatially disjoint validation regions, so they cannot be produced by train-validation overlap; rather, the crop score reflects the relative separability of the three crop classes at this scale, while the rice stage score reflects the near-deterministic relationship between the queried date and the stage label described above, together with the limited size of the validation rice subset (37 points, 259 stage-supervised queries), for which a zero-error outcome is attainable. With 259/259 correct stage predictions, a binomial rule-of-three bound places the 95% upper confidence limit of the per-query error rate at approximately 1.2% (8.1% at the point level, 37/37). The test set results corroborate rather than contradict these values: crop Macro-F1 remained 0.9941 on fully unseen data, while the stricter rice phenology metric dropped to 0.9382, quantifying the expected generalization gap. Section 3.6 probes the failure modes that these saturated headline metrics conceal.

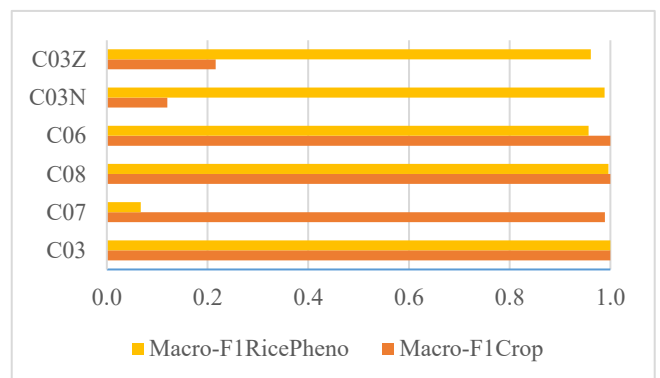


Figure 6 – Ablation results for query-date conditioning and spectral input attribution

3.4 Model complexity and generalization

The comparison between C03, C20 Fourier, and C29 TSViT suggests that increasing temporal or architectural complexity did not automatically improve the outcome. C20 Fourier introduced multi-harmonic temporal encoding and achieved strong local validation behavior, but its test set rice phenology score was lower than C03. This suggests that the richer temporal representation may have improved local fit without improving transfer to the test set.

The TSViT-style model represented a substantially larger temporal-spatial transformer alternative, with 23.17M parameters and an 88.51 MB checkpoint, compared with 3.28M parameters and 12.55 MB for C03. The final choice favored the compact query-aware CNN–transformer because it provided a better balance between performance and operational simplicity.

3.5 Runtime and inference efficiency

Runtime profiling showed that inference latency was dominated by geospatial preprocessing, not neural computation. In the submitted C03 runs, model forward time was below 0.5 s, while raster patch extraction accounted for most of the runtime. Therefore, optimization focused on the inference pipeline rather than changing the trained model.

P3 was the first optimized inference pipeline. It kept the C03 model unchanged but reduced preprocessing overhead by using faster TIFF inventory, increasing the GDAL cache, and in-memory patch tensors instead of writing an intermediate patch NPZ. Its profiling showed that the remaining bottleneck was still raster window reading.

P6 extended this idea by adding batched raster-window reads and multi-worker raster I/O. Instead of treating every point-band-date patch as an isolated read, P6 grouped compatible raster reads and executed extraction in parallel where possible. This reduced patch extraction substantially while preserving the same model weights, output format, and prediction distribution.

Table 6 – Runtime comparison of C03 inference pipelines

Pipeline/ Model	Internal runtime (s)	Patch extraction (s)	Model forward (s)	Card Time (H)
P3 (C03)	35.37	33.43	0.42	0.0119
P6 (C03)	10.15	8.37	0.39	0.0053

The runtime results indicate that system-level geospatial optimization can be as important as model architecture for deployment-oriented remote-sensing inference. The 3.48× reduction in internal runtime was achieved without changing the trained model, showing that the main efficiency gain came from reducing raster-I/O and patch-construction overhead.

