Overview of LifeCLEF 2023: evaluation of AI models for the identification and prediction of birds, plants, snakes and fungi

Alexis Joly, Christophe Botella, Lukáš Picek, Stefan Kahl, Hervé Goëau, Benjamin Deneu, Diego Marcos, Joaquim Estopinan, César Leblanc, Théo Larcher

Overview of LifeCLEF 2023: evaluation of AI models for the identification and prediction of birds, plants, snakes and fungi

Alexis Joly[1], Christophe Botella[1], Lukáš Picek[9], Stefan Kahl[6,13], Hervé Goëau[2], Benjamin Deneu[1], Diego Marcos[16], Joaquim Estopinan[1], Cesar Leblanc[1], Théo Larcher[1], Rail Chamidullin[9], Milan Šulc[12], Marek Hrúz[9], Maximilien Servajean[7], Hervé Glotin[3], Robert Planqué[4], Willem-Pier Vellinga[4], Holger Klinck[6], Tom Denton[11], 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 Insitute, 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 Computing and Mathematical Sciences, Caltech, USA
[9] Department of Cybernetics, FAV, University of West Bohemia, Czechia
[10] Department of Biological Sciences, Florida Gulf Coast University, USA
[11] Google Research, San Francisco, USA
[12] Second Foundation, Prague, Czech Republic
[13] Chemnitz University of Technology, Chemnitz, Germany

Abstract. Biodiversity monitoring through AI approaches is essential, as it enables the efficient analysis of vast amounts of data, providing comprehensive insights into species distribution and ecosystem health and aiding in informed conservation decisions. 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 LifeCLEF virtual lab has been promoting and evaluating advances in this domain since 2011. The 2023 edition proposes five data-oriented challenges related to the identification and prediction of biodiversity: (i) BirdCLEF: bird species recognition in long-term audio recordings (soundscapes), (ii) SnakeCLEF: snake identification in medically important scenarios, (iii) PlantCLEF: very large-scale plant identification, (iv) FungiCLEF: fungi recognition beyond 0-1 cost, and (v) GeoLifeCLEF: remote sensing-based prediction of species. This paper overviews the motivation, methodology, and main outcomes of that 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[1]. In 2004, Gaston and O'Neill [10] 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 [10], automated species identification has been studied in many contexts [6,12,13,20,31,49,50,58]. This area continues to expand rapidly, particularly due to advances in deep learning [2,11,32,34,35,52,53,54,56]. 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 [3]. This information is essential for making informed conservation decisions and identifying areas in need of protection.

To measure progress in a sustainable and repeatable way, the LifeCLEF[2] virtual lab was created in 2014 as a continuation and extension of the plant identification task that had been run within the ImageCLEF lab[3] since 2011 [15,16,17]. 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 [21,22,23,24,25,26,27,29,28]. Nearly a thousand researchers and data scientists register yearly to LifeCLEF to download the data, subscribe to the mailing list, benefit from the shared evaluation tools, etc. The number of participants who finally crossed the finish line by submitting runs was respectively: 22 in 2014, 18 in 2015, 17 in 2016, 18 in 2017, 13 in 2018, 16 in 2019, 16 in 2020, 1,022 in 2021 and 1146 in 2022. LifeCLEF 2023 consists of five challenges (BirdCLEF, SnakeCLEF, PlantCLEF, FungiCLEF, GeoLifeCLEF) whose methodology and main outcomes are described in this paper. Table 1 provides an overview of the data and tasks of the five challenges.

[1] `https://www.cbd.int/`
[2] `http://www.lifeclef.org/`
[3] `http://www.imageclef.org/`

Table 1: Overview of the data and tasks of the five LifeCLEF challenges

	Modality	**#species**	**#items**	**Task**	**Metric**
BirdCLEF	audio	264	16,900	Multi-Label Classification	cmAP
SnakeCLEF	images metadata	1,500	150–200K	Classification	ad-hoc metric
FungiCLEF	images metadata	1,600	300K	Classification	ad-hoc metric
PlantCLEF	images	80,000	4.0M	Classification	Macro-Average MRR
GeoLifeCLEF	images time-series tabular	10,040	5.3M	Multi-Label Classification	Micro-Average F1

The systems used to run the challenges (registration, submission, leaderboard, etc.) were the Kaggle platform for the BirdCLEF and GeoLifeCLEF challenges, the Hugging Face competition platform for SnakeCLEF and FungiCLEF challenges, and the AICrowd platform for the PlantCLEF challenge. Three of the challenges (GeoLifeCLEF, SnakeCLEF, and FungiCLEF) were organized jointly with FGVC 10, an annual workshop dedicated to Fine-Grained Visual Categorization organized in the context of the CVPR international conference on computer vision and pattern recognition.
In total, $1,226$ people/teams participated to LifeCLEF 2023 edition by submitting runs to at least one of the five challenges ($1,189$ only for the BirdCLEF challenge). Only some of them managed to get the results right, and 17 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 [30].

2.1 Objective

Recognizing bird sounds in complex soundscapes is an important sampling tool that often helps reduce the limitations of point counts. In the future, archives of recorded soundscapes will become increasingly valuable as the habitats in which they were recorded will be lost. In the past few years, deep learning approaches

have transformed the field of automated soundscape analysis. Yet, when training data is sparse, detection systems struggle to recognize bird species reliably. The goal of this competition was to establish training and test datasets that can serve as real-world applicable evaluation scenarios for endangered habitats and help the scientific community to advance their conservation efforts through automated bird sound recognition.

2.2 Dataset

We built on the experience from previous editions and adjusted the overall task to encourage participants to focus on task-specific model designs. We selected training and test data to suit this demand. As in previous iterations, Xeno-canto was the primary source for training data, and expertly annotated soundscape recordings were used for testing. We focused on bird species which are usually underrepresented in large bird sound collections, but we also included common species so that participants were able to train good recognition systems. In search of suitable test data, we considered different data sources with varying complexity (call density, chorus, signal-to-noise ratio, man-made sounds, etc.) and quality (mono and stereo recordings). We also wanted to focus on very specific real-world use cases (e.g., conservation efforts in Africa) and framed the competition based on the demand of the particular use case.

2.3 Evaluation Protocol

The challenge was held on Kaggle, and the evaluation mode resembled the test mode of previous iterations, i.e., hidden test data, code competition, etc. We used the class-wise mean average precision (cmAP) as a metric, which allowed organizers to assess system performance independent of fine-tuned confidence thresholds. Participants were asked to return a list of species for short audio segments extracted from labeled soundscape data. We used 5-second segments, which reflect a good compromise between typical signal length and sufficiently long context windows. Again, we kept the dataset size reasonably small (<50 GB) and easy to process, and we also provided introductory code repositories and write-ups to lower the entry-level of the competition.

