Overview of LifeCLEF 2024: Challenges on Species Distribution Prediction and Identification

Alexis Joly, Lukáš Picek, Stefan Kahl, Hervé Goëau, Vincent Espitalier, Christophe Botella, Diego Marcos, Joaquim Estopinan, Cesar Leblanc, Théo Larcher, et al.

Overview of LifeCLEF 2024: Challenges on Species Distribution Prediction and Identification

Alexis Joly[1], Lukáš Picek[1,8], Stefan Kahl[6,11], Hervé Goëau[2], Vincent Espitalier[2], Christophe Botella[1], Diego Marcos[1], Joaquim Estopinan[1,12] Cesar Leblanc[1], Théo Larcher[1], Milan Šulc[10], Marek Hrúz[8], Maximilien Servajean[7], Hervé Glotin[3], Robert Planqué[4], Willem-Pier Vellinga[4], Holger Klinck[6], Tom Denton[9], Ivan Eggel[5], Pierre Bonnet[2], Henning Müller[5]

[1] Inria, LIRMM, Univ Montpellier, CNRS, Montpellier, France
[2] CIRAD, UMR AMAP, Montpellier, Occitanie, France
[3] Univ. Toulon, Aix Marseille Univ., CNRS, LIS, DYNI team, Marseille, France
[4] Xeno-canto Foundation, The Netherlands
[5] Informatics Institute, HES-SO Valais , Sierre, Switzerland
[6] K. Lisa Yang Center for Conservation Bioacoustics, Cornell Lab of Ornithology, Cornell University, USA
[7] LIRMM, AMIS, Univ Paul Valéry Montpellier, Univ Montpellier, CNRS, France
[8] Department of Cybernetics, FAV, University of West Bohemia, Czechia
[9] Google Research, San Francisco, USA
[10] Second Foundation, Prague, Czech Republic
[11] Chemnitz University of Technology, Chemnitz, Germany
[12] Laboratoire d'Ecologie Alpine, Grenoble, France

Abstract. Biodiversity monitoring using machine learning and AI-based approaches is becoming increasingly popular. It allows for providing detailed information on species distribution and ecosystem health at a large scale and contributes to informed decision-making on environmental protection. Species identification based on images and sounds, in particular, is invaluable for facilitating biodiversity monitoring efforts and enabling prompt conservation actions to protect threatened and endangered species. The multiplicity of methods developed, however, makes it important to evaluate their performance on realistic datasets and using standardized evaluation protocols. The LifeCLEF lab has been setting up such evaluations since 2011, encouraging machine learning researchers to work on this topic and promoting the adoption of the technologies developed by stakeholders. The 2024 edition proposes five data-oriented challenges related to the identification and prediction of biodiversity: (i) BirdCLEF: bird call identification in soundscapes, (ii) FungiCLEF: revisiting fungi species recognition beyond 0-1 cost, (iii) GeoLifeCLEF: remote sensing based prediction of species, (iv) PlantCLEF: Multi-species identification in vegetation plot images, and (v) SnakeCLEF: revisiting snake species identification in medically important scenarios. This paper overviews the motivation, methodology, and main outcomes of those five challenges.

1 LifeCLEF Lab Overview

Accurately identifying organisms observed in the wild is an essential step in ecological studies. It forms the foundation for understanding species interactions, population dynamics, and ecological processes, allowing researchers to accurately assess biodiversity, track changes over time, and make informed management and conservation decisions. However, observing and identifying living organisms requires high levels of expertise. For instance, vascular plants alone account for more than 300,000 different species and the distinctions between them can be quite subtle. The worldwide shortage of trained taxonomists and curators capable of identifying organisms has come to be known as the *taxonomic impediment*. Since the Rio Conference of 1992, it has been recognized as one of the major obstacles to the global implementation of the Convention on Biological Diversity. In 2004, Gaston and O'Neill [20] discussed the potential of automated approaches for species identification. They suggested that if the scientific community were able to (i) produce large training datasets, (ii) precisely evaluate error rates, (iii) scale-up automated approaches, and (iv) detect novel species, then it would be possible to develop a generic automated species identification system that would open up new vistas for research in biology and related fields.

Since the publication of [20], automated species identification has been widely studied [14,40,58,65,69] and is now a key technology in most citizen science monitoring apps, e.g., iNaturalist, eBird, and Pl@ntNet [6]. Nevertheless, the development of new approaches continues to expand rapidly, in particular for processing new types of data such as passive sensors, camera traps, or autonomous vehicles [16,70,76]. Biodiversity monitoring through AI approaches is now recognized as a key solution to collect and analyze vast amounts of data from various sources, enabling us to gain a comprehensive understanding of species distribution, abundance, and ecosystem health [2,5]. This information is essential for making informed conservation decisions and identifying areas needing protection.

To measure progress of AI-assisted biodiversity monitoring in a sustainable and repeatable way, the LifeCLEF virtual lab was created in 2014 as a continuation and extension of the plant identification task that had been run within the ImageCLEF lab since 2011 [23,24,25]. Since 2014, LifeCLEF has expanded the challenge by considering animals and fungi in addition to plants and including audio and video content in addition to images [30,31,32,33,34,35,36,37,38,39]. Nearly a thousand researchers and data scientists participate yearly to LifeCLEF to analyze the data, submit predictions and benefit from the shared evaluation tools. The aim of this paper is to present the synthesis of the 2024th edition of LifeCLEF, which comprises five challenges synthesized in Table 1.

The systems used to run the challenges (registration, submission, leaderboard, etc.) were the Kaggle platform for the BirdCLEF and GeoLifeCLEF, and the Hugging Face for the PlantCLEF, SnakeCLEF, and FungiCLEF challenges. Four of the challenges (GeoLifeCLEF, SnakeCLEF, PlantCLEF, and FungiCLEF) were organized jointly with FGVC 11, an annual workshop dedicated to Fine-Grained Visual Categorization, held in conjunction with the CVPR in-

Table 1: **LifeCLEF challenges data overview**. The provided datasets vary in modality, size, and complexity as each challenge addresses different aspects of automated species identification.

	Modality	**Species**	**Items**	**Task**	**Metric**
BirdCLEF	audio	182	25K	Multi-label classification	ROC-AUC
SnakeCLEF	images metadata	1,784	190K	Classification	ad-hoc metric
FungiCLEF	images metadata	4,759	400K	Open-set classification	ad-hoc metric
PlantCLEF	images (SD+HD)	7,806	1.4M	Multi-label classification	Samples F1
GeoLifeCLEF	sat. images time-series tabular	10,358	6.6M	Multi-label classification	Sample-Average F1

ternational conference on computer vision and pattern recognition.
In total, 1277 data scientists or research teams participated in the LifeCLEF 2024 edition by submitting runs to at least one of the five challenges (1198 only for the BirdCLEF challenge). Only some of them managed to get the results right, and 18 of them went all the way through the CLEF process by writing and submitting a *working note* describing their approach and results (for publication in CEUR-WS proceedings. In the following sections, we provide a synthesis of the methodology and main outcomes of each of the five challenges. More details can be found in the extended overview reports of each challenge and in the individual working notes of the participants (references provided below).

2 BirdCLEF Challenge: Bird call identification in soundscapes

A detailed description of the challenge and a more complete discussion of the results can be found in the dedicated working note [41].

2.1 Objective

Birds are vital indicators of biodiversity change due to their mobility and diverse habitat requirements. Changes in bird species assemblage and numbers can signal the success or failure of restoration projects. Traditional observer-based bird surveys over large areas are expensive and logistically challenging. Passive acoustic monitoring (PAM) combined with machine learning enables conservationists to sample larger areas with higher temporal resolution, providing detailed insights into the relationship between restoration efforts and biodiversity.

The Western Ghats, a Global Biodiversity Hotspot along India's southwestern coast, support extraordinary biodiversity across various ecosystems, from high-elevation forest-grassland mosaics to wet-evergreen rainforests. This region also sustains large human populations relying on forest resources. The Western Ghats host a high diversity of bird species, including many endemic and endangered species. However, significant landscape and climatic changes are threatening this biodiversity, highlighting the need for advanced conservation tools to rapidly assess and monitor bird diversity. The competition aims to identify endemic bird species in the Western Ghats' sky-islands using soundscape data, detect and classify endangered bird species with limited training data, and detect and classify poorly understood nocturnal bird species.