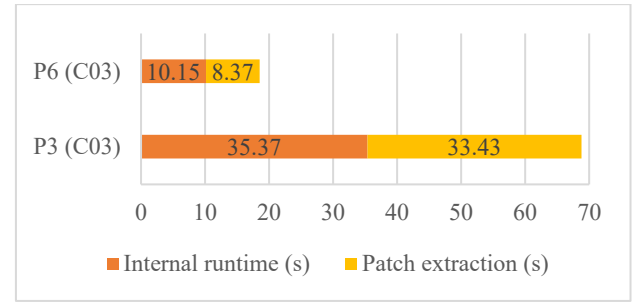


Figure 7 – Runtime breakdown of P3 and P6 inference pipelines

3.6 Class-level results and failure analysis

Because the test set returns only aggregate metrics, class-level diagnostics were computed on the validation regions. These diagnostics use the validation-phase counterparts of C03 and C20, which share the architecture and training configuration of the full data models but were trained on the 41 training regions alone, so the evaluated models have not seen the validation rows. At the nominal query date the crop confusion matrix is exactly diagonal, with all 567 corn, 259 rice, and 357 soybean queries correctly assigned, and the rice stage confusion matrix is likewise diagonal across its seven stages at 37 queries per stage. Per-class precision, recall, and F1 consequently equal 1.0 for every class. A non-parametric point-cluster bootstrap over the 169 validation points yields degenerate confidence intervals in this setting, and uncertainty at the nominal date is therefore expressed by the rule-of-three bound reported in Section 3.3.

Because the nominal date metrics are saturated, the query date was perturbed by ± 7 days at inference time in order to characterize the error behavior of the model. Under a -7 -day shift, rice stage Macro-F1 declines to 0.9807, with only sporadic confusions between adjacent stages. The $+7$ -day shift produces a larger decline, with rice stage Macro-F1 falling to 0.8546. Its confusion structure is given in Table 7. Errors are almost exclusively substitutions between adjacent or near-adjacent stages concentrated in the mid-to-late season, with maturity predicted as peak in 17 of 37 queries and peak predicted as senescence in 12 of 37, while the early stages (Greenup, MidGreenup) and the final stage (Dormancy) are largely unaffected. This pattern corresponds to the senescence-transition weakness discussed in Section 4.5: as the queried date moves toward the late season, spectrally similar drying-vegetation stages become the dominant error mode. Crop classification is insensitive to the same perturbation, with crop Macro-F1 remaining 1.0000 at all tested shifts.

Table 7 – Rice stage confusion matrix on the region-held-out validation split under a +7-day query-date shift (rows: true stage; columns: predicted). GU = Greenup, MG = MidGreenup, PK = Peak, MT = Maturity, MS = MidSenescence, SN = Senescence, DM = Dormancy; 37 supervised rice queries per stage.

True	GU	MG	PK	MT	MS	SN	DM
GU	33	4	0	0	0	0	0
MG	0	34	0	3	0	0	0
PK	0	0	25	0	0	12	0
MT	0	0	17	20	0	0	0
MS	0	0	0	0	36	0	1
SN	0	0	0	0	0	37	0
DM	0	0	0	0	0	0	37

The same protocol supports a statistical comparison between architectures. At the nominal query date, C03 and the C20 Fourier variant produce identical validation predictions, with no discordant pairs and an exact McNemar test giving $p = 1.0$, so the validation split alone does not distinguish them. This is consistent with Section 3.2, where the two models differ only on the test set. Under the +7-day query shift a difference emerges: C03 is correct on 14 rice stage queries for which C20 is not, while the reverse holds for 7, giving rice stage Macro-F1 of 0.8546 against 0.8280. The direction of this difference matches the higher test phenology score obtained by C03 (0.9382 against 0.9074), but it does not reach statistical significance on the small validation rice set (exact McNemar test, $p = 0.19$). The test set therefore remains the decisive comparison between the two architectures, and a larger held-out set would be required for a well-powered comparison.