2.4 Participants and Results

1,397 participants across 1,189 teams participated in the BirdCLEF 2023 challenge and submitted a total of 21,519 runs. In Figure 1 we report the performance achieved by the top 25 collected runs. The private leaderboard score is the primary metric and was revealed to participants after the submission deadline to avoid probing the hidden test data. Public leaderboard scores were visible to participants over the course of the entire challenge.

The baseline cmAP-score in this year's edition was 0.602 (public 0.717) with random confidence scores for all birds for all segments, and 1,165 teams managed to score above this threshold. The best submission achieved a cmAP-score

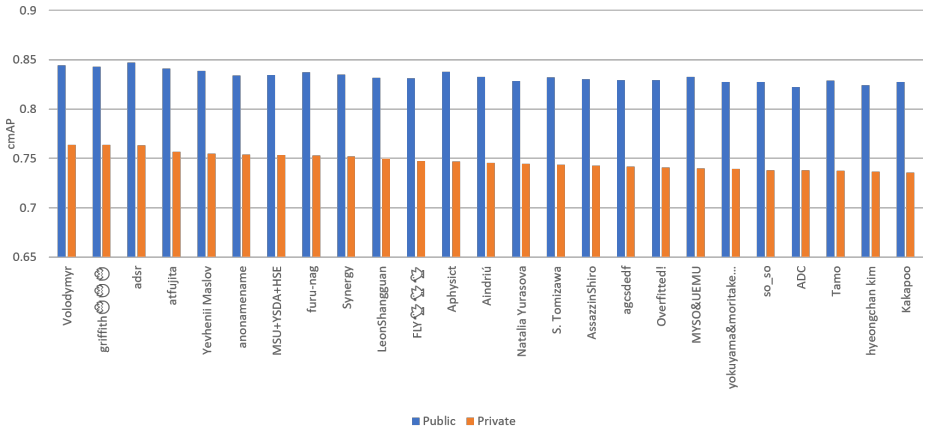

Fig. 1: BirdCLEF 2023 results of the top 25 teams.

of 0.7639 (public 0.8444) and the top 10 best performing systems were within only 1.5% difference in score. The vast majority of approaches were based on convolutional neural network ensembles and mostly differed in pre- and post-processing and neural network backbone. Interestingly, few-shot learning techniques were vastly underrepresented despite the fact that some target species only had a handful of training samples. Some teams utilized embeddings of pre-trained bird recognition models (such as BirdNET or Google Perch, both were provided as supporting models) to train on high-level features, which somewhat mitigated the need for extensive training data. Due to the limited CPU runtime for submissions, participants focused on accelerating model inference and efficient architectures, with EfficientNet backbones being the most common choice. Interestingly, participants also experimented with ONNX and openVINO to improve model inference speed.

3 SnakeCLEF challenge: Snake Identification in Medically Important scenarios

A detailed description of the challenge and a more complete discussion of the results can be found in the dedicated overview paper [37].

3.1 Motivation

Developing a robust system for identifying species of snakes from photographs is an important goal in biodiversity but also for human health. With over half a million victims of death & disability from venomous snakebite annually, understanding the global distribution of the >4,000 species of snakes and differentiating species from images (particularly images of low quality) will significantly

improve epidemiology data and treatment outcomes. We have learned from previous editions that "*machines*" can accurately recognize ($F_1^C \approx 90\%$ and Top1 Accuracy $\approx 90\%$) even in scenarios with long-tailed distributions and $\approx 1,600$ species. Thus, testing over real Medically Important Scenarios and specific countries (India and Central America) and integrating the medical importance of species is the next step that should provide a more reliable machine prediction.

3.2 Objective

The main objective of this competition is to create a machine learning model that can accurately predict snake species for given observation data, i.e., images and location, and: (i) fits limits for memory footprint (max size of 1GB), (ii) minimizes the danger to human life, i.e., the venomous $\longleftrightarrow$ harmless confusion, (iii) generalize to all countries and geographic regions.

3.3 Dataset

The dataset was constructed from observations submitted to the citizen science platforms – iNaturalist and HerpMapper – and combined roughly 110,000 reals snake specimen observations with community-verified species labels. The number of species was extended up to $\approx 1,800$ snake species from around the world. Apart from image data, we have provided information about medical importance (i.e., how venomous the species is), and country-species relevance was provided for each species. We list the dataset statistics in Table 2.

Table 2: SnakeCLEF 2023 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
Public Test	1,784	191	28,274	15,632
Private Test	182	8	8,080	3,765
↪ *India*	*76*	*1*	*2,892*	*2,395*
↪ *Central America*	*107*	*4*	*5,188*	*1,370*

Geographical bias: There is a lack of data from remote parts of developing countries that tend to lack herpetological expertise and have high snake diversity, and snakebites are common (i.e., Asia, Africa, and Central/South America).

3.4 Evaluation Protocol

To motivate research in recognition scenarios with uneven costs for different errors, such as mistaking a venomous snake for a harmless one, this year's challenge

goes beyond the 0-1 loss common in classification. We make some assumptions to reduce the complexity of the evaluation. We consider that there exists a universal antivenom that is applicable to all venomous snake bites. Furthermore, such antivenom is not lethal or seriously harmful when applied to a healthy human. Hence, we will penalize the misclassification of a venomous species with a harmless one more than the other way around. Although this solution is not perfect, it is a first step into a more complex evaluation of snake bites. We specify two metrics (T_1, T_2) reflecting these different scenarios.

$$T_1 = \frac{w_1 \mathrm{F}_1 + w_2 \mathrm{C}_{h \longrightarrow h} + w_3 \mathrm{C}_{h \longrightarrow v} + w_4 \mathrm{C}_{v \longrightarrow v} + w_5 \mathrm{C}_{v \longrightarrow h}}{w_1 + w_2 + w_3 + w_4 + w_5}, \tag{1}$$

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%.

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

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

3.5 Participants and Results

This year a total of 16 teams participated in the SnakeCLEF. However, just five teams submitted their models for private evaluation together with the working notes. Details of the best methods and systems used are synthesized in the competition overview paper [1].