2.2 Dataset

We built on the experience from previous editions and adjusted the task to encourage participants to focus on task-specific model designs. We carefully selected training and test data to match this objective. As in previous iterations, Xeno-canto was the primary source for training data, while expertly annotated soundscape recordings were used for testing. We emphasized bird species that are typically underrepresented in large bird sound collections, such as those that are ecologically important but difficult to train a classifier due to their rare or elusive nature. However, we also included common species to allow participants to train effective recognition systems. To find suitable test data, we considered various sources with differing complexities, such as call density, chorus, signal-to-noise ratio, and man-made sounds, as well as quality differences like mono and stereo recordings. This year, we also included unlabeled training data similar to the test data, enabling participants to explore alternative training methods such as self-supervised learning.

2.3 Evaluation Protocol

The challenge was hosted on Kaggle, maintaining an evaluation mode similar to previous iterations with hidden test data and a code competition format. We used a version of macro-averaged ROC-AUC that skips classes without true positive labels as the metric. This approach allowed us to assess system performance independent of fine-tuned confidence thresholds, emphasizing per-species performance rather than per-sample performance. Participants were tasked with identifying species from short audio segments extracted from labeled soundscape data. We used 5-second segments, balancing typical signal length with sufficiently long context windows. The dataset size was kept reasonably small (<50 GB) and easy to process. Additionally, we provided introductory code repositories and write-ups to lower the entry barrier for the competition.

2.4 Participants and Results

A total of 1,198 participants, organized into 974 teams, participated in the BirdCLEF 2024 challenge, submitting more than 30,000 runs. Figure 1 shows the

performance of the top 25 runs. The primary metric was the private leaderboard score, revealed after the submission deadline to prevent probing of the hidden test data. Throughout the competition, participants were able to see their public score, which was calculated based on 35% of the test data.

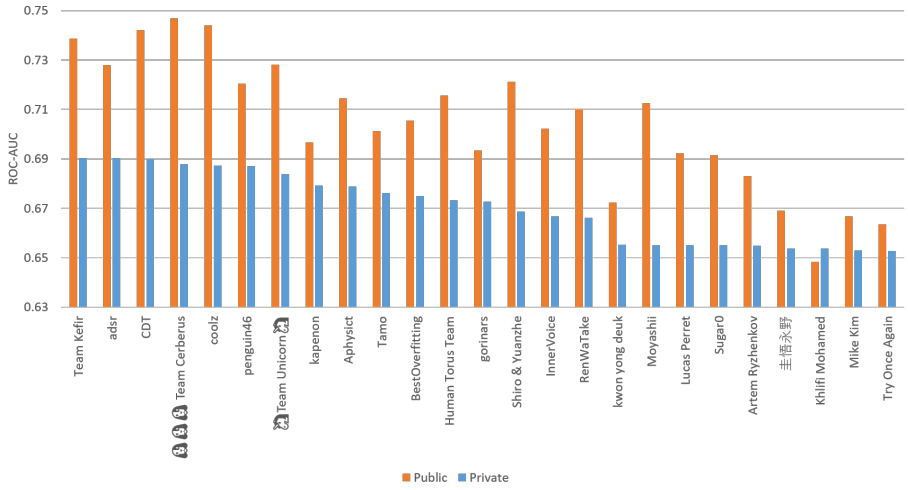

Fig. 1: BirdCLEF 2024 results. Top 25 teams sorted by private leaderboard score.

The baseline score in this year's edition was 0.5 (due to the metric) with random confidence scores for all birds for all segments. The best submission achieved a score of 0.690 (public 0.738) and the top 10 best performing systems were within only 1.5% difference in score. The majority of methods employed ensembles of convolutional neural networks, differing primarily in their pre- and post-processing techniques and the neural network backbones they used. Top participants leveraged unlabeled soundscape data to enhance their scores and adapt to the test data's acoustic domain. Given the restricted CPU runtime for submissions, participants prioritized speeding up model inference and using efficient architectures, with EfficientNet backbones being particularly popular. Additionally, participants explored ONNX and openVINO to further boost model inference speed. More details about the methods employed and the analysis of the results can be found in the detailed report of the task [41] and in the individual working notes of participants.

3 GeoLifeCLEF Challenge: Species composition prediction with high spatial resolution at continental scale using remote sensing

Comprehensive details on the challenge and an extensive discussion of the results are available in the dedicated working note [51].

3.1 Objective

Predicting species presence within specific areas is crucial for ecological research and biodiversity conservation. Accurate predictions support decisions related to protecting endangered species, land use planning, establishing protected zones, and developing sustainable agricultural practices. Nonetheless, species distributions are often influenced by intricate local variables that are difficult to quantify, such as interactions between populations, landscape connectivity, historical habitat conditions, and biases in data collection methods. Traditional ecological models often struggle with these complexities, resulting in predictions with limited spatial resolution. Furthermore, many species are underrepresented due to sampling biases. GeoLifeCLEF addresses these challenges by evaluating models on a vast scale, encompassing thousands of species, achieving spatial resolutions up to 10 meters, and leveraging millions of occurrence data points.

3.2 Dataset

The GeoLifeCLEF 2024 dataset contains species observation data, including presence-only occurrences and presence-absence surveys, alongside various environmental predictors. The dataset provides diverse environmental rasters, Sentinel2 satellite images, a 20-year climatic time series, and satellite time-series point values. Following on the work and dataset provided in the previous edition [4], we took most of the already provided Presence-Only (PO) occurrences (5 million) but tripled the Presence-Absence (PA) survey records to 90 thousand. Same as last year, the presence-absence data was split into training and test sets (95/5) using a spatial block hold-out procedure [63] with a spatial grid with 10×10km cells enabling comprehensive model evaluation. The test cells were randomly selected to ensure balance in biogeographical regions. In addition to the raw data, we have provided all the environmental predictors as pre-extracted scalar values in separated CSV files. Furthermore, the time-series data were provided in a 3d cube format (as torch tensors). More details about the dataset are available in the dedicated working note [51].

3.3 Evaluation Protocol

Same as in the previous edition [51], the evaluation metric was selected as the sample averaged F1 score (F_1). The F_1-score serves as a metric to gauge the degree of agreement between the predicted and actual species composition observed

within a specific geographical area and timeframe. In the context of ecological surveys, such as those conducted in Protected Areas (PAs), each survey instance i is associated with a ground-truth set of labels Y_i, representing the plant species identified by experts within a defined grid. Given this setup, and a list of predicted labels $\widehat{Y}_{i,1}, \widehat{Y}_{i,2}, \ldots, \widehat{Y}_{i,R_i}$, the micro F_1-score can be computed as follows:

$$\mathrm{F}_1 = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + (FP_i + FN_i)/2}, \tag{1}$$

$$\text{where} \begin{cases} TP_i = \#\text{ of correctly predicted species, i.e., } |\widehat{Y}_i \cap Y_i|. \\ FP_i = \#\text{ of species predicted but not observed, i.e., } |\widehat{Y}_i \setminus Y_i|. \\ FN_i = \#\text{ of species not predicted but present, i.e., } |Y_i \setminus \widehat{Y}_i|. \end{cases} \tag{2}$$

This formulation encapsulates the precision and recall elements crucial for assessing the accuracy of predictive models in ecological studies.

3.4 Organizer's baselines

We provide a variety of weak and strong baselines for all participants to allow a good starting point, continual performance increase, and working with different modalities. Considering the significant extent to which this baseline's performance can be enhanced, we encouraged participants to experiment with various techniques, architectures, losses, etc. Below, we briefly describe all baselines:

Naive baselines. With the dense and numerous observation data, one can naively predict the species' presence by selecting a set of the most common species within administrative or bio-geographical regions. For instance, predicting the top-25 most common species in the PA data results in a sample-averaged F_1 of 11.6%. Using the same approach but with the PO data results in an F_1 of 8.1%, showing a distribution shift between the two types of data.

