Tiles-wise Inference with Vision Transformers for Multispecies Identification in Vegetation Images*

Notebook for the LifeCLEF Lab at CLEF 2025

Andrea Menco-Tovar[1,*,†], Jairo E. Serrano[1,†], Juan Carlos Martinez-Santos[1,†] and Edwin Puertas[1,†]

[1]Universidad Tecnologica de Bolivar, Ternera km 1, Cartagena, Bolivar, Colombia

Abstract

This paper presents a method for classifying vegetation plots containing multiple species in the context of the PlantCLEF 2025 challenge. It addresses the simultaneous identification of various species in high-resolution images using a segment-based inference approach with the Vision Transformer (ViT) model, previously pre-trained using the self-supervised learning technique DINO V2. The photos were systematically divided into different patch configurations, to enable accurate classification. The results showed that the optimal configuration was the 4×2 patch, achieving a public average macro F1 score of 0.29096 and a private score of 0.28324, ranking 13th in the challenge. Errors are observed in cases involving visually similar species, unbalanced lighting conditions, and partial species presence within the evaluated tiles. Despite these limitations, the proposed methodology confirms the potential of the ViT model in complex ecological classification tasks, highlighting the importance of future developments to improve accuracy in highly complex contexts.

Keywords

Segmentation, multi-species identification, Vision Transformer, Ecological studies, Vegetation classification.

1. Introduction

Ecological monitoring systems based on biodiversity inventories are a fundamental tool for evaluating and managing ecosystems, enabling standardized sampling, long-term monitoring, and large-scale analysis. Vegetation plot inventories are essential for ecological studies, as they allow for standardized sampling, biodiversity assessment, long-term monitoring, and large-scale remote studies [1], [2]. These vegetation inventories provide key data for biological conservation and evidence-based environmental decision-making. Typically, inventories consist of multiple quadrat, each approximately 0.25 square meters in size, where botanists perform a thorough visual analysis by meticulously identifying all present species. However, these methods present notable limitations due to their high time cost and the need for specialized expertise, which restricts the frequency and coverage of ecological studies [3], [4].

The integration of Artificial Intelligence (AI) could significantly enhance the efficiency of specialists, expanding the scope and reach of ecological studies. In this context, the use of AI emerges as a promising solution to significantly improve the efficiency of environmental monitoring. It can also play a crucial role in protecting and conserving our environment. Additionally, it provides innovative solutions for biodiversity information systems, facilitating the online publication of data and information and supporting comprehensive biodiversity management in a timely and efficient manner [5]. Today, systems such as Pl@ntNet and iNaturalist play a crucial role in enabling users worldwide to generate, submit, and annotate botanical observations. They also assist scientists and resource managers in understanding where and when organisms occur. Collectively, these platforms have demonstrated

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

*Corresponding author.

†These authors contributed equally.

amenco@utb.edu.co (A. Menco-Tovar); jserrano@utb.edu.co (J. E. Serrano); jcmartinezs@utb.edu.co (J. C. Martinez-Santos); epuerta@utb.edu.co (E. Puertas)

0000-0002-6861-7547 (A. Menco-Tovar); 0000-0001-8165-7343 (J. E. Serrano); 0000-0003-2755-0718 (J. C. Martinez-Santos); 0000-0002-0758-1851 (E. Puertas)

the ability to identify individual species through isolated photographs taken by citizens and scientists, facilitating non-expert participation and large-scale data collection [6], [7].

Nonetheless, the simultaneous identification of multiple species within a single high-resolution image of vegetation plots (test images from the challenge) remains a significant technological challenge. Deep learning models applied to plant identification require large annotated datasets. The main challenge lies in the considerable disparity between available datasets. While training images typically represent single-labeled individual plant images, ecological plot images encompass diverse floristic contexts, such as those found in Pyrenean and Mediterranean floras. This challenge requires the development of robust models that can perform effective multi-label classifications in complex ecological settings.

In the PlantCLEF 2025 challenge, the evaluation focuses on multi-label prediction of plant species in high-resolution quadrat images, where multiple species may appear, though rarely dozens simultaneously. To measure participants' performance, used the average per-sample macro F1 score, which balances recall and precision, avoiding both overprediction (low precision) and underprediction (low recall). This F1 variant first computes the score for each image independently. Then, it averages the scores across all test set transects, sampling areas of approximately 10 m × 0.5 m at various sites, to mitigate bias from oversampled regions and ensure a fair comparison between approaches.

