Shadow Spark at PlantCLEF 2025 : Multi-Label Plant Species Classification using Tiling-based Inference

Notebook for the LifeCLEF Lab at CLEF 2025

Jayasree R, P Mirunalini*, Kawvya M K and Harini J

Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India

Abstract

The PlantCLEF 2025 initiative aims to advance ecological research by analyzing vegetation plot inventory photographs to support standardized sampling, biodiversity monitoring, long-term ecological assessment, and large-scale remote vegetation surveys. The challenge is framed as a multi-label classification task, requiring the identification of all visible plant species in high-resolution quadrat images. In this approach, a self-supervised Vision Transformer (DINOv2) model is employed to address the task. Both the base and fine-tuned variants of DINOv2 are utilized to extract robust and generalizable feature embeddings, which serve as inputs for training classifiers capable of predicting multiple plant species present in a single image. To handle high-resolution input efficiently, the proposed processing framework divides each image into a grid of overlapping tile individually, performs classification on each tile individually and then aggregates tile-level predictions into a unified set of probabilities per image. This tiling-based strategy enhances spatial context and improves species detection, especially in complex and densely populated vegetation plots achieving a macro-averaged F1 score of 0.00106 per sample on the official test set.

Keywords

LifeCLEF, DINOv2, fine-grained classification, species identification, vegetation plot images, multi-label classification, biodiversity informatics

1. Introduction

The PlantCLEF 2025 challenge [1], organized as part of the LifeCLEF lab [2] under the Conference and Labs of the Evaluation Forum (CLEF), focuses on the automated identification of plant species in high-resolution images of vegetation plots. This task is framed as a multi-label classification problem, where each image can contain multiple co-occurring species. A major challenge introduced in this edition is the domain shift between the training data—comprised of single-plant images—and the test data—vegetation plots containing several species within a single frame. This setup reflects real-world ecological monitoring scenarios, aiming to facilitate standardized biodiversity assessment, long-term environmental monitoring, and scalable vegetation surveys. The scale and complexity of the dataset, with over 7,800 species and more than 1.4 million images, make the task computationally demanding and crucial for automated biodiversity monitoring and ecological research.

2. Background

The PlantCLEF challenge series has played a vital role in advancing automated plant species identification, especially under real-world ecological conditions. The 2025 edition builds upon the foundation laid in 2024, continuing the focus on multi-label classification of high-resolution vegetation plot images. While the 2024 task [3] already introduced the challenge of identifying multiple co-occurring plant species in a single image, the 2025 edition raises the complexity further by increasing the test set size and emphasizing scalability and robustness of solutions.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

jayasree2370002@ssn.edu.in (J. R); miruna@ssn.edu.in (P. Mirunalini); kawvya2370039@ssn.edu.in (K. M. K); harini2370052@ssn.edu.in (H. J)

In prior editions of the challenge, top-performing teams explored a variety of innovative strategies to handle the scale and complexity of the data. Leading approaches included combining multiple tiling configurations for inference, applying false positive reduction techniques like Segment Anything (SAM), and aggregating predictions across different temporal views of the same plot [4].Other successful methods employed ensemble learning techniques, Vision Transformer (ViT) backbones — including the official pre-trained model ViTD2PC24All, which was fine-tuned on the PlantCLEF 2024 dataset using supervised learning across all layers — and probabilistic aggregation methods like Bayesian Model Averaging [5]. Most approaches used tile-wise prediction strategies to extract localized species predictions from smaller grid patches [6].

These prior works demonstrate that dense tiling, temporal aggregation, and ensemble modeling are highly effective in managing the high variability and label density present in plot-level plant classification. Inspired by these advancements, the present work utilizes a transfer learning approach based on DINOv2 [7] —a self-supervised Vision Transformer—to extract rich and generalizable visual embeddings. Images are tiled into overlapping patches, each of which is classified using a linear classifier trained on DINOv2 embeddings. The predictions are then aggregated to generate final multi-label outputs for each plot.

