Overview of LifeCLEF 2022: an evaluation of Machine-Learning based Species Identification and Species Distribution Prediction

Alexis Joly[1], Hervé Goëau[2], Stefan Kahl[6], Lukáš Picek[9], Titouan Lorieul[1], Elijah Cole[9], Benjamin Deneu[1], Maximilien Servajean[7], Andrew Durso[10], Hervé Glotin[3], Robert Planqué[4], Willem-Pier Vellinga[4], Amanda Navine[13], Holger Klinck[6], Tom Denton[11], Ivan Eggel[5], Pierre Bonnet[2], Milan Šulc[12], Marek Hrúz[9]

[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] HES-SO, Sierre, Switzerland
[6] KLYCCB, Cornell Lab of Ornithology, Cornell University, USA
[7] LIRMM, AMI, 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 LLC, San Francisco, USA
[12] Rossum.ai, Prague, Czech Republic
[13] Listening Observatory for Hawaiian Ecosystems, Univ. of Hawai'i at Hilo, USA

Abstract. Building accurate knowledge of the identity, the geographic distribution and the evolution of species is essential for the sustainable development of humanity, as well as for biodiversity conservation. However, the difficulty of identifying plants, animals and fungi is hindering the aggregation of new data and knowledge. Identifying and naming living organisms is almost impossible for the general public and is often difficult even for professionals and naturalists. Bridging this gap is a key step towards enabling effective biodiversity monitoring systems. The LifeCLEF campaign, presented in this paper, has been promoting and evaluating advances in this domain since 2011. The 2022 edition proposes five data-oriented challenges related to the identification and prediction of biodiversity: (i) PlantCLEF: very large-scale plant identification, (ii) BirdCLEF: bird species recognition in audio soundscapes, (iii) GeoLifeCLEF: remote sensing based prediction of species, (iv) SnakeCLEF: snake species identification on a global scale, and (v) FungiCLEF: fungi recognition as an open set classification problem. 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. Unfortunately, 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 world-wide 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 [17] 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 [17], automated species identification has been studied in many contexts [4,19,20,32,50,75,76,86]. This area continues to expand rapidly, particularly due to advances in deep learning [3,18,51,60,78,79,80,81]. In order to measure progress in a sustainable and repeatable way, the LifeCLEF[2] research platform 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 [22,23,24]. Since 2014, LifeCLEF expanded the challenge by considering animals and fungi in addition to plants, and including audio and video content in addition to images [33,34,35,36,37,38,39,40]. Nearly a thousand researchers and data scientists register yearly to LifeCLEF in order to either 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 (including the 1,004 participants of the BirdCLEF Kaggle challenge). The 2022 edition proposes five data-oriented challenges: three in the continuity of the 2021 edition (BirdCLEF, GeoLifeCLEF and SnakeCLEF), one new challenge related to fungi recognition with a focus on the combination of visual information with meta-data on an open species set (FungiCLEF), and a considerable expansion of the PlantCLEF challenge towards the identification of the world's flora (about 300K species).

The system used to run the challenges (registration, submission, leaderboard, etc.) was the AICrowd platform[4] for the PlantCLEF challenge and the Kaggle platform[5] for the GeoLifeCLEF, BirdCLEF, SnakeCLEF and FungiCLEF challenges. Three of the challenges (GeoLifeCLEF, SnakeCLEF, and FungiCLEF)

[1] `https://www.cbd.int/`
[2] `http://www.lifeclef.org/`
[3] `http://www.imageclef.org/`
[4] `https://www.aicrowd.com`
[5] `https://www.kaggle.com`

were organized jointly with FGVC[6], an annual workshop dedicated to Fine-Grained Visual Categorization organized in the context of the CVPR[7] international conference on computer vision and pattern recognition.
In total, 951 people/teams participated to LifeCLEF 2022 edition by submitting runs to at least one of the five challenges (802 only for the BirdCLEF challenge). Only some of them managed to get the results right, and about 30 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[8]). 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 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 [21].

2.1 Objective

Automated identification of plants has recently improved considerably thanks to the progress of deep learning and the availability of training data with more and more photos in the field. In the context of LifeCLEF 2018, we measured a top-1 classification accuracy over 10K species up to 90 % and we showed that automated systems were not so far from human expertise [33]. However, these very high performances are far from being reached at the scale of the world flora. It is estimated that there are about 391,000 vascular plant species currently known to science and new plant species are still discovered and described each year. This plant diversity is a major element in the functioning of ecosystems as well as for the development of human civilization. Unfortunately, the vast majority of these species are very poorly known and the number of training images available is extremely low for the majority of them [66].
The goal of the 2022 edition of PlantCLEF was to take another step towards identifying the world's flora. Therefore, we have built a training set of unprecedented size covering 80K species and containing 4M images. It was shared with the community through a challenge[9] hosted on the AIcrowd platform.

2.2 Dataset

The training set is composed of two subsets: a trusted training set coming from the GBIF[10] portal (the world's largest biodversity data portal) and a web-based

[6] http://www.fgvc.org/
[7] https://cvpr2022.thecvf.com/
[8] http://ceur-ws.org/
[9] https://www.aicrowd.com/challenges/lifeclef-2022-plant
[10] https://gbif.org/

training set containing images collected via web search engines and containing several kinds of noise.

More precisely, the GBIF training dataset is based on a selection of more than 2.9M images covering 80k plant species shared and collected mainly via GBIF (and Encyclopedia Of Life [11] to a lesser extent). These images come mainly from academic sources (museums, universities, national institutions) and collaborative platforms such as inaturalist or Pl@ntNet, implying a fairly high certainty of determination quality (collaborative platforms only share their highest quality data qualified as "research graded"). To limit the size of the training set and limit class imbalance, the number of images was limited to around 100 images per species, favouring types of views adapted to the identification of plants (close-ups of flowers, fruits, leaves, trunks, ...).

The web dataset, on the other side, is based on a collection of web images provided by commercial search engines (Google and Bing). The raw downloaded data has a significant rate of species identification errors and a massive presence of (near)-duplicates and images not adapted for the identification of plant photographs (e.g. herbarium sheets, landscapes, microscopic views, ...). It even contains completely off-topic images such as portrait photos of botanists, maps, graphs, other kingdoms of the living, manufactured objects, etc. Thus, the raw data was cleaned up using a semi-automatic filtering (iterations of CNNs training, inference and human labelling). This filtering process drastically reduced the number of irrelevant pictures and also improved the overall image quality by favoring close-ups of flowers, fruits, leaves, trunks, etc. The web dataset finally contains about 1.1 million images covering about 57k 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.

Lastly, the test set was built from multi-image plant observations collected on the Pl@ntNet platform during the year 2021 (observations not yet shared through GBIF, and thus 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.

[11] `https://eol.org/`

2.3 Evaluation Protocol

The primary metrics used for the evaluation of the task is be the Mean Reciprocal Rank. The MRR is a statistic measure for evaluating any process that produces a list of possible responses to a sample of queries ordered by probability of correctness. The reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct answer. The MRR is the average of the reciprocal ranks for the whole test set:

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

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) was used because of the long tail of the data distribution to rebalance the results between under- and over-represented species in the test set.

2.4 Participants and Results

Eight participants registered to the PlantCLEF challenge hosted on AICrowd but only four of them managed to perform well. The four others encountered difficulties mainly related to the very large scale of the challenge (both in terms of the number of images and number of classes) and the need of high ended GPUs for resource-intensive experiments. Details of the methods and systems used are synthesized in the extended overview working note of the challenge [21] and further developed in the individual working notes of participants ([5,8,46,58,67,85]. We report in Figure 1 the performance achieved by the different runs of the participants.

