Adaptive Quadrat-Domain Fine-Tuning and Plot-Level Frequency Aggregation for PlantCLEF 2026

Working Note for the LifeCLEF Lab at CLEF 2026

Haitong CHEN[1]

Abstract

This working note presents a PlantCLEF 2026 system for multi-species plant identification in high-resolution vegetation quadrat images. The task requires image-level species sets in a target domain of complex field scenes, whereas the available supervised visual resources are dominated by single-plant observations. To bridge this mismatch, this system combines a DINOv2-based PlantCLEF classifier, tile-level quadrat inference, exact-plot frequency aggregation across repeated observations, and transductive multiple-instance pseudo-label adaptation on quadrat imagery. The aggregation stage selectively expands species sets for the CBN subsets using repeated support within exact plot groups. The adaptation stage reuses unlabeled competition quadrat images: plot-aware teacher predictions provide pseudo labels, several tiles are sampled from each quadrat, tile logits are max-pooled at image level, and the classification head together with a small backbone region is updated. No quadrat ground-truth labels are used in this adaptation stage. The final submission, associated with the Kaggle team name *Josepha Jostar*, obtained a Public Score of 0.37681, ranking 12th on the public leaderboard, and a Private Score of 0.35532, ranking 13th on the private leaderboard. Diagnostics from the final run show a controlled prediction-set construction process: tile inference keeps preliminary sets bounded, while repeated-plot support drives the targeted set expansion used for submission.

Keywords

multi-label image classification, DINOv2, multiple-instance learning, pseudo labels, domain adaptation

1. Introduction

Vegetation plots are a practical unit for ecological monitoring: a quadrat image can contain several plant species, occlusion, bare soil, acquisition artifacts, and repeated views of the same physical plot over time. PlantCLEF 2026 is part of the LifeCLEF benchmark campaign for biodiversity-oriented machine learning systems [1, 2]. It focuses on a difficult multi-species setting while maintaining a strong source–target mismatch. The available trained visual classifier is naturally suited to single-plant recognition, whereas the test instances are high-resolution scenes in which several species may appear in different patches. This mismatch is central to the PlantCLEF 2026 task description and continues the design of the 2025 task.

This work starts from the practical observation that a strong single-plant model can be used as a local recognizer when it is evaluated on quadrat tiles. The remaining problem is not only visual recognition but also deciding how to merge uncertain tile evidence into an image-level species set. A predicted set that is too small misses species while a set that is too large quickly accumulates false positives. Repeated plot identifiers provide an additional source of weak structure. Frequency evidence across repeated images can support a species that appears consistently within the same physical plot, but it does not justify an unrestricted union of uncertain tile candidates. This observation motivates a controlled species-set construction rule that uses repeated support only where it is available.

The proposed system follows a restrained teacher–student design. A DINOv2-based PlantCLEF checkpoint serves as a teacher. It performs patch-level inference and produces image-level predictions. These predictions are aggregated by exact plot identifier to create pseudo labels. A student copy of the teacher is then adapted on quadrat tiles using a multiple-instance objective. The adapted model is

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

josephajostar@gmail.com (H. CHEN)
0009-0000-8431-918X (H. CHEN)

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

ensembled with the teacher at inference time, and the resulting predictions are post-processed with the same plot-level frequency rule.

This working note makes three contributions:

1. It formulates PlantCLEF 2026 inference as controlled species-set construction under a single-plant to quadrat-domain source–target mismatch.
2. It proposes an exact-plot aggregation rule that acts as a calibrated weak prior, expanding species sets only when repeated plot-level support is available.
3. It analyzes a transductive teacher–student adaptation pipeline that combines pseudo-label Multiple-Instance Learning with a conservative teacher pathway, and documents its limitations for prospective monitoring.

The submitted system corresponds to the Kaggle team name *Josepha Jostar*, which links the reported scores to the official PlantCLEF 2026 leaderboard.

2. Overview

2.1. The PlantCLEF 2026 competition

PlantCLEF 2026 is the multi-species plant identification task of LifeCLEF [1]. Its target images are vegetation plot images, while the supervised training resources are single-plant observations and associated species identifiers. The task overview describes the target as high-resolution vegetation quadrats, typically covering 50×50 cm, and states that the 2026 edition keeps the 2025 datasets and Kaggle setting. The resulting source–target gap is therefore deliberate: large single-plant collections support visual recognition training, while the target requires scene-level species inventory in complex quadrat imagery [2].