Small Residual Convolutional Neural Networks for data cubes. Starting from a Resnet18 architecture, we have developed an even lighter model adapted to the small input size of GLC's cube data (respectively 19×12×4 for the climatic time series and 21×4×6 for the Landsat time series). When trained on the PA data with the Binary Cross Entropy loss (BCE), they achieved a sample-averaged F_1 score of respectively 0.259 and 0.266.

Swin tranformer for the Sentinel2 images. We slightly modified the architecture of a Swin-v2-t to allow input of all 4 modalities of Sentinel2 data (RGB+IR) rather than just three. It was also trained with the BCE loss on the PA data but resulted in a lower F_1 score of 0.235.

Multi-modal model. A multimodal model merging all three individual models mentioned above was developed using an MLP (Multi-Layer Perceptron) for the fusion head. It allows reaching an F_1 score of 0.316, demonstrating the task's inherent multimodality.

3.5 Participants and Results

51 Kaggle registrants participated in the GeoLifeCLEF 2024 challenge with at least one valid submission (submissions duplicated from the organizers's baselines were filtered out). A total of 1184 entries (i.e., *runs*) were submitted with an average of 23 entries per participant and a maximum of 175 for the participant who ranked first on the leaderboard. Details of the methods and systems used by the participants who submitted a working note are synthesized in the overview paper of the task [51] and described in more detail in the participant's working notes [8,9,11,44,47,67]. In Figure 2, we report the performance achieved by all participant's methods as well as the baseline methods developed by the organizers. Hereafter, we provide a short overview of the methods of the two best teams who submitted a working note (top2, top3, and top5 on the leaderboard):

AI2Lab team (Top2) [8]: This team started from the multi-modal model provided as the baseline by the organizers and made several significant improvements: (i) addition of a fourth modality (i.e., tabular environmental data encoded with an MLP), (ii) use of PO data samples through a pseudo-labeling procedure, (iii) use of an improved encoder for the Sentinel2 images (pre-trained with self-supervised learning on an external dataset), (iv) use of an ensemble of models optimized on different folds and, (v) optimization of the detection threshold. They finally got an F_1 score of 0.368 on the private leaderboard.

Miss Qiu (Top3) [44]: This team also started from the multi-modal model provided as a baseline but used an alternative fusion method based on cross-attention rather than MLP (which slightly improved the performance). They also introduced a number of improvements, some similar to the AI2Lab team (e.g., the use of an ensemble of k-fold models) and some different, such as (i) the enrichment of predictions with species frequent in neighboring PA and PO samples, (ii) the optimization of the number of returned species, or (iii) the use of various data augmentation techniques (including mixup). They finally got an F_1 score of 0.353 on the private leaderboard.

BernIgen (Top5) [11]: This team first worked on a model using only tabular data based on the XGBoost method (known to work very well on classical species distribution models). They have previously reduced the dimensionality of the input data with a PCA (Principal Component Analysis) and also the number of output species by keeping only the most likely species (about 10% of the species). This model alone already delivers pretty good performance (F_1 score of 0.31). They improved prediction performance by adaptively predicting the number of species to return for each test plot using a regression model (also based on XGBoost). This allowed gaining one more point of F_1 score. Finally, they combined this model with the multi-modal model provided by the organizers and got an F_1 score of 0.349 on the private leaderboard.

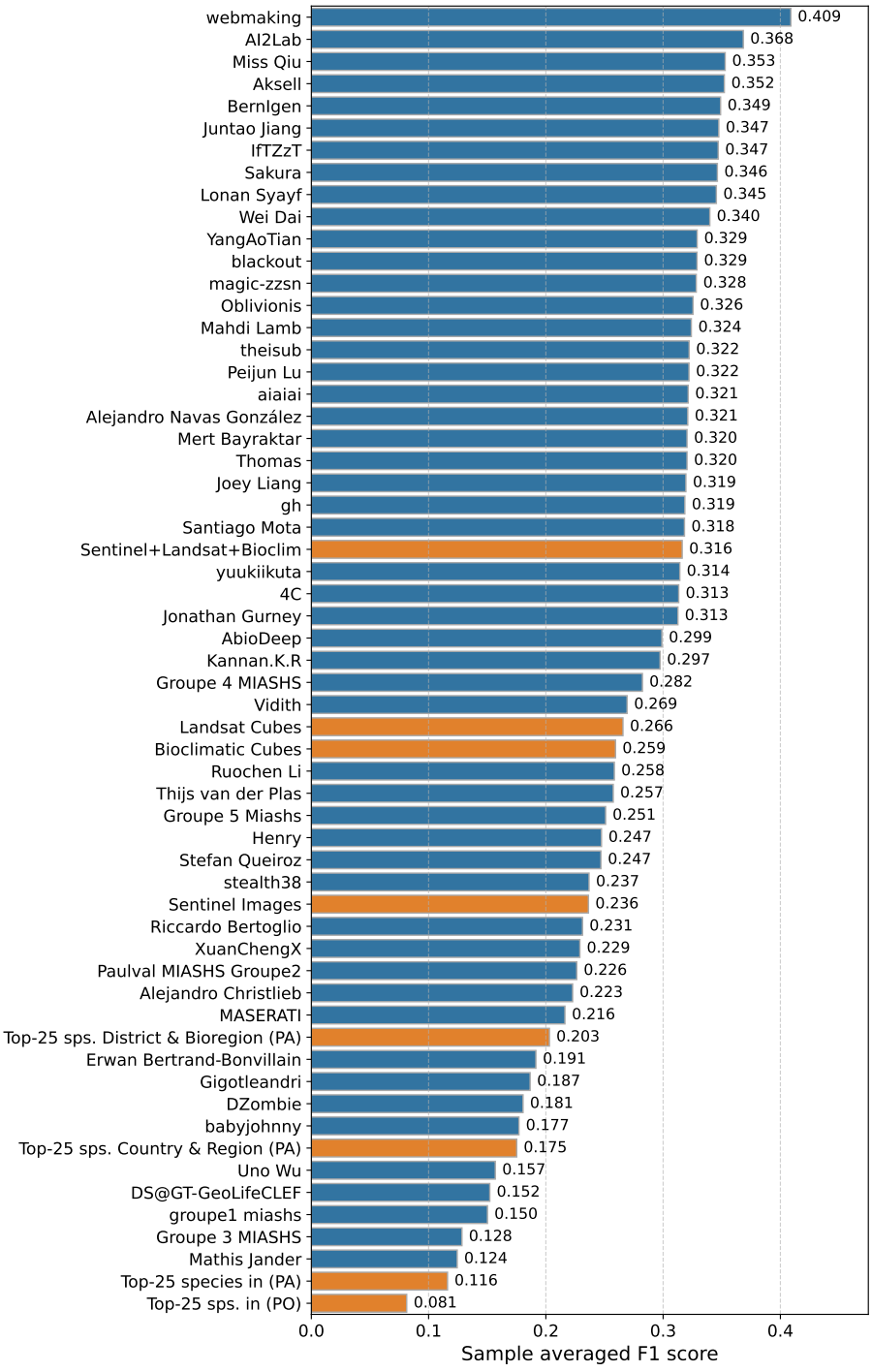

Fig. 2: GeoLifeCLEF 2024 results. All 51 teams. Orange depicts baselines.

3.6 Outcomes

The main outcomes we can derive from the GeoLifeCLEF 2024 are the following:

- Provided baselines had a positive impact on overall performance.
- Proactive engagement with the community and continual release of better baselines increases the impact.
- The use of multi-modal models with specific encoders for each modality is the main key to success.
- The provision of more PA data in the training dataset enabled much higher performance than last year's challenge (for which the best F_1 score was 0.27).
- Reciprocally, the use of PO data proves less beneficial, with only minor gains compared to models trained solely on presence/absence data.

For the future, it seems important to understand why improving performance with presence-only data is difficult, even though it is much larger. The presence of observation bias is clearly a plausible reason (some species are observed more than others), but it seems the spatial scale of the test set's plots may also be an issue. They are indeed quite small (10x10m on average) and do not necessarily reflect the presence of all the species in larger areas such as the one considered by the models. Moreover, the locations of these plots themselves follow specific protocols, which may introduce observation biases different from those of presence-only data.