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

- the best results were obtained by the only team which used vision transformers [85] contrary to the others which used convolutional neural networks, i.e. the traditional approach of the state-of-the-art for image-based plant identification. However, this gain in identification quality is paid for by a significant increase of the training time. The winning team reported that they had to stop the training of the model in order to submit their run to the challenge. Thus, better results could have surely been obtained with a few more days of training (as demonstrated through post-challenge evaluations reported in the their working note [85].
- One of the main difficulties of the challenge was the very large number of classes (80K). For most of the models used, the majority of the weights to be trained are those of the last fully connected layer of the classifier. This was an important consideration for all participants in their model selection strategy. Some teams have tried to limit this cost through specific approaches. The BioMachina team [5], in particular, used a two-level hierarchical softmax

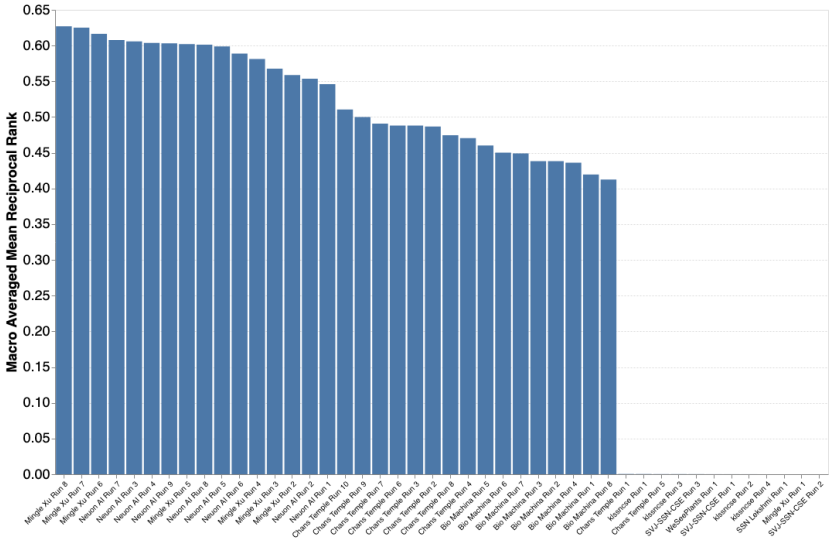


Fig. 1: PlantCLEF 2022 results

to reduce the number of weights drastically. They reported an considerable training time reduction while maintaining almost the same identification quality.

3 BirdCLEF Challenge: Bird call identification in soundscape recordings

A detailed description of the challenge and a more complete discussion of the results can be found in the dedicated working note [43].

3.1 Objective

The *LifeCLEF Bird Recognition Challenge* (BirdCLEF) was launched in 2014 and has since become the largest bird sound recognition challenge in terms of dataset size and species diversity, with multiple tens of thousands of recordings covering up to 1,500 species [25,41,42,44]. Birds are ideal indicators to identify early warning signs of habitat changes that are likely to affect many other species. They have been shown to respond to various environmental changes over many spatial scales. Large collections of (avian) audio data are an excellent resource to conduct research that can help to deal with environmental challenges of our time. The community platform Xeno-canto[12] in particular was launched in 2005 and hosts bird sounds from all continents. It receives new recordings every day

[12] `https://www.xeno-canto.org/`

from some of the remotest places on Earth. The Xeno-canto archive currently consists of more than 700,000 focal recordings covering over 10,000 species of birds, making it one of the most comprehensive collections of bird sound recordings worldwide, and certainly the most comprehensive collection shared under Creative Commons licenses. Xeno-canto data was used for BirdCLEF in all past editions to provide researchers with large and diverse datasets for training and testing.

In recent years, research in the domain of bioacoustics shifted towards deep neural networks for sound event recognition [45,72]. In past editions, we have seen many attempts to utilize convolutional neural network (CNN) classifiers to identify bird calls based on visual representations of these sounds (i.e., spectrograms) [26,48,59]. Despite their success for bird sound recognition in focal recordings, the classification performance of CNN on continuous, omnidirectional soundscapes remained low. Passive acoustic monitoring can be a valuable sampling tool for habitat assessments and the observation of environmental niches which often are endangered. However, manual processing of large collections of soundscape data is not desirable and automated attempts can help to advance this process [83]. Yet, the lack of suitable validation and test data prevented the development of reliable techniques to solve this task. Bridging the acoustic gap between high-quality training recordings and soundscapes with high ambient noise levels is one of the most challenging tasks in the domain of audio event recognition. This is especially true when sufficient amounts of training data are lacking. This is the case for many rare and endangered bird species around the globe and despite the vast amounts of data collected on Xeno-canto, audio data for endangered birds is still sparse. However, it is those endangered species that are most relevant for conservation, rendering acoustic monitoring of endangered birds particularly difficult.

The main goal of the 2022 edition of BirdCLEF was to advance automated detection of rare and endangered bird species that lack large amounts of training data. The competition was hosted on Kaggle[13] to attract machine learning experts from around the world to participate and submit. The overall task design was consistent with previous editions, but the focus was shifted towards species with very few training samples.

3.2 Dataset and Evaluation Protocol

As the "extinction capital of the world," Hawai'i has lost 68% of its bird species, the consequences of which can harm entire food chains. Researchers use population monitoring to understand how native birds react to changes in the environment and conservation efforts. But many of the remaining birds across the islands are isolated in difficult-to-access, high-elevation habitats. With physical monitoring difficult, scientists have turned to sound recordings. This approach could provide a passive, low labor, and cost-effective strategy for studying endangered bird populations.

[13] `https://www.kaggle.com/c/birdclef-2022`

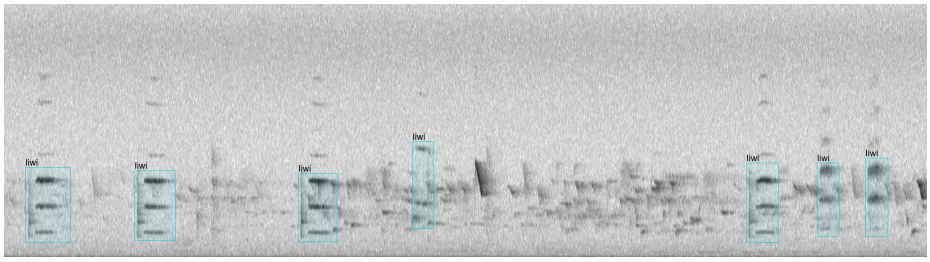


Fig. 2: Expert ornithologists provided bounding box labels for all soundscape recordings indicating calling of 21 target species. In this example, all ‘I‘iwi calls were annotated, while vocalizations of other species were not labeled. This labeling scheme was applied to all test data soundscapes.

Current methods for processing large bioacoustic datasets involve manual annotation of each recording. This requires specialized training and prohibitively large amounts of time. Thankfully, recent advances in machine learning have made it possible to automatically identify bird songs for common species with ample training data. However, it remains challenging to develop such tools for rare and endangered species, such as those in Hawai’i.

Deploying a bird sound recognition system to a new recording and observation site requires classifiers that generalize well across different acoustic domains. Focal recordings of bird species form an excellent base to develop such a detection system. However, the lack of annotated soundscape data for a new deployment site poses a significant challenge. As in previous editions, training data was provided by the Xeno-canto community and consisted of more than 14,800 recordings covering 152 species. Participants were allowed to use metadata to develop their systems. Most notably, we provided detailed location information on recording sites of focal and soundscape recordings, allowing participants to account for migration and spatial distribution of bird species.