2.1. DINOv2 Model Overview

DINOv2 is a self-supervised Vision Transformer (ViT) model trained on the large-scale LVD-142M dataset containing 142 million images. Similar to BERT [8] in NLP, it learns general-purpose image representations without labeled data. Images are divided into fixed-size patches, and each patch is linearly embedded into vector. A special classification [CLS] token is prepended [9] to the sequence of patch embeddings. Positional embeddings are added to retain spatial information. The sequence (patches + [CLS]) is passed through the Transformer encoder. The output corresponding to the [CLS] token at the final layer serves as a global representation of the image.

DINOv2 is available in various sizes (from small to giant), the proposed work utilizing the ViT-B/14 (distilled) variant, which offers a balance between performance and efficiency. This model produces a fixed output of shape 257×768, comprising 256 patch embeddings and one [CLS] token embedding [10]. The challenge organizers provided two pretrained ViT-B/14 DINOv2 models: one with a frozen backbone, and another fine-tuned end-to-end [11]. The fine-tuned model is used to extract [CLS] embeddings from each image, leveraging its stronger feature representations for the multi-label classification task

3. System Overview

3.1. Dataset Overview

The dataset used in the PlantCLEF 2024 challenge consists of two main components: a large single-plant training set and an unlabeled test set of vegetation plot images. The training metadata, provided in PlantCLEF2024_single_plant_training_metadata.csv, contains over 1.4 million images associated with $1,408,033$ unique entries. Each entry has the following key attributes:

- **image_name:** name of the image file
- **species_id:** numeric ID representing the plant species
- **organ:** visible plant part in the image (flower, habit, or other)
- **obs_id:** identifier for the observation instance
- **Metadata:** Geo-location (latitude, longitude, altitude) and contributor-related fields (author, partner, license).

A separate file provides the complete list of $7,800$ unique species IDs used for classification.The test set includes $2,105$ high-resolution images of vegetation plots. Unlike the training set, which

contains single-species images, the test set presents a multi-label setting, where each image may contain multiple species. This domain shift makes the task significantly more complex, requiring robust feature extraction and effective aggregation strategies.

3.2. Proposed System

The overall flow diagram of the proposed system is given in Figure 1. It consists of four main stages: data downloading, preprocessing, modelling, and inference. The input images are first collected from the PlantCLEF 2025 dataset. These images then undergo preprocessing to standardize the format and ensure consistency. In the modelling phase, a pretrained DINOv2 ViT-B/14 model is used to extract features, and a classifier is trained using the [CLS] token representation. During inference, test images are passed through the trained classifier using two approaches — full image classification and grid-based classification.

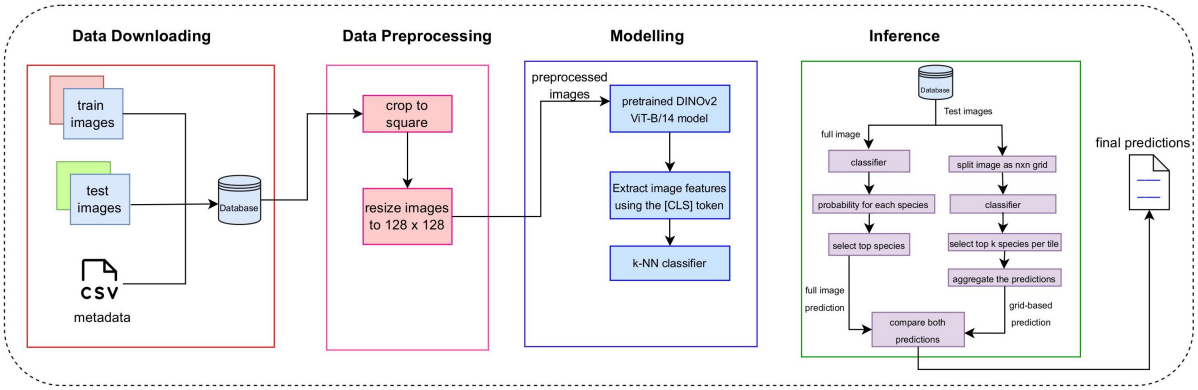

Figure 1: Overall flow diagram of the proposed solution

3.3. Data Preprocessing

Before training, the images in the single-plant dataset were preprocessed to reduce computational cost and ensure consistency in input format.