In general, vegetation classification requires careful examination and often involves a lengthy process. The PlantCLEF 2025 challenge aims to predict all plant species in high-resolution plot images, making it a multi-label classification task. This paper describes our approach and results based on our submissions to PlantCLEF 2025. Employed the pre-trained ViT provided for this task. Adopted patch-based inference on the test set using pre-trained models, dividing test images into non-overlapping tiles. The patch sizes experimented with were 1×1, 4×2, 3×3, 16×16, and 14×7. Our best submission achieved a public average per-plot macro F1 score of 0.29096 and a private average per-plot macro F1 score of 0.28324, ranking us 13th in the challenge.

2. Related Work

TThis section presents relevant contributions that have employed different architectures and strategies for species conservation and identification, demonstrating the potential of these methods in diverse application contexts.

It was found that, Yang et al. [8] studied the structure and dynamics of biodiversity in forest ecosystems using forest soundscape data. They applied a deep learning-based multi-label classification approach to field recordings, successfully automating the classification of sound sources into bioacoustics, geophony, and anthrophony. For this purpose, they designed a Convolutional Neural Network (CNN) consisting of a convolutional feature extraction module and a fully connected classification module, achieving macro F1 scores around 0.8421. In turn, Brun et al. [9] jointly mapped and modeled the distributions of 2,477 plant species using Deep Neural Networks (DNNs) to assess changes in species distributions, phenology, and dominance. They trained different versions of multi-species DNNs and emphasized that multi-species DNNs predict species distributions and, especially, community composition with higher accuracy, reporting AUC values around 0.954.

In the same line, Hu et al. [10] employed four deep learning models, including a multilayer perceptron (MLP), a CNN, a Vision Transformer (ViT), and a multimodal model to predict species and entire community distributions, aiming to provide quantitative data for conservation efforts. They tested the models on multispecies plant community images, obtaining macro-TSS metrics around 69.61% for the MLP and 71.35% for the multimodal model, the latter showing better inference and fewer false positives. Subsequently, Ghasemkhani et al. [11] presented Multi-label Federated Learning (FMLL) used to understand and manage animal populations for biodiversity conservation and ecological management. The proposed FMLL adopts a binary relevance strategy to handle the multi-label nature of the data and employs the reduced error pruning tree as a classifier, achieving accuracy rates ranging from 73.24% to 94.50% on animal datasets.

Continuing with the use of ViT models, the study by Elharrouss et al. [12] offers a comprehen-

sive review of these models, emphasizing their fundamental principles, including the self-attention mechanism and multi-head attention. It also highlights the versatility of these models in applications such as image classification, medical imaging, object detection, and visual question answering while acknowledging the challenges associated with their use, including high computational demands, extensive data requirements, and generalization difficulties. Additionally, Saha and Xu [13] highlight and explore understanding methods to optimize ViT models while also addressing their limitations. This review highlights the application of ViT in various tasks, including image classification, object detection, and segmentation. On the other hand, it highlights that despite their strong performance in terms of accuracy, these models have high computational costs, memory consumption, and energy usage.

On the other hand, Lefort et al. [6] described the Pl@ntNet system, which facilitates global data collection by allowing users to upload and annotate plant observations. They also noted that their proposed label aggregation method aims to train AI models collaboratively for plant identification, emphasizing the significant support that computer vision models provide in species recognition in the field. Similarly, Van et al. [7] introduced a species classification and detection dataset called iNaturalist and conducted experiments using classification and detection computer vision models, obtaining accuracy results of 67%, which are considerably good. They concluded that state-of-the-art computer vision models still have room for improvement when applied to large and imbalanced datasets.

Overall, the reviewed works demonstrate that the use of deep architecture-based models and well structured datasets, can significantly improve performance in species classification tasks. However, challenges, such as addressing multi-label classification scenarios, remain.

3. Methodology

This section details the training set used, which consists of individual observations organized by species, as well as the test set comprising multi-species vegetation plot images captured in various floristic contexts. Regarding the architecture, described the process of image partition, data preparation for analysis, and the mechanism used to obtain and filter the final predictions.

3.1. Dataset Description

This section describes the characteristics and composition of the data used for both training and evaluation. Additionally, it highlights the differences between both sets, as well as the challenges associated with variability in capture conditions and the complexity of the evaluated scenes.

