Overview of ImageCLEF2026 AI4AGRI AgriPotential Challenge: Viticulture Edition

AI4AGRI AgriPotential at CLEF 2026

Mohammad El Sakka[1,*], Lotfi Chaari[2] and Josiane Mothe[3]

[1]Université de Toulouse, IRIT UMR5505 CNRS, Toulouse, France
[2]Toulouse INP, IRIT UMR5505 CNRS, Toulouse, France
[3]Université de Toulouse, UT2J, INSPE, IRIT UMR5505 CNRS, Toulouse, France

Abstract

We present the AgriPotential: Viticulture Edition challenge, organized as part of the ImageCLEF 2026 AI4Agri track. The task requires pixel-level prediction of viticultural potential from Sentinel-2 multispectral satellite image time series (MS SITS) over the Hérault department in southern France, framed as an ordinal classification problem over five classes. Ground truth labels are expert-derived ordinal annotations produced under the GDPA program and paired with 34 acquisitions across 10 spectral bands super-resolved to 5 m. Evaluation is based on the Accuracy±1 metric, which accounts for the ordinal structure of the label space. The challenge attracted 18 registered participants, of whom 6 actively submitted 184 runs. The best participants' submission achieved 69.08% Accuracy±1, against our benchmark of 85.87% obtained by a CSD-3D UNet, confirming substantial remaining headroom. Analysis of the three submitted working notes reveals convergence on U-Net-based architectures and ordinal-aware loss functions, while temporal modeling strategies and compute constraints emerged as key differentiating factors.

Keywords

ordinal classification, satellite image time series, remote sensing, agricultural potential, viticulture

1. Introduction

Remote sensing has emerged as a powerful tool for large-scale agricultural monitoring, enabling the analysis of land surface dynamics at spatial and temporal resolutions that were previously inaccessible. Multispectral satellite image time series (MS SITS), in particular from the Sentinel-2 constellation, have been widely applied to crop type mapping, yield estimation, and phenological monitoring [1, 2, 3, 4, 5, 6]. However, the dominant paradigm in this literature remains reactive: models are trained to characterize land that is already under cultivation, rather than to prospectively assess its intrinsic suitability for a given crop. This distinction matters in practice because of the rapid change of land potential due to climate change.

We organized the AgriPotential: Viticulture Edition challenge to benchmark the community's ability to address this prospective assessment problem. The task requires predicting the viticultural potential of land parcels from Sentinel-2 MS SITS, where agricultural potential is represented as an ordinal variable derived from expert agronomic assessments over the Hérault department in southern France [7]. The ordinal structure of the label space is a defining characteristic of the task: rank errors must be penalized proportionally, and evaluation metrics must reflect this. The challenge was hosted on CodaBench as part of the ImageCLEF 2026 AI4Agri track [8], and used a subset of the AgriPotential dataset [9], the first benchmark pairing Sentinel-2 time series with expert-derived agricultural potential labels. It attracted 18 registered participants, of whom 6 actively submitted a total of 184 runs.

The remainder of this paper is organized as follows. Section 2 gives the necessary background about the concept of agricultural potential and its assessment methods. Section 3 describes the dataset.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
mohammad.el-sakka@irit.fr (M. E. Sakka)
0009-0007-0531-7946 (M. E. Sakka)

Section 4 details the challenge setup and evaluation protocol. Section 5 reports participation statistics, and Section 6 presents the results and post-competition survey findings. Section 7 concludes with lessons learned and directions for future editions.

2. Background

Agricultural potential refers to the capacity of a land unit to support agricultural production under given environmental, technological, and socio-economic conditions, governed by the interaction of soil properties, climatic variables, and human management practices [10]. Crucially, agricultural potential is not an intrinsic or static property of the land itself: it emerges from the match between land characteristics and crop requirements and must be evaluated within specific agronomic and environmental contexts [11, 12]. It is also dynamic, evolving in response to climate variability [13, 14] and changes in technology or management [15, 16], which makes continuous monitoring essential for sustainable land use planning.

Traditional assessment relies on field surveys and threshold-based soil indicators [17]. Advances in geographic information systems (GIS) and remote sensing have since enabled spatially explicit assessments over large areas, integrating soil, climate, topography, and land-use data through multi-criteria evaluation [18, 19]. However, these approaches remain largely descriptive and diagnostic: they characterize land that is already under use, rather than prospectively predicting suitability for a target crop.