In this edition, test data, consisting of 5,356 soundscapes amounting to more than 90 hours of recordings, were hidden and only accessible to participants during the inference process. These soundscapes were collected for various research projects by the Listening Observatory for Hawaiian Ecosystems (LOHE) at the University of Hawai‘i at Hilo from 7 sites across the islands of Hawai‘i, Maui, and Kaua‘i. All soundscapes received some level of manual bird vocalization annotation by specially trained members of the LOHE lab using Raven Pro 1.5 software, however some recordings had a select few target species annotated, while others were annotated for every detectable species (see Figure 2). In light of these uneven annotation strategies, only the subset of species for which every vocalization was annotated were scored for any given file. This resulted in a total of 21 scored bird species in the contest, 15 species endemic to the Hawaiian Islands and 6 introduced species.

The goal of the task was to localize and identify 21 target bird species within the provided soundscape test set. Each soundscape was divided into segments of

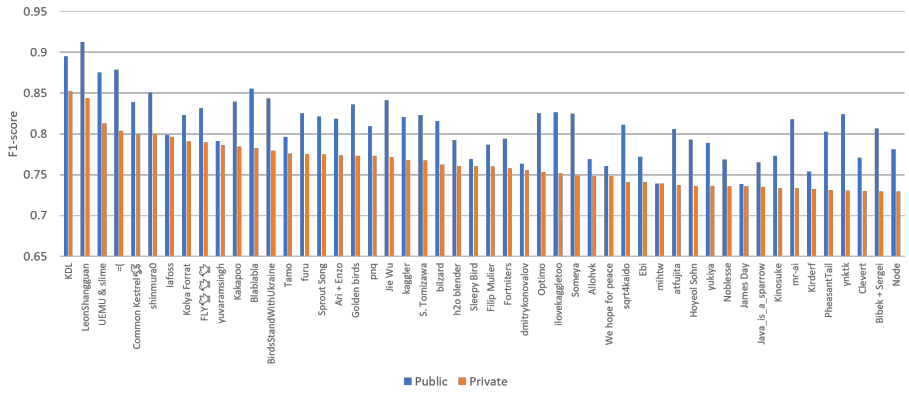


Fig. 3: Scores achieved by the best systems evaluated within the bird identification task of LifeCLEF 2022.

5 seconds, and a list of audible species had to be returned for each segment. The used evaluation metric was a weighted variant of the macro-averaged F1-score. In previous editions, ranking metrics were used to assess the overall classification performance. However, when applying bird call identification systems to real-world data, confidence thresholds have to be set in order to provide meaningful results. The F1-score as balanced metric between recall and precision appears to better reflect this circumstance. For each 5-second segment, a binary call indication for all 21 scored species had to be returned. Participants had to apply a threshold to determine if a species is vocalizing during a given segment (True) or not (False).

3.3 Participants and Results

1,019 participants from 62 countries on 807 teams entered the BirdCLEF 2022 competition and submitted a total of 23,352 runs. Details of the best methods and systems used are synthesized in the overview working notes paper of the task [43] and further developed in the individual working notes of participants. In Figure 3 we report the performance achieved by the top 50 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 F1-score in this year's edition was 0.5112 (public 0.4849) with all scored birds marked as silent (False) for all segments, and 665 teams managed to score above this threshold. The best submission achieved a F1-score of 0.8527 (public 0.9128) and the top 10 best performing systems were within only 7% 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. Participants employed various sophisticated post-processing schemes, most notably a percentile based thresholding approach that was established during the 2021 edition [28]. Some participants experimented with different loss functions, especially focal loss being the most notable. However, results were inconsistent across teams. Some teams used audio transformers, but again, results were inconsistent and led to discussions about whether these methods were appropriate for the task of bird call identification.

4 GeoLifeCLEF Challenge: Predicting Species Presence From Multi-Modal Remote Sensing, Bioclimatic and Pedologic Images

A detailed description of the challenge and a more complete discussion of the results can be found in the dedicated working note [57].

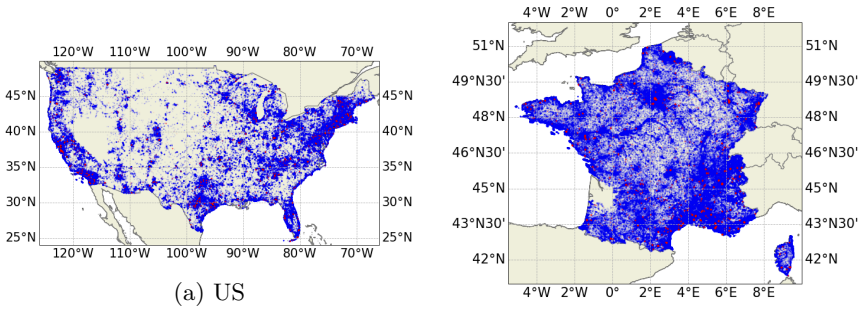
4.1 Objective

Automatic prediction of the list of species most likely to be present at a given location is useful for many scenarios related to biodiversity management and conservation. First, it can improve species identification tools (whether automatic, semi-automatic or based on traditional field guides) by reducing the list of candidate species observable at a given site. Moreover, it can facilitate decision making related to land use and land management with regard to biodiversity conservation obligations (e.g., to determine new constructible areas or new natural areas to be protected). Last but not least, it can be used in the context of educational and citizen science initiatives, e.g., to determine regions of interest with a high species richness or vulnerable habitats to be monitored carefully.

4.2 Data Set and Evaluation Protocol

Data collection. The data for this year's challenge is a cleaned-up version of the data from previous years, essentially removing species integrated by error and those observed less than 3 times. A detailed description of the GeoLifeCLEF 2020 dataset is provided in [9] and a complete changelog of the cleaning process is available on the Kaggle page[14]. In a nutshell, the dataset consists of over 1.6 million observations covering $17,037$ plant and animal species distributed across US and France (as shown in Figure 4). Each species observation is paired with high-resolution covariates (RGB-NIR imagery, land cover and altitude data) as illustrated in Figure 5. These high-resolution covariates are resampled to a spatial resolution of 1 meter per pixel and provided as 256×256

[14] `https://www.kaggle.com/c/geolifeclef-2022-lifeclef-2022-fgvc9/data`



(a) US

(b) France

Fig. 4: Observations distribution over the US and France in GeoLifeCLEF 2022. Blue dots represent training data, red dots represent test data.

images covering a 256m × 256m square centered on each observation. RGB-NIR imagery come from the 2009-2011 cycle of the National Agriculture Imagery Program (NAIP) for the US[15], and from the BD-ORTHO® 2.0 and ORTHO-HR® 1.0 databases from the IGN for France[16]. Land cover data originates from the National Land Cover Database (NLCD) [30] for the U.S. and from CESBIO[17] for France. All elevation data comes from the NASA Shuttle Radar Topography Mission (SRTM)[18]. In addition, the dataset also includes traditional coarser resolution covariates: bio-climatic rasters ($1km^2$/pixel, from WorldClim [29]) and pedologic rasters ($250m^2$/pixel, from SoilGrids [27]).

Train-test split. The full set of occurrences is split in a training and testing set using a spatial block holdout procedure to limit the effect of *spatial auto-correlation* in the data [69]. Using this splitting procedure, a model cannot achieve a high performance by simply interpolating between training samples. The split was based on a global grid of 5km × 5km quadrats. 2.5% of these quadrats were randomly sampled and the observations falling in those formed the test set. 10% of those observations were used for the public leaderboard on Kaggle while the remaining 90% allowed to compute the private leaderboard providing the final results of the challenge. Similarly, another 2.5% of the quadrats were randomly sampled to provide an official validation set. The remaining quadrats and their associated observations were assigned to the training set.