In Figure 2 and Figure 3 we report the public and private leaderboard performance achieved by individual teams using: (i) Track 1 Metric (T_1), (ii) Track 2 Metric (T_2) and (iii) the macro F_1 score. The main outcomes we can derive from the achieved results are as follows:

NLP model encoded metadata might be the next big thing. Same as in previous years, most of the teams used the provided metadata and showed that by doing so the competition metric improves. CLIP [44] – a strong multi-modal descriptor, was used for the first time in this competition to encode the metadata. This trend may lead to the utilization of bigger NLP models.

Transformers for the win. But do not rule out the CNNs yet. On the vision part, convolutional models (ResNet [18], EfficientNet [48], ConvNext [57])

and Transformer models (MetaFormer [8], Swin [33], VOLO [59]) were used to extract the visual features. When teams compared the architectures side-by-side, most of the times the Transformer architecture performed better. However, the winning team used ConvNextv2. Due to the lack of a fair and exhaustive ablation study, it is not clear how a Transformer model would fare.

Task-tailored losses and self-supervision are the key to learning. Traditionally, Seesaw loss [55] and SimCLR [7] were used to cope with the long-tailed data. Some teams introduced a weighted version of the loss functions tackling the different penalization for different errors. Multi-Instance Learning [19] was applied to make use of more images per observation.

Medically important scenarios might be on to something. The final results on the private dataset show an interesting behavior of the models. The best team (named *word2vector*) achieved macro F_1 score of 53.58% with the competition score of 91.31%. The runner-up (BBracke) actually achieved a much better F_1 score of 61.39% but had a lower competition score of 90.19%. We hypothesize that this was possible due to the post-processing step of team *word2vector*. When they observed that the top-5 results contained a venomous species, the observation was classified as such.

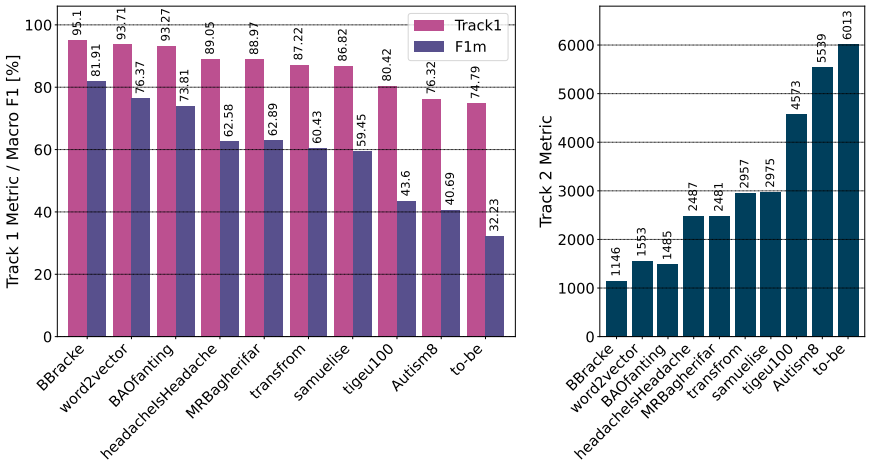

Fig. 2: **Public Leaderboard** – SnakeCLEF 2023 competition – Top10 teams.

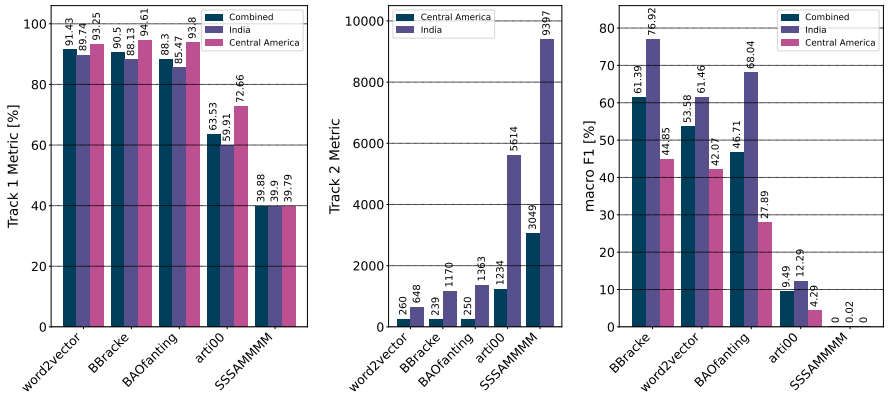

Fig. 3: **Private Leaderboard** – SnakeCLEF 2023 competition – 5 teams.

4 FungiCLEF Challenge: Fungi Recognition Beyond 0-1 Cost

A detailed description of the challenge and a more complete discussion of the results can be found in the dedicated working note [42].

4.1 Objective

Automatic recognition of species at scale, such as in popular citizen-science projects [47,39], requires efficient prediction on limited resources. In practice, species identification typically depends not solely on the visual observation of the specimen but also on other information available to the observer, e.g., habitat, substrate, location, and time. Thanks to rich metadata, precise annotations, and baselines available to all competitors, the challenge aims at providing a major benchmark for combining visual observations with other observed information. Additionally, the 2023 competition considers decision processes for different usage scenarios, which go beyond the commonly assumed 0/1 cost function – e.g., cost for misclassification of edible and poisonous mushrooms is an important practical aspect to be evaluated.

4.2 Dataset

The challenge builds upon the Danish Fungi 2020 dataset [40], which comes from a citizen science project, the Atlas of Danish Fungi, where all samples went through an expert validation process, guaranteeing a high quality of labels. Rich metadata (Habitat, Substrate, Timestamp, GPS, EXIF etc.) are provided for most samples. The training set will be the union of the training and public-test set (without out-of-scope samples) from the 2022 challenge [41] – i.e., 295,938 training images belonging to 1,604 species observed mostly in Denmark.

The validation and test sets include all the expert validates observations with species labels collected in 2021 and 2022. respectively. Both the validation and test set cover roughly 3,000 fungi species and include a high number of observations with "*unknown*" species. The test set was further split (50/50 ratio) to provide different data for a public and private evaluation. We list the dataset statistics in Table 3.

Table 3: FungiCLEF 2023 dataset statistics for each subset.

Subset	**Species** → **Known/Unknown**	**Images**	**Observations**
Training	1,604 1,604 / –	295,938	177,170
Validation	2,713 1,084 / 1,629	60,832	30,131
Public Test	2,650 1,085 / 1,565	60,225	30,130
Private Test	3,299 1,116 / 2,183	91,231	45,021

4.3 Evaluation Protocol

Given the set of real fungi species observations and corresponding metadata, the goal of the task is to create a classification model that predicts a species for each given observation. The classification model must fit limits for memory footprint (max size of 1GB) and should have to consider and minimize the danger to human life, i.e., the confusion between poisonous and edible species.