The Hérault department in southern France illustrates the need for prospective assessment. Its vineyards define the landscape and local economy, yet rising temperatures, shifting precipitation, and increasing drought frequency are actively reshaping parcel-level viticultural suitability [7]. To address this problem using remote sensing, we developed the AgriPotential dataset [9], pairing Sentinel-2 MS SITS acquired over the Hérault department with expert-derived ordinal labels produced under the GDPA program [1]. The dataset provides pixel-level annotations across multiple crop types; this challenge focuses on the viticulture label, encoding the intrinsic capacity of each parcel to support vine cultivation.

3. Dataset

The challenge dataset is a subset of the AgriPotential dataset [9], covering diverse agricultural landscapes in the Hérault department of southern France. It pairs Sentinel-2 multispectral satellite imagery with expert-derived ordinal labels from the GDPA program introduced in Section 2.

3.1. Satellite Imagery

The imagery consists of 34 Sentinel-2 Level-2A images acquired over tile T31TEJ between 2017 and 2019, each comprising 10 spectral bands (B2–B8A, B11, B12). We selected the images for low cloud coverage (generally below 2%) and approximate monthly regularity, though the resulting temporal sampling is irregular due to cloud contamination and revisit constraints. All bands were super-resolved to 5 m spatial resolution using the CARN architecture [20] pretrained on the Sen2Venμs dataset, yielding a uniform spatial grid of $18\,770 \times 15\,286$ pixels in the EPSG:32631 coordinate system. This pre-processing is further detailed in our paper [9].

3.2. Ground Truth Labels

Ground truth annotations are pixel-level ordinal labels representing viticultural potential across five ordered classes: very low, low, average, high, and very high. These labels were produced under the GDPA program from the BD SOL Hérault soil database, in which experts evaluated each soil unit based on field

[1] https://data.laregion.fr/explore/dataset/bd-sol-gdpa-herault%40data-herault-occitanie/information/

observations of pH, texture, depth, and available water capacity, combined with agronomic knowledge. Labels reflect the intrinsic capacity of each land parcel to support vine cultivation, independently of current land use [2]. Only GDPA polygons with sufficient confidence indices were retained. The viticulture label exhibits the most balanced class distribution of the three crop types covered by the dataset, with class proportions ranging from 16% to 24% in the training set.

3.3. Data Partitioning and Access

The full raster was divided into non-overlapping 256×256 pixel blocks, and only blocks containing at least one labeled pixel were retained. Blocks were then randomly split following an 80/10/10 ratio into training, validation, and test subsets. Each block was subdivided into overlapping 128×128 patches using a 50% stride, retaining only patches where at least 5% of pixels carry a label. Table 1 summarizes the resulting partitioning.

Split	Patches	Total Pixels	Labeled Pixels (%)
Train	6 329	103.6 M	65.1 M (62%)
Validation	781	12.7 M	7.4 M (58%)
Test	800	13.1 M	7.3 M (56%)
Overall	7 910	129.5 M	80.0 M (61%)

Table 1
Dataset partitioning into training, validation, and test subsets. The block-level split prevents spatial data leakage between subsets.

Not all pixels within a retained patch are labeled: unlabeled pixels arise from areas not covered by GDPA polygons or from polygons discarded during confidence filtering. On average, 61% of pixels carry a label across the full dataset. True labels for the test set were withheld from participants and retained by the organizers for evaluation.

Participants accessed the data through a dedicated Google Colab tutorial notebook providing PyTorch `Dataset` and `DataLoader` objects, though custom data loading pipelines were permitted. The full dataset is publicly available on HuggingFace [3]. Input tensors are four-dimensional, with axes corresponding to time, spectral bands, height, and width, capturing the full spatio-temporal structure of the MS SITS. The task presents two modeling challenges: (1) *spectral-spatio-temporal modeling*, requiring the joint exploitation of spatial and spectral context with temporal dynamics across irregularly sampled acquisitions; and (2) *ordinal classification*, ensuring that larger errors between predicted and true ranks are penalized more severely than smaller ones.