All the input images were cropped to a square and then resized to 128×128 pixels. This standardization helps maintain focus on the main subject while ensuring compatibility with batch processing and feature extraction. While this resizing may lead to the loss of fine details and reduce the effectiveness of advanced data augmentation techniques, it was chosen to reduce computational cost and training time for this initial exploration.

During the training phase, the preprocessed images are further used by the proposed framework. During the test phase, the vegetation plot images were divided into smaller square tiles using a grid-based approach (e.g., 3×3 or 8×8), and the same preprocessing steps were applied to each tile. Features were extracted from each tile individually, enabling the model to capture localized plant features within the high-resolution test plots. The proposed preprocessing pipeline significantly reduced the size of the dataset and made the training and inference process more efficient without compromising the quality of visual information.

3.4. Methodology

The overall process of the proposed framework is illustrated in Figure 1.
The proposed framework utilizes the preprocessed images which is of 128×128 pixel size. This helps standardize the input size for our model and reduces the overall storage and processing time. The fine-tuned DINOv2 ViT-B/14 model is used to extract image embeddings. This model produces a

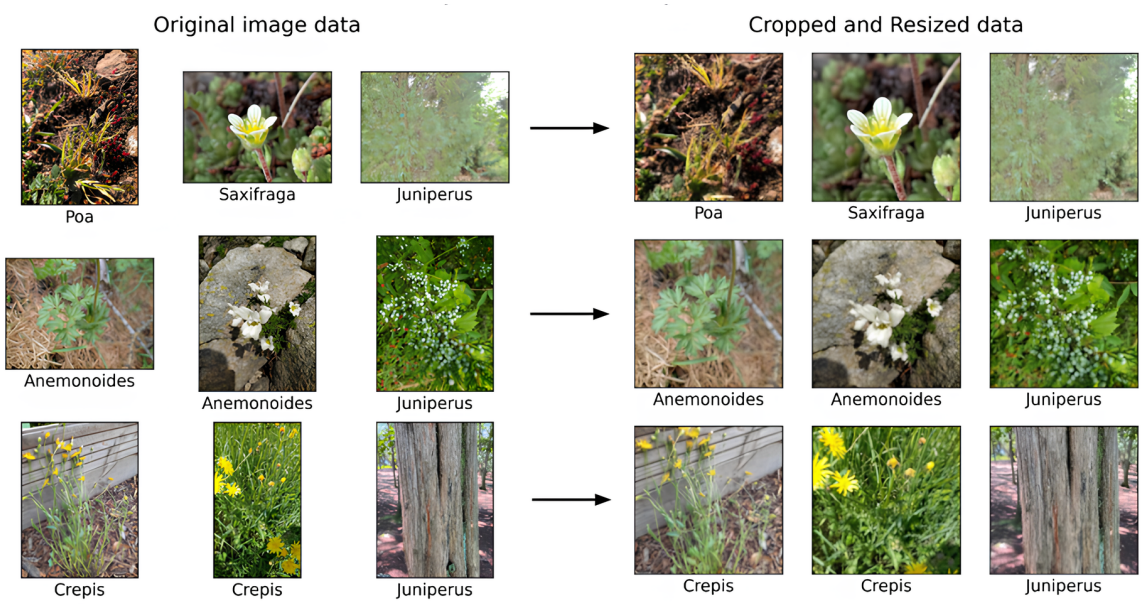

Figure 2: Comparison of original images with 128×128 cropped and resized squared images. The original images have a minimum resolution of 800 pixels on the longest side, allowing for the use of high-resolution classification models and potentially improving the prediction of small plants in large vegetative plots. Note: The labels shown represent plant genera, not exact species names.

768-dimensional embedding vector from the special [CLS] token, which summarizes the overall content of the image. Although the DINOv2 model was trained on 518×518 images, compatibility with our 128×128 inputs was ensured by enabling the dynamic_img_size option from the timm library.

All 1.4 million image embeddings from the training dataset were used without applying any indexing optimizations. Although this increased memory usage and inference time, it remained feasible for our experimentation. A simple k-NN classifier [12] is used by storing the extracted features to recognize the plant species. The k-NN approach was chosen because it can effectively identify multiple possible species in each tile based on similarity to training images. This also improves the chances of detecting smaller or less dominant plants present in localized regions of the image. A value of $k = 5$ was used as per the default setting of the scikit-learn k-NN classifier.