Evaluation metric. For each occurrence in the test set, the goal of the task was to return a candidate set of species likely to be present at that location. To measure the precision of the predicted sets, top-30 error rate was chosen as the main evaluation criterion. Each observation i is associated with a single ground-truth label y_i corresponding to the observed species. For each observation, the

[15] `https://www.fsa.usda.gov`
[16] `https://geoservices.ign.fr`
[17] `http://osr-cesbio.ups-tlse.fr/~oso/posts/2017-03-30-carte-s2-2016/`
[18] `https://lpdaac.usgs.gov/products/srtmgl1v003/`

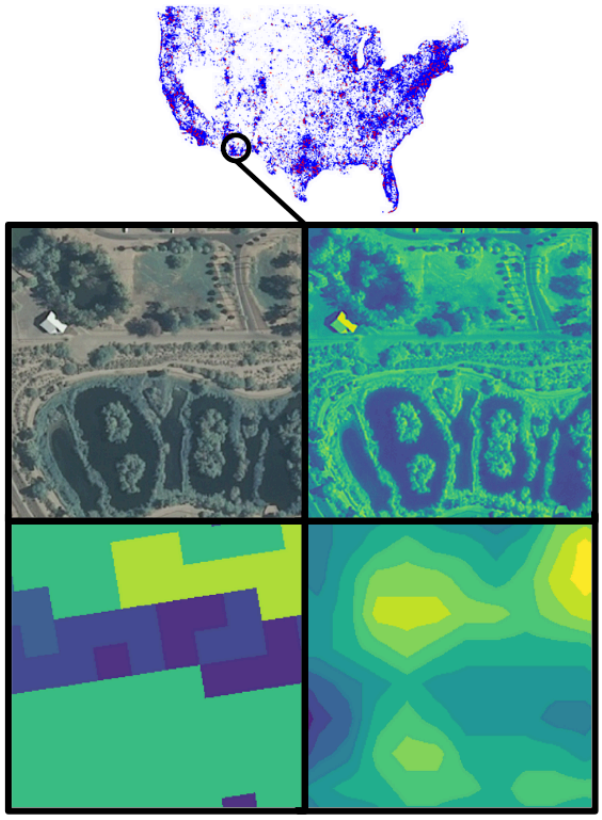


Fig. 5: In the GeoLifeCLEF dataset, each species observation is paired with high-resolution covariates (clockwise from top left: RGB imagery, IR imagery, altitude, land cover).

submissions provided 30 candidate labels $\hat{y}_{i,1}, \hat{y}_{i,2}, \ldots, \hat{y}_{i,30}$. The top-30 error rate was then computed using

$$\text{Top-30 error rate} = \frac{1}{N} \sum_{i=1}^{N} e_i, \tag{2}$$

where

$$e_i = \begin{cases} 1 & \text{if } \forall k \in \{1, \ldots, 30\},\ \hat{y}_{i,k} \neq y_i \\ 0 & \text{otherwise} \end{cases}. \tag{3}$$

Note that this evaluation metric does not try to correct the sampling bias inherent to present-only observation data (linked to the density of population, etc.). The absolute value of the resulting figures should thus be taken with care. Nevertheless, this metric does allow to compare the different approaches and

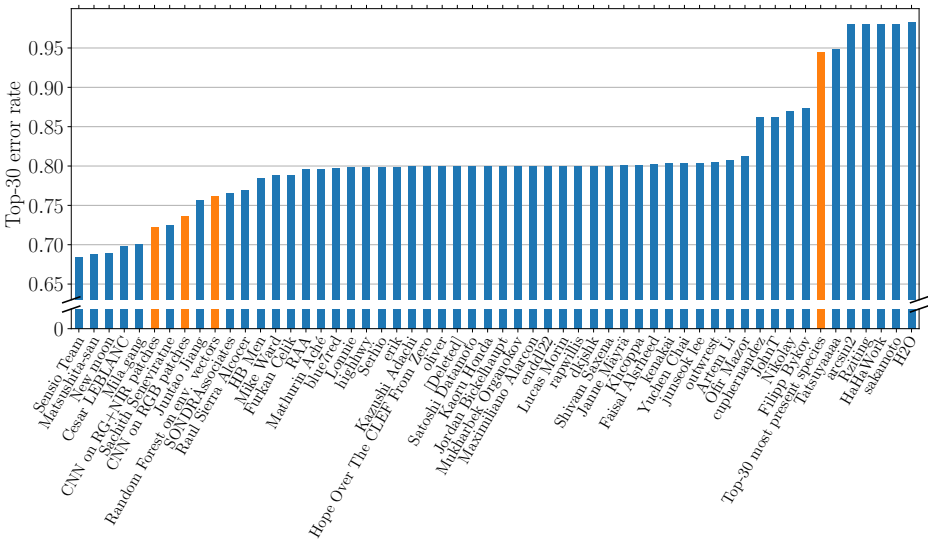


Fig. 6: Results of the GeoLifeCLEF 2022 task. The top-30 error rates of the best submission of each participant are shown in blue. The provided baselines are shown in orange.

to determine which type of input data and of models are useful for the species presence detection task.

4.3 Participants and Results

52 teams participated and submitted at least one prediction file through the Kaggle[19] page of the GeoLifeCLEF 2022 challenge for a total number submission in the course of the competition of 261. Out of these teams, 7 managed to beat the weakest (non-constant) baseline provided and 5 the strongest one. Details of the baselines provided and of the methods used in the submitted runs are synthesized in the overview working note paper for this task [57]. The runs of 5 of those participants are further developed in their individual working notes [31,47,49,74,87]. Figure 6 shows the final standings given by the Kaggle's private leaderboard. We briefly highlight the main methods used by the participants.

Multi-modal data. The main challenge of this competition is to find a proper way to aggregate the heterogeneous sources of data and to deal with their respective characteristics: while RGB and NIR patches are standard images, other data is not directly provided in this format. For instance, altitude can not be casted in `uint8` without loss of information, land cover data is a categorical variable, bioclimatic and pedologic data have a resolution and range of their own, and, localisation (GPS coordinates) is a punctual information. Interestingly, the participants did try different means of aggregating this heterogeneous

[19] `https://www.kaggle.com/c/geolifeclef-2022-lifeclef-2022-fgvc9`

data with more or less success and conflicting result. For instance, [49] tried the most straight-forward and easy to implement approach: train separate models and average their predictions. The winning team and [47,74] used complete networks as feature extractors for each chosen modality separately, concatenated the resulting representation and fed it to a final classifier (single or multiple linear layers). This is the approach which was chosen by GeoLifeCLEF 2021 winning solution [71]. [87] used single-layer features extractors which outputs are summed before being fed to a Swin transformer [55]. Finally, [74] used early aggregation by directly feeding the network with aggregated patches with more than 3 channels.

Species imbalance. Another important trait of the dataset is its imbalance: a few species account for most of the observations, while a lot of them have only been observed a handful of times. [47,31] tried to use specialized method for this type of data such as focal loss [52], balanced softmax [68] or more advanced methods. These did not help improve their scores, most likely because the test set shares the same imbalance as the training set and the evaluation metric did take it into account (the fixed list of metrics implemented by Kaggle did not allow us to use a class-averaged top-30 error rate).

