Spatio-Temporal Deep Learning for Viticulture Potential Prediction and CNN-Based Crop Identification Using Sentinel-2 Satellite Imagery

Innovware at ImageCLEF 2026 AI4AGRI

Janani S P[1], Ishana B[1] and Mirunalini P[1]

[1]*Department of Computer Science and Engineering,*
Sri Sivasubramaniya Nadar College of Engineering, Rajiv Gandhi Salai, Chennai 603110, TamilNadu, India

Abstract

The AI4AGRI challenge is motivated by the increasing need for sustainable and data-driven agricultural management through advanced artificial intelligence and remote sensing technologies. By leveraging multi-source Sentinel satellite imagery, the challenge aims to support applications such as land suitability assessment and early crop identification, enabling efficient agricultural planning, optimized resource utilization, and informed decision-making in precision agriculture. This paper presents the systems developed by Team Innovware for the ImageCLEF 2026 AI4AGRI challenge, which comprises two subtasks. Subtask 1, AgriPotential Viticulture, focuses on predicting grape-growing suitability at the pixel level using multi-temporal Sentinel-2 imagery and is formulated as an ordinal segmentation problem with five classes ranging from Very Low to Very High viticulture potential. To address this task, we propose a framework combining per-band normalization, a 3D convolutional temporal encoder for capturing seasonal vegetation dynamics across 34 acquisition dates, and a U-Net segmentation architecture with a pretrained ResNet-18 backbone. The model was trained using a combination of cross-entropy and ordinal mean squared error losses, achieving a validation Accuracy±1 score of 0.7284 and a competition test score of 56.15%. In Subtask 2, DACIA5 Crop Identification, involves classifying 32×32 Sentinel-2 image patches into one of seven crop categories under a temporal domain shift setting, where models trained on 2020–2023 data are evaluated on unseen 2024 data. For this task, we developed a lightweight spectral CNN with residual connections trained using weighted cross-entropy loss to address class imbalance. The proposed system achieved an Overall Accuracy of 37.31%, an Average Accuracy of 26.93%, and a Q1 score of 32.12.

Keywords

viticulture potential, crop identification, multi-spectral images, satellite image time series, spatio-temporal deep learning, 3D convolution, U-Net, ordinal classification, Sentinel-2, DACIA5, ImageCLEF 2026

1. Introduction

Predicting agricultural potential and identifying crop types from satellite imagery are practically important problems for precision agriculture and land use planning. The ImageCLEF 2026 AI4AGRI task [1] addresses both challenges through two complementary subtasks. These subtasks are motivated by the need to enhance sustainable and efficient agricultural practices through advanced artificial intelligence.

Subtask 1 focuses on viticulture potential prediction, which aims to estimate how suitable the land is for grapevine cultivation rather than predicting crop yield or production. Unlike yield prediction, which estimates the quantity of agricultural output, viticulture potential prediction classifies land based on its capability to support grapevine cultivation. Viticulture, the cultivation of grapevines, is especially sensitive to local environmental conditions such as soil type, drainage, sun exposure, and microclimate. These factors interact over time, meaning a single satellite image is often insufficient to characterize a location. A time series of images captured across multiple seasons can reveal patterns of vegetation growth, stress response, and phenological change that are diagnostic of different land use types and agricultural potentials. This task makes use of multi-spectral Sentinel-2 imagery from Southern France

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany
janani2410063@ssn.edu.in (J. S. P); ishana2410340@ssn.edu.in (I. B); miruna@ssn.edu.in (M. P)

collected between 2017 and 2019. The task is formalized as a pixel-level ordinal classification problem where each pixel is represented with one of the 5 possible classes of viticulture potential.

Subtask 2 addresses crop identification using the DACIA5 dataset, which contains Sentinel-2 multi-spectral patches from a region in Braşov, Romania, collected between 2020 and 2024. The challenge simulates a real-world scenario where a model trained on historical data (2020–2023) must generalize to unseen data from 2024. Each patch must be classified into one of seven crop categories. This paper describes our approaches to both subtasks, the issues encountered during development, and the results obtained.

2. Task and Dataset Description

2.1. Subtask 1: AgriPotential Viticulture

The AgriPotential Viticulture task requires participants to predict, for each pixel of a 128×128 patch of agricultural land, its viticulture potential class. The model should be able to estimate the agricultural potentials for viticulture (grapevine cultivation) on a agricultural land based on multi-temporal multi-spectral satellite imagery. The five ordinal classes are defined in Table 1.