4. Challenge Design

The AgriPotential: Viticulture Edition challenge was organized as Subtask 1 of the ImageCLEF 2026 AI4Agri track [8] and hosted on the CodaBench platform [21] [4]. The AI4Agri track comprises two complementary subtasks: our AgriPotential challenge, which targets prospective land suitability assessment from MS SITS, and DACIA5, which addresses in-season crop identification using multimodal Sentinel-1/Sentinel-2 imagery from Romania [8]. Participants could register for one of the two sub-tasks or for both. This section describes the task formulation, evaluation metric, and submission protocol for Subtask 1.

4.1. Task Formulation

Participants were required to produce pixel-level predictions of viticultural potential for each patch in the test subset, submitted as a ZIP archive of PNG files named after their corresponding `patch_id`.

[2] https://data.laregion.fr/explore/dataset/bd-sol-gdpa-herault%40data-herault-occitanie/information/
[3] https://huggingface.co/datasets/m-sakka/agripotential
[4] https://www.codabench.org/competitions/12055/

Predicted values must be integers in the range $[0, 4]$, mapping to the five ordinal classes from very low to very high respectively. Participants were also encouraged to include a `report.pdf` describing their methodology.

4.2. Evaluation Metric

Evaluation was conducted using the Accuracy± 1 metric, which measures the proportion of predictions falling within one rank of the true label:

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

where N is the total number of pixels across all test patches. This metric accounts for the ordinal structure of the label space by treating near-miss predictions as partially correct, while still penalizing large rank errors. The final score reported on the leaderboard (Table 2) is the Overall Accuracy± 1 computed across all pixels in the test set.

4.3. Submission Protocol and Scoring

The challenge ran under a single competitive phase, with submissions and leaderboard management handled entirely through CodaBench. Scoring was automated: the evaluation script ingests each submitted ZIP archive, extracts predicted PNG files, and aligns each file to its corresponding ground truth patch via the `patch_id` filename. Predicted pixel values are read as integers in $[0, 4]$ and compared against withheld test annotations using Equation 1. The score is aggregated over all pixels across all 800 test patches and reported on the leaderboard in real time. Submissions containing fewer than the expected 800 patches were accepted but penalized through a lower overall score, as the metric is computed over all pixels.

Submissions failing to meet format requirements (incorrect file naming, values outside the valid range, or the presence of subdirectories within the ZIP archive) were partially or completely rejected by the scoring system. Extra files, with the exception of the optional `report.pdf`, were silently ignored. Test set labels were retained server-side throughout the competition, ensuring evaluation integrity. All participating teams were invited to submit working notes papers for publication in the CLEF 2026 proceedings.

5. Participation

The challenge attracted 18 registered participants and groups, of whom 6 actively submitted predictions, producing a total of 184 runs across the competition phase. The disparity between registrations and active submissions is consistent with trends observed in other ImageCLEF tasks [8] and likely reflects the technical barrier posed by the spatio-temporal nature of the task, which requires complex data loading and modeling pipelines beyond standard image classification frameworks.

Among the 6 active participants, submission activity was unevenly distributed, with some teams submitting iteratively throughout the phase and others submitting only a small number of runs. At the close of the competition, 3 working notes papers were submitted, providing methodological descriptions of the participating systems.

6. Results

Table 2 reports the final leaderboard standings based on each team's best submission.

The top four participants achieved scores between 65% and 70%, indicating broad methodological convergence around similar performance levels. The two lowest-scoring submissions fell notably below this cluster. As a reference, we benchmarked 12 combinations of four architectures (2D UNet, 3D

Participant / Team	Submission ID	Accuracy±1 (%)
ours [22]	-	85.87
snspace	700105	69.08
DS@GT AI4Agri	712441	68.32
ronny_toro	699815	65.66
titouann	516652	65.38
Innovware: janani_sp	713987	56.15
vb7132	713812	51.28

Table 2
Final leaderboard of the AgriPotential: Viticulture Edition challenge, reporting each participant's best Accuracy±1 score on the test set.

UNet, ViViT, UTAE) and three prediction strategies (categorical, K-rank ordinal, continuous space discretization) on the same dataset [22]. The best-performing configuration, CSD-3D UNet -which treats agricultural potential as a continuous variable and applies joint spatio-temporal modeling via 3D convolutions - achieved 85.87% Accuracy±1 on the viticulture test set, establishing a meaningful performance ceiling and confirming substantial remaining headroom above the best participant result.

6.1. Participant Approaches

Three working notes papers were submitted describing systems for Subtask 1. Table 3 provides a comparative overview of the architectures, and Table 4 summarises the training configurations. We describe each approach below.