For each test image, the system returns a list of integer species identifiers. The official metric computes an F1 score for each image, averages image scores within each transect, and finally averages across transects. This hierarchy rewards balanced per-image species sets and prevents oversampled transects from dominating the final score. In practice, image-level cardinality is a sensitive hyperparameter: high-confidence local evidence from a tile should not automatically expand to a large species list for the whole quadrat [2].

2.2. Resources

The system uses the PlantCLEF 2026 test images, the official mapping between model class indices and competition species identifiers, and a PlantCLEF-trained DINOv2 ViT-B/14 classifier from the released PlantCLEF DINOv2 model family [3, 4]. The optional PlantCLEF 2024 single-plant metadata file is used only for robust data loading and inspection, not as a direct geographic or taxonomic prior. The method therefore relies on the trained visual checkpoint and on pseudo labels derived from target-domain quadrat predictions rather than on manually specified metadata priors.

2.3. Problem Formulation and Metric Implications

For each quadrat image x_i, the system predicts a set of plant species identifiers

$$\hat{Y}_i = \{s_1, s_2, \ldots, s_{k_i}\},$$

rather than a single class label. This distinction is important for the PlantCLEF 2026 setting: the source model is trained for single-plant recognition, whereas the target output is a multi-species inventory of a vegetation quadrat.

The official score is based on per-image F1 values that are averaged within transects and then across transects [2]. Consequently, the prediction-set cardinality k_i is part of the decision problem. A set that is too small omits occluded, small, or weakly recognized species and reduces recall. A set that

Trained PlantCLEF DINOv2 checkpoint → tile inference on quadrats → teacher image predictions → exact-plot frequency aggregation → pseudo labels → multiple-instance adaptation on random quadrat tiles → teacher–student ensemble → final aggregation and prediction set

Figure 1: Overview of the proposed PlantCLEF 2026 pipeline.

is too large accumulates false positives and reduces precision. This trade-off motivates the structure of this system: tile inference extracts local visual evidence, plot-level aggregation adds only species supported across repeated observations, and target-domain adaptation is applied without removing the conservative teacher pathway at final inference.

3. Method

The final method was designed around three observations.

- First, a single-plant classifier is more naturally applied to local quadrat regions than to an entire high-resolution vegetation scene. This system therefore uses overlapping tiles to obtain local species evidence.
- Second, tile-level evidence does not directly define a reliable quadrat-level species set: a high-scoring local candidate may still be a false positive, while blind unioning of local predictions can expand the output set too aggressively. This system therefore separates local inference from final set construction.
- Third, repeated observations of the same plot provide weak but useful support for species presence. This support is used both to construct plot-aware pseudo labels and to regularize the final submission.

The adaptation stage follows the same logic. Quadrat-level pseudo labels do not imply that every sampled crop contains every labeled species. The multiple-instance formulation therefore treats the sampled tiles as a bag and uses image-level supervision after max pooling over tile logits. Figure 1 summarizes the complete pipeline.

3.1. Teacher model

The teacher is the trained kaggle model `dinov2_patch14_reg4_onlyclassifier_then_all` [4], loaded as a classifier with architecture string `vit_base_patch14_reg4_dinov2.lvd142m`. DINOv2 provides strong generic visual representations for downstream recognition [5]; the Vision Transformer backbone makes tile-level processing natural [6]. This system uses the trained PlantCLEF classifier checkpoint instead of a foundation-only DINOv2 weight file. This choice is necessary because the later set-construction stages assume species logits from a PlantCLEF classifier head rather than generic foundation-model features.

3.2. Patch-level quadrat inference

For a quadrat image, the teacher model is evaluated over overlapping patches with the model input resolution. The main inference configuration of the final system is listed in Table 1. For each tile, softmax probabilities are computed over species classes. Only the top three tile predictions with probability at least 0.12 are retained. For each species, tile evidence is aggregated with a maximum-probability score. The preliminary image prediction keeps at most five species and uses an adaptive cutoff equal to the maximum of 0.10 and 35% of the best image-level species score. Boundary regions are retained by reflect-padding the image on the bottom and right before extracting overlapping patches. Thus, the system avoids zero padding while still allowing the final patch grid to cover the full quadrat image.