3.1.1. Training Set

The training dataset consists of observations of individual plants, as shown in the Figure 1. These are a subset of the Pl@ntNet training data, focusing on southwestern Europe, and cover approximately 7,800 plant species. The challenge organizers state that some reliable labels for underrepresented species are completed using data from the GBIF platform. The images have relatively high resolution, with a minimum or maximum of 800 pixels on the longest side, which allows for the use of classification models capable of handling relatively high-resolution inputs and can reduce the difficulty of predicting small plants in large vegetative plot images. The images are pre-organized into subfolders by species to facilitate the training of individual plant classification models.

There is also a complementary set of images, composed of high-resolution pseudo-square photos, as shown in the Figure2, without labels, made available to participants to improve model adaptation to quadrat images of multi-species vegetation.

Figure 1: Shows representative examples of individual plant images used in the training set, organized by species.

Figure 2: Examples of high-resolution, unlabeled complementary images available to improve model adaptation to square images of multi-species plant plots.

3.1.2. Test Set

The test set comprises various datasets of plot images in different floristic contexts, including Pyrenean and Mediterranean flora, as illustrated in Figure 3. Experts created these datasets, which comprise a total of 2,105 high-resolution images. The capture protocol may vary significantly depending on the context, including the use of wooden frames or measuring tape to delimit the plot, as well as the viewing angles relative to the ground. Moreover, image quality may vary depending on weather conditions, which can result in more or less pronounced shadows, blurred areas, and other effects. The main challenge of the task lies in the domain shift between the training and test datasets. Unlike the training images, which consist of observations of individual plants, the test images include multiple species within a single image. Additionally, there are images containing withered plants or plants at different stages of growth, as well as images with rocks, sand, and moss.

Figure 3: Examples of images of plant plots from the test set containing multiple species in different floristic contexts and capture conditions.

3.1.3. Architecture

We worked with the pre-trained model `vit_base_patch14_reg4_dinov2.lvd142m` provided by the competition, which they based on a ViT architecture pre-trained using the self-supervised learning approach DINO V2 with 142 million images. Initially, photos from the test set, composed of high-resolution vegetation plot images, were loaded. Subsequently, files such as `class_mapping`, which maps the model's output identifiers to species identifiers, and `load_species_mapping`, which maps species identifiers to recognized scientific names, were loaded. Figure 4 illustrates the entire workflow.

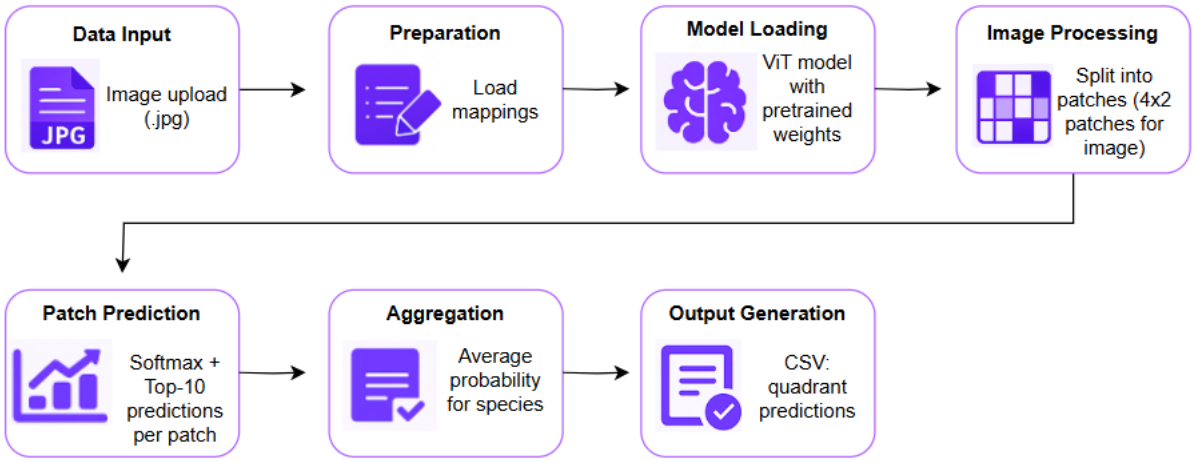

Figure 4: General diagram of the system workflow, showing the steps from image upload to generating and filtering predictions by species.

In our approach, each image underwent a systematic spatial division process. A uniform partition was applied to divide each original image into eight parts, arranged in four rows by two columns. Performed this partition to obtain a more detailed and precise representation of the relevant visual features present in each image, aiming to leverage the approach the authors used to train the ViT model. Individually processed each resulting partition through a specific transformation that prepared

the images for further analysis. These transformations ensured that each portion met the appropriate dimensional and normalization requirements for the ViT model.

