

Overview of PlantCLEF 2026: Identify Multi-species Plants in Images of Vegetation Plots

LifeCLEF at CLEF 2026

Giulio Martellucci¹, Ilyass Moummad², Hervé Goëau^{3,2}, Pierre Bonnet³, Fabrice Vinatier¹ and Alexis Joly^{3*,†}

¹LISAH, Univ Montpellier, INRAE, IRD, Montpellier, France

²Inria, LIRMM, Univ Montpellier, CNRS, Montpellier, France

³AMAP, Univ Montpellier, CIRAD, CNRS, INRAE, IRD, Montpellier, France

Abstract

Vegetation plot images are essential for ecological studies, as they allow standardized sampling, assessment of plant biodiversity, long-term monitoring, and large-scale field campaigns. These images typically cover an area of fifty centimeters or one square meter, where botanists carefully identify all the species present. Integrating AI could help specialists accelerate their inventories and expand the spatial coverage of ecological studies.

In order to evaluate advances in this field, the PlantCLEF 2026 challenge relies on a test set of 2,105 high-resolution, top-down vegetation plot images from South-Western Europe. These images are annotated by experts and contain around 400 species. In contrast, the challenge provides a large training set of 1.4 million single-label images sourced from citizen science platforms. This training data covers the entire regional flora of 7,800 species, which requires the plant identification model to learn all classes without knowing in advance which species will be included in the test set.

This setup implies a significant domain shift. The opportunistic training images focus on centered, individual plants and often feature blooming flowers. On the other hand, the expert-annotated test images are complex multi-species scenes captured from above, displaying different phenological stages where flowers are much less prominent. The task is thus formulated as a highly challenging multi-label classification problem under strong domain shift. This paper provides a detailed description of the data, the evaluation methodology, the methods and models used by the participants, and the results achieved.

Now in its third edition, the challenge continues to attract a growing number of participants despite its difficulty. Furthermore, this year's results reveal promising new methodological advances, showing that the community has progressed well beyond the stage of simply identifying species using pre-trained models applied to crop-out tiles.

Keywords

LifeCLEF, fine-grained classification, species identification, vegetation plots, vegetation quadrat images, multi-label classification, biodiversity informatics, evaluation, benchmark

1. Introduction

Vegetation plot inventories are essential for ecology. They allow scientists to assess biodiversity, sample plant life, and monitor environments over time. This data helps us understand ecosystems and make better conservation decisions.

In ecology, a "vegetation plot" is a generic term for an area of variable size and shape where botanists inventory all the plant species present. In open environments, such as meadows, agricultural fields, alpine pastures, or sand dunes, these surveys are often organized along "transects" (straight sampling lines). Along these transects, botanists place small, square frames called "quadrats", typically 0.5×0.5

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

†These authors contributed equally.

✉ giulio.martellucci@inrae.fr (G. Martellucci); ilyass.moummad@inria.fr (I. Moummad); herve.goeau@cirad.fr (H. Goëau); pierre.bonnet@cirad.fr (P. Bonnet); fabrice.vinatier@inrae.fr (F. Vinatier); alexis.joly@inria.fr (A. Joly)

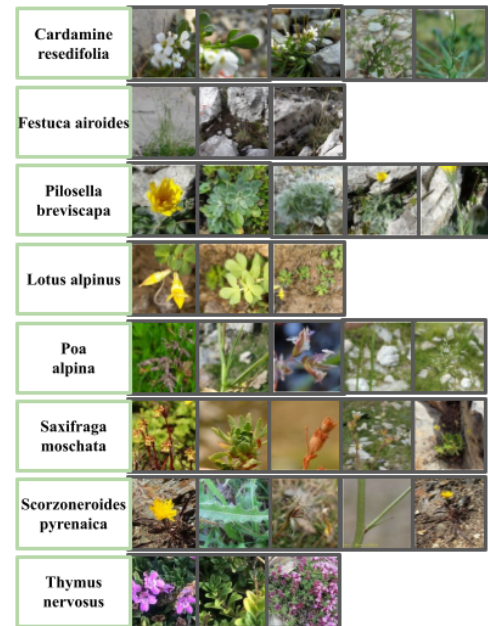
ORCID 0009-0008-1804-0920 (G. Martellucci); 0009-0003-2925-2500 (I. Moummad); 0000-0003-3296-3795 (H. Goëau); 0000-0002-2828-4389 (P. Bonnet); 0000-0003-3693-4422 (F. Vinatier); 0000-0002-2161-9940 (A. Joly)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



(a) Vegetative quadrat (test)



(b) Individual plant images (training)

Figure 1: Illustration of the visual gap between test and training data. (a) The test set consists of vertical top views of quadrats containing multiple species. (b) The training set contains images of individual plants, mostly focusing on specific organs (flowers, fruits, leaves) taken from various angles by the Pl@ntNet community.

meters in size, to focus mostly on herbaceous plants. Inside each quadrat, they carefully identify every species and measure its abundance using metrics like biomass or the percentage of the ground it covers.