Table 1
Teacher inference settings in the final system.

Parameter	Value
Architecture	`vit_base_patch14_reg4_dinov2.lvd142m`
Tile batch size	32
Stride ratio	0.70
Top classes per tile	3
Tile probability threshold	0.12
Tile aggregation	maximum probability
Maximum preliminary species	5
Final minimum score	0.10
Adaptive margin	0.35

Patch-level evaluation is computationally inexpensive compared with re-training from scratch on the complete single-plant image collection. More importantly, it aligns the local receptive field of the single-plant classifier with subregions of the target quadrat scene.

3.3. Exact-Plot Frequency Aggregation

The competition test identifiers expose repeated observations of a physical plot by a date suffix. This system defines an exact plot key by removing a terminal date string from the image identifier. For each exact plot group, this system counts how often each species is present in preliminary image predictions. The group counts are used only for the CBN data subsets, where repeated acquisition dates provide enough within-plot support for this strategy to be meaningful.

The aggregation rule was designed as a conservative set-expansion heuristic rather than as a general ecological occurrence model. It is based on three constraints. First, an added species should have repeated support within the same exact plot, because a single low-ranking tile candidate is not reliable enough to expand an image-level prediction set. Second, support should be measured jointly by an absolute count and a relative frequency: the count threshold prevents isolated detections from being promoted, while the frequency threshold prevents large groups from adding species that appear only sporadically. Third, the final prediction set should remain bounded, because the official image-level F1 score penalizes excessive false positives. The subset-specific target cardinalities therefore act as caps on expansion rather than as estimates of the true number of species in a plot.

The selected rule is:

- aggregate only groups containing at least eight repeated images;
- require at least three group occurrences and at least 10% group frequency;
- target up to 11 predicted species for `CBN-PdlC`;
- target up to 14 predicted species for `CBN-Pla`.

Listing 1 describes the rule. This work refers to the selected final aggregation as `rulePlus`. The rule acts as a weak plot-consistency prior: a species repeatedly recognized at the same plot has stronger support than a one-off low-ranking tile candidate. The numerical thresholds were selected from development runs as submission-time hyperparameters. The rule should be interpreted as a calibrated weak prior for the PlantCLEF batch setting, not as a universal ecological occurrence model.

3.4. Pseudo-label Multiple-Instance Learning Adaptation

The frequency-aggregated teacher output is reused as a pseudo-label set for quadrat-domain adaptation on unlabeled competition images. This stage is therefore a transductive learning, or test-time adaptation, procedure: unlabeled target-domain test images are used for model adaptation, while no quadrat ground-truth labels are introduced. It can also be viewed as a self-training strategy in which model predictions

Listing 1: Exact-plot rulePlus aggregation.

```
for image in test_images:
    S = preliminary_species(image)
    key = remove_terminal_date(image.id)
    if image.dataset == "CBN" and group_size(key) >= 8:
        target = 11 if image.subset == "PdlC" else 14
        for species in frequency_rank(key):
            if count(key, species) >= 3
               and count(key, species) / group_size(key) >= 0.10:
                S.add(species)
            if len(S) >= target:
                break
    emit(image.id, S)
```

become training targets in an unlabeled target domain [7]. The adaptation stage uses a Multiple-Instance Learning (MIL) formulation because quadrat-level pseudo labels do not specify which sampled tile contains each species. Let $X_i = \{x_{ij}\}_{j=1}^{T}$ be $T = 8$ tiles sampled from quadrat image i, and let Y_i be the pseudo-positive species set. The student starts as a copy of the teacher. This system evaluates tile logits with the student and max-pools over the tile dimension:

$$z_i(c) = \max_{1 \le j \le T} z_{ij}(c).$$

The max operation implements a simple multiple-instance assumption: a species pseudo label may be visible in only some tiles of a quadrat. This bag-level view avoids asserting that every crop contains every pseudo-positive species and is consistent with deep Multiple-Instance Learning formulations for weakly labeled images [8].