Presence-only observation data. One last major characteristic of the dataset is that the observation data provided is presence-only data: at a given location, we only know that one species is present and do not have access of the complete list of species present nor the ones absent. The winning team and [47] tried to address this by using a grid of squared cells to aggregate the species observed into each cell. They then used this information in a different manner. The winning team tried to map the 30 species closest to each training point falling into its cell and used this list as the new label. Unfortunately, in the given time, this approach only resulted in overfitting. On the other hand, [47] successfully used the aggregated observations as a regularization method by replacing the label assigned to each training observation by another species from its cell 10% of the time.

Other methods were also tried out such as different architectures, different approaches for model pretraining (no pretraining, pretraining on ImageNet, on another dataset closer to GeoLifeCLEF 2022, etc.), multi-task learning, and a lot more. These are more exhaustively listed in the GeoLifeCLEF 2022 overview working note paper [57] along with a more detailed description of the methods presented above and further analyses.

5 SnakeCLEF challenge: Automated Snake Species Identification on a Global Scale

A detailed description of the challenge and a more complete discussion of the results can be found in the dedicated overview paper [64].

5.1 Objective

Building an automatic and robust image-based system for snake species identification is an important goal for biodiversity, conservation, and global health. With over half a million victims of death and disability from venomous snakebite annually, such a system could significantly improve eco-epidemiological data and treatment outcomes (e.g. based on the specific use of antivenoms) [2,6]. Importantly, most herpetological expertise and most snake images are concentrated in developed countries in areas of the world where snake diversity is relatively low and snakebite is not a major public health concern. In contrast, remote parts of developing countries tend to lack expertise and images, even in areas where snake diversity is high and snakebites are common [15]. Thus, snake species identification assistance has a bigger potential to save lives in areas with the least information.

A primary difficulty of snake species identification lies in the high intra-class and low inter-class variance in appearance, which may depend on geographic location, color morph, sex, or age. At the same time, many species are visually similar to other species – mimicry (Figure 7). Furthermore, our knowledge of which snake species occur in which countries is incomplete, and it is common that most or all images of a given snake species might originate from a small handful of countries or even a single country. Furthermore, many snake species resemble species found on other continents, with which they are entirely allopatric. Incorporating metadata on the geographic origin of an unidentified snake can narrow down the possible correct identifications considerably because only about 125 of the approximately 3,900 snake species co-occur in any given location [70]. It is known that more widespread species with more images are over-predicted relative to rare species with few images [16], and this can be a particularly vexing problem when trying to predict the identity of species that are widespread across areas of the world with few images.

The main goal of the SnakeCLEF 2022 competition was to provide a reliable evaluation ground for automatic snake species recognition. Like other LifeCLEF competitions, the SnakeCLEF 2022 competition was hosted on Kaggle[20] primarily to attract machine learning experts to participate and present their ideas.

5.2 Dataset and Evaluation Protocol

For this year, the dataset used in previous editions [62,?] has been extended with new and rare species. The number of species was doubled and the number of images from remote geographic areas with none or just a few samples was increased considerably, i.e., the uneven species distributions across all the countries was straightened. The SnakeCLEF 2022 dataset is based on 187,129 snake observations – multiple images of the same individual (refer to Figure 8) – with 318,532 photographs belonging to 1,572 snake species and observed in

[20] `https://www.kaggle.com/competitions/fungiclef2022`

Fig. 7: Harmless mimic species *Cemophora coccinea ssp. coccinea* (top row) and poisonous lookalike species. *Micrurus pyrrhocryptus*, *Micrurus ibiboboca*, and *Micrurus nigrocinctus* (left to right, bot. row). *©roadmom–iNaturalist, ©Anthony Damiani–iNaturalist, ©Adam Cushen–iNaturalist, ©Alexander Guiñazu–iNaturalist, ©Tarik Câmara–iNaturalist, and ©Cristhian Banegas–iNaturalist.*

208 countries. The dataset has a heavy long-tailed class distribution, where the most frequent species (*Natrix natrix*) is represented by 6,472 images and the least frequent species just by 5 samples. The difference in the number of images between the species with the most and fewest was reduced by an order of magnitude relative to SnakeCLEF2021. All the data was gathered from the online biodiversity platform – iNaturalist[21].

For testing, two sets were created: (i) the full test set for a machine evaluation, with 48,280 images from 28,431 observations, and (ii) the subset from the full test set with 150 observations, tailored for the human performance evaluation. Unlike in other LifeCLEF competitions, where the final testing set remained undisclosed, we provided the test data without labels to the participants. To prevent over fitting to the leaderboard, the evaluation method was composed of two stages; the first being the public leaderboard where the user scores were calculated on an unknown 20% of the test set, and the second a private leaderboard where participants were scored on the remaining part of the test set. In addition to image data, we provide:

[21] `https://www.inaturalist.com/`

Fig. 8: Two snake observations from SnakeCLEF2022 dataset – three images for each individual. ©*André Giraldi – iNaturalist,* ©*Harshad Sharma – iNaturalist.*

- human verified species labels that allows up-scaling to higher taxonomic ranks,
- the country-species mapping file describing species-country presence to allow better regularization towards all geographical locations, based on The Reptile Database [77], and
- information about endemic species – species that occur only in one geographical region, e.g., Australia or Madagascar.

The geographical information, e.g., state and country labels, was included for approximately 95% of the training and test images. Additionally, we provide a mapping matrix (MM_{cs}) describing country-species presence to allow better worldwide regularization.

$$\mathrm{MM}_{cs} = \begin{cases} 1 & \text{if species } S \in \text{country } C, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Unlike last year's dataset, where the vast majority (77%) of all images came from the United States and Canada, the SnakeCLEF 2022 dataset includes just a fraction of the data (28.3%) from the United States and Canada. The rest of the data is distributed across remaining regions, e.g., Europe, Asia, Africa, Australia and Oceania.

Evaluation: The main goal of this challenge was to build a system that is capable of recognizing 1,572 snake species based on the given snake observation – unseen set of images – and relevant geographical location. As a main metric, we use the macro F1 score (F_1^m). The F_1^m is defined as the mean of class-wise F1 scores:

$$\mathrm{F}_1^m = \frac{1}{N}\sum_{s=0}^{N} F_{1_s}, \quad F_{1_s} = 2 \times \frac{P_s \times R_s}{P_s + R_s}, \tag{5}$$

where s is species index, N equals to the number of classes in a training set. The F1 score for each class represents harmonic mean of the class precision P_s and recall R_s.

5.3 Participants and Results

A total of 29 teams participated in the SnakeCLEF 2022 challenge and contributed with 648 submissions. Everyone who submitted a solution better than baseline submission, i.e., random predictions, was considered a participant. The number of participants quadrupled since last year, primarily as Kaggle was used as an evaluation platform. The best performing team achieved F_1^m of 86.47% on a private part of a test set and 94.01% accuracy on the full test set. On the expert set, the best performing team achieved an F_1^m of 90.28%. The performance evaluation for top-20 Teams is provided in Figure 9. At the time of writing, the organisers could not reproduce any score from the leaderboard, even though most teams provided code.