4 FungiCLEF Challenge: Revisiting Fungi Species Recognition Beyond 0-1 Cost

Comprehensive details on the challenge and an extensive discussion of the results are available in the dedicated working note [55].

4.1 Objective

Efficient and scalable species recognition is crucial for large-scale initiatives like citizen science projects [56,66], which often operate with limited computational resources. In practice, accurate species identification relies on visual observations of the specimen and additional contextual data such as habitat, substrate, GPS coordinates, and temporal factors. This challenge sets a significant benchmark by integrating visual and contextual information, leveraging rich metadata, precise annotations, and standardized baselines available to all participants. Given that mushrooms are frequently foraged for consumption, the competition also addresses scenarios related to edible$\leftarrow\rightarrow$poisonous misclassifying.

The task requires participants to develop a classification model that generates a ranked list of predicted fungi species for each observation. Each observation includes multiple photographs of the same specimen and geographical location data. The classification model must comply with stringent constraints on memory usage and inference (prediction) time, specifically within a maximum of 120 minutes, using a dedicated HuggingFace server instance (Nvidia T4, 4 vCPUs, 15GB RAM, 16GB VRAM).

Table 2: FungiCLEF 2024 dataset statistics for each subset.

Subset	Species	→	Known/Unknown	Images	Observ.
Training	1,604		1,604 / –	295,938	177,170
Validation	3,299		1,084 / 1,629	91,231	45,021
Test	1,398		749 / 649	41,177	22,412
↪ *CzechFungi App*	137		94 / 43	393	215
↪ *Atlas of Danish Fungi*	1,261		721 / 540	40,784	22,197

4.2 Dataset

The FungiCLEF 2024 dataset builds upon the previous editions of the FungiCLEF [60,61] and the Danish Fungi 2020 dataset [57]. All the data is derived from a citizen science platform – the Atlas of Danish Fungi. Each fungi observation in this dataset has undergone expert validation, ensuring high-quality species labels. The dataset features rich observation metadata, i.e., information about habitat, substrate, timestamp, location, etc. Provided subsets (i.e., training, validation, and test) are briefly described below, and their statistics in detail are listed in Table 2.

The training set is based on 295,938 training images (177,170 observations) of 1,604 species. The dataset is built exclusively from the Danish Fungi 2020 data by combining the training and public test sets. This results in 295,938 training images across 1,604 species primarily observed in Denmark.

The validation set comprises expert-validated observations with species labels collected solely in 2022. This subset includes around 3,299 fungi species and contains 45,021 observations with many "unknown" species.

The test set is based on two subsets originating from two sources (e.g., Atlas of Danish Fungi and CheckFungi Application). and two countries, e.g., Denmark and Czechia. The CheckFungi is a small subset containing just around 200 submissions and is included primarily as a control set to prevent cheating. The test set was split 80/20 for public and private evaluation, respectively.

4.3 Evaluation Protocol

The task involves developing a classification model to predict species from a given set of real fungi observations accompanied by metadata. This model should adhere to a memory footprint constraint of a maximum of 1GB and prioritize minimizing risks to human safety, mainly by reducing misclassification between poisonous and edible species. The FungiCLEF 2024 challenge employed 2 decision-making scenarios, focusing on minimizing the empirical loss $L = \sum_i W(y_i, q(x_i))$, where $q(x)$ represents the decision rule for observations x, and y denotes the true labels. The cost function $W(y, q(x))$ was tailored for each scenario:

- **Track 1**: Standard classification incorporating an "*unknown*" category;
- **Track 2**: Penalization for edible and poisonous species confusion;
- **Track 3**: A user-centric loss combining Track1 and Track2;

4.4 Participants and Results

Seven teams participated in the FungiCLEF 2024 challenge; of these, six outperformed the baseline with EfficientNet-B1, and five submitted working notes. Details of the best methods and systems used are synthesized in the challenge overview paper [55] and further developed in participants working notes [7,10,18,68,73]. Achieved performance is reported in Figure 3. This year, the three tracks of FungiCLEF have three different best-performing submissions by three different teams:

The best-performing submission in Track 1 by *Jack Etheredge* [18] combined visual information with metadata using MetaFormer-0 and MetaFormer-2 [15] and further improved the ensemble by a vision-only CAFormer-S18 [75], and proposed a novel application of openGAN [43] for open-set recognition of fine-grained images utilizing WGAN-GP [28].

The best scores in Track 2 were achieved by team *upupup* [68], using Dynamic MLP [74] for the fusion of image features and metadata, identifying unknown classes using an entropy-based approach, training with a marginal expected loss for recognizing poisonous mushrooms while maintaining accuracy.

Finally, the best score for Track 3 was achieved by team *IES* [73], utilizing a Swin Transformer V2 Base [45] for image feature extraction, encoding metadata similarly to the approach of Ren et al. [62] from the previous edition of FungiCLEF, and introducing 1. a poisonous re-ranking that prevents predicting an edible species when there is a significant chance of the sample being poisonous, and 2. a genus loss improves the feature space's regularization.

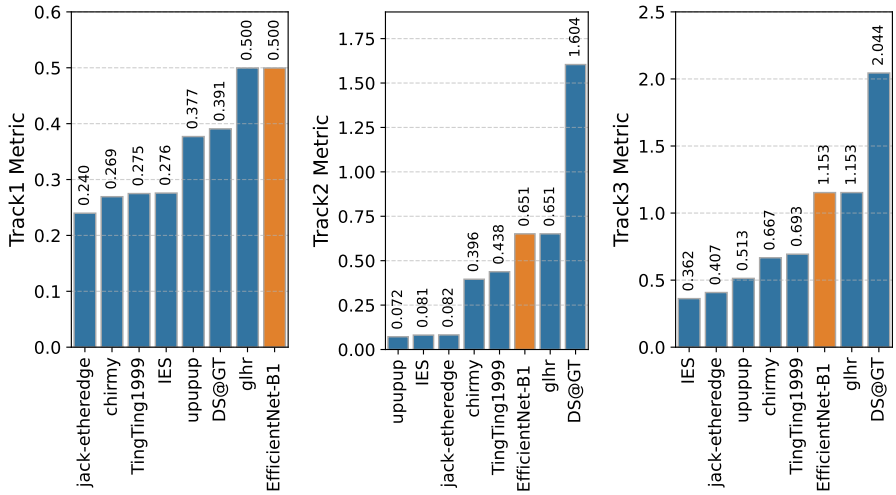

Fig. 3: **Private Leaderboard** – FungiCLEF 2024 competition – All 7 teams. The orange color depicts baseline performance.

5 PlantCLEF Challenge: multi-species plant identification in vegetation plot images

A detailed description of the challenge and a more complete discussion of the results can be found in the dedicated working note [26] and the working note participants [12,19,29].

5.1 Objective

Vegetation plot inventories are crucial for ecological studies, enabling standardized sampling, biodiversity assessment, long-term monitoring, and large-scale remote surveys. They provide valuable data on ecosystems, biodiversity conservation, and evidence-based environmental decision-making. Plot images, typically 0.5×0.5 meters in size, are meticulously analyzed by botanists who identify all species present. They also quantify species abundance using indicators like biomass, qualification factors, and areas occupied in photographs. AI could greatly improve the efficiency of surveys (with, for example, the participation of non-specialists), thereby increasing the frequency and coverage of ecological studies.

While it is now possible to access very large volumes of images of individual plants and to train very large classification models [21,22], a multi-label declination on large plot images would require complete annotation of all visible species to consider supervised learning of classification models. Unfortunately, such data doesn't exist nowadays and would require considerable efforts to be produced. The PlantCLEF 2024 challenge aims instead to evaluate approaches using classical observations of individual plants as training data, despite the discrepancies between training and test data, as shown in the figure 4. Specifically, the challenge is a weakly-supervised multi-label classification task aimed at predicting all plant species visible in high-resolution plot images but with single-label plant images as training data. One of the main difficulties lies in the domain shift between the high-resolution test images of vegetation plots with potentially many species and the training data, which primarily consists of close-up images of individual plants collected through the collaborative platform Pl@ntNet [1].