FungiCLEF 2023 considered five different decision scenarios, minimizing the empirical loss $L = \sum_i W(y_i, q(x_i))$ for decisions $q(x)$ over observations x and true labels y, given a cost function $W(y, q(x))$. Five cost functions were given for the following scenarios:

- Track 1: Standard classification with "unknown" category;
- Track 2: Cost for confusing edible species for poisonous and vice versa;
- Track 3: An application user-focused loss composed of both the classification error (e.g., accuracy) and the poisonous $\longleftrightarrow$ edible confusion;
- Track 4: Cost for missing *"unknown"* species is higher; misclassifying for *"unknown"* is cheaper than confusing species;

Baseline procedures of how metadata can help the classification, pre-trained baseline classifiers, and code submission example were provided to all participants as part of the task description.

4.4 Participants and Results

Twelve teams participated in the FungiCLEF 2023 challenge; four provided their models for a private evaluation, and three submitted working notes. Details of the best methods and systems used are synthesized in the overview working note paper of the task [38] and further developed in the individual working notes of participants (see references in [38]). In Figure 4 and Figure 5, we report the

performance achieved by the participants. Interestingly, none of the teams that submitted working notes optimized decision-making for each of the five tasks.

The best-performing team – *meng18* – combined visual information with metadata using MetaFormer [8], tackled class imbalance with the Seesaw loss [55], proposed an entropy-guided recognition of unknown species, and introduced an additional poisonous-classification loss.

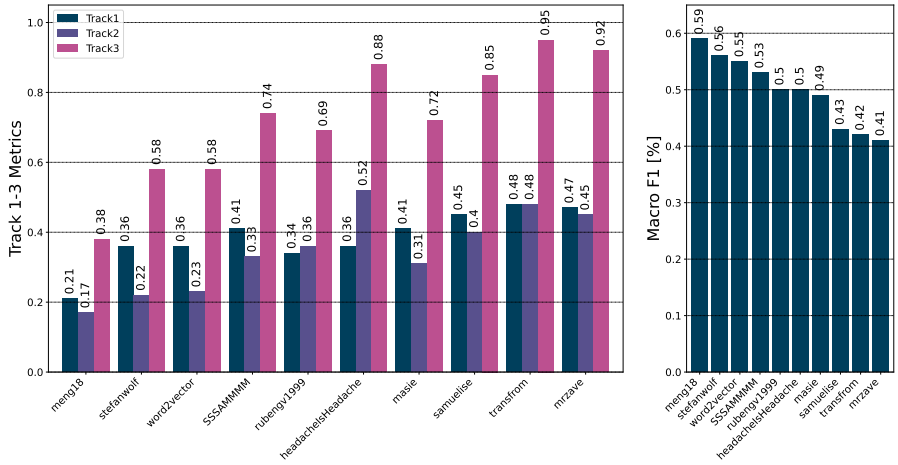

Fig. 4: **Public Leaderboard** – FungiCLEF 2023 competition – Top10 teams.

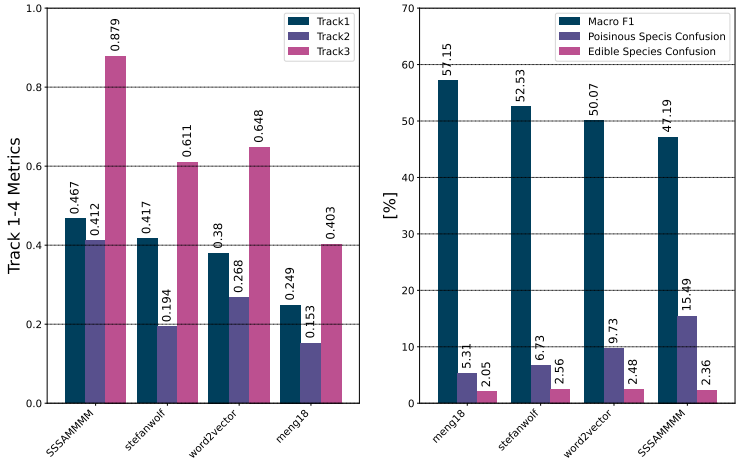

Fig. 5: **Private Leaderboard** – FungiCLEF 2023 competition – 4 teams.

5 PlantCLEF Challenge: Identify the World's Flora

A detailed description of the challenge and a more complete discussion of the results can be found in the dedicated working note [14].

5.1 Objective

Advancements in deep learning and the growing abundance of field photographs have significantly enhanced the automated identification of plants. A notable milestone was achieved during LifeCLEF 2018, where a top-1 classification accuracy of up to 90% was attained for over 10k species. This demonstrated that automated systems had made remarkable progress and are approaching human expertise in this domain [21]. However, it is crucial to recognize that such impressive performance levels are still a long way off from encompassing the vastness of the world's flora. Presently, science has identified approximately 391,000 vascular plant species, with new discoveries and descriptions being made each year. The significance of this plant diversity extends beyond the mere existence of species; it plays a pivotal role in ecosystem functioning and the advancement of human civilization. Regrettably, the majority of these species remain poorly understood, and there is an acute scarcity of training images available for the vast majority of them [43].
The objective of the PlantCLEF challenges in 2022 and 2023 was to advance the field of plant identification on a global scale. To achieve this, a training dataset was curated, encompassing a remarkable 80,000 species and comprising 4 million images. This expansive dataset was made accessible to the community through a challenge hosted on the AIcrowd platform[4], providing an opportunity for researchers and enthusiasts to contribute to the development of plant recognition.

5.2 Dataset

The training set consists of two distinct subsets. The first subset, referred to as the trusted training dataset, is derived from the GBIF (Global Biodiversity Information Facility) portal[5], which is the largest biodiversity data portal globally. This subset comprises over 2.9 million images encompassing 80,000 plant species. These images have been shared and collected primarily through GBIF, with some contributions from the Encyclopedia of Life[6](EOL). The sources of these images include academic institutions such as museums, universities, and national institutions, as well as collaborative platforms like iNaturalist and Pl@ntNet, implying a fairly high certainty of determination quality (collaborative platforms only share their highest quality data qualified as "research graded"). To maintain a

[4] https://www.aicrowd.com/challenges/lifeclef-2022-23-plant/
[5] https://gbif.org/
[6] https://eol.org/

manageable training set size and address class imbalance, the number of images per species was restricted to approximately 100. Additionally, the selection process favored specific views that are conducive to plant identification, such as close-ups of flowers, fruits, leaves, trunks, and other relevant features. This approach ensures that the training dataset comprises informative and relevant images for accurate plant recognition.