DS@GT AI4Agri (68.32%). This team proposed a weighted logit ensemble of two independently trained models: a U-Net with residual connections and a fine-tuned Prithvi-EO-2.0-100M [23] geospatial foundation model. The U-Net treats all 34 timesteps as additional input channels by stacking $34 \times 15 = 510$ channels (10 spectral bands plus 5 derived spectral indices: NDVI, NDMI, NDWI, NDRE, and NBR). The Prithvi branch aggregates the 34 timesteps into four seasonal means over the 6 bands used during pretraining, incorporating temporal and location coordinate encodings. Both models share the same ordinal prediction head: a single-channel output followed by $K-1 = 4$ learnable thresholds computed via cumulative softplus, trained with an ordinal binary cross-entropy loss. Ensemble weights ($w_{\text{U-Net}} = 0.65$, $w_{\text{Prithvi}} = 0.35$) were calibrated on the validation set. Additionally, the team employed a semi-supervised teacher-student strategy in which the trained U-Net generated pseudo-labels for unlabeled pixels with confidence above 0.7, incorporated into the Prithvi training loss with weight $\lambda = 0.3$. The team also explored TSViT, U-TAE, Swin-V2, and Presto as alternative temporal backbones, but none outperformed the stacked-channel U-Net on the test set.

ronny_toro (65.66%). This submission used a standard U-Net architecture trained from scratch, with encoder channel widths of $64 \to 128 \to 256$ and a 512-channel bottleneck (31.6M parameters). Temporal information was handled by stacking all $34 \times 10 = 340$ channels along the input dimension. The key contribution is the loss function: a composite of standard cross-entropy and a squared Earth Mover's Distance (EMD) penalty ($\lambda = 0.5$), which penalises large ordinal errors quadratically by comparing cumulative distribution functions of the predicted softmax and the one-hot target. Training used the AdamW optimiser with a OneCycleLR schedule over 30 epochs and completed in approximately 90 minutes on a single NVIDIA RTX 4090.

Innovware: janani_sp (56.15%). This team proposed a two-stage architecture combining a 3D convolutional temporal encoder with a U-Net segmentation network using a ResNet-18 backbone pretrained on ImageNet. Unlike the two preceding approaches, this system preserves the temporal ordering of the 34 acquisitions: two 3D convolutional blocks ($10 \to 32 \to 64$ channels with $3 \times 3 \times 3$ kernels) operate jointly across time and space, followed by 3D adaptive average pooling that collapses

the time dimension into a 2D feature map of shape $(B, 64, 128, 128)$. This feature map is then passed to the U-Net decoder. The loss combines cross-entropy with an ordinal MSE term ($\lambda = 0.05$) computed from the expected predicted class. Training was severely constrained by compute limitations: the 200GB dataset was streamed from HuggingFace via GDAL VSI on a Kaggle T4 GPU with only 20GB of local disk, limiting training to 5 epochs at approximately 2 hours per epoch. Despite these constraints, the team reported that switching from channel-stacking to 3D convolutions improved validation accuracy by 4 percentage points and that per-band z-score normalisation (with clipping to $[-6, 6]$ and zero-imputation of missing values) was essential for training stability.

Team	Architecture	Pretrained weights	Temporal strategy	Bands
DS@GT AI4Agri	U-Net + Prithvi-EO-2.0 ensemble (logit avg.)	Prithvi-EO-2.0-100M (geospatial FM)	Stack (U-Net); seasonal agg. (Prithvi)	10 + 5 indices (U-Net); 6 (Prithvi)
ronny_toro	U-Net (from scratch)	None	Stack (34×10)	10
Innovware	3D CNN + U-Net	ResNet-18 (ImageNet)	3D convolutions	10

Table 3
Architectural comparison of submitted systems for Subtask 1.

Team	Loss	Normalisation	Augmentation	Epochs	GPU	Val Acc± 1
DS@GT AI4Agri	Ordinal BCE ($K-1$ thresholds)	$\div 10{,}000$ (U-Net); z-score (Prithvi)	Flips, rotations	50 / 40	PACE cluster	80.26%
ronny_toro	CE + EMD^2 ($\lambda = 0.5$)	—	Flips, rotations	30	RTX 4090	78.57%
Innovware	CE + ordinal MSE ($\lambda = 0.05$)	z-score, clip $[-6, 6]$, zero-impute	None	5	Kaggle T4	72.84%

Table 4
Training configuration comparison. Val Acc± 1 is the best validation Accuracy± 1 reported by each team.