Furthermore, different weather conditions and shooting angles, along with varying phenological stages, can increase data disparity. Collaborative data might be overrepresented by opportunistic views of flowers, which facilitate identification. In contrast, vegetation plots are typically observed multiple times over one or several years without prior assumptions about the plants' phenological stages (some may be flowering, others fruiting, some in seedling stage, and others senescent or affected by disease).

5.2 Dataset

The training set is composed of observations of individual plants, similar to those used in previous editions of PlantCLEF. More precisely, it is a subset of

(a) Vegetative plot (test) (b) Single plant images (training)

Fig. 4: PlantCLEF 2024: illustration of the visual discrepancy between (a) the test set, composed exclusively of vertical top-down views potentially showing many plant species, and (b) the training set, based on images of individual plants, primarily focusing on specific organs (flowers, fruits, leaves, stems).

the Pl@ntNet training data focusing on south western Europe and covering 7,806 plant species. It contains about 1.4 million images extended with some images with trusted labels aggregated from the GBIF platform to complete the less illustrated species. Links to original images are provided in the 'url' column of the metadata csv file. The images have a relatively high resolution (the minimum side is 800 pixels) to allow the use of classification models that can handle relatively large resolution inputs and may reduce the difficulty of predicting small plants in large vegetative plot images. Images are pre-organized into subfolders by class (i.e., by species) and split into a predefined train-validation-test sets to facilitate the training of individual plant classification models.

The test set is a compilation of several image datasets of plots in different floristic contexts, including Pyrenean and Mediterranean floras. These datasets are all produced by experts and consist of a total of 1,695 high-resolution images. The shooting protocol can vary significantly from one context to another: the use of wooden frames or measuring tape to delimit the plot or not, angles of view more or less perpendicular to the ground. Additionally, the quality of the images may vary depending on the weather, which can result in more or less pronounced shadows, blurry areas, etc.

For participants who may have difficulty finding the computational power necessary to train a plant image identification model on such a large volume of data, or to enable direct work with a pre-trained backbone, two pretrained models are shared through Zenodo [27]. Both are based on a vision transformer ar-

chitecture initially pretrained with the dinov2 self-supervised learning approach [13,49] and fine-tuned on PlantCLEF 2024 training data (with a classical softmax and cross-entropy loss function).

5.3 Evaluation Protocol

The aim of the challenge is to exhaustively list the presence of every plant species on each high-resolution vegetation plot image, from among more than 7,800 species, bearing in mind that plots are generally 50x50cm in size, and that it's rare for there to be dozens and dozens of species simultaneously.

The metric chosen to differentiate the runs of the participants is the F1 score, adapted to finding a good compromise between recall and precision, i.e. not proposing too many species at the risk of being imprecise, but at the same time not proposing too few species at the risk of being incomplete. Among the several variants of F1 score, the sample-average version is selected as the primary evaluation metric of the challenge (i.e. the average of the F1 scores calculated individually for each vegetation plot). Two other F1 scores variants, namely the micro-average and macro-average, are also shown for information purposes (noticing that the macro-average is difficult to interpret because of missing species in the test set and that the micro-average is known to be biased by data imbalance).

The use of the metadata (image names, EXIF data, licenses) is authorised provided that, for each run using metadata, an equivalent run using only the visual information without metadata in submitted in order to assess the raw contribution of a purely visual analysis. The use of additional data is permitted provided that an equivalent run with only the data provided is submitted to enable more accurate and fair comparisons.

5.4 Participants and Results

The challenge is hosted on the hugging face platform, providing an opportunity for researchers and enthusiasts to contribute to the development of plant recognition in such new context.

Of the 83 teams officially registered on the CLEF registration system for LifeCLEF, 34 registered specifically for the PlantCLEF challenge. On the Hugging Face platform hosting the challenge, 9 teams attempted to submit runs, and in the end 7 teams were able to submit a total of 181 runs. Details of the best methods and systems used are synthesized in the overview working notes paper of the task [26]. In Figure 5 we report the best performance achieved for each team.

The main outcomes we can derive from that results are the following:

- Despite the sharing of pre-trained and finetuned state-of-the-art models on a large volume of data specifically on the flora studied, overall performance is low and does not exceed an F1 score of 29%.
- Highest scores were achieved by combining tiling of the high definition images and Vision Transformers models.

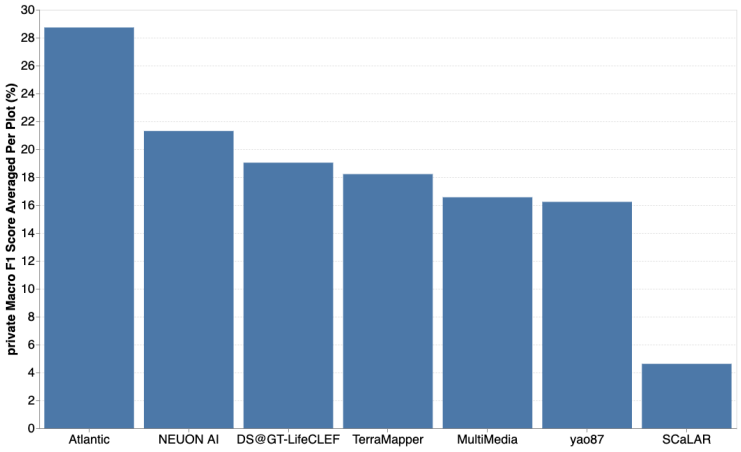

Fig. 5: PlantCLEF 2024: top samples F1 scores for each team.

- A direct method based on the supplied dinov2 pre-trained model and a tiling approach achieves a F1 score of 22.19% at best, according to participants' working notes.
- The use of an additional background analysis method, based on zero-shot learning with segment-anything [42], proves effective to gain a few extra points, but at the cost of significant computing time.
- Metadata can reveal plots photographed repeatedly over the years, enabling combined predictions for better accuracy. This approach reflects botanists' method of refining identifications through ongoing photo series analysis.

6 SnakeCLEF Challenge: Revisiting Snake Species Identification in Medically Important Scenarios

Comprehensive details on the challenge and an extensive discussion of the results are available in the dedicated working note [53].

6.1 Objective

Given the significant impact of venomous snakebites, creating a robust system to identify snake species from photos is crucial for biodiversity and global health. With over half a million annual deaths and disabilities, understanding the global distribution of 4,000+ snake species through image differentiation enhances epidemiology and treatment outcomes. Despite machines showing accuracy in predictions, especially with long-tailed distributions and 1800 species [3], challenges persist in neglected regions. The next step involves testing in specific tropical and subtropical countries while considering species' medical importance for more reliable machine predictions.

The SnakeCLEF challenge [50,52,54,59] aims to be a major benchmark for observation-based snake species identification. The goal of the task is to create a classification model that returns a ranked list of predicted species for each set of images and location (i.e., snake observation) and minimize the danger to human life and the waste of antivenom if a bite from the snake in the image were treated as coming from the top-ranked prediction. The classification model will have to (i) fit memory footprint limits and a prediction time limit (60 minutes) within a given HuggingFace server instance (Nvidia T4 small 4vCPU, 15GB RAM, 16GB vRAM), (ii) minimize the danger to human life, i.e., the venomous $\longleftrightarrow$ harmless confusion, (iii) generalize well to all geographic regions.

6.2 Dataset

The training dataset was constructed from observations submitted to the citizen science platforms iNaturalist and HerpMapper and includes around 110,000 real snake observations with community-verified species labels. While constructing the dataset, the species records were sampled based on the country of origin in order to lower the bias towards North America and Europe. Apart from image data, we have provided information about medical importance (i.e., how venomous the species is) and country-species relevance for each snake observation. We list the dataset statistics in Table 3.

Table 3: SnakeCLEF 2024 dataset statistics for each subset.