Artificial Intelligence (AI) could make this process much faster and allow for more frequent ecological surveys. It would also help non-experts and citizen scientists contribute to these monitoring programs. Current apps like Pl@ntNet or iNaturalist can identify plants when photographed individually. However, taking a single high-resolution picture of the entire quadrat and identifying all species at once would be much more efficient (see Figure 1(a)). Building AI models to solve this multi-species classification task is a major challenge.

Ideally, we would train AI models using a massive dataset of quadrat images, where every visible species is fully labeled. However, manually annotating every species in these complex images demands a massive and unrealistic amount of expert labor. On the other hand, platforms like Pl@ntNet already have millions of labeled images of individual plants. These large datasets are highly effective for training standard plant classifiers [1, 2].

To bridge this gap, the PlantCLEF challenge focuses on this specific problem for the third consecutive year. Since 2024 [3], the task has been to predict multiple species in quadrat images using only single-plant images for training. Because this challenge is far from being solved, the focus remains firmly on this topic. The 2026 edition reuses the diverse test set introduced in 2025 [4] to accurately measure the community's encouraging progress over time.

As shown in Figure 1, the main difficulty is the visual gap (or "domain shift") between training and test data. The test data are high-resolution images of dense vegetation plots with multiple species. In contrast, the training data from Pl@ntNet [5] are mostly close-up images of single plants or specific organs like flowers. Besides different viewpoints and lighting, the plants also appear in different life stages. Training images often focus on mature flowers. Quadrat images, however, are taken in the field across different seasons. The plants inside can be young, without flowers, fruiting, dying, or even diseased.

Table 1

Training dataset statistics.

Split	Images	Observations	Species	Genera	Families
All	1,408,033	1,151,904	7,806	1,446	181
Train	1,308,899	1,052,927	7,806	1,446	181
Val	51,194	51,045	6,670	1,415	181
Test	47,940	47,932	5,912	1,375	181

2. Dataset

2.1. Training dataset

The training data consists of images of individual plants, similar to those provided in previous PlantCLEF editions. It is a subset of the Pl@ntNet database covering 7,806 species from South-Western Europe. The dataset includes approximately 1.4 million images, supplemented with GBIF records for less common species.

To facilitate model development, images are pre-organized into species-specific subfolders with predefined training, validation, and test splits (see Table 1). Each image has a minimum resolution of 800 pixels on its shortest side. It is important to note that this specific split is strictly a single-plant test set intended for evaluating individual plant identification models, and must not be confused with the multi-species quadrat images used for the main challenge.

2.2. Complementary training dataset: unlabeled pseudo-quadrats

To help participants adapt their models to the complex field scenes of the challenge, we provided a supplementary unlabeled dataset. These images are framed similarly to the test quadrats but lack species-level annotations. They are derived from the LUCAS Cover Photos 2006–2018 archive [6], resulting in 212,782 images. Although introduced during the 2025 edition, no participant has yet successfully leveraged this data to improve their results. However, this dataset remains available for self-supervised or weakly supervised learning to help reduce the visual domain shift between the single-plant training data and the actual test quadrats.

2.3. Test dataset

The challenge test set (hereafter simply referred to as the "test set") comprises 2,105 high-resolution images (typically 3000x3000 pixels) of 0.5 x 0.5m vegetation quadrats. Collected by botanists across various floristic environments, ranging from alpine mountain areas to Mediterranean beaches, these images capture significant real-world variability. Factors such as weather conditions, shadows, shooting angles, and the presence of measurement tools (e.g., wooden frames or tapes) vary considerably between plots (see Figure 2).

2.4. Pre-trained models

To assist participants with limited computational resources, two DINOv2-based Vision Transformer models, fine-tuned on the challenge training set [7], were made available. Both models utilize the ViT-Base architecture extended with 4 registers ((ViT-B/14 reg4, 12 blocks, 768-dimensional embeddings) and were ported using the timm library [8].