After being transformed, evaluated the images using the selected ViT model, which generated predictions in the form of probabilities associated with various possible species for each segment. From these probabilities, selected only those with the highest values, and performed a statistical aggregation to determine the relative relevance of each predicted species within the total set of evaluated tiles. Finally, it was established as a criterion that only those species whose average probabilities exceeded 5% would be considered valid for the final classification. This average probability threshold of 5% was deliberately selected as a low value to avoid discarding species that, although appearing with low confidence in the predictions, could actually be present in the images. This approach aims to maximize the model's sensitivity in the multi-species context, where partial presence or limited visibility of some plants may generate low but relevant probabilities. These species were subsequently ranked according to the confidence obtained.

4. Experimental Results

The evaluation metric used by the organizers was the F1 score, designed to balance recall and precision, ensuring that models neither overpredict nor underpredict species. Used the average per-sample macro F1 score. Divided the results into public and private scores, and based the final rankings on the private split. Table 1 shows the official public and private results of our submission. Report the average per-sample macro F1 score and the overall ranking of our system, as well as the top-performing team for comparison.

Table 1
Public and private F1 scores and ranks for teams

Team	F1 Public	Rank	F1 Private	Rank
Task Best System	0.35900	3/45	0.36479	1/45
VerbaNexAI (4X2)	0.29096	13/45	0.28324	13/45

Our system achieved a public average per-sample macro F1 score of 0.29096 and a private score of 0.28324, ranking 13th out of 45 participating teams. Compared these results with the baseline system, which led the competition with a public average per-sample macro F1 score of 0.35900, ranking 3rd out of 45 teams, and a private score of 0.36479, ranking 1st out of 45 teams.

Table 2
Impact of patch partitioning on model performance.

Tiles	F1 Public	F1 Private
3x3	0.28371	0.25970
14x7	0.26733	0.26275
4x2	0.29096	0.28324

An increase in performance was observed when using the same model with varying numbers of tiles 4×2, 3×3, and 14×7 as shown in Table 2. This behavior can be attributed to the fact that the 4×2 patch division allows capturing sufficient distinct regions of the image while maintaining an adequate resolution in each sub-image processed by the model. In contrast, the 14×7 sampling could introduce redundancy or a loss of useful resolution in smaller tiles, and the 3×3 division might be insufficient to efficiently represent areas with higher species density. It is important to highlight that the images were divided into tiles through a uniform partition in rows and columns, resulting in rectangular tiles whose pixel dimensions depend on the original size of each image and the chosen division configuration.

No additional cropping or padding was applied in order to preserve the spatial proportion and content of each patch. The tile configuration was primarily illustrated using the 4×2 format, chosen for offering the best performance in the conducted experiments. Although this configuration produces

rectangular tiles, the spatial integrity of the images was maintained by avoiding cropping, padding, or stretching to force square tiles. While the main comparison reported results with non-overlapping patch configurations, overlapping patch configurations were also explored. However, these configurations did not improve the model's performance in preliminary experiments and, in some cases, introduced redundancy that negatively affected generalization.

During inference on the test set, the average time per image was approximately 1.67 seconds, measured in a computing environment with a GPU on the Google Compute Engine backend using Python 3. Throughout this process, resource usage remained stable, with an average system RAM consumption of 4.4 GB out of 12.7 GB available, 3.6 GB of GPU memory used from 15 GB available, and a steady disk usage of approximately 43.5 GB out of a total 112.6 GB. This indicates that the employed methodology is feasible for applications with moderate demands in terms of processing time and computational resource consumption. However, it is noteworthy that partitioning images into multiple tiles increases computational cost, making it relevant to explore optimization techniques in future work to improve efficiency without sacrificing accuracy.

Heatmaps

The heatmaps Figure 5 reveal the model's maximum confidence in each patch of the evaluated images, allowing spatial identification of areas where the model shows higher or lower certainty in its predictions. It is observed that regions with high confidence, represented by warm colors, generally coincide with areas where species are clearly visible and less occluded, while low confidence zones, indicated by cool colors, correspond to sectors with shadows, dense vegetation, or visual noise factors that hinder precise detection. These findings open the possibility of focusing future improvements on problematic areas through specific techniques aimed at increasing the model's robustness under adverse conditions.