Subset	**#Species**	**#Countries**	**#Images**	**#Observations**
Training	1,784	212	168,144	95,588
↪ *iNaturalist*	*1,784*	*210*	*154,301*	*85,843*
↪ *HerpMapper*	*889*	*119*	*13,843*	*9,745*
Validation	1,599	177	14,117	7,816
Private Test	199	12	8,865	4,226
↪ *India*	*76*	*1*	*2,892*	*2,395*
↪ *Central America*	*107*	*4*	*5,188*	*1,370*
↪ *Central Africa*	*80*	*4*	*786*	*462*

6.3 Evaluation Protocol

To motivate research in recognition scenarios with uneven costs for different errors, such as mistaking a venomous snake for a harmless one, we again went beyond the 0-1 cost common in image classification. In addition to Accuracy and macro averaged F_1, we use two metrics (introduced last year) that consider venomous $\longleftrightarrow$ harmless confusion and different error costs, i.e., penalizing misclassification of a venomous species with a harmless one more than the other way around. We also calculated two standard metrics, macro averages F1 Score and Accuracy.

The two above-mentioned metrics (T_1 and T_2) are then defined as follows:

$$T_1 = \frac{w_1 \mathrm{F}_1 + w_2 \mathrm{C}_{h \longrightarrow h} + w_3 \mathrm{C}_{h \longrightarrow v} + w_4 \mathrm{C}_{v \longrightarrow v} + w_5 \mathrm{C}_{v \longrightarrow h}}{w_1 + w_2 + w_3 + w_4 + w_5}, \tag{3}$$

where C is equal to 1–ratio of misclassified samples, confusing h-armless and v-enomous species. This metric has a lower bound of 0% and an upper bound of 100%. The lower bound is achieved when all species are misclassified, including misclassifications of harmless species as venomous and vice versa. On the other hand, if the F1-score reaches 100%, indicating the correct classification of all species, each C value must be zero, leading to an overall score of 100%.

$$T_2 = \sum_i L(y_i, \hat{y}_i), \qquad L(y, \hat{y}) = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{if } y \neq \hat{y} \text{ and } p(y) = 0 \text{ and } p(\hat{y}) = 0 \\ 2 & \text{if } y \neq \hat{y} \text{ and } p(y) = 0 \text{ and } p(\hat{y}) = 1\,, \\ 2 & \text{if } y \neq \hat{y} \text{ and } p(y) = 1 \text{ and } p(\hat{y}) = 1 \\ 5 & \text{if } y \neq \hat{y} \text{ and } p(y) = 1 \text{ and } p(\hat{y}) = 0 \end{cases} \tag{4}$$

where the function p returns 0 if y is a harmless species and 1 if it is venomous.

6.4 Participants and Results

This year, a total of 14 teams participated in the SnakeCLEF. However, just nine teams submitted solutions different from the baseline, and four submitted working notes. Details of the best methods and systems used are synthesized in the competition overview paper [53], with further elaboration available in the individual working notes submitted by the participants [17,48,64,71].

In Figure 6, we report the private leaderboard performance achieved by individual teams using (i) Track 1 Metric (T_1) and (ii) Track 2 Metric (T_2). Hereafter, we provide a short overview of the methods of the two best teams who submitted a working note (top1, top2, and top5 on the leaderboard)

upupup (Top1) [71]: The team uses a branching mechanism with gating. In total, there are three branches. They all share the first three stages of a ConvNeXt model [46], but then each branch uses different weights for the fourth stage of the ConvNeXt model. The first branch is used for the classification of all the species and is also responsible for the computation of the gating parameter. The second branch focuses on venomous snakes, while the third one focuses on harmless species. The gating parameter is used to decide which of these branches will be used. This approach is used only in the training phase and is omitted for the inference. However, the authors show that this training setup helps the model perform overall well. A combination of Seesaw loss and CE loss is used for optimization.

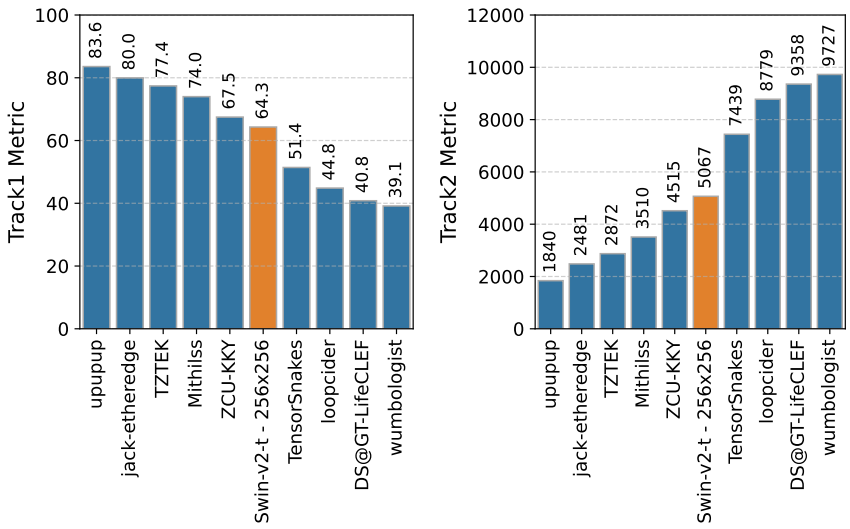

Fig. 6: **Private Leaderboard** – SnakeCLEF 2024 competition – All 9 teams. The orange color depicts baseline performance.

jack-etheredge (Top2) [17]: The team uses a CAFormer [75] model in the final solution. They introduce a new venom loss, which considers the different penalties for misclassification. A cost matrix between the predicted class and the misclassification penalty is used to reweight the softmax values of the prediction. The addition of the venom loss significantly improves the performance of the tested models across all metrics, even the F1 score. The team uses an ensemble of models trained on different data splits. Contrary to open set problems, the LogitNorm [72] did not improve the recognition rate.

ZCU-KKY (Top5) [64]: The team uses a Swin-v2 Tiny [45] model for the recognition. The reasoning behind it is so that the model can be used on mobile devices for fast and practical inference. The team combines two heads - one is for the species classification, and the other one is for venomous/harmless classification. They combine the Seesaw loss with a binary cross entropy loss. Even though the results are not as good as the results of other teams, they show an improvement over the baseline model by introducing the head responsible for venomousness recognition.

The main outcomes we can derive from the achieved results are as follows:

– An introduction of a custom loss that takes the different penalties for misclassification into account always helps. It seems to be the leading factor in improving the results.

– Branching or multi-head approach to classification of venomous vs. harmless species is another important factor in achieving better results. Although the mechanism aims to optimize the competition metric, it also improves the F1 scores. This is interesting because it shows that there are recognizable visual queues for venomousness, and it is best to model them explicitly.
– The architecture of the model (CNN vs. Transformer) is not a major cause of the success. Choosing the architecture according to other factors, such as run time or memory limitations, might be possible.

7 Conclusions and Perspectives

This new edition of LifeCLEF delivers a unique view of state-of-the-art performance on species identification and prediction problems, thanks to realistic datasets and controlled evaluation methodologies. One important conclusion is that domain shift problems remain a major problem for the emergence of new techniques such as passive acoustic sensors, HD images of plant cover, or remote sensing monitoring. The lack of annotated data for these new domains considerably hinders the progress of supervised methods, and alternative cross-domain methods are struggling to emerge. A great hope may lie in the use of unlabeled data, which will become increasingly available and whose use for domain adaptation or self-supervised learning is beginning to emerge as an effective solution (notably in BirdCLEF and GeoLifeCLEF). Another very promising prospect is multi-modal model learning, which was the key to the success of the best methods for the GeoLifeCLEF challenge and has enabled improvements in other tasks, including FungiCLEF and PlantCLEF. As far as model architectures are concerned, there is a wide disparity between the use of large-scale foundation models such as DinoV2 in PlantCLEF, SnakeCLEF, and FungiCLEF and a certain trend towards frugal architectures in GeoLifeCLEF, FungiCLEF, SnakeCLEF, and BirdCLEF. Finally, it's important to note the strength of collaborative work in the progression of the challenges. The sharing of knowledge, models, or codes, whether by the organizers or the participants themselves, has a direct impact on their subsequent developments and promotes co-construction rather than sole competition.