During test phase, each test image is divided into a nn grid of smaller parts. Each part is processed through the DINOv2 model to get predictions for the top k most likely species in that part. Finally, combine all the probability values from the n tiles by filtering the duplicates of the species ID and the final list of species for each image have been generated. This final output is submitted for evaluation.

Two distinct approaches such as full-image and grid-based image prediction were employed to evaluate the performance of the proposed framework as shown in Figure 3. The test dataset is not cropped and resized to preserve the high quality of multi-label images.

Full-Image Prediction: In this approach, the entire test image is processed in its original dimension. The fine-tuned ViT model evaluates the image and outputs probabilities for each of the 7806 plant species classes. Top 20 probabilities was then taken out, representing the most likely species present, and map these probabilities to their corresponding species IDs.

Grid-based Image Prediction: The test image was divided into an $N \times N$ grid of tiles, resulting in M tiles. Each tile is independently processed using the fine-tuned ViT model, which outputs probabilities for each species class. The top K probabilities (default $K = 10$) are mapped to their respective species classes for each tile. For example, a 3×3 grid yields nine tiles, each with ten top probabilities, selecting the top five species in each tile, totaling 45 species IDs and probability mappings.

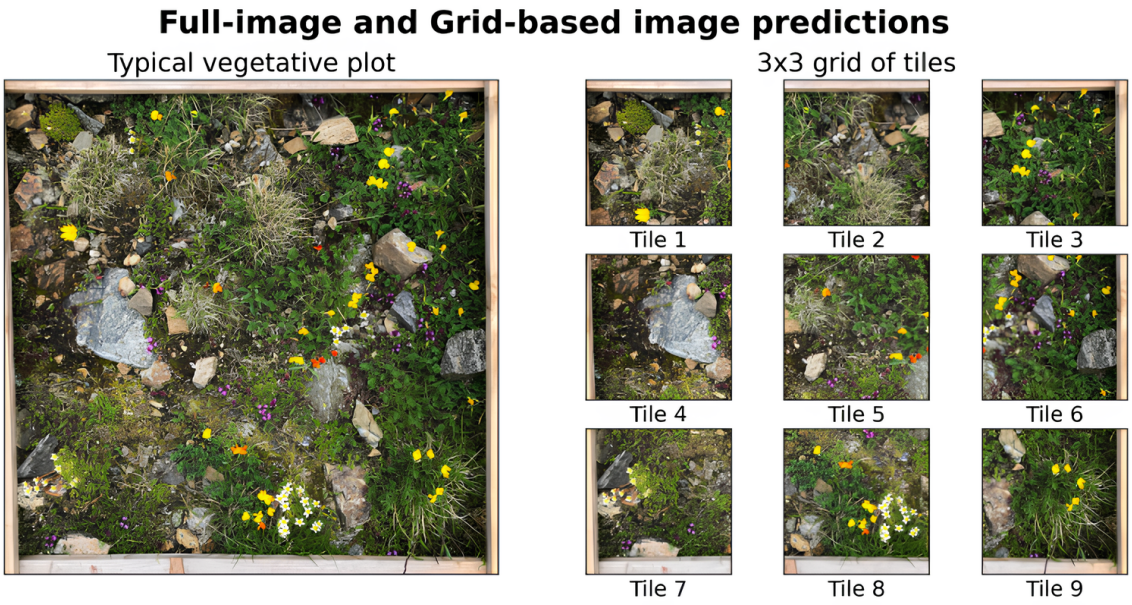

Figure 3: Comparison of full-image prediction and grid-based image prediction. The left plot shows a typical vegetative plot from the test set. The right plot illustrates the same image divided into a 3×3 grid, demonstrating the grid-based approach for species classification by processing each tile independently.

4. Results

The predictions obtained using the official pretrained DINOv2 model resulted in a macro-averaged F1 score of only 0.00012 per plot, which is relatively low. In contrast, the multi-class full-image inference method, using a linear classifier trained on fine-tuned DINOv2 embeddings, achieved a macro-F1 score of 0.00098, representing an $8\times$ improvement over the base pretrained model. For a more advanced approach, a grid-based inference method was adopted to enhance multi-label classification performance. Among the grid-based methods, using the top k species per tile achieved the highest scores across all metrics.