Table 1
Viticulture potential class definitions.

Value	Label	Description
0	Very Low	Very poor suitability for grapevine cultivation
1	Low	Below-average suitability
2	Average	Moderate suitability
3	High	Good suitability for grapevines
4	Very High	Excellent suitability for viticulture

2.1.1. Input Data

The input data consists of 34 Sentinel-2 [5] multi-spectral images of the T31TEJ tile, located in Southern France, acquired between January 2017 and August 2019 at irregular intervals, drawn from the AgriPotential benchmark dataset [2]. Each image has 10 spectral bands: B2 (Blue), B3 (Green), B4 (Red), B5, B6, B7 (Red Edge), B8 (NIR), B8A (Narrow NIR), B11, and B12 (SWIR), unlike normal cameras which has only 3 spectral bands (RGB). All bands are super-resolved to a common spatial resolution of 5 metres per pixel. Patches of 128×128 pixels are extracted from the full tile. The full input tensor for a single patch has shape $(34, 10, 128, 128)$.

2.1.2. Dataset Splits

The dataset is partitioned into three subsets as shown in Table 2. Labels for test patches are withheld by the organisers and used only for evaluation.

Table 2
Subtask 1 dataset split sizes.

Split	Patches	Labels available
Training	6329	Yes
Validation	781	Yes
Test	800	No (hidden)

The full dataset totals approximately 200 GB. The compute environment (a Kaggle T4 GPU kernel) provided only around 20 GB of usable local disk space. The proposed system therefore streamed data directly from the HuggingFace repository during training using the GDAL VSI interface with caching enabled. This reduced disk requirements to near zero but added substantial latency: each training epoch took approximately 2 hours, limiting the number of experiments we could run.

2.2. Subtask 2: DACIA5 Crop Identification

The DACIA5 Crop Identification task [8] requires classifying each 32×32 Sentinel-2 patch into one of seven crop categories: Wheat, Corn, Peas, Rapeseed, Potato, Sugarbeet, and Alfalfa. The DACIA5 dataset provides both Sentinel-1 Synthetic Aperture Radar (SAR) imagery and corresponding Sentinel-2 multispectral imagery. In this work, only the Sentinel-2 multispectral patches were used for crop classification.Several original DACIA5 classes are merged: winter wheat and spring wheat into Wheat; corn and corn silage into Corn; late potato and other potato into Potato. Soybean is excluded from the test set.

2.2.1. Input Data

Each patch is a $32 \times 32 \times 12$ Sentinel-2 multispectral image. The dataset covers a region north of Braşov, Romania, with images collected from 2020 to 2024. Training data consists of patches from 2020–2023; the test set consists of patches from 2024, simulating generalisation to unseen future data.

3. System Description

3.1. Subtask 1: Viticulture Potential System

Figure 1 provides an overview of the complete Subtask 1 pipeline, from input data acquisition through to final pixel-level prediction.

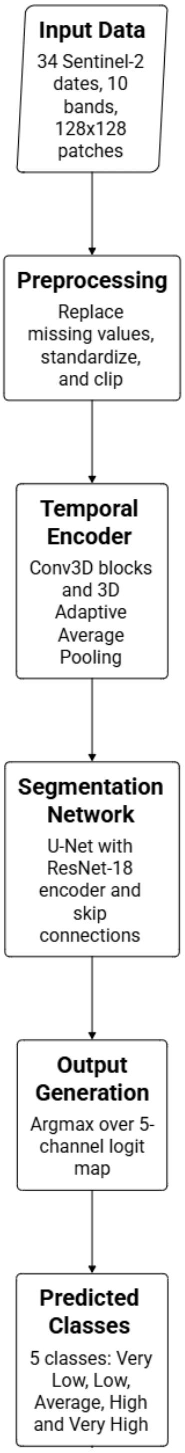

Figure 1: Overview of the Sentinel-2 data processing and segmentation pipeline for Subtask 1.