Acknowledgements

The research described in this paper was partly funded by the European Commission via the GUARDEN and MAMBO projects, which have received funding from the European Union's Horizon Europe research and innovation program under grant agreements 101060693 and 101060639. The opinions expressed in this work are those of the authors and are not necessarily those of the GUARDEN or MAMBO partners or the European Commission.

References

1. Affouard, A., Goeau, H., Bonnet, P., Lombardo, J.C., Joly, A.: 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)
2. Besson, M., Alison, J., Bjerge, K., Gorochowski, T.E., Høye, T.T., Jucker, T., Mann, H.M., Clements, C.F.: Towards the fully automated monitoring of ecological communities. Ecology Letters **25**(12), 2753–2775 (2022)
3. Bolon, I., Picek, L., Durso, A.M., Alcoba, G., Chappuis, F., Ruiz de Castañeda, R.: An artificial intelligence model to identify snakes from across the world: Opportunities and challenges for global health and herpetology. PLoS neglected tropical diseases **16**(8), e0010647 (2022)
4. Botella, C., Deneu, B., Marcos Gonzalez, D., Servajean, M., Larcher, T., Estopinan, J., Leblanc, C., Bonnet, P., Joly, A.: The GeoLifeCLEF 2023 dataset to evaluate plant species distribution models at high spatial resolution across europe. XXXX (2023)
5. Buchelt, A., Adrowitzer, A., Kieseberg, P., Gollob, C., Nothdurft, A., Eresheim, S., Tschiatschek, S., Stampfer, K., Holzinger, A.: Exploring artificial intelligence for applications of drones in forest ecology and management. Forest Ecology and Management **551**, 121530 (2024)
6. Ceccaroni, L., Oliver, J.L., Roger, E., Bibby, J., Flemons, P., Michael, K., Joly, A.: Advancing the productivity of science with citizen science and artificial intelligence. Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research (2023)
7. Chai, J., Ma, Q.: Technical report for fungusclef2024 competition. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
8. Chen, Y., Peng, T., Li, W., Chen, C.S.: Combining present-only and present-absent data with pseudo-label generation for species distribution modeling. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
9. Cheng, Z., Dai, W., Sun, J.: Multi-modal feature fusion networks for geolifeclef 2024. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
10. Chiu, C., Heil, M., Kim, T., Miyaguchi, A.: Fine-grained classification for poisonous fungi identification with transfer learning. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
11. Chopard, T., Rawlings, D.: Exploring biodiversity: A multi-model approach to multi-label plant species prediction. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
12. Chulif, S., Ishrat, H.A., Chang, Y.L., Lee, S.H.: Patch-wise inference using pretrained vision transformers: Neuon submission to plantclef 2024. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
13. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers (2024)
14. Das, N., Mondal, A., Chaki, J., Padhy, N., Dey, N.: Machine learning models for bird species recognition based on vocalization: A succinct review. Information Technology and Intelligent Transportation Systems pp. 117–124 (2020)
15. Diao, Q., Jiang, Y., Wen, B., Sun, J., Yuan, Z.: Metaformer: A unified meta framework for fine-grained recognition. arXiv preprint arXiv:2203.02751 (2022)
16. Dyrmann, M., Mortensen, A.K., Linneberg, L., Høye, T.T., Bjerge, K.: Camera assisted roadside monitoring for invasive alien plant species using deep learning. Sensors **21**(18), 6126 (2021)

17. Etheredge, J.: Generalizable training techniques for fine-grained long-tailed image recognition: Transferring methods optimized for fungiclef 2024 to snakeclef 2024. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
18. Etheredge, J.: openwgan-gp for fine-grained open-set fungi classification. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
19. Foy, S., McLoughlin, S.: Utilizing dino v2 for domain adaptation in vegetation plot analysis. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
20. Gaston, K.J., O'Neill, M.A.: Automated species identification: why not? Philosophical Transactions of the Royal Society of London B: Biological Sciences **359**(1444), 655–667 (2004)
21. Goëau, H., Bonnet, P., Joly, A.: 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)
22. Goëau, H., Bonnet, P., Joly, A.: 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)
23. Goëau, H., Bonnet, P., Joly, A., Bakic, V., Barthélémy, D., Boujemaa, N., Molino, J.F.: The imageclef 2013 plant identification task. In: CLEF task overview 2013, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2013, Valencia, Spain. Valencia (2013)
24. Goëau, H., Bonnet, P., Joly, A., Boujemaa, N., Barthélémy, D., Molino, J.F., Birnbaum, P., Mouysset, E., Picard, M.: The imageclef 2011 plant images classification task. In: CLEF task overview 2011, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2011, Amsterdam, Netherlands. (2011)
25. Goëau, H., Bonnet, P., Joly, A., Yahiaoui, I., Barthélémy, D., Boujemaa, N., Molino, J.F.: Imageclef2012 plant images identification task. In: CLEF task overview 2012, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2012, Rome, Italy. Rome (2012)
26. Goëau, H., Espitalier, V., Bonnet, P., Joly, A.: 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)
27. Goëau, H., Lombardo, J.C., Affouard, A., Espitalier, V., Bonnet, P., Joly, A.: 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 (Mar 2024). https://doi.org/10.5281/zenodo.10848263, `https://doi.org/10.5281/zenodo.10848263`
28. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. Advances in neural information processing systems **30** (2017)
29. Gustineli, M., Miyaguchi, A., Stalter, I.: 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)
30. Joly, A., Botella, C., Picek, L., Kahl, S., Goëau, H., Deneu, B., Marcos, D., Estopinan, J., Leblanc, C., Larcher, T., et al.: Overview of lifeclef 2023: evaluation of ai models for the identification and prediction of birds, plants, snakes and fungi. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 416–439. Springer (2023)

31. Joly, A., Goëau, H., Botella, C., Glotin, H., Bonnet, P., Vellinga, W.P., Müller, H.: Overview of LifeCLEF 2018: a large-scale evaluation of species identification and recommendation algorithms in the era of ai. In: Jones, G.J., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. (eds.) CLEF: Cross-Language Evaluation Forum for European Languages. Experimental IR Meets Multilinguality, Multimodality, and Interaction, vol. LNCS. Springer, Avigon, France (Sep 2018)
32. Joly, A., Goëau, H., Botella, C., Kahl, S., Servajean, M., Glotin, H., Bonnet, P., Planqué, R., Stöter, F.R., Vellinga, W.P., Müller, H.: Overview of LifeCLEF 2019: Identification of Amazonian Plants, South & North American Birds, and Niche Prediction. In: Crestani, F., Brascher, M., Savoy, J., Rauber, A., Müller, H., Losada, D.E., Bürki, G.H., Bürki, G.H., Cappellato, L., Ferro, N. (eds.) CLEF 2019 - Conference and Labs of the Evaluation Forum. Experimental IR Meets Multilinguality, Multimodality, and Interaction, vol. LNCS, pp. 387–401. Lugano, Switzerland (Sep 2019). https://doi.org/10.1007/978-3-030-28577-7_29, https://hal.umontpellier.fr/hal-02281455
33. Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.P., Champ, J., Planqué, R., Palazzo, S., Müller, H.: LifeCLEF 2016: Multimedia Life Species Identification Challenges. In: Fuhr, N., Quaresma, P., Gonçalves, T., Larsen, B., Balog, K., Macdonald, C., Cappellato, L., Ferro, N. (eds.) CLEF: Cross-Language Evaluation Forum. Experimental IR Meets Multilinguality, Multimodality, and Interaction, vol. LNCS, pp. 286–310. Springer, Évora, Portugal (Sep 2016). https://doi.org/10.1007/978-3-319-44564-9_26, https://hal.archives-ouvertes.fr/hal-01373781
34. Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.P., Lombardo, J.C., Planque, R., Palazzo, S., Müller, H.: LifeCLEF 2017 Lab Overview: Multimedia Species Identification Challenges. In: Jones, G.J., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. (eds.) CLEF: Cross-Language Evaluation Forum. Experimental IR Meets Multilinguality, Multimodality, and Interaction, vol. LNCS, pp. 255–274. Springer, Dublin, Ireland (Sep 2017). https://doi.org/10.1007/978-3-319-65813-1_24, https://hal.archives-ouvertes.fr/hal-01629191
35. Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.P., Planque, R., Rauber, A., Fisher, B., Müller, H.: LifeCLEF 2014: Multimedia Life Species Identification Challenges. In: CLEF: Cross-Language Evaluation Forum. Information Access Evaluation. Multilinguality, Multimodality, and Interaction, vol. LNCS, pp. 229–249. Springer International Publishing, Sheffield, United Kingdom (Sep 2014). https://doi.org/10.1007/978-3-319-11382-1_20, https://hal.inria.fr/hal-01075770
36. Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.P., Planqué, R., Rauber, A., Palazzo, S., Fisher, B., et al.: Lifeclef 2015: multimedia life species identification challenges. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction, pp. 462–483. Springer (2015)
37. Joly, A., Goëau, H., Kahl, S., Deneu, B., Servajean, M., Cole, E., Picek, L., De Castaneda, R.R., Bolon, I., Durso, A., et al.: Overview of lifeclef 2020: a system-oriented evaluation of automated species identification and species distribution prediction. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 342–363. Springer (2020)
38. Joly, A., Goëau, H., Kahl, S., Picek, L., Lorieul, T., Cole, E., Deneu, B., Servajean, M., Durso, A., Glotin, H., Planqué, R., Vellinga, W.P., Navine, A., Klinck, H.,