The original classifier is a softmax-style single-plant recognizer. To avoid immediately replacing its calibration with an unrestricted sigmoid objective, the default adaptation loss distributes probability mass uniformly across pseudo-positive labels:

$$q_i(c) = \begin{cases} 1/|Y_i|, & c \in Y_i, \\ 0, & c \notin Y_i. \end{cases}$$

After label smoothing $\epsilon = 0.02$, the loss is

$$\mathcal{L}_i = -\sum_c \tilde{q}_i(c) \log \mathrm{softmax}(z_i)_c.$$

A weighted BCE alternative was considered during development, but the final default is the multi-positive softmax objective.

The adaptation stage samples at most 1200 quadrat images, trains for two epochs with mixed precision when a GPU is available, updates the classification head with learning rate 8×10^{-5}, and updates the last transformer block plus normalization parameters with learning rate 8×10^{-6}. The short two-epoch schedule was chosen as a conservative heuristic to limit overfitting to noisy pseudo labels while still allowing limited target-domain adjustment. Table 2 gives the configuration.

3.5. Teacher–student inference

After adaptation, three inference variants are evaluated: teacher-only, adapted-only, and teacher–adapted ensemble. The final submission uses the teacher–adapted ensemble followed by the same exact-plot aggregation rule. The teacher-only variant is retained as a sanity control: if this control does not match the expected teacher behavior, checkpoint selection or inference configuration has changed and adaptation should not be interpreted in isolation.

Table 2
Pseudo-label Multiple-Instance Learning adaptation settings.

Parameter	Value
Epochs	2
Images sampled for adaptation	at most 1200
Images per training batch	2
Tiles per quadrat	8
Tile crop scale	1.35
Loss	multi-positive softmax with smoothing 0.02
Head learning rate	8×10^{-5}
Backbone learning rate	8×10^{-6}
Weight decay	0.02
Trainable backbone region	last transformer block and norms