6.2. Cross-Cutting Observations

Several methodological patterns emerge from the submitted approaches.

Ordinal modelling was universal. All three teams explicitly modelled the ordinal structure of the labels, though through different mechanisms: ordinal binary cross-entropy with learnable thresholds (DS@GT), cross-entropy with a squared EMD penalty (ronny_toro), and cross-entropy with an ordinal MSE term (Innovware). This confirms the inadequacy of standard categorical cross-entropy for this task and suggests that ordinal-aware losses are a prerequisite for competitive performance.

Temporal modelling yielded mixed results. Two of the three teams collapsed the temporal dimension into channels (stacking), while one used 3D convolutions to preserve temporal ordering. Interestingly, DS@GT reported that explicit temporal modelling via architectures such as TSViT, U-TAE, and Presto did not outperform the simpler channel-stacking strategy on the test set, and hypothesised that stacking introduces a useful inductive bias for this fixed-period (2017–2019) task. In contrast, Innovware reported a 4-point validation improvement when switching from stacking to 3D convolutions, though their test score was substantially lower, likely due to severe compute constraints limiting training to only 5 epochs.

Pretrained weights helped but were not decisive. DS@GT's ensemble used Prithvi-EO-2.0 pretrained on 4.2 billion global time series samples, and Innovware used ImageNet-pretrained ResNet-18, while ronny_toro trained entirely from scratch and achieved competitive results (65.66%). This suggests that pretraining provides useful initialisation but does not dominate performance when the downstream task is sufficiently distinct from the pretraining distribution.

Compute constraints were a significant bottleneck. The gap between Innovware's validation score (72.84%) and their test score (56.15%) was the largest among the three teams and can be attributed in part to training for only 5 epochs due to data streaming latency on Kaggle. More broadly, the 200GB dataset size posed practical challenges for all participants: DS@GT required access to a university HPC cluster, ronny_toro rented cloud GPUs, and Innovware was forced to stream data in real time. These constraints likely depressed overall participation rates and limited hyperparameter exploration even for active teams.

Generalization gap. All three teams reported a gap between validation and test performance. DS@GT observed a drop from 80.26% to 68.32%, ronny_toro from 78.57% to 65.66%, and Innovware from 72.84% to 56.15%. The source of this gap remains an open question: DS@GT hypothesised a train–test distribution shift (e.g., different terrain types or class distributions), while Innovware attributed it partly to insufficient training. Diagnosing this gap would require releasing per-class or per-region test set statistics, which we plan for future editions.

Comparison with organizers' benchmark (ours [22]). Our systematic benchmark offers additional context for interpreting participant results. The best participant approaches used ordinal binary cross-entropy (DS@GT) or hybrid CE+EMD losses (ronny_toro), which correspond to ranking-based strategies in the benchmark taxonomy. However, the benchmark showed that continuous space discretization (CSD)—treating potential as a regression target before discretizing—consistently outperforms ranking-based and categorical approaches across all architectures. No participant explored this strategy. Similarly, the benchmark found that 3D convolutions with joint spatio-temporal modeling outperform channel-stacking, yet participants generally favoured channel-stacking due to its simplicity and competitive validation scores. These gaps between participant choices and the benchmark's findings suggest that future editions could benefit from providing more detailed baseline information to guide participants toward promising directions.

6.3. Post-Competition Survey

Following the competition, we conducted a voluntary post-competition survey to collect additional methodological details and participant feedback. These are not necessarily the same teams as the three working notes analyzed in Section 6.1. We present the three responses that we received below.

Respondent profile. All three respondents identified as academic researchers with no prior experience with remote sensing data. This profile suggests the challenge successfully attracted participants from the broader machine learning community rather than exclusively remote sensing specialists and our task is approachable to newcomers to the domain.

Model design. On primary architecture, two participants reported hybrid temporal models (in the UTAE / CNN-LSTM family) and one a pure CNN (ResNet/U-Net/3D-CNN). All three used pretrained weights: two initialized from the Prithvi-EO geospatial foundation model [23] and one from an ImageNet-pretrained ResNet-18 encoder [24]. Every participant answered that they explicitly modeled the ordinal nature of the labels through ordinal binary cross-entropy with four thresholds for the five classes in two cases, and through a cross-entropy term combined with a custom ordinal MSE computed from the expected (weighted-sum) predicted class in the third. All three also reported using temporal modeling, via temporal attention (two teams) or 3D convolutions (one team).