3.7 Summary

The results show that the proposed query-aware CNN–transformer achieves strong test set performance while remaining compact and efficient. The final C03 + P6 system obtained the best total score among the evaluated candidates, primarily because it preserved high crop classification performance while achieving the strongest rice phenology score. The ablation study confirms that query-date conditioning is essential, removing the queried day-of-year leaves crop classification largely intact but causes rice phenology prediction to fail. The complementary ablation in Section 3.3 shows the converse: removing the imagery leaves the rice stage score largely intact, so the phenology metric reflects the queried date rather than the spectral input. Finally, runtime profiling shows that inference efficiency is limited more by raster patch extraction than by neural forward computation, and that the P6 pipeline substantially reduces this overhead.

4. DISCUSSION

This section interprets the experimental findings and positions the proposed approach within the broader context of remote sensing and deep learning literature. The results

demonstrate not only strong predictive performance but also important insights into model design, temporal reasoning, and deployment constraints.

4.1 Impact of spatial-temporal fusion

One of the key design choices in this study is the use of spatial patches (15×15) instead of single-pixel inputs. This spatial-temporal fusion strategy significantly improves model robustness by incorporating local contextual information.

Single-pixel approaches are highly sensitive to noise, atmospheric effects, and sensor inconsistencies, particularly in satellite imagery. By contrast, the use of spatial patches allows the model to capture textural patterns and neighborhood-level relationships, effectively smoothing spectral noise and improving class separability.

The experimental results confirm that this design contributes to stable and high-accuracy predictions across most phenological stages. This finding is consistent with prior studies demonstrating that CNN-based spatial feature extraction enhances remote sensing classification performance [13].

4.2 Multitask learning and the ‘Double Hit’ strategy

The proposed framework adopts a multitask learning paradigm, jointly predicting crop type and phenological stage. A key innovation in this setup is the ‘Double Hit’ strategy, where phenology loss is selectively emphasized for rice samples, aligning directly with the total scoring objective.

This targeted learning mechanism encourages the model to develop shared representations that capture both structural (crop identity) and temporal (growth stage) characteristics. As a result, the model achieves strong generalization while maintaining high sensitivity to rice-specific phenological transitions.

The effectiveness of this approach is reflected in the near-perfect validation performance and high test set scores. These findings support the hypothesis that multitask learning, when properly aligned with task priorities, can significantly improve performance through feature sharing and inductive bias [41].

4.3 Query-aware learning as a paradigm shift

A central contribution of this work is the introduction of query-aware prediction, where the model explicitly conditions its output on a requested Day-of-Year (DOY).

Traditional multi-temporal models implicitly learn seasonal patterns but lack the ability to answer time-specific queries. The ablation results clearly demonstrate that removing query awareness leads to a catastrophic drop in phenology prediction performance.

This suggests that phenological modeling should not be treated purely as a sequence classification problem, but

rather as a conditional prediction task. The query-aware design bridges this gap by enabling the model to dynamically interpret temporal signals based on user-defined queries.

This paradigm aligns with recent advances in attention-based architectures, where conditioning mechanisms and positional encoding play a critical role in temporal reasoning [25].

4.4 Computational efficiency for deployment-oriented inference

Runtime analysis showed that the main execution bottleneck was not the neural model itself, but the geospatial preprocessing required to extract multi-temporal patches from raster data. In the earlier C03 pipeline, the total internal runtime was 35.37 s, of which 33.43 s was spent on patch extraction, while the model forward pass required only 0.42 s. This indicates that most of the computational cost came from raster-window reading and patch construction rather than from the CNN–transformer model.

The optimized P6 pipeline reduced this bottleneck without changing the trained C03 weights or prediction outputs. Total internal runtime decreased from 35.37 s to 10.15 s, corresponding to a $3.48\times$ speed-up. Patch extraction time decreased from 33.43 s to 8.37 s, while model forward time remained nearly unchanged, from 0.43 s to 0.39 s. The reported card time also decreased from 0.0119 H to 0.0053 H, showing that the optimization improved the practical execution cost of the system.

These results suggest that, in operational remote-sensing workflows, model accuracy alone is not sufficient; efficient raster access and preprocessing can be equally important for real-world deployment. By reducing the data-loading bottleneck while preserving predictive performance, P6 makes the C03 framework more suitable for time-sensitive and resource-constrained agricultural monitoring scenarios.