In contrast, a second "web" training dataset comprises images obtained from commercial search engines like Google and Bing. This dataset comes with its own set of challenges. The raw downloaded data from these search engines contains a notable number of species identification errors and a substantial presence of (near)-duplicates and images that are not well-suited for plant identification purposes. For instance, the dataset includes images of herbarium sheets, landscapes, microscopic views, and various other non-relevant visuals. Moreover, the web dataset contains a significant amount of unrelated images, such as portraits of botanists, maps, graphs, images from other kingdoms of living organisms, and even manufactured objects. To address these issues, a semi-automatic filtering approach was adopted. This process involved multiple iterations of training Convolutional Neural Networks (CNNs), conducting inference, and human labeling. Through this iterative process, the raw data was as best as possible cleaned up, leading to a drastic reduction in the number of irrelevant pictures. Furthermore, the image quality was improved by prioritizing close-ups of flowers, fruits, leaves, trunks, and other relevant plant features. As a result of this filtering process, the web dataset consists of approximately 1.1 million images, covering approximately 57k plant species.

Participants were allowed to use complementary training data (e.g. for pre-training purposes) but at the condition that (i) the experiment is entirely reproducible, i.e. that the used external resource is clearly referenced and accessible to any other research group in the world, (ii) the use of external training data or not is mentioned for each run, and (iii) the additional resource does not contain any of the test observations. External training data was allowed but participants had to provide at least one submission that used only the provided data.

The test set used in the PlantCLEF challenge was constructed using multi-image plant observations obtained from the Pl@ntNet platform during the year 2021. These observations had not been shared through GBIF, meaning they were not present in the training set. Only observations that received a very high confidence score in the Pl@ntNet collaborative review process were selected for the challenge to ensure the highest possible quality of determination. This process involves people with a wide range of skills (from beginners to world-leading experts), but these have different weights in the decision algorithms. Finally, the test set contains about 27k plant observations related to about 55k images (a plant can be associated with several images) covering about 7.3k species.

5.3 Evaluation Protocol

The evaluation of the task in the PlantCLEF challenge primarily relies on the Mean Reciprocal Rank (MRR) metric. MRR is a statistical measure used to

assess processes that generate a list of potential responses to a set of queries, ordered by the probability of correctness. It quantifies the performance of a system by considering the reciprocal rank of the first correct answer for each query. The reciprocal rank of a query response is calculated as the multiplicative inverse of the rank of the first correct answer. In other words, if the correct answer is ranked first, the reciprocal rank is 1. If it is ranked second, the reciprocal rank is 1/2, and so on. To determine the MRR for the entire test set, the reciprocal ranks for all the queries are averaged together:

$$MRR = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{\text{rank}_q} \tag{3}$$

where Q is the total number of query occurrences (plant observations) in the test set. However, the macro-average version of the MRR (average MRR per species in the test set - MA-MRR) was used because of the long tail of the data distribution to re-balance the results between under- and over-represented species in the test set.

5.4 Participants and Results

Although over a hundred participants signed up for the challenge, in the end only 3 participants from 3 countries participated to the PlantCLEF 2023 challenge and submitted a total of 22 runs. Details of the best methods and systems used are synthesized in the overview working notes paper of the task [14]. In Figure 6 we report the performance achieved by the different runs of the participants.

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

- The most impressive outcomes were achieved by vision transformer-based approaches, particularly the vision-centric foundation model EVA [9], that was the state-of-the-art position during the challenge in the first quarter of 2023. While CNN-based approaches also produced respectable results, with a maximum MA-MRR of 0.618 (Neuon AI Run 9), they still fell notably short of the highest score attained by an EVA approach. The best EVA approach, Mingle Xu Run 8, achieved a remarkable MA-MRR of 0.674.
- Utilizing the complete PlantCLEF training dataset, comprising both the trusted and web datasets, proved advantageous, despite the added training time and the residual noise inherent in the web dataset. The inclusion of the web training dataset resulted in a noticeable improvement, with the MA-MRR reaching 0.674, compared to a maximum of 0.65 without it.
- The reduction of the training set by removing the classes with the fewest images (Mingle Xu Run 1-4-2-6 vs Run 5) implies a significant drop in performance. This demonstrates that there might not always be a direct connection between the training data and the test data, emphasizing the importance of considering all classes, including those linked to uncommon species, when addressing the task of monitoring plant biodiversity

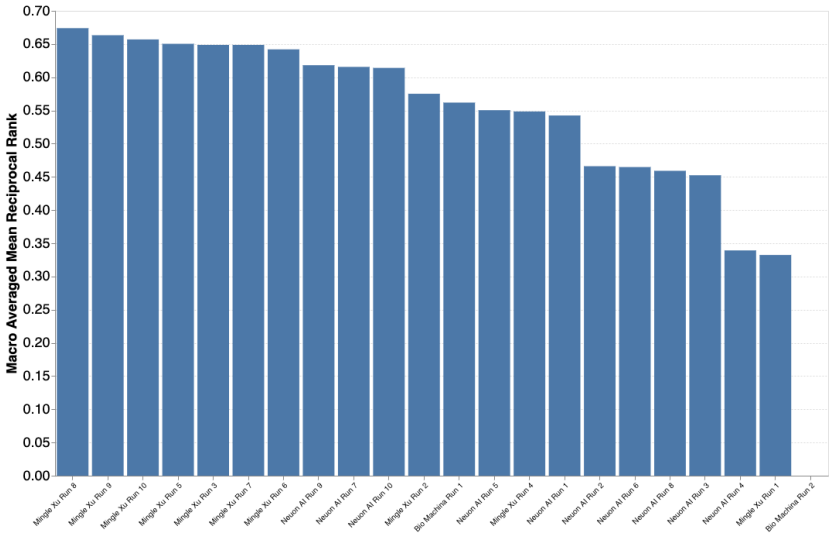

Fig. 6: PlantCLEF 2023 results

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

A detailed description of the challenge and a more complete discussion of the results can be found in the dedicated working note [5]. A graphical abstract of the challenge is provided in Figure 7.