- **ViTD2PC24OC**: A model where only the classification head was trained, keeping the pretrained DINOv2-based backbone frozen.
- **ViTD2PC24All**: A model where both the backbone and the classification head were fully fine-tuned on the training set.



Figure 2: Illustration of the diversity of quadrat images in the test set.

Table 2

Top-1 and top-5 accuracies of the pre-trained models on individual plant images.

Model	Short name	Top1	Top5
vit_base_patch14_reg4_dinov2_lvd142m_onlyclassifier	ViTD2PC24OC	63.69	83.88
vit_base_patch14_reg4_dinov2_lvd142m_onlyclassifier_then_all	ViTD2PC24All	75.91	92.82

As shown in Table 2, the fully fine-tuned model (ViTD2PC24All) achieves higher accuracy on individual plant identification.

2.5. Metadata Availability and Design Rationale

To fully understand the constraints of the challenge, it is necessary to clarify the metadata made available to the participants. The training and test sets present a clear asymmetry regarding auxiliary information:

- **Training Set Metadata:** The single-plant training CSV file contains rich contextual attributes

for each image. This includes taxonomic hierarchy (family, genus), the plant organ observed (leaf, flower, fruit, stem/bark, entire view), and precise geographical coordinates (latitude, longitude, altitude), when available, along with observer names.

- **Challenge Test Set Metadata:** In contrast, the test CSV file is strictly minimalist. It provides only four fields: `quadrat_id`, `author`, `date`, and `license`. Exact GPS coordinates, specific locality names, and EXIF location data were completely removed from the test set. While the structure of the `quadrat_id` implicitly hints at grouped plots along transects, no explicit spatial description was provided.

This restriction was a deliberate experimental design choice. By withholding precise geographic localization for the target vegetation plots, the organizers aimed to prevent participants from solving the task through geographic filtering or fetching external GIS data. This constraint forced the community to focus on visual cues and domain adaptation strategies.

3. Task description

The objective of the challenge is to detect every plant species present in high-resolution quadrat images, selected from a pool of more than 7,800 species. As introduced earlier, these images capture 50×50 cm quadrats sampled along transects. Within a single quadrat frame, it is rare to observe more than a few dozen species simultaneously. Formally, this is framed as a weakly supervised multi-label classification task. The "multi-label" aspect is due to the need to predict all species co-occurring in a single quadrat. The supervision is considered "weak" because the models must achieve this using only single-label training images of individual plants, without any explicit spatial annotations (such as bounding boxes or segmentation masks) to guide localization in dense scenes. Consequently, this involves a severe distribution shift, as models must generalize to complex multi-species field scenes using this limited supervision.

3.1. Metric

The challenge uses the F1 score for evaluation, as it balances precision and recall. This metric avoids rewarding excessive predictions that would reduce precision, while also penalizing missing species that would lower recall.

As described earlier, quadrats are sampled along transects to capture biodiversity patterns across time and space. To avoid evaluation bias caused by uneven sampling density or oversampled locations, the final metric incorporates a two-level aggregation. First, F1 scores are averaged across all quadrat images within the same transect. Then, these per-transect averages are averaged across the entire test set.

Formally, the final transect-averaged sample F1 score is calculated as:

$$F1_{\text{avg-transect}} = \frac{1}{T} \sum_{k=1}^T \left(\frac{1}{Q_k} \sum_{j=1}^{Q_k} F1_j^{(k)} \right) \quad (1)$$

where:

- T is the total number of unique transects in the test set.
- Q_k is the number of quadrat images inside transect k .
- $F1_j^{(k)}$ is the image-level F1 score for the j^{th} quadrat image in transect k , defined as:

$$F1_j^{(k)} = \frac{2 \cdot \text{Precision}_j \cdot \text{Recall}_j}{\text{Precision}_j + \text{Recall}_j} \quad (2)$$