Details of the best submitted methods and systems are synthesized in the overview working notes paper [64] and further developed in the individual working notes. The main outcomes we can derive from the achieved results are as follows:

Transformer-based architectures outperformed CNNs. This year various deep neural network architectures – Convolutional Neural Networks and Transformers – were evaluated; ConvNext [56], EfficientNet [73], Vision Transformer [14], Swin Transformer [55], and MetaFormer [13]. Unlike last year, where the CNN architectures overwhelmed the performance, Vision Transformer architectures were a vital asset for most methods submitted this year. The second best method with F_1^m score of 84.56% was based on an ensemble of exclusively ViT models and performed slightly worst (-0.9%) than the best performing system that used a combination of Transformer and CNN models. An ensemble of MetaFormer models achieved the third-best score of 82.65%. It seems that Transformers and CNNs benefit from each other in an ensemble, while a standalone Transformer ensemble performs better than a pure CNN ensemble which achieved an F_1^m score of "only" 70.8%

Loss Function matters. Several loss functions were evaluated: Label Aware Smoothing [88], (modified) Categorical Cross-Entropy, and Seesaw [82]. Overall, any Loss function if used is better than standard CrossEntropy. The wining team used Label Aware Smoothing. The runner-up used an Effective Logit Adjustment Loss and showed an improvement of around 2% of F_1^m score when compared to Cross Entropy, reducing the error rate by 15%. The the third team used Logit adjustment to outperform the Seesaw loss from an F_1^m score of 76.5% to 78.6%.

Self-supervision has potential. Adding unlabeled data to the train set is a welcome option when not many observations of a species are available. The third team used the SimCLR [7] method with InfoNCE [61] loss function to increase the F_1^m score from 63.76% to 68.83% when compared to an ImageNet-1k pretrained models.

Geographical metadata improves classification performance. Most teams report accuracy improvement when adding the metadata into the learning process. The second team achieved an improvement of 10.9% in terms of the F_1^m score using a simple location filtering approach. The third team described an absolute improvement of 7.5% when adding the metadata into the MetaFormer.

Ensemble helps, but at what cost? Most teams used ensembling to increase the accuracy of classification. The standard approach was to compute an average of the individual models' decisions. Some teams used a late fusion of deep features by concatenation as an ensemble technique. Even though the improvement in accuracy is observable (around 1 percentage point of F_1^m across the board), it would be interesting to measure the added computational complexity vs the added accuracy. In the case of snakebite, the system's inference time plays a crucial role.

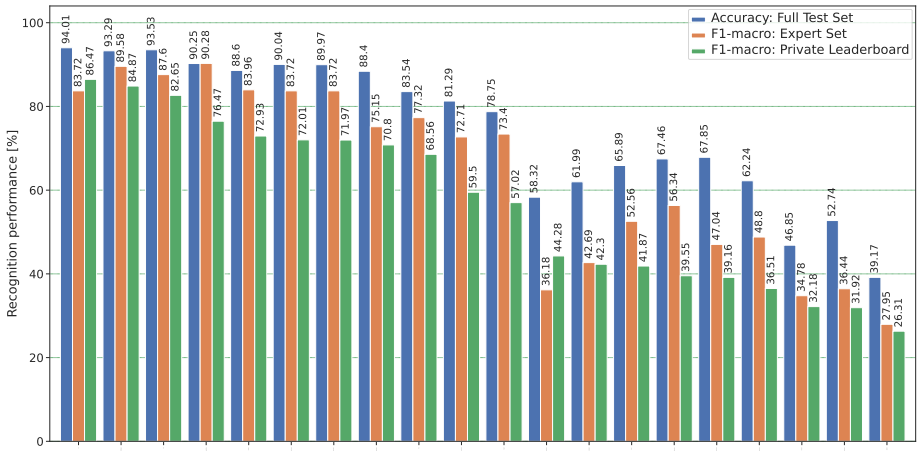


Fig. 9: SnakeCLEF 2022 competition: Top20 teams Performance. Accuracy on Full Test set, and Macro F_1 score on private part of the test set and Expert set. Sorted by performance on the private leaderboard.

6 FungiCLEF Challenge: Fungi Recognition as an Open Set Classification Problem

A detailed description of the challenge and a more complete discussion of the results can be found in the dedicated overview paper [65].

Fig. 10: Two fungi specimen observations from the Danish Fungi 2020 dataset. *Atlas of Danish Fungi: ©Jan Riis-Hansen and ©Arne Pedersen.*

6.1 Objective

Automatic recognition of fungi species assists mycologists, citizen scientists and nature enthusiasts in species identification in the wild. Its availability supports the collection of valuable biodiversity data. In practice, species identification typically does not depend solely on the visual observation of the specimen but also on other information available to the observer – such as habitat, substrate, location and time. Thanks to rich metadata, precise annotations, and baselines available to all competitors, the challenge provides a benchmark for image recognition with the use of additional information.

The main goal for the new FungiCLEF competition was to provide an evaluation ground for automatic methods for fungi recognition in an open class set scenario, i.e, the submitted methods have to handle images of unknown species. Similarly to previous LifeCLEF competitions, The competition was hosted on Kaggle[22] primarily to attract machine learning experts to participate and present their ideas.

6.2 Dataset and Evaluation Protocol

Data collection: The FungiCLEF 2022 dataset is based on data collected through the Atlas of Danish Fungi mobile (iOS[23] and Android[24]) and Web[25] applications.

[22] `https://www.kaggle.com/competitions/fungiclef2022`
[23] `https://apps.apple.com/us/app/atlas-of-danish-fungi/id1467728588`
[24] `https://play.google.com/store/apps/details?id=com.noque.svampeatlas`
[25] `https://svampe.databasen.org/`

The Atlas of Danish Fungi is a citizen science platform with more than 4,000 actively contributing volunteers and with more than 1 million content-checked observations of approximately 8,650 fungi species.

For training, the competitors were provided with the DanishFungi 2020 (DF20) dataset [63]. DF20 contains 295,938 images – 266,344 for training and 29,594 for validation – belonging to 1,604 species. All training samples passed an expert validation process, guaranteeing high quality labels. Furthermore, rich observation metadata about habitat, substrate, time, location, EXIF etc. are provided.

The test dataset is constructed from all observations submitted in 2021, for which expert-verified species labels are available. It includes observations collected across all substrate and habitat types. The test set contains 59,420 observations with 118,676 images belonging to 3,134 species: 1,165 known from the training set and 1,969 unknown species covering approximately 30% of the test observations. The test set was further split into public (20%) and private (80%) subsets – a common practice for Kaggle competitions to prevent participants from overfitting to the leaderboard.

Task description: The goal of the task is to return the correct species (or *"unknown"*) for each test observation, consisting from a set of images and metadata. Photographs of unknown fungi species should be classified into an *"unknown"* class with label id -1. A baseline procedure to include meta-data in the decision problem and baseline pre-trained image classifiers were provided as part of the task description to all participants.

Evaluation Protocol: The evaluation process consisted of two stages: (i) a public evaluation, which was available during the whole competition with a limit of two submissions a day, and (ii) a private evaluation used for the final leaderboard. The main evaluation metric for the competition was the F_1^m, defined as the mean of class-wise F_1 scores:

$$\mathrm{F}_1^m = \frac{1}{N}\sum_{s=1}^{N} F_{1_s}\,, \tag{6}$$

where N represents the number of classes – in case of the Kaggle evaluation, $N = 1,165$ (#classes in the test set) – and s is the species index. The F_1 score for each class is calculated as a harmonic mean of the class precision P_S and recall R_S:

$$F_{1_s} = 2 \times \frac{P_s \times R_s}{P_s + R_s}\,, \quad P_s = \frac{tp_s}{tp_s + fp_s}\,, \quad R_s = \frac{tp_s}{tp_s + fn_s} \tag{7}$$

In single-label multi-class classification, the True Positives (tp) of a species represents the number of correct Top1 predictions of that species, False Positive (fp) denotes how many times was different species predicted instead of the (tp), and False Negatives (fn) indicates how many images of species s have been wrongly classified.