6.1 Objective

Predicting which species are present in a given area through Species Distribution Models (SDM) is a central problem in ecology and a crucial issue for biodiversity conservation. Such predictions are a fundamental element of many decision-making processes, whether for land use planning, the definition of protected areas, or the implementation of more ecological agricultural practices. Classical SDMs are well-established but have the drawback of covering only a limited number of species at spatial resolutions often coarse in the order of kilometers, or hundreds of meters at best. In addition, while the use of the massive presence-only data arising from large citizen science platforms has grown, the SDM built from such data are affected by many sampling biases, as, for instance, species detection bias or species set size bias. Developing scalable methods suited to account and correct for these biases is a necessary step to update regularly species distributions maps by capitalizing on the massive flow of citizen science data. The objective of GeoLifeCLEF is to evaluate models with orders of magnitude hitherto unseen, whether in terms of the number of species covered (thousands),

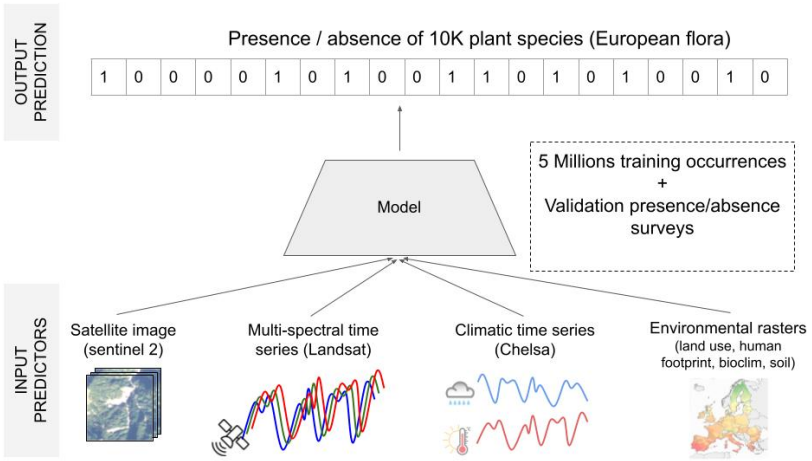

Fig. 7: GeoLifeCLEF 2023 graphical abstract

spatial resolution (on the order of 10 meters), or the number of occurrences used as training data (several million). These models have the potential to greatly improve biodiversity management processes, especially at the local level (e.g. municipalities), where the need for spatial and taxonomic precision is greatest.

6.2 Training Dataset

A brand new dataset was built for the 2023 edition of GeoLifeCLEF in the framework of a large-scale European project on biodiversity monitoring (MAMBO, Horizon EU program). It contains about 5 million plant species presence-only records (single positive labels, hereafter PO) covering 10 thousand species extracted from thirteen selected datasets of the Global Biodiversity Information Facility (GBIF) and covers the whole EU territory (38 countries including E.U. members). We also provided the participants with a validation set of 5 thousand standardized presence-absence (hereafter PA) surveys of small spatial plots (multi-label) to help calibrate the models, and specifically to correct for sampling biases. For the explanatory variables (to be used as inputs of the models), the dataset contains both high-resolution remote sensing data (10m resolution Sentinel-2 satellite images and Landsat multi-spectral time-series at each data location, along with elevation) and coarser resolution environmental raster data (land cover, human footprint, bioclimatic and soil variables). The geocoordinates and date of the species observations are also provided and can also be used as one modality. Participants are free to use one, several, or all available modalities in their models. The detailed description of the GeoLifeCLEF 2023 dataset is provided in [4].

6.3 Evaluation Protocol

The challenge is as a multi-label (/set) classification task. Given a test set of locations (i.e., geo-coordinates) and corresponding remote sensing data and environmental covariates, the goal of the task is to return for each location the set of plant species truly present in a small spatial plot (of area 10-400m²) as reported in a standardized presence-absence survey carried by botanical experts (same type as the validation PA data). This test set includes 22,404 PA surveys. Thus, one of the major difficulties of the challenge is to predict the presence or absence of all species from a dataset mostly made of PO data (i.e. the 5 million GIF records). As noted earlier, to enable participants to calibrate their models, and specifically to correct for sampling biases, we also provided a validation set of only 5 thousand PA surveys, spatially separated from the test set. Indeed, following the recommendations of [46], the split of the validation and test set was done using a spatial blocking strategy that enables a more robust estimation of the model's performance (based on a 50x50km spatial grid). Moreover, we have excluded the PO records located less than 500m from the test plots to avoid the risk that some may have originated from these plots. The detailed protocol is described in [4].
The evaluation metric is the F1 score. It measures the precision and recall score for each test plot x and computes their harmonic mean:

$$F1(x) = \frac{2}{\frac{1}{Precision(x)} + \frac{1}{Recall(x)}}$$

It is equivalent to the Sørensen-Dice coefficient defined as the size of the intersection between the predicted and true set of species, divided by the mean of their respective size.
The final global metric is calculated by averaging the F1 score of all plots in the test set.

6.4 Participants and Results

Six participants from four countries participated in the GeoLifeCLEF 2023 challenge and submitted a total of 121 entries (i.e. /textitruns). Details of the best methods and systems used are synthesized in the overview paper of the task [5] and the winning team methodology is explained in details in their working note ([51]). In Table 4 we report the performance achieved by the best performing methods of the participants as well as the baseline methods developed by the organizers. Hereafter, we briefly describe those different methods:

Participant's methods

- **KDDI research**: This team trained various convolutional neural networks, all based on the ResNet backbone (ResNet34 and 50). One of the CNN was trained solely on the 19 bioclimatic rasters while others were multi-modal

networks with a late fusion layer to merge the different modalities used (see Table 4). The best performing run was an ensemble of the best models based on a simple average of their output. The best models were trained in three steps, firstly on the PA plots with a binary cross-entropy loss, then fine-tuned on the PO records with a cross-entropy loss, and finally fine-tuned again on the PA with the binary-cross-entropy loss. This team carried an ablation showing the importance of these three steps.
- **Jiexun Xu**: This researcher focused on the tabular environmental data only, i.e he didn't use the spatial structure of the environmental co-variates nor the remotes sensing images and times series. The model used is XGBoost and it was trained on the PA plots. He also added the one-hot encoded species presences in GBIF in a 1km radius of these plots as input variables.
- **Lucas Morin**: This researcher optimized a K-Nearest Neighbor predictor using only the spatial coordinates and the PA plots.
- **QuantMetry**: This team trained various models on the PA data, and their best scoring model was a ResNet50 using only the Sentinel2 satellite images (RGB+NIR) as input. The model was pre-trained on the satellite images in a prior work ([60]) and fine-tuned to the PA data in the challenge.
- **Nina van Tiel**: This researcher used a small CNN, with two convolutional layers and two fully connected layers on the RGB images, along the bioclimatic, soil and land-cover rasters, trained on the PA plots.
- **Ousmane Youme**: This researcher focused solely on the Landsat time series data at the location of the PA plots. He used a Conv1D neural network model with a binary-cross entropy loss. A common probability threshold was used to convert the predicted species-wise presence probabilities into a set of predicted species.