Data handling. No team used external data, despite it being permitted by the challenge rules, which could indicate many possibilities. For example, participants may have found the provided MS SITS sufficient, judged the integration overhead not worth it, or lacked the experience, time, or resources to incorporate additional sources. Spectral preprocessing converged on normalization:

dividing reflectances by 10 000 followed by z-score standardization, with one team additionally clipping per-band standardized values to [-6,6] and replacing missing values with 0. Two teams applied data augmentation (image flips and rotations). None of the three teams reported addressing class imbalance, which confirms our observation (Section 3) that the viticulture label has the most balanced class distribution of the AgriPotential dataset, ranging from 16% to 24% per class.

Training and validation strategy. The free-text answers here exposed two contrasting regimes. One team described a primarily leaderboard-driven strategy, using validation accuracy only as a secondary signal. Another was constrained to only five training epochs (roughly two hours each) because the 200GB dataset had to be streamed live from HuggingFace under a 20GB local-disk limit on Kaggle.

Class difficulty. When we asked which classes their models predicted most easily, responses diverged: one team cited the three middle classes (low, average, high), while the other two cited the extremes, with *very low* named twice.

Feedback and open science. All three teams indicated willingness to release their code under an MIT license for reproducibility. On the biggest challenge faced, every team pointed to the same two issues already visible in their results: the generalization gap between validation and test performance, and compute/data constraints tied to the dataset's size.

7. Conclusion

This paper presented the AgriPotential: Viticulture Edition challenge, the first benchmark to address prospective agricultural potential assessment from Sentinel-2 time series. The task attracted diverse participation, with the best system reaching 69.08% Accuracy± 1 against the organizers' benchmark of 85.87% achieved by a CSD-3D UNet, confirming substantial remaining headroom.

Analysis of the three submitted working notes reveals several methodological convergences: all teams adopted U-Net-based segmentation architectures and explicitly modeled the ordinal label structure through different loss formulations, including ordinal binary cross-entropy with learnable thresholds, cross-entropy combined with a squared EMD penalty, and cross-entropy with an ordinal MSE term. The role of temporal modeling proved more nuanced: the best individual model on the test set used channel stacking rather than explicit temporal encoders, though 3D convolutions showed promise when training was not compute-limited. Pretrained weights from geospatial foundation models or ImageNet provided useful initialization but were not decisive, as a U-Net trained from scratch achieved competitive results.

A consistent generalization gap between validation and test performance was observed across all teams, with drops ranging from 12 to 17 percentage points. The 200GB dataset size also emerged as a practical barrier, constraining both participation rates and the depth of hyperparameter exploration for active teams. From an organizational standpoint, managing test set concealment while maintaining a clear submission protocol proved non-trivial.

For future editions, we plan to release per-class and per-region test set statistics to support generalization gap diagnosis, provide more granular baseline information, and make submission limits explicit. We also aim to expand the range of crop types beyond viticulture and investigate semi-supervised and unsupervised learning to leverage unlabeled time series. Through this challenge, we aim to stimulate progress on a high-impact, underexplored task linking advances in machine learning with sustainable land use decision-making.

Acknowledgments

We would like to thank the CLEF 2026 organizers for their support throughout the challenge organization, and in particular Cristian Stanciu for his availability and responsiveness to our questions.

This work was partially funded by the European Union through the AI4AGRI project entitled "Romanian Excellence Center on Artificial Intelligence on Earth Observation Data for Agriculture", which received funding from the European Union's Horizon Europe research and innovation programme under the grant agreement no. 101079136. The Défi Région Occitanie "Observation de la Terre et Territoire en Transition" also supported this work.

Our own results on the task were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) for grammar and spelling checks and LaTeX formatting assistance. The authors reviewed and edited the generated content when needed; they take full responsibility for the publication's content.

References

[1] M. Rußwurm, M. Körner, Multi-temporal land cover classification with sequential recurrent encoders, ISPRS International Journal of Geo-Information 7 (2018) 129.

[2] V. S. F. Garnot, L. Landrieu, Panoptic segmentation of satellite image time series with convolutional temporal attention networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4872–4881.