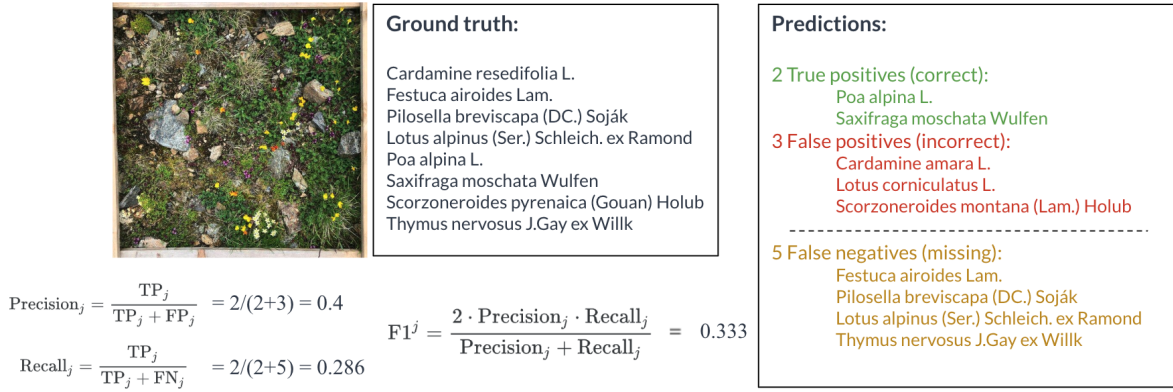


Figure 3: Example of the image-level F1 score calculation for a single quadrat.

For each image j , Precision and Recall are computed based on the predicted and ground-truth species sets:

$$\text{Precision}_j = \frac{\text{TP}_j}{\text{TP}_j + \text{FP}_j}$$

$$\text{Recall}_j = \frac{\text{TP}_j}{\text{TP}_j + \text{FN}_j}$$

where TP_j represents true positives (correctly predicted species), FP_j represents false positives (predicted species that are absent), and FN_j represents false negatives (present species that were missed).

Figure 3 provides a detailed visual example of this per-sample calculation on a test quadrat image.

This hierarchical averaging ensures that the final leaderboard ranking is not dominated by a few oversampled sites. It closely reflects the actual ecological design of the field surveys.

3.2. Rules regarding metadata and external data

Participants were allowed to use image metadata (such as licenses or EXIF data) under one condition: for every run using metadata, teams had to submit an equivalent run relying exclusively on visual information. This rule was designed to isolate the baseline contribution of pure visual analysis.

Similarly, the use of external data was permitted, provided that participants submitted a comparative pure-vision run trained strictly on the provided datasets to ensure a fair methodological comparison.

3.3. Challenge platform

The PlantCLEF2026 challenge was hosted on the [Kaggle platform](#), providing an opportunity for researchers and enthusiasts to contribute to the development of plant recognition in such an original context of quadrat analysis. To prevent overfitting during the competition, the test set was split into a public leaderboard (e.g., 12% of the test quadrats) used for real-time feedback, and a private leaderboard (e.g., 88% of the test quadrats) used to determine the final rankings.

4. Participants and methods

Out of the 557 initial entrants who expressed interest in participating, 123 participants grouped into 103 teams submitted 16,106 runs. Eight teams produced and shared working notes [9, 10, 11, 12, 13, 14, 15, 16].

To address the single-plant to multi-species domain gap, the vast majority of participants explored a common same approach, using a Vision Transformer (primarily the provided PlantCLEF DINOv2 pre-trained models) combined with a tiling extraction strategy to process the high-resolution quadrats. However, teams distinguished themselves through their preprocessing techniques (e.g., multi-scale tiling,

explicit segmentation for rejecting areas without plants, image compression adjustments), their training strategies (e.g., transductive learning, alternative foundation models), and their post-hoc aggregation methods (e.g., kNN retrieval, ecological metadata filtering). Regarding the use of metadata and external data, the decision not to provide precise geographical coordinates in the test set drastically reduced the scope for exploiting external resources (e.g., fetching localized spatial, soil, or climate data). Consequently, the vast majority of submissions relied exclusively on visual cues. Rather than fetching external data, top-performing teams creatively exploited the few available intrinsic metadata—specifically the `date` and the `quadrat_id` structure—to group images and apply post-hoc ecological filters (such as temporal fusion). Because metadata usage was confined to these specific, late-stage filtering steps by a handful of teams, a global quantitative chart comparing metadata versus pure-vision runs across all participants was not deemed relevant. Instead, the specific performance gains of these metadata-driven post-processing steps are highlighted within the individual team method descriptions below. The following paragraphs summarize the approaches tested by these participants. In the section headers, the primary rank reflects their position among the teams that submitted working notes, while the number in parentheses indicates their global rank across all participating teams.