3.1.1. Data Preprocessing

Sentinel-2 pixels with missing data are represented with the sentinel value $-10{,}000$. The proposed system replaced all such values with 0 before any further processing. Each Sentinel-2 spectral band was then independently standardized using fixed per-band normalization parameters according to:

$$x'_b = \frac{x_b - \mu_b}{\sigma_b + \epsilon} \tag{1}$$

where μ_b and σ_b are the mean and standard deviation for band b, and $\epsilon = 10^{-6}$ prevents division by zero. The same normalization procedure was applied consistently to the training, validation, and test sets before being provided as input to the network. Finally, normalized values were clipped to $[-6, 6]$ to reduce the influence of extreme outliers. This preprocessing places all spectral bands on a comparable scale, improving numerical stability and facilitating efficient model optimization. Without normalization, gradient norms exceeded 40–80 during the initial training iterations, causing numerical instability and NaN loss from the third training epoch onward.

The final model consists of two sequentially connected components: a temporal encoder and a spatial segmentation network.

3.1.2. Temporal Feature Extraction using 3D CNN Encoder

The temporal encoder uses 3D convolutions that apply filters simultaneously across the time axis and both spatial axes, allowing the model to detect temporal patterns while respecting acquisition date ordering. The encoder consists of two 3D convolutional blocks:

- **Block 1:** Conv3D($10 \rightarrow 32$, kernel $3 \times 3 \times 3$, padding $1 \times 1 \times 1$) + BatchNorm3D + ReLU
- **Block 2:** Conv3D($32 \rightarrow 64$, kernel $3 \times 3 \times 3$, padding $1 \times 1 \times 1$) + BatchNorm3D + ReLU

After temporal encoding, a 3D adaptive average pooling layer collapses the time dimension to produce a 2D feature map of shape $(B, 64, 128, 128)$. A 1×1 convolution with batch normalization and ReLU refines the features before passing them to the segmentation network.

3.1.3. Spatial Segmentation using U-Net

The temporal features extracted by the 3D convolutional encoder are then passed to a U-Net [3] segmentation architecture with a ResNet-18 [4] encoder pretrained on ImageNet. The overall architecture is illustrated in Figure 2. U-Net consists of an encoder, decoder and skip connections. ResNet-18 (Encoder) extracts multi-scale spatial features while leveraging ImageNet pretrained weights for improved convergence. The Decoder upsamples feature maps to recover spatial resolution and generate dense predictions. Skip connections transfer low-level spatial information from encoder layers to decoder layers, preserving fine spatial details and improving pixel-level localization.

The final decoder layer produces a 5-channel logit map of shape $(B, 5, 128, 128)$.

The full architecture is summarized in Table 3.

Table 3
Subtask 1 model architecture summary.

Component	Details
Input shape	$(B,\ 34,\ 10,\ 128,\ 128)$
Temporal encoder	$2\times$ 3D Conv (kernel 3^3) + AvgPool3D
Temporal output	$(B,\ 64,\ 128,\ 128)$
Segmentation network	U-Net + ResNet-18 (ImageNet pretrained)
Output shape	$(B,\ 5,\ 128,\ 128)$
Total parameters	≈ 15.4 million

3.1.4. Loss Function

The system used a combined cross-entropy and ordinal MSE (Mean Squared Error) loss. The ordinal loss penalizes predictions that are numerically far from the true class, encouraging smoother class transitions. Let $p_c^{(i)}$ be the softmax probability for class c at pixel i, and let $\mathcal{M}$ be the set of labelled