Organizer's baselines

- **MAXENT**: the MAXENT method is a modeling approach widely used in ecology to predict the distribution of a given species based on tabular environmental variables. It is not adapted to handle complex input data such as the Sentinel images or Landsat time series. The model creates a predefined set of non-linear transformations of the input environmental variables consistent with the theoretical ecological response of species to environmental gradients (e.g. quadratic and threshold responses, see [36]). The statistical model is equivalent to a Poisson regression modeling the count of a species per location ([45]). We fitted one Maxent model per species present in the PA plots. The species count was set to one when present or zero otherwise. The environmental input variables included were the climate, soil, land cover and human footprint variables, but only a subset of these variables were included for species with a smal number of observations. One random subset of the PA plots was used to train all species models while the other was used to assess the predictive accuracy of each species model. We thus determined that it was optimal to keep only the 391 most trustable species, in terms of validation score, for the final prediction, the left-out species being always

predicted absent. A run including on all species models was also submitted, achieving a much lower performance due to an over-prediction of rare species in extrapolation (see [5] for details).
- **Environmental Random Forest**: Random forests are also widely used in ecology to predict the distribution of species based on a set of environmental variables. As for Maxent, the Env. Random Forest models were trained only on the environmental tabular variables at the location of the PA plots. One Random Forest was trained per species in the PA plots and its hyper-parameters were optimized through a cross-validation grid search.
- **Spatial Random Forest**: Contrary to the two previous baseline, this Random Forest were trained solely on the spatial coordinates of the PA plots, regardless the environmental variables.
- **Species co-occurrence**: Conditionally to the presence of each species, we computed the proportion of presences of all other species among the PA plots. Then, for each test location, we combined the species probabilities conditionally to the species observed in the PO data in a 1km-radius into a predicted species set through a weighted average. Therefore, this method doesn't use any input variable except the spatial coordinates.
- **Constant predictor**: this baseline always predict the same set of species, i.e. the ones that are the K most frequent in the PA plots, where K maximizes the F1-micro score over these PA plots ($K = 25$ species).

Outcomes

The main outcomes we can derive from the challenge are the following:

- The problem remains very difficult and the best model only achieves a F1-score of 0.27
- The MAXENT method remains a strong baseline when considering only the tabular environmental data, regardless the spatial structure of the environment or the more complex data such as remote sensing images.
- Training a model on the PO data (with a cross-entropy loss) and fine-tuning it on PA (with a binary cross-entropy loss) resulted in a considerable performance gain. This shows the wealth of information that can be mobilised in the PO data, provided that the learning strategy avoids sampling biases.
- The best model was based on a Convolutional Neural Network which confirms that this kind of model is relevant for the task. It allows capturing complex patterns in the input data while allowing elaborated training strategy such as transfer learning.
- Making use exclusively of PO data remains a major hurdle, and all the methods that did so had a very low performance. Most participants used only the PA validation data in the training of their models, and the best method succeeded by combining both. Much work lies ahead to extract the information from PO without complementary standardized data, if that is even possible.

7 Conclusions and Perspectives

The main outcome of this collaborative evaluation is a new snapshot of the performance of state-of-the-art computer vision, bioacoustic, and machine learning techniques toward building real-world biodiversity monitoring systems. Overall, this study shows that the field continues to progress year after year, and that, although the challenges that are most closely related to common tasks, such as multi-class classification based on images, are able to profit from the most recent advances in computer vision, certain problems are still wide open, such as the prediction of species as a function of location (as part of the GeoLifeCLEF challenge). In terms of the methods used, the results show that convolutional neural networks are still a very powerful method for image and sound processing. In 4 of the 5 challenges, the best results were obtained using CNNs. Only the PlantCLEF challenge obtained much better results (for the identification of plants from images) with the use of foundation vision transformer models such as EVA [9]. The best submission to FungiCLEF was based on MetaFormer [8], utilizing both a convolutional backbone and a transformer to fuse visual and meta information. Complementary to vision-based models, NLP models were also used successfully, in particular hybrid models such as CLIP [44] that efficiently learn visual concepts from natural language supervision. We believe that this principle of combining different modalities in the training of deep learning models will be a key to future progress in AI for biodiversity.

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. Overview of SnakeCLEF 2023: Snake identification in medically important scenarios. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
2. Banan, A., Nasiri, A., Taheri-Garavand, A.: Deep learning-based appearance features extraction for automated carp species identification. Aquacultural Engineering **89**, 102053 (2020)
3. 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)
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. Botella, C., Deneu, B., Marcos Gonzalez, D., Servajean, M., Larcher, T., Leblanc, C., Estopinan, J., Bonnet, P., Joly, A.: Overview of GeoLifeCLEF 2023: Species

composition prediction with high spatial resolution at continental scale using remote sensing. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
6. Cai, J., Ee, D., Pham, B., Roe, P., Zhang, J.: Sensor network for the monitoring of ecosystem: Bird species recognition. In: Intelligent Sensors, Sensor Networks and Information, 2007. ISSNIP 2007. 3rd International Conference on (2007). https://doi.org/10.1109/ISSNIP.2007.4496859
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
8. 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)
9. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale (2022)
10. 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)
11. Ghazi, M.M., Yanikoglu, B., Aptoula, E.: Plant identification using deep neural networks via optimization of transfer learning parameters. Neurocomputing **235**, 228–235 (2017)
12. Glotin, H., Clark, C., LeCun, Y., Dugan, P., Halkias, X., Sueur, J.: Proc. 1st workshop on Machine Learning for Bioacoustics - ICML4B. ICML, Atlanta USA (2013), `http://sabiod.org/ICML4B2013_book.pdf`
13. Glotin, H., LeCun, Y., Artières, T., Mallat, S., Tchernichovski, O., Halkias, X.: Proc. Neural Information Processing Scaled for Bioacoustics, from Neurons to Big Data. NIPS Int. Conf. (2013), `http://sabiod.org/nips4b`
14. 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)
15. 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)
16. 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)
17. 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)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
19. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
20. Joly, A., Goëau, H., Bonnet, P., Bakić, V., Barbe, J., Selmi, S., Yahiaoui, I., Carré, J., Mouysset, E., Molino, J.F., et al.: Interactive plant identification based on social image data. Ecological Informatics **23**, 22–34 (2014)