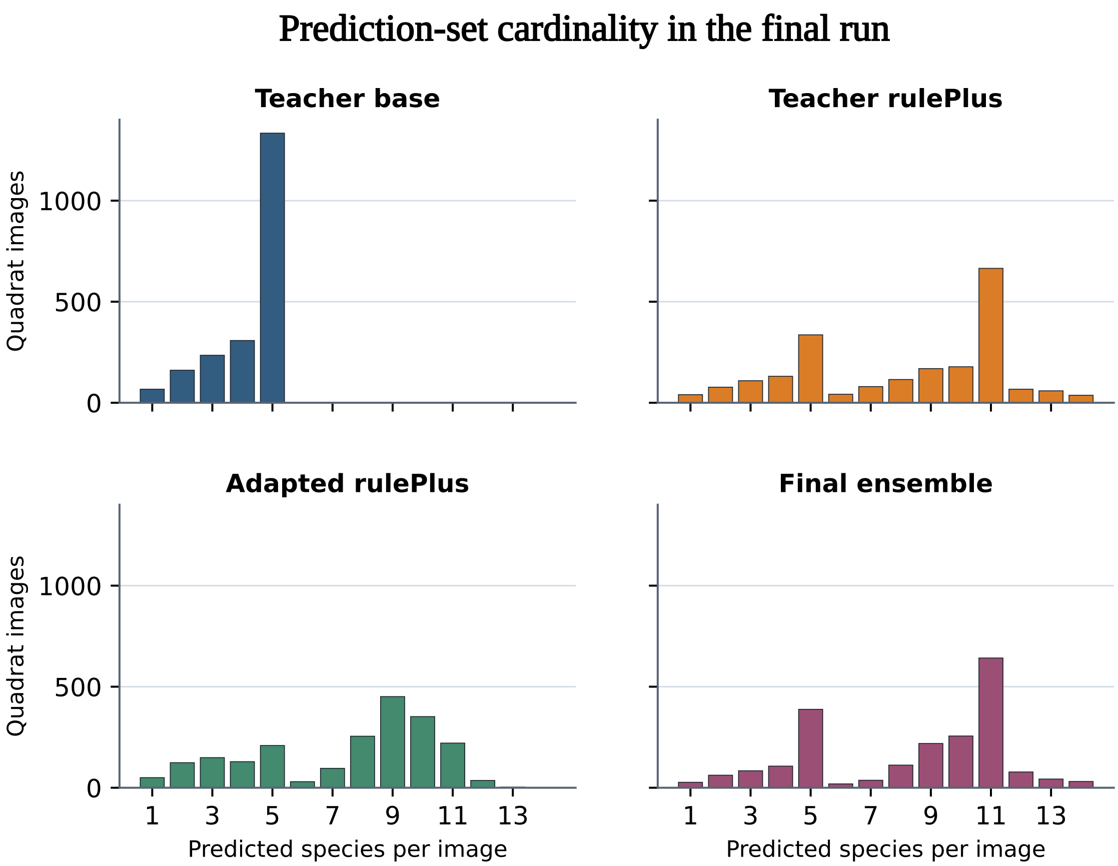

Figure 2: Prediction-set cardinality distributions for the final system variants. Base teacher inference is conservative, while aggregation expands the species sets for plot-level submission.

4. Results

The final system records diagnostic summaries for the teacher, adapted, and teacher–adapted inference paths. These summaries make it possible to inspect how the prediction set evolves inside the final run without mixing in earlier exploratory post-processing experiments.

4.1. Prediction-Set evolution

Table 3 summarizes the output-set cardinality of the six final variants. Figure 2 visualizes the corresponding cardinality distributions. Before aggregation, all three base inference paths remain conservative:

Table 3
Prediction-set diagnostics generated by the final system run.

Variant	Mean	Median	P90	Max
Teacher base	4.274	5	5	5
Teacher rulePlus	8.153	9	11	14
Adapted base	4.445	5	5	5
Adapted rulePlus	7.448	9	11	13
Teacher–adapted base	4.573	5	5	5
Teacher–adapted rulePlus	8.343	9	11	14

Table 4
Effect of final aggregation by dataset subset.

Subset	Rows	Changed rows	Added species entries
Non-CBN	629	0	0
CBN-other	32	0	0
CBN-PdlC	816	816	4444
CBN-Pla	628	628	3491

their 90th percentile and maximum prediction counts are both five species per image. Plot-level aggregation then produces larger quadrat-level species sets, especially for the teacher and final teacher–adapted paths.

This pattern clarifies the role of each stage. Tile inference provides a bounded preliminary set, while exact-plot aggregation performs the main species-set expansion. The final submission is therefore not an unrestricted union of tile predictions; it is the output of a conservative local recognizer followed by a plot-aware aggregation rule.

4.2. Effect of Exact-Plot aggregation

Removing the terminal date suffix from quadrat identifiers produced 656 exact plot groups in the final run. Only 72 groups contained at least eight repeated images, covering 1444 test rows. The `rulePlus` aggregation therefore acts on a specific repeated-plot subset rather than on the entire test set.

Table 4 reports the change from the teacher–adapted base output to the final teacher–adapted rulePlus output. The aggregation stage adds species only for the two targeted CBN subsets. No non-CBN row is modified by this rule.

The added entries also show repeated support within the preliminary predictions. For final rule additions, the supporting occurrence count has minimum 3, median 5, and 90th percentile 12. The corresponding within-group support frequency has minimum 0.12, median 0.25, and 90th percentile 0.556. Thus, the aggregation rule converts repeated plot-level evidence into additional species predictions while avoiding one-off candidate additions. Figure 3 provides a qualitative example of this repeated-plot expansion, and Figure 4 summarizes the occurrence-count and frequency support behind the added species.

4.3. Effect of Quadrat-Domain adaptation

The adaptation loss decreased from 4.142 after the first epoch to 3.641 after the second epoch. This training history indicates that the pseudo-label objective was optimized during the final run, although loss reduction alone is not a substitute for labeled quadrat-domain validation.