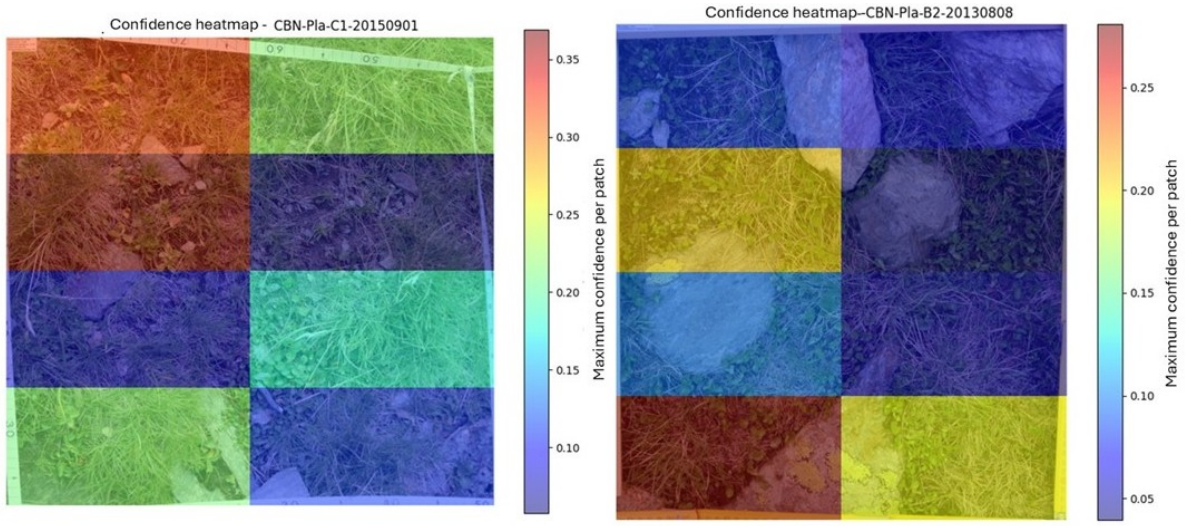

Figure 5: Heatmap visualization of the model's confidence across image tiles.

5. Conclusions

This paper presents our work on multi-species vegetation plot classification using the PlantCLEF 2025 dataset. Considering the challenging task involving high-resolution test images of vegetation plots, implemented patch-based inference by dividing each plot into sizes of 1×1, 4×2, 3×3, 16×16, and 14×7, reducing the task from multi-species identification to the prediction of one or a few classes. Achieved our best submission using the 4×2 patch, reaching a public average per-sample macro F1 score of 0.29096 and a private score of 0.28324, ranking 13th.

Additionally, observed that the model's errors tend to concentrate on morphologically similar species or those partially present within the analyzed tiles. In other words, the model tends to confuse species that share close visual features, especially in contexts of high vegetation density where leaves overlap or appear only partially. Also identified errors in images with unbalanced lighting conditions or slight blurriness, which affects the quality of the extracted representations. In some cases, there was an overestimation of dominant species from the training set, indicating a possible imbalance in the class distribution learned by the model. Nevertheless, ViT has once again proven to be a competitive option for plant identification in the context of PlantCLEF. However, further research is needed to address resource limitations and fully leverage this robust architecture in vegetation-related tasks such as plot classification.

Finally, in future work, propose expanding the availability of labeled datasets with images of vegetation plots that include diverse floristic contexts and varying capture conditions. It would enable the training of more robust models adapted to complex multi-species scenarios. Additionally, it would be relevant to investigate advanced techniques in semantic segmentation and object-based segmentation to improve the accurate identification of individual species within each patch. It would also be feasible to evaluate adaptive image patching methods based on criteria such as plant density or spatial distribution to maximize the quality of extracted visual features and, consequently, model accuracy. Finally, it is essential to examine the impact of various semi-supervised and collaborative labeling strategies, which could significantly contribute to reducing the cost of generating annotated data and enhancing prediction quality in ecological classification tasks.

Generative AI Declaration

During the preparation of this work, ChatGPT was used to review translation, grammar, and spelling. After using this tool, the content, coherence, and cohesion were reviewed and edited as necessary, and full responsibility for the content of the publication was ultimately assumed.

Acknowledgments

The authors express their gratitude to the Call 933 "Training in National Doctorates with a Territorial, Ethnic and Gender Focus in the Framework of the Mission Policy — 2023" of the Ministry of Science, Technology and Innovation (Minciencia). In addition, we thank the team of the Artificial Intelligence Laboratory VerbaNex [1], affiliated with the UTB, for their contributions to this project.