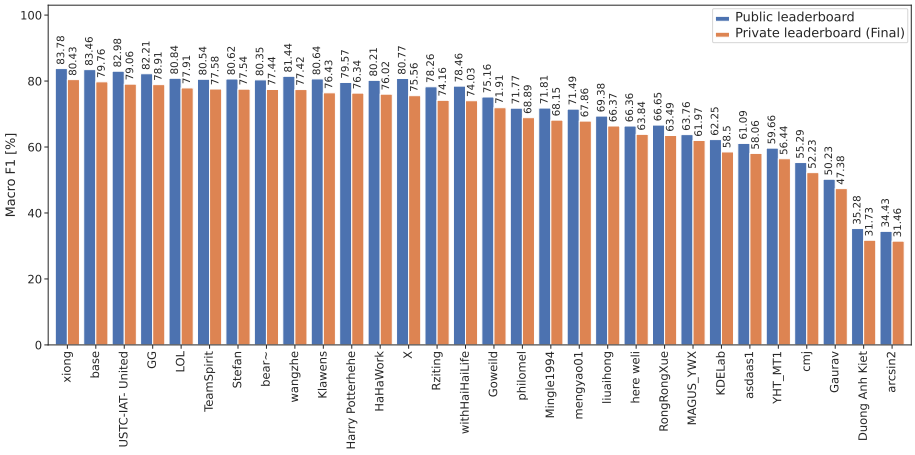


Fig. 11: Results of the FungiCLEF 2022 competition on Kaggle, sorted by performance on the final (private) test set.

6.3 Participants and Results

In total, 38 teams contributed with 701 valid submissions to the challenge evaluation on Kaggle. A detailed description of the methods used in the submitted runs is available in the overview working note paper [65] and further developed in the individual working notes. The results on the public and private test sets (leaderboards) are displayed in Figure 11.

All submissions that shared their working notes were based on modern Convolutional Neural Network (CNN) or transformer-inspired architectures, such as Metaformer [13], Swin Transformer [55], and BEiT [1]. The best performing teams used ensembles of both CNNs and Transformers. The winning team [84] achieved 80.43% accuracy with a combination of ConvNext-large [56] and MetaFormer [13]. The results were often improved by combining predictions belonging to the same observation and by both training-time and test-time data augmentations.

Participants experimented with a number of different training losses to battle the long tail distribution and fine-grained classification with small inter-class differences and large intra-class differences: besides the standard Cross Entropy loss function, we have seen successful applications of the Seesaw loss [82], Focal loss [52], Arcface loss [11], Sub-Center loss [10] and Adaptive Margin [53].

We were happy to see the participants experimented with different use of the provided observation metadata, which often lead to improvements in the recognition scores. Besides the probabilistic baseline published with the dataset [63], we have seen hand-crafted encoding of the metadata into feature vectors, as well as encoding of the meta-data with a multilingual BERT model [12] and RoBERTa [54]. The meta-data were then combined with image features extracted from a CNN or Transformer image classifier, or directly used as an input to Metaformer [13].

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, bio-acoustic and machine learning techniques towards building real-world biodiversity monitoring systems. This study shows that recent deep learning techniques still allow some consistent progress for most of the evaluated tasks. One of the main new insights of this edition of LifeCLEF is that vision transformers performed better than CNNs in some tasks, in particular in the PlantCLEF task for which the best model is a vision transformer whose training was not yet completed at the time of the challenge closure. This shows the potential of these techniques on huge datasets such as the one of PlantCLEF (4M images of 80K species). However, training those models requires more computational resources that only participants with access to large computational clusters can afford. In the other challenges, what seems to best explain the best performances is the model selection methodology employed given the time constraints and the available computational resources. Participants must carefully prioritize the approaches they want to test with a compromise between novelty and efficiency. New methods are typically more risky than that well-known recipes. However, when they work they can make a real difference to the other participants. The challenge where there were the most methodological novelty is probably the GeoLifeCLEF challenge. It is indeed quite unusual due to its multi-modal nature (mixing very different types) and the originality of the task itself (set-valued classification based on presence-only data). The way all the modalities were combined was clearly one of the main driver of success. Moreover, the set-valued classification problem has encouraged the implementation of an original label swapping strategy that has proven to be effective. In the FungiCLEF challenge, several participants utilized the provided metadata in the decision process of a fine-grained image classification task – either by combining image and metadata embeddings in a classifier, or by directly feeding the image and the metadata in a transformer / MetaFormer [13] architecture. Finally, the long-tail distribution problem (common to all tasks) has also been one of the most explored research topics through the different challenges (in particular the SnakeCLEF and FungiCLEF challenges). While it is difficult to draw a simple conclusion about the superiority of some approaches over others, many participants showed that substantial gains could be made by taking the long tail problem into account (including alternative loss functions to cross-entropy or self-supervision on unlabeled data).

Acknowledgements This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No° 863463 (Cos4Cloud project), and the support of #DigitAG.

References

1. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)

2. Bolon, I., Durso, A.M., Botero Mesa, S., Ray, N., Alcoba, G., Chappuis, F., Ruiz de Castañeda, R.: Identifying the snake: First scoping review on practices of communities and healthcare providers confronted with snakebite across the world. PLoS one **15**(3), e0229989 (2020)
3. Bonnet, P., Goëau, H., Hang, S.T., Lasseck, M., Šulc, M., Malécot, V., Jauzein, P., Melet, J.C., You, C., Joly, A.: Plant identification: experts vs. machines in the era of deep learning. In: Multimedia Tools and Applications for Environmental & Biodiversity Informatics, pp. 131–149. Springer (2018)
4. 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
5. Carranza-Rojas, J., Gonzalez-Villanueva, R., Jimenez-Morales, K., Quesada-Montero, K., Esquivel-Barboza, E., Carvajal-Barboza, N.: Extreme automatic plant identification under constrained resources. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (2022)
6. Ruiz de Castañeda, R., Durso, A.M., Ray, N., Fernández, J.L., Williams, D.J., Alcoba, G., Chappuis, F., Salathé, M., Bolon, I.: Snakebite and snake identification: empowering neglected communities and health-care providers with ai. The Lancet Digital Health **1**(5), e202–e203 (2019)
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. Chulif, S., Lee, S.H., Chang, Y.L.: A global-scale plant identification using deep learning: Neuon submission to plantclef 2022. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (2022)
9. Cole, E., Deneu, B., Lorieul, T., Servajean, M., Botella, C., Morris, D., Jojic, N., Bonnet, P., Joly, A.: The GeoLifeCLEF 2020 dataset. arXiv preprint arXiv:2004.04192 (2020)
10. Deng, J., Guo, J., Liu, T., Gong, M., Zafeiriou, S.: Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In: European Conference on Computer Vision. pp. 741–757. Springer (2020)
11. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
13. 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)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Durso, A.M., Ruiz de Castañeda, R., Montalcini, C., Mondardini, M.R., Fernandez-Marques, J.L., Grey, F., Müller, M.M., Uetz, P., Marshall, B.M., Gray, R.J., et al.: Citizen science and online data: Opportunities and challenges for snake ecology and action against snakebite. Toxicon: X **9**, 100071 (2021)
16. Durso, A.M., Moorthy, G.K., Mohanty, S.P., Bolon, I., Salathé, M., Ruiz De Castañeda, R.: Supervised learning computer vision benchmark for snake species iden-