Denton, T., Eggel, I., Bonnet, P., Šulc, M., Hruz, M.: Overview of lifeclef 2022: an evaluation of machine-learning based species identification and species distribution prediction. In: International Conference of the Cross-Language Evaluation Forum for European Languages. Springer (2022)
39. Joly, A., Goëau, H., Kahl, S., Picek, L., Lorieul, T., Cole, E., Deneu, B., Servajean, M., Durso, A., Bolon, I., et al.: Overview of lifeclef 2021: An evaluation of machine-learning based species identification and species distribution prediction. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 371–393. Springer (2021)
40. Júnior, T.D.C., Rieder, R.: Automatic identification of insects from digital images: A survey. Computers and Electronics in Agriculture **178**, 105784 (2020)
41. Kahl, S., Denton, T., Klinck, H., Ramesh, V., Joshi, V., Srivathsa, M., Anand, A., Arvind, C., CP, H., Sawant, S., Robin, V.V., Glotin, H., Goëau, H., Vellinga, W.P., Planqué, R., Joly, A.: Overview of BirdCLEF 2024: Acoustic identification of under-studied bird species in the western ghats. Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
42. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
43. Kong, S., Ramanan, D.: Opengan: Open-set recognition via open data generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 813–822 (2021)
44. Liu, H., Tao, Z., Jiang, P., Sun, Q., Wan, M.: Plant species prediction task based on graph neural networks and cross attention methods. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
45. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12009–12019 (2022)
46. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. arXiv preprint arXiv:2201.03545 (2022)
47. Miyaguchi, A., Aphiwetsa, P., McDuffie, M.: Tiled raster compression and embeddings for multilabel classification in geolifeclef 2024. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
48. Miyaguchi, A., Gustineli, M., Fischer, A., Lundqvist, R.: Transfer learning with self-supervised vision transformer for snake identification. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
49. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
50. Picek, L., Ruiz De Castañeda, R., Durso, A.M., Sharada, P.M.: Overview of the snakeclef 2020: Automatic snake species identification challenge. In: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum (2020)
51. Picek, L., Botella, C., Servajean, M., Deneu, B., Marcos Gonzalez, D., Palard, R., Larcher, T., Leblanc, C., Estopinan, J., Bonnet, P., Joly, A.: Overview of GeoLifeCLEF 2024: Species presence prediction based on occurrence data and high-resolution remote sensing images. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)

52. Picek, L., Durso, A.M., Bolon, I., de Castañeda, R.R.: Overview of snakeclef 2021: Automatic snake species identification with country-level focus. In: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum (2021)
53. Picek, L., Hruz, M., Durso, A.M.: Overview of SnakeCLEF 2024: Revisiting snake species identification in medically important scenarios. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
54. Picek, L., Šulc, M., Chamidullin, R., Durso, A.M.: Overview of snakeclef 2023: Snake identification in medically important scenarios. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
55. Picek, L., Sulc, M., Matas, J.: Overview of FungiCLEF 2024: Revisiting fungi species recognition beyond 0-1 cost. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
56. Picek, L., Šulc, M., Matas, J., Heilmann-Clausen, J., Jeppesen, T.S., Lind, E.: Automatic fungi recognition: Deep learning meets mycology. Sensors **22**(2), 633 (2022)
57. Picek, L., Šulc, M., Matas, J., Jeppesen, T.S., Heilmann-Clausen, J., Læssøe, T., Frøslev, T.: Danish fungi 2020-not just another image recognition dataset. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1525–1535 (2022)
58. Picek, L., Šulc, M., Patel, Y., Matas, J.: Plant recognition by ai: Deep neural nets, transformers, and knn in deep embeddings. Frontiers in plant science **13**, 787527 (2022)
59. Picek, L., Durso, A.M., Hrúz, M., Bolon, I.: Overview of SnakeCLEF 2022: Automated snake species identification on a global scale. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (2022)
60. Picek, L., Šulc, M., Heilmann-Clausen, J., Matas, J.: Overview of FungiCLEF 2022: Fungi recognition as an open set classification problem. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (2022)
61. Picek, L., Šulc, M., Heilmann-Clausen, J., Matas, J.: Overview of FungiCLEF 2023: Fungi recognition beyond 0-1 cost. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
62. Ren, H., Jiang, H., Luo, W., Meng, M., Zhang, T.: Entropy-guided open-set fine-grained fungi recognition. Aliannejadi et al.[1] pp. 2122–2136 (2023)
63. Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., et al.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography **40**(8), 913–929 (2017)
64. Sieber, M., Železný, T.: Do not lose to losses for snakeclef2024. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
65. Šulc, M., Matas, J.: Fine-grained recognition of plants from images. Plant Methods **13**, 1–14 (2017)
66. Sulc, M., Picek, L., Matas, J., Jeppesen, T., Heilmann-Clausen, J.: Fungi recognition: A practical use case. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2316–2324 (2020)
67. Syayfetdinov, A.: Multimodal networks for species distribution modeling. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
68. Tan, B.F., Li, Y.Y., Wang, P., Zhao, L., Wei, X.S.: Say no to the poisonous: An effective strategy for mitigating 0-1 cost in fungiclef2024. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)

69. Wäldchen, J., Rzanny, M., Seeland, M., Mäder, P.: Automated plant species identification—trends and future directions. PLoS computational biology **14**(4), e1005993 (2018)
70. Wan, F., Wan, H., Zhang, Z., Gao, J., Sun, C., Wang, Y.: The application potential of unmanned aerial vehicle surveys in grassland plant diversity. Biodiversity Science **32**(3), 23381 (2024)
71. Wang, P., Li, Y., Tan, B.F., Zhou, Y.C., Li, Y., Wei, X.S.: Multibranch co-training to mine venomous feature representation: A solution to snakeclef2024. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
72. Wei, H., Xie, R., Cheng, H., Feng, L., An, B., Li, Y.: Mitigating neural network overconfidence with logit normalization. In: International conference on machine learning. pp. 23631–23644. PMLR (2022)
73. Wolf, S., Thelen, P., Beyerer, J.: Open-set fungi classification focused on reducing risk of poisonous confusion. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024)
74. Yang, L., Li, X., Song, R., Zhao, B., Tao, J., Zhou, S., Liang, J., Yang, J.: Dynamic mlp for fine-grained image classification by leveraging geographical and temporal information. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10945–10954 (2022)
75. Yu, W., Si, C., Zhou, P., Luo, M., Zhou, Y., Feng, J., Yan, S., Wang, X.: Metaformer baselines for vision. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
76. Zwerts, J.A., Stephenson, P., Maisels, F., Rowcliffe, M., Astaras, C., Jansen, P.A., van Der Waarde, J., Sterck, L.E., Verweij, P.A., Bruce, T., et al.: Methods for wildlife monitoring in tropical forests: Comparing human observations, camera traps, and passive acoustic sensors. Conservation Science and Practice **3**(12), e568 (2021)