4.1. DS@GT ARC, Georgia Institute of Technology, USA, 1,494 runs, Rank: 1 (Global: 3, kaggle: DS@GT LifeCLEF) [10]

Team DS@GT ARC developed a highly optimized pipeline built around a DINOv2 ViT-L/14 backbone, which they fine-tuned specifically on the PlantCLEF dataset to surpass the provided ViT-B/14 reg 4. Their inference strategy processes quadrats through a multi-scale tiling grid (ranging from 3x3 to 6x6 tiles) and aggregates the predictions using max-pooling to preserve the strongest visual evidence of species presence. A key to their success lies in sophisticated post-hoc adjustments: the system fuses tile-level predictions with a kNN-based retrieval classifier, applies source-dependent temporal fusion for quadrats with multiple visits, and enforces a "habitat-fit" penalization that integrates geographic and altitude priors. The team also documented valuable negative results, demonstrating that specific exploratory approaches, such as a Tile, based Query Decoder, cross-region distillation on the LUCAS dataset, and SAM3 instance-crop augmentations, did not yield improvements under their configuration.

4.2. ANU, Australian National University, Australia, 617 runs, Rank: 2 (Global: 7, kaggle: ARM Wision) [15]

The ANU team relied on a larger foundational model, integrating the recent BioCLIP 2.5 (ViT-H/14) encoder inherently aligned with biological taxonomy. To prevent catastrophic forgetting and maintain the biological "Tree-of-Life" knowledge embedded in the foundation model, they employed a partial fine-tuning strategy, unfreezing only the last four blocks. Their inference pipeline utilizes a 4x4 grid of tiles evaluated at dual resolutions (224 and 336 pixels), followed by a softmax-mean aggregation and a class-prior logit adjustment. They tested several complementary strategies that ultimately proved inconclusive, such as the explicit addition of phenology priors, or enriching the BioCLIP 2.5 features with those extracted from DINOv3 or ConvNeXt-V2-L pre-trained models. This suggests that their partially fine-tuned Bioclip2.5 encoder had already implicitly absorbed most of the biological signals that could be useful for the task.

4.3. SUT, Swinburne University of Technology, Malaysia, 193 runs, Rank: 3 (Global: 12, kaggle: kiwimckiwiman) [12]

Instead of focusing solely on metric optimization, the SUT team proposed a rigorous "three-tier diagnostic framework" (analyzing patch, plot, and site levels) to systematically evaluate model error modes. While their visual pipeline relies on a fixed multi-scale tiling strategy using the provided pre-trained DINOv2 model, their primary contribution lies in characterizing specific prediction failures. They explicitly distinguished between "attentional failures" (where the model is distracted by irrelevant

background objects) and “representational failures” (where the model attends to the correct region but lacks the fine-grained discriminative power to identify the exact taxon). They also provided an in-depth analysis of visual similarity “contamination”, highlighting how morphologically similar co-occurring plants within a single plot can easily mislead the classifier’s confidence.

4.4. HG, Haitong Chen, 62 runs, Rank: 4 (Global: 13, kaggle: Josepha Jostar) [9]

To bridge the gap between single-plant training data and complex field scenes, Haitong Chen implemented a two-stage strategy rooted in transductive learning. First, the pipeline refines species predictions by aggregating visual evidence across repeated visits. Second, it employs a self-training mechanism using pseudo-labels: by framing the quadrat-level prediction as a Multiple-Instance Learning (MIL) task where the quadrat is the bag of instances, the model iteratively fine-tunes its performance. By combining the original model with this adapted version in a teacher-student ensemble, the final submission achieves a more robust identification, effectively leveraging the structure of the test data without requiring any additional human-labeled examples.

4.5. MicroNova AG, Germany, 316 runs, Rank: 5 (Global: 16, kaggle: MicroNovians) [16]

The MicroNova AG team built their pipeline around the pre-trained DINOv2 model, using hierarchical multi-scale tiling and max-pooling aggregation to improve identification. Rather than introducing a new discovery, their work provides a rigorous evaluation of how JPEG compression artifacts affect model performance. They extended previous research [17] by analyzing specific differences in color encoding (chroma subsampling) between the training and test datasets, confirming that matching these compression statistics is a reliable way to reduce the gap between different data domains. Additionally, the authors transparently reported that more complex methods—specifically LoRA fine-tuning and retrieval-augmented classification—did not perform better than their optimized tiling baseline under their limited computational resources. This highlights that, for this challenge, well-tuned preprocessing and tiling strategies remain highly effective and robust solutions.