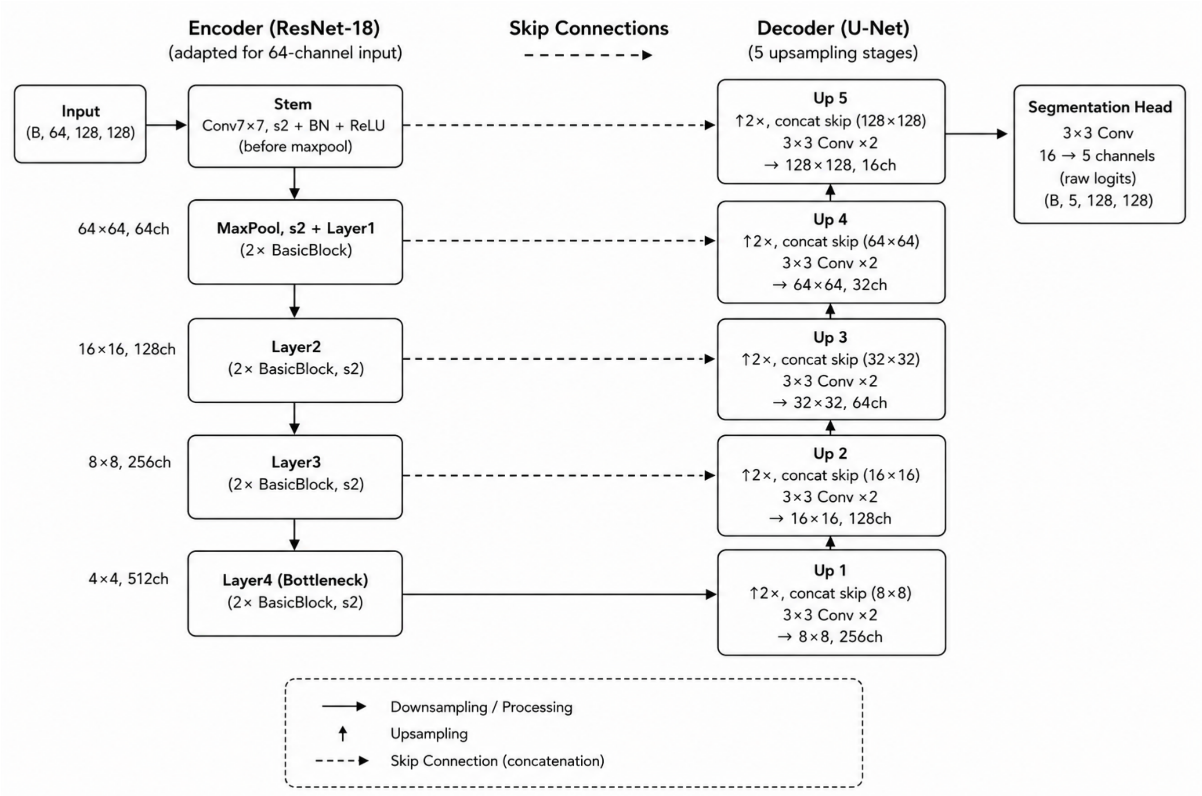

Figure 2: Overview of the proposed U-Net segmentation architecture with a ResNet-18 encoder. The encoder extracts hierarchical spatial features from the fused temporal feature map, while the decoder progressively reconstructs the spatial resolution using skip connections to generate pixel-wise predictions for the five viticulture potential classes.

pixels. The expected predicted class is:

$$\hat{y}_i = \sum_{c=0}^{4} c \cdot p_c^{(i)} \tag{2}$$

The ordinal loss is:

$$\mathcal{L}_{\text{ord}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} (\hat{y}_i - y_i)^2 \tag{3}$$

The combined loss is:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{ord}} \tag{4}$$

where $\lambda = 0.05$. A hard clamp was applied to the total loss (maximum 10.0) as a safety measure.

3.1.5. Training Configuration

The proposed system used AdamW [6] using PyTorch [7] with a learning rate of 10^{-5} and weight decay of 10^{-4}, with a `ReduceLROnPlateau` scheduler (patience=2, factor=0.5). Global gradient norm clipping was applied at 5.0. PyTorch AMP was used for mixed precision training. Batch size was fixed at 2 and training ran for 5 epochs ($\sim$2 h per epoch). The full hyperparameter configuration is given in Table 4.

3.2. Subtask 2: Crop Identification System

Figure 3 illustrates the complete CropCNN architecture for Subtask 2.

Table 4
Subtask 1 training hyperparameters.

Hyperparameter	**Value**
Optimiser	AdamW
Learning rate	1×10^{-5}
Weight decay	1×10^{-4}
Batch size	2
Gradient clip norm	5.0
Ordinal loss weight	0.05
LR scheduler	ReduceLROnPlateau (patience=2, factor=0.5)
Mixed precision	torch.amp
Training epochs	5
Time per epoch	$\sim$2 h (data streaming)

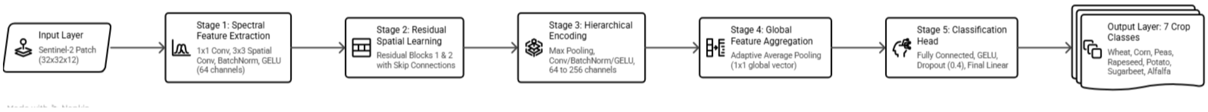

Figure 3: Overview of the CropCNN architecture for Subtask 2 crop identification.