tification from photographs: Implications for herpetology and global health. Frontiers in Artificial Intelligence **4**, 17 (2021)
17. 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)
18. Ghazi, M.M., Yanikoglu, B., Aptoula, E.: Plant identification using deep neural networks via optimization of transfer learning parameters. Neurocomputing **235**, 228–235 (2017)
19. 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`
20. 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`
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., 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)
23. 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)
24. 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)
25. Goëau, H., Glotin, H., Planqué, R., Vellinga, W.P., Stefan, K., Joly, A.: Overview of birdclef 2018: monophone vs. soundscape bird identification. In: CLEF task overview 2018, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2018, Avignon, France. (2018)
26. Grill, T., Schlüter, J.: Two convolutional neural networks for bird detection in audio signals. In: 2017 25th European Signal Processing Conference (EUSIPCO). pp. 1764–1768 (Aug 2017). https://doi.org/10.23919/EUSIPCO.2017.8081512
27. Hengl, T., de Jesus, J.M., Heuvelink, G.B., Gonzalez, M.R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., et al.: Soilgrids250m: Global gridded soil information based on machine learning. PLoS one **12**(2) (2017)
28. Henkel, C., Pfeiffer, P., Singer, P.: Recognizing bird species in diverse soundscapes under weak supervision. In: CLEF Working Notes 2021, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2021, Bucharest, Romania (2021)
29. Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A.: Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology: A Journal of the Royal Meteorological Society **25**(15), 1965–1978 (2005)
30. Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N., Wickham, J., Megown, K.: Completion of the 2011 national land cover database for the conterminous united states – representing a decade of land cover

change information. Photogrammetric Engineering & Remote Sensing **81**(5), 345–354 (2015)

31. Jiang, J.: Localization of plant and animal species prediction with convolutional neural networks. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (2022)
32. 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)
33. 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)
34. 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`
35. 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`
36. 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`
37. 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`
38. 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)
39. 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)
40. 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)
41. Kahl, S., Clapp, M., Hopping, A., Goëau, H., Glotin, H., Planqué, R., Vellinga, W.P., Joly, A.: Overview of birdclef 2020: Bird sound recognition in complex acoustic environments. In: CLEF task overview 2020, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2020, Thessaloniki, Greece. (2020)
42. Kahl, S., Denton, T., Klinck, H., Glotin, H., Goëau, H., Vellinga, W.P., Planqué, R., Joly, A.: Overview of BirdCLEF 2021: Bird call identification in soundscape recordings. In: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum (2021)
43. Kahl, S., Navine, A., Denton, T., Klinck, H., Hart, P., Glotin, H., Goëau, H., Vellinga, W.P., Planqué, R., Joly, A.: Overview of BirdCLEF 2022: Endangered bird species recognition in soundscape recordings. Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (2022)
44. Kahl, S., Stöter, F.R., Glotin, H., Planqué, R., Vellinga, W.P., Joly, A.: Overview of birdclef 2019: Large-scale bird recognition in soundscapes. In: CLEF task overview 2019, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2019, Lugano, Switzerland. (2019)
45. Kahl, S., Wood, C.M., Eibl, M., Klinck, H.: Birdnet: A deep learning solution for avian diversity monitoring. Ecological Informatics **61**, 101236 (2021)
46. Karun, A., Divyasri, K., Balasundaram, P., Sella Veluswami, J.R.: Plant species identification using probability tree approach of deep learning models. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (2022)
47. Kellenberger, B., Devis, T.: Block label swap for species distribution modelling. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (2022)
48. Lasseck, M.: Audio-based bird species identification with deep convolutional neural networks. In: CLEF working notes 2018, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2018, Avignon, France. (2018)
49. Leblanc, C., Lorieul, T., Servajean, M., Bonnet, P., Joly, A.: Species distribution modeling based on aerial images and environmental features with convolutional neural networks. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (2022)
50. 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)
51. 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)

52. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
53. Liu, H., Zhu, X., Lei, Z., Li, S.Z.: Adaptiveface: Adaptive margin and sampling for face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11947–11956 (2019)
54. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
55. 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)
56. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. arXiv preprint arXiv:2201.03545 (2022)
57. Lorieul, T., Cole, E., Deneu, B., Servajean, M., Joly, A.: Overview of GeoLifeCLEF 2022: Predicting species presence from multi-modal remote sensing, bioclimatic and pedologic data. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (2022)
58. Min Ong, J., Yang, S.J., Ng, K.W., Chan, C.S.: Image-based plant identification with taxonomy aware architecture. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (2022)
59. Mühling, M., Franz, J., Korfhage, N., Freisleben, B.: Bird species recognition via neural architecture search. In: CLEF working notes 2020, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2020, Thessaloniki, Greece. (2020)
60. 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)
61. Van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv e-prints pp. arXiv–1807 (2018)
62. Picek, L., Ruiz De Castañeda, R., Durso, A.M., Sharada, P.M.: Overview of the snakeclef 2020: Automatic snake species identification challenge. In: CLEF task overview 2020, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2020, Thessaloniki, Greece. (2020)
63. 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)
64. 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)
65. 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)
66. 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)
67. Pravinkrishnan, K., Sivakumar, N., Balasundaram, P., Kalinathan, L.: Classification of plant species using alexnet architecture. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (2022)

68. Ren, J., Yu, C., Ma, X., Zhao, H., Yi, S., et al.: Balanced meta-softmax for long-tailed visual recognition. Advances in Neural Information Processing Systems **33**, 4175–4186 (2020)
69. 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)
70. Roll, U., Feldman, A., Novosolov, M., Allison, A., Bauer, A.M., Bernard, R., Böhm, M., Castro-Herrera, F., Chirio, L., Collen, B., et al.: The global distribution of tetrapods reveals a need for targeted reptile conservation. Nature Ecology & Evolution **1**(11), 1677–1682 (2017)
71. Seneviratne, S.: Contrastive representation learning for natural world imagery: Habitat prediction for 30,000 species. In: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum (2021)
72. Shiu, Y., Palmer, K., Roch, M.A., Fleishman, E., Liu, X., Nosal, E.M., Helble, T., Cholewiak, D., Gillespie, D., Klinck, H.: Deep neural networks for automated detection of marine mammal species. Scientific reports **10**(1), 1–12 (2020)
73. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
74. Teng, M., Elkafrawy, S.: Participation to the GeoLifeCLEF challenge 2022 working notes. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (2022)
75. Towsey, M., Planitz, B., Nantes, A., Wimmer, J., Roe, P.: A toolbox for animal call recognition. Bioacoustics **21**(2), 107–125 (2012)
76. 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)
77. Uetz, P., Freed, P., Hošek, J., et al.: The reptile database (2020), `https://reptile-database.reptarium.cz/advanced_search`
78. 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)
79. 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)
80. Wäldchen, J., Mäder, P.: Machine learning for image based species identification. Methods in Ecology and Evolution **9**(11), 2216–2225 (2018)
81. Wäldchen, J., Rzanny, M., Seeland, M., Mäder, P.: Automated plant species identification—trends and future directions. PLoS computational biology **14**(4), e1005993 (2018)
82. 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)
83. Wood, C.M., Kahl, S., Chaon, P., Peery, M.Z., Klinck, H.: Survey coverage, recording duration and community composition affect observed species richness in passive acoustic surveys. Methods in Ecology and Evolution **12**(5), 885–896 (2021)
84. Xiong, Z., Ruan, Y., Hu, Y., Zhang, Y., Zhu, Y., Guo, S., Zhu, W., Han, B.: An empirical study for fine-grained fungi recognition with transformer and convnet.

In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (2022)
85. Xu, M., Yoon, S., Lee, J., Park, D.S.: Vision transformer-based unsupervised transfer learning for large scale plant identification. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (2022)
86. 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)
87. Zhang, X., Zhou, Y.: A multimodal model for predict the localization of plant and animal species. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (2022)
88. Zhong, Z., Cui, J., Liu, S., Jia, J.: Improving calibration for long-tailed recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16489–16498 (2021)