These runtime figures were measured on an A100 GPU on the ground-based evaluation platform. Transfer to on-orbit hardware would require dedicated benchmarking on power-constrained, radiation-tolerant processors, which lies outside the scope of this study.

4.5 Limitations and failure modes

Despite its strong performance, the proposed model exhibits certain limitations. The most notable challenge arises in late-stage phenology (senescence), where spectral signatures of dry vegetation closely resemble bare soil. This causes senescence to be over-predicted, reducing its precision and the recall of the preceding maturity and peak stages, and indicates a fundamental limitation of spectral-based classification approaches.

Additionally, the model relies on the availability of sufficient temporal coverage. In scenarios with sparse acquisition dates or heavy cloud contamination, performance may degrade despite the use of valid mask channels. Future work could address these limitations by incorporating additional data sources (e.g., SAR imagery) or more advanced temporal interpolation techniques.

The failure analysis in Section 3.6 places quantitative bounds on these limitations: under a +7-day query perturbation, 29 of the 37 stage errors on validation rice queries fall within the transition zone spanning peak, maturity and senescence. Two further constraints apply to the results reported here. The validation rice subset comprises only 37 points, so the perfect nominal date stage scores carry a non-negligible upper error bound, as quantified in Section 3.3. Because the test set returns only aggregate metrics, per-class test metrics and test set confusion matrices cannot be reported, and the diagnostic conclusions drawn in this work therefore rest on validation data. All results likewise remain limited to a single regional dataset covering three crop classes.

4.6 Comparison with alternative architectures

The experimental study also evaluated more complex architectures, including Fourier-based temporal encoding and TSViT models. While these approaches demonstrated strong validation performance, they did not outperform the C03 model on the test set.

Moreover, these alternatives introduce significantly higher computational costs and model complexity. For example, TSViT models require substantially more parameters and exhibit slower inference speeds, making them less suitable for deployment under strict resource constraints.

These findings highlight an important trade-off: model complexity does not always translate to better real-world performance, especially in constrained deployment scenarios. The success of the C03 model demonstrates the value of balanced design, where simplicity, efficiency, and robustness are prioritized.

4.7 Implications for future remote sensing systems

The results of this study have broader implications for the design of next generation remote sensing systems. The combination of query-aware modeling, spatial-temporal feature fusion, and efficient deployment pipelines provides a scalable framework for satellite-based intelligence.

Such systems can move beyond passive observation and enable interactive, query-driven Earth monitoring, where users request specific temporal insights rather than static classifications. This paradigm shift has the potential to significantly enhance applications in agriculture, climate monitoring, and food security analysis.

5. CONCLUSION

This study presents a comprehensive framework for multi-temporal agricultural monitoring using Sentinel-2 satellite imagery, addressing the joint tasks of crop type classification and rice phenology prediction. By combining spatial feature extraction with temporal attention mechanisms and introducing a query-aware prediction paradigm, the proposed approach achieves both high accuracy and deployment efficiency.

5.1 Summary of contributions

The primary contribution of this work is the development of a query-aware CNN-transformer architecture that effectively models both spatial and temporal dependencies in multispectral satellite data. Unlike conventional approaches, the proposed model explicitly incorporates the requested query date, enabling accurate point-in-time phenological predictions rather than static seasonal classification.

A second key contribution lies in the integration of multitask learning with a task-aligned optimization strategy. The proposed ‘Double Hit’ mechanism prioritizes rice phenology prediction in accordance with the evaluation metric, resulting in strong performance on both crop classification and stage prediction tasks.

In addition, the study demonstrates the importance of spatial-temporal data representation, showing that patch-based inputs and valid mask channels significantly improve robustness against noise, missing data, and cloud contamination. From a systems perspective, the P6 inference pipeline represents a practical contribution toward real-world deployment. By optimizing raster I/O operations and memory usage, the framework achieves substantial reductions in runtime without compromising predictive performance. This makes it a candidate for constrained deployment environments, although operation on flight hardware would first require validation on flight-representative processors.