3.2.1. Data Preprocessing

The DACIA5 dataset was accessed from Zenodo [8]. Patches were loaded using `rasterio` and a fixed random seed of 42 was used for all experiments. Each patch was normalized independently per spectral band using min-max normalization to scale values into the range [0,1].Data augmentation was applied during training using random rotations and horizontal/vertical flipping to improve spatial robustness.

3.2.2. Proposed CropCNN Architecture

The proposed CropCNN architecture is designed to classify Sentinel-2 multispectral image patches into seven crop categories: Wheat, Corn, Peas, Rapeseed, Potato, Sugarbeet, and Alfalfa. The architecture is lightweight and designed to handle temporal domain variations between training and test years.

The network takes Sentinel-2 image patches of size $32 \times 32 \times 12$ as input, where the 12 channels correspond to spectral bands.

The architecture consists of five major stages:

- Spectral feature extraction
- Residual spatial learning
- Hierarchical feature encoding
- Global feature aggregation
- Classification

The architecture is summarised in Table 5.

Stage 1: Spectral Feature Extraction
The first stage employs a *SpectralBlock*, which consists of a 1×1 convolution followed by a 3×3 convolution, batch normalization, and GELU activation. The 1×1 convolution combines spectral channels and enables the model to learn relationships between Sentinel-2 bands before spatial processing.

This stage transforms the input from 12 spectral channels to 64 feature channels while preserving the spatial resolution of 32×32.

Stage 2: Residual Spatial Learning
The extracted spectral features are passed through two residual blocks. Each residual block contains two 3×3 convolutional layers with batch normalization and GELU activation. Skip connections preserve low-level information and improve gradient flow during training, allowing the network to learn spatial representations more effectively while reducing vanishing gradient problems.

Stage 3: Hierarchical Feature Encoding
Max-pooling layers progressively reduce the spatial dimensions from 32×32 to 16×16, and then to 8×8. Simultaneously, the number of feature channels increases from 64 to 128 and finally to 256.

Convolutional and residual layers at deeper stages extract higher-level semantic representations related to crop texture, spatial arrangement, and spectral characteristics.

Stage 4: Global Feature Aggregation
A global average pooling layer compresses the final feature map into a compact 1×1 representation. This operation aggregates information across the entire image patch and produces a global feature vector that is spatially invariant.

Stage 5: Classification Head
The aggregated feature vector is passed through a fully connected classification head consisting of:

- A linear layer
- GELU activation
- Dropout regularization
- A final linear classifier

The final layer outputs logits corresponding to the seven crop classes: Wheat, Corn, Peas, Rapeseed, Potato, Sugarbeet, and Alfalfa.

Dropout with probability 0.4 is used to reduce overfitting and improve generalization to unseen 2024 test data.

Table 5
Subtask 2 CropCNN architecture summary.

Component	Details
Input shape	$(B,\ 12,\ 32,\ 32)$
Stem	SpectralBlock $(12 \rightarrow 64)$
Encoder 1	2× ResBlock(64) + MaxPool2D → 16×16
Encoder 2	Conv$(64 \rightarrow 128)$ + 2× ResBlock(128) + MaxPool2D → 8×8
Encoder 3	Conv$(128 \rightarrow 256)$ + ResBlock(256)
Head	GAP → Linear(256,128) → Dropout(0.4) → Linear(128,7)
Output	$(B,\ 7)$ logits

3.2.3. Loss Function and Class Imbalance

A weighted cross-entropy loss was used to handle class imbalance, with per-class weights computed as:

$$w_c = \frac{N}{C \cdot n_c} \tag{5}$$

where N is the total number of training samples, C is the number of classes, and n_c is the count for class c.

3.2.4. Training Configuration

The model was trained using AdamW with a learning rate of 10^{-3} and weight decay of 10^{-4}, with a CosineAnnealingLR scheduler over 50 epochs. The dataset was split into 85% training and 15% validation. Batch size was 32. The full configuration is given in Table 6.

Table 6
Subtask 2 training hyperparameters.