To inspect whether adaptation changes the predicted sets, this work compares the outputs produced by the final system. Table 5 reports row-level set differences between the teacher, adapted, and final teacher–adapted variants. Mean Jaccard is the row average of $\left|\hat{Y}_i^{(a)} \cap \hat{Y}_i^{(b)}\right| / \left|\hat{Y}_i^{(a)} \cup \hat{Y}_i^{(b)}\right|$, so larger

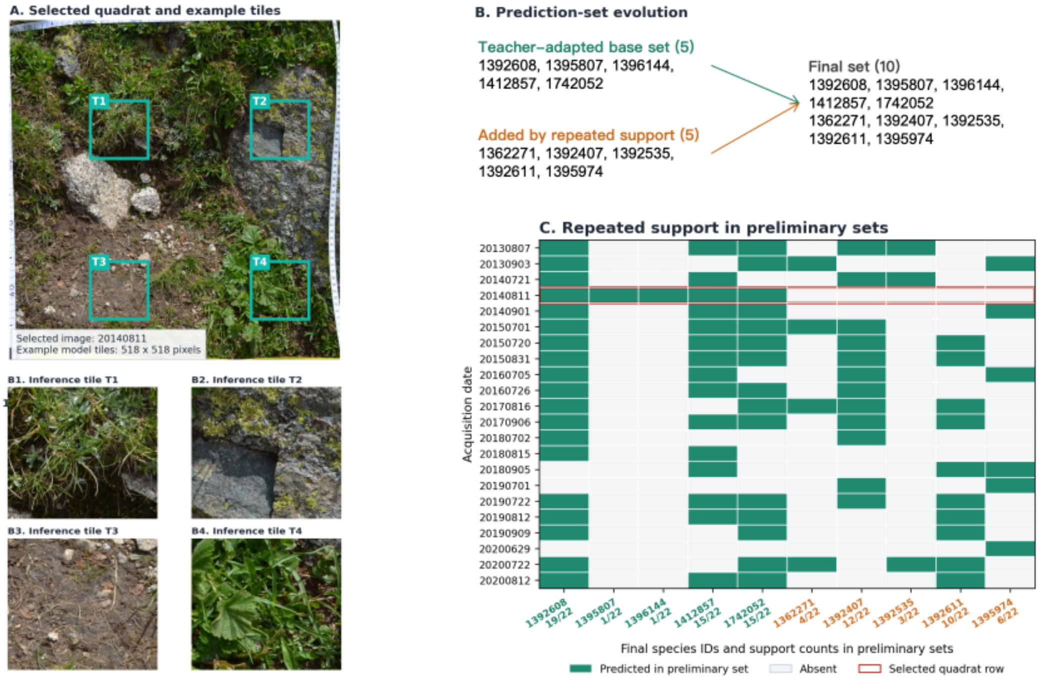

Figure 3: Qualitative prediction-set evolution for a repeated CBN plot. Panel A shows a representative quadrat and example inference tiles. Panel B compares the conservative teacher–adapted base prediction set with the final `rulePlus` prediction set for the selected image. Panel C shows repeated preliminary-set support across acquisition dates in the same exact plot group. Species IDs are predictions on hidden test images; no quadrat ground-truth labels are shown.