Overall, the proposed approach achieves near-perfect performance on region-held-out validation data and a high test set score for the region and crop set studied here. Establishing scalability to other regions and crop taxonomies remains a matter for future work.

5.2 Future work

Despite its strong performance, several avenues for future research remain open.

One promising direction is the integration of Synthetic Aperture Radar (SAR) data, which can provide complementary information under cloud-covered or low-visibility conditions. Combining optical and SAR data could significantly enhance robustness and enable continuous monitoring across all weather scenarios [49].

Another important extension involves geographical and taxonomic generalization. The present results derive from three crops in a single regional dataset, with validation regions drawn from the same agro-ecological zone as the training regions, so cross-zone behavior remains open. A systematic evaluation would proceed in stages. The first stage is cross-region transfer within the same sensor and season, in which the trained model is applied zero-shot to geographically distant areas. The second is cross-year transfer, where phenological calendars shift with weather. The third is evaluation on public multi-region crop-mapping benchmarks, which would enable direct comparison with prior work. The fourth is extension of the crop taxonomy beyond corn, rice, and soybean, which would require revisiting the class-imbalance behavior of the model.

Domain adaptation techniques such as feature alignment, phenological-calendar normalization of the day-of-year encodings, and lightweight regional fine-tuning could then be applied to close the transfer gaps that these evaluations expose, with statistical comparisons of the kind introduced in Section 3.6 applied at each stage.

Future work may also investigate model compression and optimization techniques, such as pruning and quantization, to further reduce computational requirements for edge deployment. Similar to previous work on AI-based performance prediction in photovoltaic systems [50], the integration of data-driven models with real-world constraints highlights the importance of efficient and scalable architectures. Additionally, incorporating more advanced temporal modeling strategies, including adaptive attention or hybrid architectures, could improve performance in challenging phenological stages.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the International Telecommunication Union (ITU) and Zhejiang Lab (ZJ Lab) for organizing the AI and Space Computing Challenge 2026 and for providing the Sentinel-2 dataset and the evaluation service used in this study.

REFERENCES

- [1] H. Lee and J. Romero. "Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change". In: IPCC (2023).
- [2] S. Bargiel. "A new method for crop classification using time-series of Sentinel-1 images and hierarchical multi-temporal conditional random fields". In: Remote Sensing of Environment (2017).
- [3] R. Saini and S. K. Ghosh. "Crop classification on single date Sentinel-2 imagery using random forest and support vector machine". In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (2018).
- [4] R. Sonobe et al. "Crop classification from Sentinel-2-derived vegetation indices using ensemble learning". In: Journal of Applied Remote Sensing (2018).
- [5] Neetu and S. S. Ray. "Exploring machine learning classification algorithms for crop classification using Sentinel-2 data". In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (2019).
- [6] N. Kobayashi et al. "Crop classification using spectral indices derived from Sentinel-2A imagery". In: Journal of Information and Telecommunication (2020).
- [7] Z. Yi et al. "Crop classification using multi-temporal Sentinel-2 data in the Shiyang River Basin of China". In: Remote Sensing (2020).