Hyperparameter	**Value**
Optimiser	AdamW
Learning rate	1×10^{-3}
Weight decay	1×10^{-4}
Batch size	32
Training epochs	50
LR scheduler	CosineAnnealingLR
Train/Val split	85% / 15%

4. Experimental Results

4.1. Subtask 1: Viticulture Potential Results

Table 7 shows training and validation metrics per epoch. Validation Accuracy$\pm$1 improved from 0.6803 to 0.7284 through epoch 4. The model was trained for five epochs due to computational and time constraints associated with processing multi-temporal Sentinel-2 imagery. The checkpoint corresponding to Epoch 4 was selected because it achieved the highest validation accuracy among the trained epochs.We note that a longer training schedule may yield further improvements and leave this to future work.

The challenge is evaluated using the Accuracy$\pm$1 metric, which counts a prediction as correct if its absolute difference from the true class is at most 1:

$$\text{Accuracy}_{\pm 1} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left(\left| y_i^{\text{true}} - y_i^{\text{pred}} \right| \leq 1 \right) \tag{6}$$

where N is the total number of labelled pixels across all patches. Pixels with a raw ground truth value of 0 are unlabelled and excluded.

Table 7
Subtask 1 training and validation metrics per epoch.

Epoch	**Train Loss**	**Val Loss**	**Val Acc$\pm$1**	**Notes**
1	1.7267	1.5870	0.6803	Saved
2	1.6093	1.5087	0.6847	Saved
3	1.5787	1.4943	0.7098	Saved
4	1.5568	1.4561	**0.7284**	Saved (best)
5	1.5307	1.4848	0.7034	Not saved

Table 8 shows the predicted class distribution across the 800 test patches. Class 0 (Very Low) is the most frequently predicted at 36.7%, while Class 1 (Low) is markedly underrepresented at 3.1%, reflecting training set class imbalance. In this model no explicit class-balancing strategy such as weighted cross-entropy, focal loss, or class-aware sampling was incorporated, allowing us to evaluate the proposed architecture using a standard loss formulation. Incorporating imbalance-aware loss functions represents an important direction for future work.

The submission achieved a test Accuracy$\pm$1 of **56.15%**.

Table 8
Subtask 1 predicted class distribution on 800 test patches.

Class	Label	% of Pixels
0	Very Low	36.7%
1	Low	3.1%
2	Average	31.7%
3	High	9.8%
4	Very High	18.6%

4.2. Subtask 2: Crop Identification Results

The CropCNN model was evaluated on the 2024 test set. Results are summarised in Table 9.

The task is evaluated using:

$$Q_1 = 0.5 \times AA + 0.5 \times OA \tag{7}$$

where AA is the Average Accuracy (mean per-class accuracy) and OA is the Overall Accuracy, both expressed as percentages.

Table 9
Subtask 2 evaluation results on 2024 test set.

Metric	Value
Overall Accuracy (OA)	37.31%
Average Accuracy (AA)	26.93%
Q1 Score	32.12

The Overall Accuracy achieved by the proposed CropCNN is lower than that reported in the original DACIA5 study. This difference can be attributed to several factors. First, our approach uses only Sentinel-2 multispectral imagery, whereas the DACIA5 dataset also provides Sentinel-1 SAR data, which offers complementary structural information that can improve crop discrimination. Second, the evaluation protocol involves a temporal domain shift, where the model is trained on data from 2020–2023 and evaluated on unseen 2024 imagery, making generalization more challenging. Finally, the proposed lightweight CNN prioritizes computational efficiency over model complexity and therefore has limited capacity to distinguish spectrally similar crop classes. Despite these limitations, the model demonstrates the feasibility of lightweight optical-only crop classification under realistic temporal domain shift conditions.

Figure 4 shows the confusion matrix across the seven crop classes. Crops with similar spectral characteristics such as Wheat and Corn are frequently misclassified. Figure 5 shows the per-class accuracy distribution; Peas achieves the highest accuracy while Potato and Sugarbeet show near-zero accuracy, likely due to their low representation in the training set.

These classes are present in the 2024 test set; therefore, the zero accuracy indicates that none of their samples were correctly classified rather than their absence from the dataset. This behaviour is mainly attributed to class imbalance, temporal domain shift between the training and test years, and spectral similarity with other crop classes.