4.6. Neuon AI, Malaysia, 30 runs, Rank: 6 (Global: 20, kaggle: NEUON AI - Desmond) [14]

The Neuon AI team introduced an innovative three-stage inference pipeline featuring explicit zero-shot spatial reasoning. Rather than blindly tiling the image, they first applied the SAM3 segmentation model to filter out irrelevant background elements (such as soil and stones) and isolate living plant regions. In the second stage, a local Vision-Language Model (VLM) was used as a spatial reasoning module: guided by structured prompts on grid overlays, the VLM identified cells containing visually distinct plant groups without attempting taxonomic identification. Finally, these isolated cells were extracted and passed to the provided pretrained DINOv2 classifier. Their robust pipeline was rounded out by JPEG simulation preprocessing, Test-Time Augmentation (TTA), and MAX pooling aggregation.

4.7. Atlantic, Atlantic Technological University & Technological University Dublin, Ireland, 84 runs, Rank: 7 (Global: 22, kaggle: Atlantic) [11]

Instead of focusing solely on leaderboard performance, the Atlantic team conducted a diagnostic study to measure the gap between the single-plant training domain and the complex target domain of vegetation quadrats. To audit this representation gap, they implemented a lightweight “morphotype” classification head atop the frozen DINOv2 model, which categorizes plants into five broad biological forms (graminoid, broadleaf forb, woody, fern, and succulent). Their analysis revealed that source and target images are nearly perfectly separable in the model’s embedding space, confirming that direct transfer is insufficient without addressing this representation-level shift. Their final submission pipeline built upon this diagnostic foundation using multi-scale sliding windows, mask-guided and

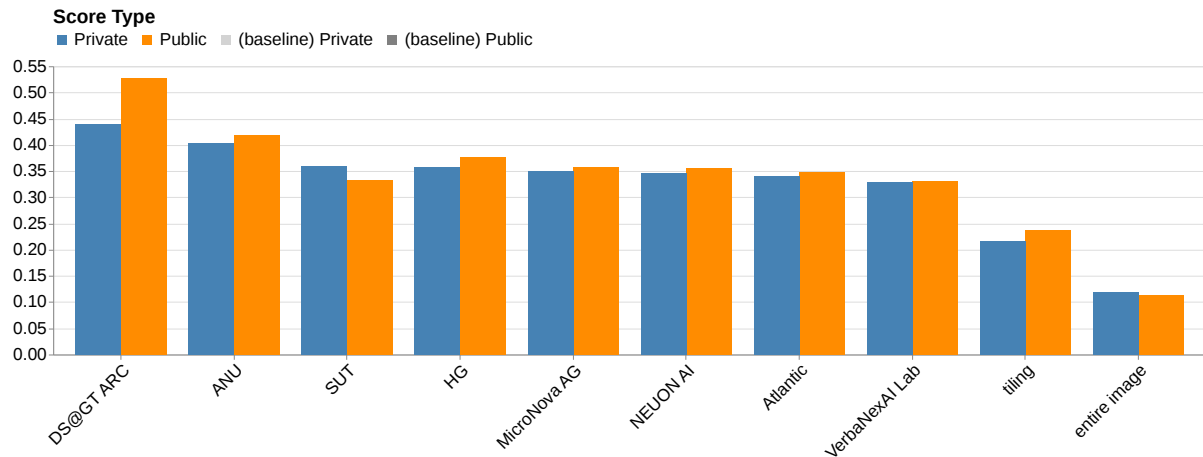


Figure 4: Best run for each team; Mean Sample-Averaged F1 Score Per Transect (each transect includes multiple quadrats from the same site).

"flower-aware" crop selection to improve feature extraction, and a final domain-critic module to refine predictions.

4.8. VerbaNexAI Lab, Universidad Tecnologica de Bolivar, Colombia, 53 runs, Rank: 8 (Global: 27, kaggle: ANDREAMENCOTOVAR) [13]

The VerbaNexAI Lab team implemented a hybrid pipeline that combines multi-view visual inference with non-parametric retrieval and contextual ecological constraints. Rather than relying on full-image classification, they decomposed each quadrat into multiple spatial views—including full scenes, central crops, and overlapping tiles—to capture both community-level context and localized evidence for small or occluded species. Their visual inference was complemented by a global prototype-based retrieval system, which matched test embeddings against species-specific visual prototypes to recover candidates that might be missed by the classifier. Finally, they introduced a site-date "ecological consensus" post-processing layer. By grouping quadrats sharing spatial and temporal identifiers, the system refined predictions to ensure ecological plausibility, reinforcing recurrent species while suppressing inconsistent false positives.