- [8] A. Chakhar et al. "Improving the accuracy of multiple algorithms for crop classification by integrating Sentinel-1 observations with Sentinel-2 data". In: Remote Sensing (2021).
- [9] L. Pan et al. "Mapping winter crops using a phenology algorithm, time-series Sentinel-2 and Landsat-7/8 images, and Google Earth Engine". In: Remote Sensing (2021).
- [10] T. Rusňák et al. "Crop mapping without labels: Investigating temporal and spatial transferability of crop classification models using a 5-year Sentinel-2 series and machine learning". In: Remote Sensing (2023).
- [11] H. Xue et al. "Object-oriented crop classification using time series Sentinel images from Google Earth Engine". In: Remote Sensing (2023).
- [12] S. T. Seydi et al. "A dual attention convolutional neural network for crop classification using time-series Sentinel-2 imagery". In: Remote Sensing (2022).
- [13] F. Feng et al. "A deep learning framework for crop mapping with reconstructed Sentinel-2 time series images". In: Computers and Electronics in Agriculture (2023).
- [14] A. N. A. A. Bandar and A. Coşkunçay. "Crop classification with attention based BI-LSTM and temporal convolution neural network combination for remote sensing Breizhcrop time series data". In: Yuzuncu Yil University Journal of the Institute of Natural & Applied Sciences (2024).
- [15] X. Han et al. "MCTNet: A lightweight multi-stage CNN-transformer network for pixel-based crop mapping using Sentinel-2 time series". In: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2024).
- [16] Z. Li et al. "Mamba for remote sensing: Architectures, hybrid paradigms, and future directions". In: Preprints.org (2025).
- [17] A. Alimagambetova et al. "Early crop type classification based on seasonal spectral features and machine learning methods". In: Technologies (2026).
- [18] J. Wang et al. "Deep learning-based crop phenology extraction and classification in an arid oasis using multi-temporal Sentinel-2 imagery". In: Journal of Zhejiang University-SCIENCE B (2026).
- [19] M. Drusch et al. "Sentinel-2: ESA's optical high-resolution mission for operational services". In: Remote Sensing of Environment (2012).
- [20] International Telecommunication Union (ITU) and Zhejiang Lab. "AI and Space Computing Challenge". (2026).
- [21] V. S. F. Garnot et al. "Satellite image time series classification with pixel-set encoders and temporal self-attention". In: CVPR Workshops (2019).
- [22] M. Rußwurm and M. Körner. "Multi-temporal land cover classification with sequential recurrent encoders". In: ISPRS International Journal of Geo-Information (2018).
- [23] A. Taravat et al. "Automatic cloud masking for Sentinel-2 imagery using deep learning structures". In: Remote Sensing (2021).
- [24] S. Ioffe and C. Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: ICML (2015).
- [25] D. Hendrycks and K. Gimpel. "Gaussian error linear units (GELUs)". In: arXiv (2016).
- [26] A. Vaswani et al. "Attention is all you need". In: NeurIPS (2017).
- [27] Z. Zhang et al. "Facial landmark detection by deep multi-task learning". In: ECCV (2014).
- [28] A. Kendall et al. "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics". In: CVPR (2018).
- [29] S. Vandenhende et al. "Multi-task learning for dense prediction tasks: A survey". In: IEEE TPAMI (2021).
- [30] T.-Y. Lin et al. "Focal loss for dense object detection". In: IEEE ICCV (2017).
- [31] H. He and E. A. Garcia. "Learning from imbalanced data". In: IEEE Transactions on Knowledge and Data Engineering (2009).
- [32] K. He et al. "Deep residual learning for image recognition". In: CVPR (2016).
- [33] M. Tan and Q. Le. "EfficientNet: Rethinking model scaling for convolutional neural networks". In: ICML (2019).
- [34] M. Lin et al. "Network in network". In: arXiv (2013).
- [35] A. Howard et al. "Searching for MobileNetV3". In: ICCV (2019).
- [36] R. Xiong et al. "On layer normalization in the transformer architecture". In: ICML (2020).
- [37] L. Zeng et al. "A review of crop-specific condition monitoring by integrating remote sensing and phenological modeling". In: Remote Sensing (2019).
- [38] N. Reimers and I. Gurevych. "Sentence-BERT: Sentence embeddings using Siamese BERT-networks". In: EMNLP (2019).
- [39] Y. Dai et al. "Attentional feature fusion". In: WACV (2021).
- [40] R. Caruana. "Multitask learning". In: Machine Learning (1997).
- [41] J. Baxter. "A model of inductive bias learning". In: Journal of Artificial Intelligence Research (2000).
- [42] M. Guo et al. "Dynamic task prioritization for multitask learning". In: ECCV (2018).