21. 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)
22. 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
23. 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
24. 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
25. 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
26. 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)
27. 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)
28. 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)
29. 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)
30. Kahl, S., Denton, T., Klinck, H., Reers, H., Cherutich, F., Glotin, H., Goëau, H., Vellinga, W.P., Planqué, R., Joly, A.: Overview of BirdCLEF 2023: Automated bird species identification in eastern africa. Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
31. Lee, D.J., Schoenberger, R.B., Shiozawa, D., Xu, X., Zhan, P.: Contour matching for a fish recognition and migration-monitoring system. In: Optics East. pp. 37–48. International Society for Optics and Photonics (2004)
32. Lee, S.H., Chan, C.S., Remagnino, P.: Multi-organ plant classification based on convolutional and recurrent neural networks. IEEE Transactions on Image Processing **27**(9), 4287–4301 (2018)
33. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
34. Norouzzadeh, M.S., Morris, D., Beery, S., Joshi, N., Jojic, N., Clune, J.: A deep active learning system for species identification and counting in camera trap images. Methods in Ecology and Evolution **12**(1), 150–161 (2021)
35. Ovalle, J.C., Vilas, C., Antelo, L.T.: On the use of deep learning for fish species recognition and quantification on board fishing vessels. Marine Policy **139**, 105015 (2022)
36. Phillips, S.J., Dudík, M.: Modeling of species distributions with maxent: new extensions and a comprehensive evaluation. Ecography **31**(2), 161–175 (2008)
37. Picek, L., Šulc, M., Chamidullin, R., Durso, A.M.: Overview of snakeclef 2023: Snake identification in medically important scenarios. In: CLEF 2023-Conference and Labs of the Evaluation Forum (2023)
38. Picek, L., Šulc, M., Chamidullin, R., Matas, J.: Overview of fungiclef 2023: Fungi recognition beyond 1/0 cost. In: CLEF 2023-Conference and Labs of the Evaluation Forum (2023)
39. 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)
40. 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)
41. 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)
42. 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)

43. Pitman, N.C., Suwa, T., Ulloa Ulloa, C., Miller, J., Solomon, J., Philipp, J., Vriesendorp, C.F., Derby Lewis, A., Perk, S., Bonnet, P., et al.: Identifying gaps in the photographic record of the vascular plant flora of the americas. Nature plants **7**(8), 1010–1014 (2021)
44. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
45. Renner, I.W., Warton, D.I.: Equivalence of maxent and poisson point process models for species distribution modeling in ecology. Biometrics **69**(1), 274–281 (2013)
46. 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)
47. 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)
48. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
49. Towsey, M., Planitz, B., Nantes, A., Wimmer, J., Roe, P.: A toolbox for animal call recognition. Bioacoustics **21**(2), 107–125 (2012)
50. Trifa, V.M., Kirschel, A.N., Taylor, C.E., Vallejo, E.E.: Automated species recognition of antbirds in a mexican rainforest using hidden markov models. The Journal of the Acoustical Society of America **123**, 2424 (2008)
51. Ung, H.Q., Kojima, R., Wada, S.: Leverage samples with single positive labels to train cnn-based models for multi-label plant species prediction. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
52. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. CVPR (2018)
53. Villon, S., Mouillot, D., Chaumont, M., Subsol, G., Claverie, T., Villéger, S.: A new method to control error rates in automated species identification with deep learning algorithms. Scientific reports **10**(1), 1–13 (2020)
54. Wäldchen, J., Rzanny, M., Seeland, M., Mäder, P.: Automated plant species identification—trends and future directions. PLoS computational biology **14**(4), e1005993 (2018)
55. Wang, J., Zhang, W., Zang, Y., Cao, Y., Pang, J., Gong, T., Chen, K., Liu, Z., Loy, C.C., Lin, D.: Seesaw loss for long-tailed instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9695–9704 (2021)
56. Wang, Y., Zhang, Y., Feng, Y., Shang, Y.: Deep learning methods for animal counting in camera trap images. In: 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI). pp. 939–943. IEEE (2022)
57. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. arXiv preprint arXiv:2301.00808 (2023)
58. Yu, X., Wang, J., Kays, R., Jansen, P.A., Wang, T., Huang, T.: Automated identification of animal species in camera trap images. EURASIP Journal on Image and Video Processing (2013)

59. Yuan, L., Hou, Q., Jiang, Z., Feng, J., Yan, S.: Volo: Vision outlooker for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
60. Zheng, Z., Ma, A., Zhang, L., Zhong, Y.: Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 15193–15202 (2021)

Table 4: Overview of the results of GeoLifeCLEF 2023 challenge - the acronyms PA and PO respectively stand for (PA) Presence/Absence: meaning that the plots in the validation set were used to fit the model, and (PO) Presence-Only which means that the GBIF occurrences of the training set were used to fit the model.

Team/scientist	Method	Used data	Used modalities	Score
KDDI research	Ensemble of CNN models Multimodal & bioclim	PO & PA	Sentinel-2 RGB-NIR, soil, bio-climatic, human footprint	0.270
KDDI research	Multi-modal CNN (3 x ResNet-50)	PO & PA	Sentinel-2 RGB-NIR, soil, bio-climatic, human footprint	0.249
KDDI research	Bioclim CNN (ResNet-50 w/ 19 channels)	PO & PA	bio-climatic	0.239
Organizer (baselines)	MAXENT (391 most confident species)	PA	soil, bio-climatic, human footprint	0.224
Jiexun Xu	XGBoost	PO & PA	soil, bio-climatic, human footprint	0.223
Lucas Morin	K-Nearest Neighbors (K=500)	PA	lat. / long.	0.208
Quantmetry	ResNet50	PA	Sentinel-2 RGB-NIR	0.206
Organizer (baselines)	Spatial Random Forest	PA	lat. / long.	0.191
Organizer (baselines)	Env. Random Forest	PA	soil, bio-climatic, human footprint	0.188
Organizer (baselines)	Species co-occurrence	PA/PO	lat./long.	0.167
Organizer (baselines)	Constant predictor	PA	NONE	0.160
Nina van Tiel	Small CNN	PA	Sentinel-2 RGB, soil, bio-climatic, human footprint	0.158
Ousmane Youm	Conv1d CNN	PA	Landsat time series	0.134