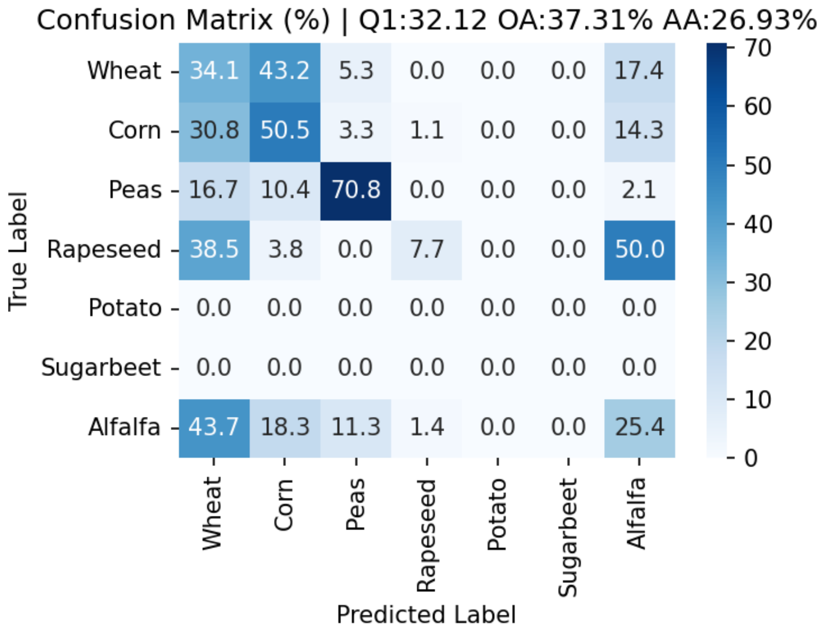

Figure 4: Row-normalized confusion matrix (%) for Subtask 2 crop identification on the 2024 test set.

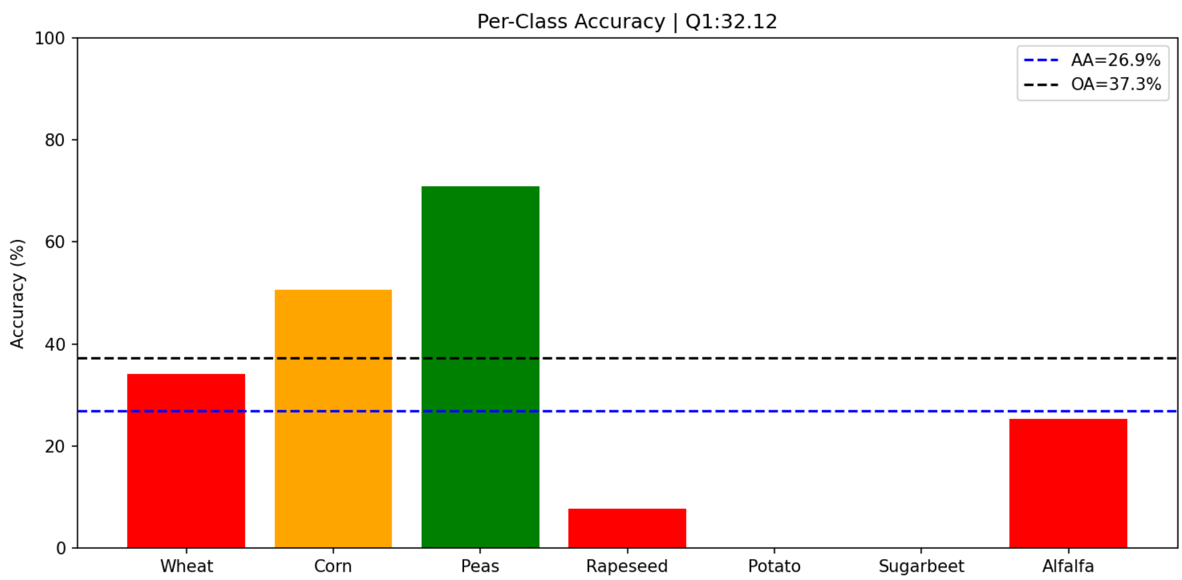

Figure 5: Per-class accuracy distribution for Subtask 2.

5. Discussion

5.1. Subtask 1

Two changes had the largest impact. Per-band normalization was essential. Without it, training was not viable. Raw values were too large for neural network calculations and Normalization transformed each spectral band to approximately zero mean and unit variance. In the first attempt, the system took all 34 satellite images and stacked their 10 bands together, giving 340 input channels. Flattening the time dimension destroyed the order in which the images were taken and the model could not learn seasonal patterns. By switching to 3D convolutional encoder the validation accuracy improved by 4 percentage points by learning seasonally structured features. The combined ordinal loss further reduced large prediction errors because the ordinal loss penalizes predictions that are far from the true class.