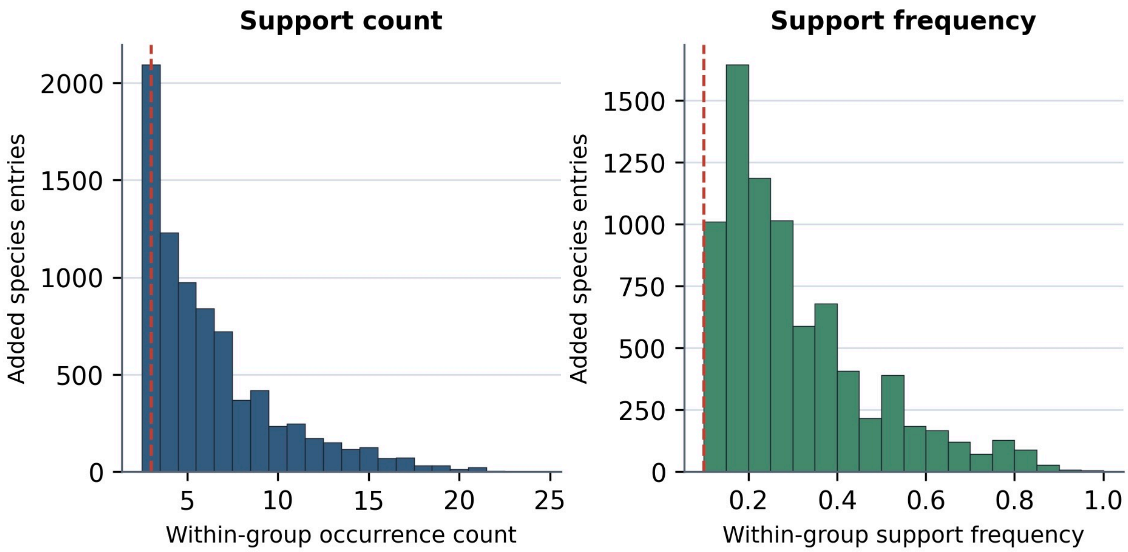

Figure 4: Repeated-support statistics for species added by final aggregation. The `rulePlus` rule requires multiple preliminary predictions within an exact plot group before a species can be added to a quadrat-level prediction set.

values indicate more similar species sets.

The adapted-only output differs substantially from the teacher output, showing that pseudo-label fine-tuning changes the quadrat-domain prediction behavior. The final teacher–adapted ensemble remains closer to the teacher than the adapted-only variant, while still changing a large fraction of rows. This supports the use of the ensemble as a conservative way to incorporate target-domain adaptation without fully discarding the original trained PlantCLEF classifier.

Table 5
Prediction-set differences among final `rulePlus` variants.

Comparison	Changed rows	Mean Jaccard	Added	Removed
Teacher vs adapted	2029	0.484	5308	6791
Teacher vs final	1750	0.759	2778	2378
Adapted vs final	1922	0.640	4898	3015

Table 6
Final PlantCLEF 2026 Kaggle results.

Leaderboard split	Score	Rank
Public	0.37681	12th
Private	0.35532	13th

4.4. Analysis

The final submission was associated with the Kaggle team name *Josepha Jostar*. Table 6 reports the leaderboard results provided for this submission.

The private score is lower than the public score and establishes the official result on the hidden split. The diagnostic tables explain how the final prediction sets are formed, but they should not be read as an isolated leaderboard ablation for each stage. The method remains constrained by pseudo-label quality: if the teacher omits rare species or repeats systematic false positives, adaptation may reinforce these errors. The teacher–student ensemble partly mitigates this risk by retaining teacher evidence at inference time.

5. Reproducibility

The implementation is written in PyTorch and `timm`. It uses the PlantCLEF 2026 test images, the official species-identifier mapping, and the released PlantCLEF-trained DINOv2 classifier checkpoint described in Section 3.1. The random seed is set to 42 for Python, NumPy, and PyTorch. Exact reproduction can still depend on GPU kernels, image order, and Kaggle package versions.

For reproducibility and diagnostic inspection, the released system produces three categories of outputs: teacher-only predictions, adapted-model predictions, and teacher–student ensemble predictions. It also records diagnostic summaries used to compare prediction-set cardinality, aggregation effects, and set overlap between model variants. The default submitted output corresponds to the teacher–student ensemble followed by the exact-plot aggregation rule.

6. Discussion

PlantCLEF 2026 illustrates a practical problem in ecological computer vision: the strongest labels are often attached to individual organisms, but field deployment requires scene-level inventory. The patch-level teacher resolves part of this gap by locally matching image statistics of the supervised source domain. Exact-plot aggregation introduces a low-cost repeated-plot prior. Transductive Multiple-Instance Learning adaptation then lets the model update its representation on quadrat imagery without requiring additional human labels.

These stages address different errors and should not be conflated. Tiling reduces spatial dilution by exposing the single-plant teacher to local evidence; exact-plot aggregation regularizes the final set by requiring repeated support before adding species; adaptation changes the representation using weak target-domain supervision. The final score reflects their interaction, not a causal ablation of each stage. The main remaining scientific risk is confirmation bias: the same teacher produces the pseudo labels and supplies half of the final ensemble, so repeated false positives can survive both aggregation and adaptation.