5. Discussion

The 2026 edition of PlantCLEF shows significant methodological changes. Participants have moved beyond simple visual optimization, adopting rigorous diagnostic workflows that distinguish between attentional and representational failures while bridging computer vision with ecological principles. The main trends are summarized below:

1. Diagnostic Research and Model Scaling

From Optimization to Scientific Diagnostic: Instead of focusing only on leaderboard scores, top-performing teams used the challenge to audit the limits of current foundation models [14, 11]. SUT analyzed inference failures by separating attentional from representational modes [14], while Team Atlantic used "morphotype" probes to quantify the domain shift between single-plant training data and community-level test images [11]. These studies provide a mechanistic explanation for why vision transformers struggle with cluttered ecological scenes, offering paths toward future improvements like concept-bottleneck models.

Scaling Model Capacity: Leading teams shifted from base-sized backbones (ViT-B) to larger architectures (ViT-L and ViT-H). Larger backbones capture fine-grained botanical features more effectively, maintaining discriminative power even in fragmented or cluttered scenes where small species might otherwise be missed.

Biology-Specific Foundation Models: The use of domain-specific models, such as BioCLIP 2.5, marks a major architectural shift [18]. By using partial fine-tuning—freezing early layers and updating taxonomic heads—these models leverage "Tree-of-Life" pre-training to outperform generalist models on specialized ecological tasks [18].

2. Pipeline Engineering for Scene Understanding

Supremacy of Hierarchical Multi-Scale Tiling: Full-image classification was largely abandoned due to the loss of detail. The standard strategy is now hierarchical multi-scale tiling (e.g., 3×3 to 8×8 grids) combined with max-pooling, which is essential for detecting small or partially occluded species [10, 14, 11].

Spatial Reasoning and Domain Shift Mitigation: Teams integrated explicit spatial reasoning to handle overlapping scenes, using SAM3 segmentation to filter background noise or local Vision-Language Models (VLM) to localize distinct plant groups [14]. Additionally, teams mitigated domain shift through controlled compression simulation (e.g., JPEG, chroma subsampling) [16] and transductive learning via teacher-student frameworks to better align training and test data [9].

3. Contextual and Retrieval-Augmented Inference

Retrieval-Augmented Classification: To improve parametric models, several teams integrated non-parametric retrieval modules [13, 10]. By comparing tile features against kNN indices or global prototypes built from training data, these systems blended visual similarity scores with model logits to resolve local ambiguities, effectively leveraging the reference database to guide classification [13, 10].

Context-Aware Ecological Inference: Pure visual recognition has reached its limits, leading to the widespread adoption of ecological and temporal priors [13, 10]. Competitive approaches exploited site-date metadata to group related quadrats, implementing robust post-processing—such as temporal fusion and "habitat-fit" penalizations—to reinforce ecologically plausible species and suppress false positives [13, 10]. This represents a definitive move from isolated visual classification to community-level ecological inference, where predictions are validated against the known ecological constraints of the study sites.

6. Conclusion and Perspectives

The PlantCLEF 2026 results mark a significant turning point. While the 2025 edition saw incremental refinements of standard tiling pipelines, this year demonstrated a substantial progress in both leaderboard performance and methodological maturity. The structural difficulty of multi-label, fine-grained classification under severe domain shift remains, yet the community has successfully transitioned from pure visual optimization to context-aware ecological reasoning.

Several contrasts with the 2025 edition are notable. Last year, strategies attempting to isolate vegetative regions faced challenges with consistency, and ecological priors were quite ignored. In 2026, explicit spatial reasoning, driven by segmentation models like SAM3 and Vision-Language Models (VLMs), proved effective. Furthermore, the integration of ecological priors emerged as a key strategy of top-performing systems. By leveraging geographic and temporal metadata, participants successfully filtered out visually plausible but ecologically impossible false positives. Crucially, this year also saw

the rise of diagnostic research, where teams moved beyond "black-box" optimization to explicitly audit and categorize model failure modes, such as attentional versus representational gaps.

Domain adaptation also underwent a paradigm shift. Unlabeled target data was actively exploited through transductive learning and Multiple-Instance Learning (MIL). Additionally, the transition toward biology-specific foundation models, such as BioCLIP, demonstrated that architectures specifically aligned with the "Tree of Life" offer a competitive advantage over generalist backbones.

Looking ahead, while hierarchical multi-scale tiling remains standard, it is computationally intensive. A promising avenue for future research involves designing end-to-end architectures capable of natively capturing multi-scale relationships without relying on brute-force crop aggregation. Furthermore, deepening the integration of ecological data directly into the representation learning phase, rather than merely as a post-hoc filter, could unlock even greater robustness. The increasing sophistication observed this year confirms that PlantCLEF continues to drive meaningful, deployable AI innovations that transform computer vision from a stand-alone task into a powerful support tool for biodiversity monitoring and ecosystem management.