For inference with the fine-tuned model, a 3×3 grid size was found to strike a good balance between computational efficiency and spatial coverage of species distribution. We also experimented with 2×2 and 5×5 grid sizes but observed no substantial performance improvements, as shown in Table 1. A total of 5 species per tile were selected to maximize the macro-averaged F1 score per sample, focusing on capturing the most represented species in the dataset. Although this approach predicts up to 45 species per image, it was chosen to prioritize recall, even at the cost of lower precision.

Our final submission achieved a macro-averaged F1 score of 0.00106 on 89% of the test dataset, and 0.00293 on the remaining 11% of the dataset.

Table 1
An overview of inference methods and their Macro-F1 performance (averaged per test sample)

Inference Method	Prediction Strategy	Macro-F1 Score
Full-image	Top 5 species	0.00098
	Top 20 species	0.00078
Grid-based	Top 5 species, 3×3 grid	**0.00106**
	Top 3 species, 3×3 grid	0.00101
	Top 5 species, 5×5 grid	0.00083
	Top 5 species, 2×2 grid	0.00051

5. Conclusion

This work presents a lightweight multi-label plant species classification approach based on self-supervised Vision Transformer (DINOv2) models. A scalable solution is proposed for multi-label image classification utilizing only single-label training data. Future improvements may involve enhancing model training and inference by incorporating additional data augmentation techniques, experimenting with various grid sizes, exploring alternative dimensionality reduction methods, and employing loss functions such as binary cross-entropy and asymmetric loss [13]. Furthermore, the development of more sophisticated aggregation strategies for multi-label prediction holds potential to further improve classification performance.

Acknowledgments

We extend our gratitude to the developers of the CLEF2025 team for their pre-trained DINOv2 model and the organizers of PlantCLEF and LifeCLEF for hosting the competition.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: check grammar and spelling, paraphrase, and reword. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] 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.

[2] L. Picek, S. Kahl, H. Goëau, L. Adam, T. Larcher, C. Leblanc, M. Servajean, K. Janoušková, J. Matas, V. Čermák, K. Papafitsoros, R. Planqué, W.-P. Vellinga, H. Klinck, T. Denton, J. S. Cañas, G. Martellucci, F. Vinatier, P. Bonnet, A. Joly, Overview of lifeclef 2025: Challenges on species presence prediction and identification, and individual animal identification, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2025.

[3] A. Joly, L. Picek, S. Kahl, H. Goëau, V. Espitalier, C. Botella, B. Deneu, D. Marcos, J. Estopinan, C. Leblanc, T. Larcher, M. Šulc, M. Hrúz, M. Servajean, J. Matas, Overview of lifeclef 2024: Challenges on species distribution prediction and identification, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2024.

[4] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al., Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4015–4026.

[5] M. Gustineli, A. Miyaguchi, I. Stalter, Transfer learning for multi-label plant species classification with self-supervised vision transformers, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.

[6] S. Chulif, H. A. Ishrat, Y. L. Chang, S. H. Lee, Patch-wise inference using pre-trained vision transformers: Neuon submission to plantclef 2024, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.

[7] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., Dinov2: Learning robust visual features without supervision, arXiv preprint arXiv:2304.07193 (2023).

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[9] L. Wu, W. Zhang, T. Jiang, W. Yang, X. Jin, W. Zeng, [cls] token is all you need for zero-shot semantic segmentation, arXiv preprint arXiv:2304.06212 (2023).

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[11] 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, https://doi.org/10.5281/zenodo.10848263, 2024. doi:10.5281/zenodo.10848263.

[12] G. Guo, H. Wang, D. Bell, Y. Bi, K. Greer, Knn model-based approach in classification, in: R. Meersman, Z. Tari (Eds.), On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE (OTM 2003), volume 2888 of *Lecture Notes in Computer Science*, Springer, 2003, pp. 986–996. doi:10.1007/978-3-540-39964-3_64.

[13] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, L. Zelnik-Manor, Asymmetric loss for multi-label classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 82–91.