The transductive setting is also a practical limitation. The submitted system is best interpreted as a retrospective archive annotation method, where a batch of unlabeled target-domain quadrat images is available before final prediction. This setting differs from prospective ecological monitoring, where observations arrive sequentially and future images are unavailable when earlier observations must be annotated. In such a deployment scenario, using all images from the same exact plot would introduce temporal information leakage, because later observations could influence predictions for earlier ones. Continuous ecological monitoring would therefore require incremental aggregation and adaptation procedures that operate only on past and current observations.

The exact-plot aggregation rule is another source of method-specific risk. It relies on manually selected thresholds for group size, occurrence count, frequency support, and target prediction-set cardinality. Although these thresholds were effective in the competition setting, they may partially reflect the structure of the CBN subsets and would need recalibration before being transferred to other ecological datasets or monitoring protocols.

Several improvements remain open. First, future work could filter tile sampling with a vegetation or artifact mask so that pseudo-label updates are less influenced by frame borders, labels, and soil. Second, explicit taxonomic or occurrence priors could be added after careful calibration. Third, pseudo labels could be weighted by agreement between teacher tiles, repeated plot frequency, and adapted-model confidence instead of treating all pseudo-positive species equally. Finally, future systems could replace the current batch-level adaptation strategy with continual or incremental learning procedures designed for sequential biodiversity monitoring. These directions are compatible with observations in recent PlantCLEF systems that use tiling, metadata, and ecological context.

7. Conclusion

This work presented a PlantCLEF 2026 system that combines DINOv2-based tile inference, exact-plot frequency aggregation, and transductive pseudo-label Multiple-Instance Learning adaptation on vegetation quadrats. The final run reached a Private Score of 0.35532 and a Public Score of 0.37681. The reported diagnostics connect the method choices to the final output: base tile inference keeps preliminary species sets bounded, `rulePlus` expands predictions only when repeated plot support is available, and teacher–student comparison shows that adaptation changes target-domain predictions while the final ensemble retains teacher evidence. This design addresses the single-plant to quadrat-domain gap with controlled set construction rather than treating every local candidate as a final image-level species prediction.

Acknowledgments

The author thanks the PlantCLEF and LifeCLEF organizers for the challenge resources and the Kaggle platform used for submission and evaluation.

Declaration on Generative AI

During the process of editing this working note, the author utilized OpenAI Codex to review and correct the grammatical and LaTeX formatting issues that arose. After using this tool, the author conducted a review and editing of the content as needed, and takes full responsibility for the published content.

References

[1] L. Picek, L. Adam, S. Kahl, R. Bossy, L. Chrobak, H. Go"eau, K. Papafitsoros, H. Klinck, W.-P. Vellinga, R. Planqu'e, T. Denton, K. Barnard, C. N'edellec, L. Del'eger, M. Courtin, G. Martellucci, I. Moummad, F. Vinatier, P. Bonnet, A. Joly, Overview of lifeclef 2026: Ai challenges for biodiversity understanding and ecosystem management, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2026.

[2] G. Martellucci, I. Moummad, H. Goëau, P. Bonnet, F. Vinatier, A. Joly, Overview of PlantCLEF 2026: Identify multi-species plants in images of vegetation plots, in: Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, 2026.

[3] H. Goëau, J.-C. Lombardo, A. Affouard, V. Espitalier, P. Bonnet, A. Joly, Plantclef 2024 pretrained models on flora of southwestern europe based on pl@ntnet images and vit base dinov2, 2024. URL: https://doi.org/10.5281/zenodo.10848263. doi:10.5281/zenodo.10848263.

[4] G. Martellucci, dinov2_patch14_reg4_onlyclassifier_then_all, Kaggle Models, 2025. URL: https://www.kaggle.com/models/juliostat/dinov2_patch14_reg4_onlyclassifier_then_all/PyTorch/default/3, pyTorch model resource, version 3. Accessed: 2026-05-21.

[5] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, et al., Dinov2: Learning robust visual features without supervision, Transactions on Machine Learning Research (2024). URL: https://openreview.net/forum?id=a68SUt6zFt.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[7] D.-H. Lee, Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, in: Workshop on Challenges in Representation Learning, ICML, 2013.

[8] M. Ilse, J. M. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, 2018, pp. 2127–2136. URL: https://proceedings.mlr.press/v80/ilse18a.html.