Acknowledgments

The research described in this paper was partly funded by the European Commission via the MAMBO project, which has received funding from the European Union's Horizon Europe research and innovation program under grant agreement 101060639. The opinions expressed in this work are those of the authors and are not necessarily those of the MAMBO partners or the European Commission.

This work was granted access to the HPC resources of IDRIS under the allocation 2025-A0191011389 made by GENCI.

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini 3.1 Pro and DeepL in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

- [1] H. Goëau, P. Bonnet, A. Joly, Overview of PlantCLEF 2022: Image-based plant identification at global scale, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, 2022.
- [2] H. Goëau, P. Bonnet, A. Joly, Overview of PlantCLEF 2023: Image-based plant identification at global scale, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.
- [3] H. Goëau, V. Espitalier, P. Bonnet, A. Joly, Overview of PlantCLEF 2024: multi-species plant identification in vegetation plot images, Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024).
- [4] G. Martellucci, H. Goëau, P. Bonnet, F. Vinatier, A. Joly, Overview of PlantCLEF 2025: Multi-species plant identification in vegetation quadrat images, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, 2025.
- [5] A. Affouard, H. Goeau, P. Bonnet, J.-C. Lombardo, A. Joly, Pl@ntnet app in the era of deep learning, in: 5th International Conference on Learning Representations (ICLR 2017), April 24-26 2017, Toulon, France, 2017.
- [6] R. d'Andrimont, M. Yordanov, L. Martinez-Sanchez, P. Haub, O. Buck, C. Haub, B. Eiselt, M. van der Velde, Lucas cover photos 2006–2018 over the eu: 874 646 spatially distributed geo-tagged close-up photos with land cover and plant species label, *Earth System Science Data* 14 (2022) 4463–4472. URL: <https://essd.copernicus.org/articles/14/4463/2022/>. doi:10.5194/essd-14-4463-2022.

- [7] H. Goëau, J.-C. Lombardo, A. Affouard, V. Espitalier, P. Bonnet, A. Joly, PlantCLEF 2024 pretrained models on the flora of the south western Europe based on a subset of Pl@ntNet collaborative images and a ViT base patch 14 dinoV2, 2024. URL: <https://doi.org/10.5281/zenodo.10848263>. doi:10.5281/zenodo.10848263.
- [8] R. Wightman, Pytorch image models, <https://github.com/rwightman/pytorch-image-models>, 2019. doi:10.5281/zenodo.4414861.
- [9] H. Chen, Adaptive quadrat-domain fine-tuning and plot-level frequency aggregation for plantclef 2026, in: Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, 2026.
- [10] A. Erten, M. Gustineli, A. Cheung, Multi-scale vit inference with habitat-fit priors and knn retrieval for multi-species plant identification, in: Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, 2026.
- [11] S. Foy, S. McLoughlin, Quantifying domain shift in foundation models for ground vegetation imagery, in: Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, 2026.
- [12] H. A. Ishrat, S. H. Lee, Beyond domain shift: A three-tier diagnostic framework for multi-species plant identification in vegetation quadrats, in: Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, 2026.
- [13] A. Menco-Tovar, J. E. Serrano, J. C. Martinez-Santos, E. Puertas, Contextual multi-label identification of plant species in multi-species plots, in: Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, 2026.
- [14] D. K. T. Oh, A. Y. H. Chai, S. H. Lee, Vision-language guided plant localization for multi-label species prediction in quadrat images, in: Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, 2026.
- [15] A. Raj, M. d. Mel, R. Wasif, W. Brake, Fine-tuning of bioclip 2.5 with taxonomic heads for multi-species plant identification, in: Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, 2026.
- [16] C. A. Rickert, T. Haar, Bridging single-label training and multi-label inference with vision transformers for vegetation monitoring, in: Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, 2026.
- [17] V. Espitalier, Preprocessing is all you need, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, 2025.
- [18] J. Gu, S. Stevens, E. G. Campolongo, M. J. Thompson, N. Zhang, J. Wu, A. Kopanev, Z. Mai, A. E. White, J. Balhoff, W. M. Dahdul, D. Rubenstein, H. Lapp, T. Berger-Wolf, W.-L. Chao, Y. Su, BioCLIP 2.5 Huge, 2026. URL: <https://huggingface.co/imageomics/bioclip-2.5-vith14>.