[3] X. Jin, L. Kumar, Z. Li, H. Feng, X. Xu, G. Yang, J. Wang, A review of data assimilation of remote sensing and crop models, European journal of agronomy 92 (2018) 141–152.

[4] M. El Sakka, J. Mothe, M. Ivanovici, Images and cnn applications in smart agriculture, European Journal of Remote Sensing 57 (2024) 2352386.

[5] M. Ivanovici, G. Olteanu, C. Florea, R.-M. Coliban, M. Ștefan, K. Marandskiy, Digital transformation in agriculture, in: Digital Transformation: Exploring the Impact of Digital Transformation on Organizational Processes, Springer, 2024, pp. 157–191.

[6] M. El Sakka, M. Ivanovici, L. Chaari, J. Mothe, A review of cnn applications in smart agriculture using multimodal data, Sensors 25 (2025) 472.

[7] G. V. Jones, M. A. White, O. R. Cooper, K. Storchmann, Climate change and global wine quality, Climatic change 73 (2005) 319–343.

[8] B. Ionescu, H. Müller, D.-C. Stanciu, A. Radzhabov, A. G. S. 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, Springer, 2026, pp. 336–344. doi:doi.org/10.1007/978-3-032-21321-1_44.

[9] M. El Sakka, C. De Pourtales, L. Chaari, 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), IEEE, 2025, pp. 1–6.

[10] E. N. Ferdon Jr, Agricultural potential and the development of cultures, Southwestern Journal of Anthropology 15 (1959) 1–19.

[11] T. Snelder, L. Lilburne, D. Booker, A. Whitehead, S. Harris, S. Larned, A. Semadeni-Davies, D. Plew, R. McDowell, Land-use suitability is not an intrinsic property of a land parcel, Environmental Management 71 (2023) 981–997.

[12] R. McDowell, T. Snelder, S. Harris, L. Lilburne, S. Larned, M. Scarsbrook, A. Curtis, B. Holgate, J. Phillips, K. Taylor, The land use suitability concept: Introduction and an application of the concept to inform sustainable productivity within environmental constraints, Ecological Indicators 91 (2018) 212–219.

[13] A. De la Casa, G. Ovando, Climate change and its impact on agricultural potential in the central region of argentina between 1941 and 2010, Agricultural and Forest Meteorology 195 (2014) 1–11.

[14] V. Alexandrov, J. Eitzinger, V. Cajic, M. Oberforster, Potential impact of climate change on selected agricultural crops in north-eastern austria, Global change biology 8 (2002) 372–389.
[15] S. H. Wittwer, Maximizing agricultural production, Research Management 13 (1970) 89–110.
[16] P. R. Jennings, The amplification of agricultural production, Scientific American 235 (1976) 180–195.
[17] C. K. Ajaero, L. C. Mba, C. Okeke, An analysis of the factors of agricultural potentials and sustainability in adani, nigeria, Greener Journal of social sciences 3 (2013) 286–291.
[18] A. Ceballos-Silva, J. López-Blanco, Evaluating biophysical variables to identify suitable areas for oat in central mexico: a multi-criteria and gis approach, Agriculture, ecosystems & environment 95 (2003) 371–377.
[19] S. Bandyopadhyay, R. Jaiswal, V. Hegde, V. Jayaraman, Assessment of land suitability potentials for agriculture using a remote sensing and gis based approach, International journal of remote sensing 30 (2009) 879–895.
[20] Y. Li, E. Agustsson, S. Gu, R. Timofte, L. Van Gool, Carn: Convolutional anchored regression network for fast and accurate single image super-resolution, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 0–0.
[21] Z. Xu, S. Escalera, A. Pavão, M. Richard, W.-W. Tu, Q. Yao, H. Zhao, I. Guyon, Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform, Patterns 3 (2022).
[22] M. E. Sakka, L. Chaari, J. Mothe, Benchmarking Deep Learning for Agricultural Potential Estimation from Multispectral Satellite Time Series, 2026. URL: https://hal.science/hal-05669822, accepted at ECML PKDD 2026, awaiting publication.
[23] D. Szwarcman, S. Roy, P. Fraccaro, O. E. Gíslason, B. Blumenstiel, R. Ghosal, P. H. De Oliveira, J. L. de Sousa Almeida, R. Sedona, Y. Kang, et al., Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications, IEEE Transactions on Geoscience and Remote Sensing (2025).
[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.