References

[1] D. R. A. de Almeida, L. B. Vedovato, M. Fuza, P. Molin, H. Cassol, A. F. Resende, P. M. Krainovic, C. T. de Almeida, C. Amaral, L. Haneda, R. W. Albuquerque, E. Gorgens, J. Romanelli, M. Ferreira, C. Salk, N. Espinoza, C. Silva, E. Broadbent, P. H. S. Brancalion, Remote sensing approaches to monitor tropical forest restoration: Current methods and future possibilities, Journal of Applied Ecology 62 (2025) 188–206. doi:10.1111/1365-2664.14830.

[2] D. Sánchez-Fernández, A. Jiménez-Jiménez, R. Fox, R. L. H. Dennis, J. M. Lobo, Identifying biodiversity hotspots over time: Stability, sampling bias, and conservation implications, Global Ecology and Conservation (2025) e03586. doi:10.1016/j.gecco.2025.e03586.

[3] 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.

[1] https://github.com/VerbaNexAI/CLEF2025/tree/main/LifeCLEF/Notebook

[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] J. Sebastián, C. Nas, C. Parra-Guevara, M. Montoya-Castrilión, J. M. Ramírez-Mejía, G.-A. Perilla, E. Marentes, N. Leuro, J. V. Sandoval-Sierra, S. Martinez-Callejas, A. Díaz, M. Murcia, E. A. Noguera-Urbano, J. M. Ochoa-Quintero, S. Rodríguez Buriticá, J. Sebastián Ulloa, Inteligencia artificial para la conservación y uso sostenible de la biodiversidad, una visión desde colombia (artificial intelligence for conservation and sustainable use of biodiversity, a view from colombia), https://arxiv.org/pdf/2503.14543v2, 2025.

[6] T. Lefort, A. Affouard, B. Charlier, J.-C. Lombardo, M. Chouet, H. Goëau, J. Salmon, P. Bonnet, A. Joly, Cooperative learning of pl@ntnet's artificial intelligence algorithm: how does it work and how can we improve it?, 2024. doi:10.1111/2041-210X.14486.

[7] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, S. Belongie, The inaturalist species classification and detection dataset, http://arxiv.org/abs/1707.06642, 2017.

[8] C. Yang, X. Liu, Y. Li, X. Yu, Deep learning-based multi-label classification for forest soundscape analysis: A case study in shennongjia national park, Forests 2025, Vol. 16, Page 899 16 (2025) 899. URL: https://www.mdpi.com/1999-4907/16/6/899/htmhttps://www.mdpi.com/1999-4907/16/6/899. doi:10.3390/F16060899.

[9] P. Brun, D. N. Karger, D. Zurell, P. Descombes, L. C. de Witte, R. de Lutio, J. D. Wegner, N. E. Zimmermann, Multispecies deep learning using citizen science data produces more informative plant community models, Nature Communications 15 (2024) 1–15. URL: https://www.nature.com/articles/s41467-024-48559-9. doi:10.1038/S41467-024-48559-9;TECHMETA=141;SUBJMETA=158,2668,449,631,670,851;KWRD=BIODIVERSITY,MACROECOLOGY,PLANT+ECOLOGY.

[10] Y. Hu, S. Si-Moussi, W. Thuiller, Introduction to deep learning methods for multi-species predictions, Methods in Ecology and Evolution 16 (2024) 228–246. URL: /doi/pdf/10.1111/2041-210X.14466https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.14466https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14466. doi:10.1111/2041-210X.14466;REQUESTEDJOURNAL:JOURNAL:2041210X;WEBSITE:WEBSITE:BESJOURNALS;WGROUP:STRING:PUBLICATION.

[11] B. Ghasemkhani, O. Varliklar, Y. Dogan, S. Utku, K. U. Birant, D. Birant, Federated multi-label learning (fmll): Innovative method for classification tasks in animal science, Animals 2024, Vol. 14, Page 2021 14 (2024) 2021. URL: https://www.mdpi.com/2076-2615/14/14/2021/htmhttps://www.mdpi.com/2076-2615/14/14/2021. doi:10.3390/ANI14142021.

[12] O. Elharrouss, Y. Himeur, Y. Mahmood, S. Alrabaee, A. Ouamane, F. Bensaali, Y. Bechqito, A. Chouchane, Vits as backbones: Leveraging vision transformers for feature extraction, Information Fusion 118 (2025) 102951. doi:10.1016/J.INFFUS.2025.102951.

[13] S. Saha, L. Xu, Vision transformers on the edge: A comprehensive survey of model compression and acceleration strategies, Neurocomputing 643 (2025) 130417. doi:10.1016/J.NEUCOM.2025.130417.