The main limitation is the temporal pooling strategy: 3D average pooling weights all 34 acquisition dates equally, whereas a temporal attention mechanism would allow the model to focus on more informative dates during the growing season. This is because growing season images are more informative than winter images for viticulture prediction. The TIF files were streamed directly from Hugging Face rather than stored locally, which increased I/O overhead and limited the feasible training duration to five epochs. If the data were stored locally, we could have trained the model with many epochs. A noticeable reduction is observed when moving from validation set to the test set. This suggests limited generalization beyond the validation distribution. Possible contributing factors include the class imbalance of the training data, the limited training duration, and distributional differences between the validation and hidden test sets. These observations indicate that improved regularization, longer training, and imbalance-aware optimization may further enhance the performance.

5.2. Subtask 2

The proposed CropCNN achieved an Overall Accuracy (OA) of 37.31%, an Average Accuracy (AA) of 26.93%, and a Q1 score of 32.12. Although these results are lower than those reported in the original DACIA5 study published in Big Earth Data, the comparison is not directly equivalent. The original work evaluates more advanced approaches that leverage both Sentinel-1 SAR and Sentinel-2 multispectral imagery, whereas our model uses only Sentinel-2 multispectral data with a lightweight CNN architecture. The proposed model performs well for dominant classes such as Peas (70.8%) and Corn (50.5%) but struggles to generalize to minority classes. In particular, Potato and Sugarbeet were not correctly classified, while Rapeseed achieved only 7.7% accuracy. These results highlight the impact of class imbalance, temporal domain shift between the training and test years, and the spectral similarity among several crop categories. Nevertheless, the proposed approach demonstrates that a compact CNN can learn meaningful spectral-spatial representations from Sentinel-2 imagery alone and provides a useful baseline for future work incorporating multimodal data, improved imbalance handling, and transformer-based architectures.

6. Conclusion

We described the Innovware team's submissions to both subtasks of the ImageCLEF 2026 AI4AGRI task. For Subtask 1, the system combining per-band normalization, a 3D convolutional temporal encoder, and a U-Net with ResNet-18 achieved a validation Accuracy$\pm$1 of 0.7284 and a competition test score of 56.15%. For Subtask 2, a lightweight spectral CNN with residual connections achieved a Q1 score of 32.12 on the 2024 test set. In both cases, class imbalance and limited training resources were the primary bottlenecks, and addressing these remains the most promising direction for future improvement. Our experiments show the usefulness of pretrained convolutional neural networks in handling agricultural image analysis tasks. The study demonstrates the growing importance of AI in supporting sustainable farming practices.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) for writing assistance and grammar checking. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content. All experimental results, methodology, findings, and conclusions are the authors own work.

Acknowledgments

We thank the ImageCLEF 2026 organisers and the AI4AGRI task organisers for providing the datasets and evaluation infrastructure. The AgriPotential dataset is released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Compute resources were provided through Kaggle's free GPU program (Tesla T4, 15 GB).

References

[1] B. Ionescu, H. Müller, D.-C. Stanciu, A. Radzhabov, A. G. Seco de Herrera, A.-G. Andrei, A. Băicoianu, A. Neacşu, A. Storås, A. B. Abacha et al., "ImageCLEF 2026: Multimodal Challenges in Medicine, Science, Agritech, and Security," in *European Conference on Information Retrieval*, pp. 336–344, Springer, 2026.

[2] M. El Sakka, C. De Pourtales, L. Chaari, and J. Mothe, "AgriPotential: A Novel Multi-Spectral and Multi-Temporal Remote Sensing Dataset for Agricultural Potentials," in *2025 International Conference on Content-Based Multimedia Indexing (CBMI)*, pp. 1–6, IEEE, 2025.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. MICCAI*, pp. 234–241, Springer, 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. CVPR*, pp. 770–778, 2016.

[5] M. Drusch et al., "Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services," *Remote Sensing of Environment*, vol. 120, pp. 25–36, 2012.

[6] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Proc. ICLR*, 2019.

[7] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[8] M. Ivanovici et al., "AI4AGRI Sentinel-2 Brasov area 2020-2024 multi-spectral dataset for crop monitoring and identification," Zenodo, Dec. 2024. https://doi.org/10.5281/zenodo.14283243