- [43] I. Goodfellow et al. "Deep learning". In: MIT Press (2016).
- [44] L. Denis et al. "Crop phenology forecasting using multi-temporal Sentinel-2 data and deep learning". In: Remote Sensing of Environment (2020).
- [45] T. Yu et al. "Gradient surgery for multi-task learning". In: NeurIPS (2020).
- [46] J. Chen and X. Ran. "Deep learning with edge computing: A review". In: Proceedings of the IEEE (2019).
- [47] A. Paszke et al. "PyTorch: An imperative style, high-performance deep learning library". In: NeurIPS (2019).
- [48] D. P. Kingma and J. Ba. "Adam: A method for stochastic optimization". In: ICLR (2015).
- [49] M. Schmitt et al. "The fusion of Sentinel-1 and Sentinel-2 data for earth observation: A review". In: ISPRS Journal of Photogrammetry and Remote Sensing (2020).
- [50] B. Durmaz and C. Karakuş. "Kırşehir yöresindeki fotovoltaik güç santrallerinin uzun dönemli performans analizi ve yapay zeka tabanlı verim tahminlemesi [Long-term performance analysis and AI-based efficiency prediction of photovoltaic power plants in the Kırşehir region]". In: Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi (2026).

AUTHORS

Yusuf Öztıp (ORCID: 0009-0002-3764-5673) is currently pursuing a B.A. degree in translation and interpreting studies at Boğaziçi University, İstanbul, Türkiye. He is the team leader of the Kybelix AI Team. His professional experience includes post-production and Turkish-Spanish translation work for international television productions at Acunmedya, as well as data annotation and reviewer for AI training pipelines in Scale AI. His academic interests lie at the intersection of language technologies and artificial intelligence, focusing on machine learning, natural language processing, computer vision, neural architectures, and resource-constrained AI deployment.

Sudenaz Şenbay (ORCID: 0009-0001-9869-4471) received a B.S. degree in computer engineering from Burdur Mehmet Akif Ersoy University, Burdur, Türkiye, in 2026. She is a developer and researcher within the Kybelix AI team, where she has contributed to award-winning satellite intelligence frameworks. Her research interests include remote sensing, computer vision, Explainable AI (XAI), and trustworthy machine learning with an emphasis on model robustness.

Cavit Karakuş (ORCID: 0009-0002-7702-3975) received his B.Sc. degree in information systems and technologies from Mersin University. He began his professional career serving as a social media specialist for two years at a Petrol Ofisi distributor in Kayseri. He is currently pursuing an interdisciplinary M.Sc. degree in advanced technologies at Kırşehir Ahi Evran University, where his research focuses

on the integration of artificial intelligence with physical sciences. Professionally serving as an AI specialist, he specializes in developing automated data pipelines, Large Language Models (LLMs), and deep learning architectures. He is the co-author of the paper titled "Long-term performance analysis and artificial intelligence-based yield prediction of photovoltaic power plants in the Kırşehir region". His recent work involves high-dimensional data modeling for material property prediction using Density Functional Theory (DFT) and Graph Neural Networks (GNNs).

Berrin Bilgin (ORCID: 0009-0006-6874-4219) received a B.S. degree in physics from Middle East Technical University (METU) in 2023. She is currently a scholar in artificial intelligence and data science at the AI and Technology Academy, a joint initiative supported by the Ministry of Industry and Technology of Türkiye, Google Türkiye, and the Turkish Entrepreneurship Foundation. Her academic background is rooted in applied physics and data-driven analysis, as evidenced by her publication, "Impact of COVID-19 on student performance in a physics laboratory course," which utilizes statistical data science techniques to evaluate educational outcomes. Her current research interests include computational physics, satellite remote sensing, machine learning for Earth observation, and high-dimensional data analysis.

Ozan Çelik (ORCID: 0009-0009-3839-8054) received his B.Sc. degree in computer engineering from Atılım University, Ankara, Türkiye, in 2025. His professional background includes a summer internship as a mobile application developer at KAM Innovation in 2023, followed by a Test and Automation Engineering internship at HAVELSAN from 2024 to 2025. He is currently an active fellow at the AI and Technology Academy. His professional and academic achievements include winning first place at the Atılım University Business Idea Competition in 2023. He has one paper currently under review regarding tactical graphics validation using deep learning. His current research interests include deep learning, computer vision, natural language processing (NLP), and applied artificial intelligence. He is a co-founder of the Atılım University R&D Community and has served in leadership roles within AIESEC.