Overview of LifeCLEF 2020: A System-Oriented Evaluation of Automated Species Identification and Species Distribution Prediction

Alexis Joly[1(✉)], Hervé Goëau[2,3], Stefan Kahl[7], Benjamin Deneu[1], Maximillien Servajean[8], Elijah Cole[10], Lukáš Picek[11], Rafael Ruiz de Castañeda[9], Isabelle Bolon[9], Andrew Durso[13], Titouan Lorieul[1], Christophe Botella[12], Hervé Glotin[4], Julien Champ[1], Ivan Eggel[6], Willem-Pier Vellinga[5], Pierre Bonnet[2,3], and Henning Müller[6]

[1] Inria, LIRMM, Montpellier, France
alexis.joly@inria.fr
[2] CIRAD, UMR AMAP, 34398 Montpellier, France
[3] AMAP, Univ Montpellier, CIRAD, CNRS, INRAE, IRD, Montpellier, France
[4] Aix Marseille Univ, Université de Toulon, CNRS, LIS, DYNI, Marseille, France
[5] Xeno-canto Foundation, The Hague, The Netherlands
[6] HES-SO, Sierre, Switzerland
[7] Cornell Lab of Ornithology, Cornell University, Ithaca, USA
[8] LIRMM, Université Paul Valéry, University of Montpellier, CNRS, Montpellier, France
[9] Institute of Global Health, Department of Community Health and Medicine, Faculty of Medicine, University of Geneva, Geneva, Switzerland
[10] Caltech, Pasadena, USA
[11] Department of Cybernetics, FAV, University of West Bohemia, Pilsen, Czechia
[12] CNRS, LECA, Grenoble, France
[13] Department of Biological Sciences, Florida Gulf Coast University, Fort Myers, 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 and animals in the field is hindering the aggregation of new data and knowledge. Identifying and naming living plants or animals 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 2020 edition proposes four data-oriented challenges related to the identification and prediction of biodiversity: (i) PlantCLEF: cross-domain plant identification based on herbarium sheets (ii) BirdCLEF: bird species recognition in audio soundscapes, (iii) GeoLifeCLEF: location-based prediction of species based on environmental and occurrence data, and (iv) SnakeCLEF: snake identification based on image and geographic location.

A. Arampatzis et al. (Eds.): CLEF 2020, LNCS 12260, pp. 342–363, 2020.
https://doi.org/10.1007/978-3-030-58219-7_23

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, plants alone account for more than 400,000 different species and the distinctions between them can be quite subtle. Since the Rio Conference of 1992, this *taxonomic gap* 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 [14] 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 [14], automated species identification has been studied in many contexts [5,16,32,42,47,51,52,57]. This area continues to expand rapidly, particularly due to recent advances in deep learning [4,15,43,53,55,56]. 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 [27] that had been run within the ImageCLEF lab[3] since 2011 [22–24]. Since 2014, LifeCLEF expanded the challenge by considering animals in addition to plants, and including audio and video content in addition to images [33–38]. Four challenges were evaluated in the context of LifeCLEF 2020 edition:

1. **PlantCLEF 2020:** Identifying plant pictures from herbarium sheets.
2. **BirdCLEF 2020:** Bird species recognition in audio soundscapes.
3. **GeoLifeCLEF 2020:** Species distribution prediction based on occurrence data, environmental data and remote sensing data.
4. **SnakeCLEF 2020:** Automated snake species identification based on images and two level geographic location data - continent and country.

The system used to run the challenges (registration, submission, leaderboard, etc.) was the AICrowd platform[4]. About 172 researchers or students registered to at least one of the four challenges of the lab and 16 of them finally crossed the finish line by completing runs and participating in the collaborative evaluation. In the following sections, we provide a synthesis of the methodology and main results of each of the four challenges of LifeCLEF2020. More details can be found in the overview reports of each challenge and the individual reports of the participants (references provided below).

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

2 PlantCLEF Challenge: Identifying Plant Pictures from Herbarium Sheets

A detailed description of the task 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. For instance, we measured in 2018 a top-1 classification accuracy over 10 K species up to 90% and we showed that automated systems are not so far from human expertise [33]. However, this profusion of field images only concerns a few tens of thousands of species, mostly located in North America and Western Europe, with fewer images from the richest regions in terms of biodiversity such as tropical countries. On the other hand, for several centuries, botanists have collected, catalogued and systematically stored plant specimens in herbaria, particularly in tropical regions. Recent huge efforts by the biodiversity informatics community such as iDigBio[5] or e-ReColNat[6] made it possible to put millions of digitized collections online. In the continuity of the PlantCLEF challenges organized in previous years [17–20,22–24,26,28], this year's challenge was designed to evaluate to what extent automated plant species identification on tropical data deficient regions can be improved by the use of herbarium sheets. Herbaria collections represent potentially a large pool of data to train species prediction models, but they also introduce a difficult and interesting problem of cross domain classification because typically a same plant photographed in the field takes on a different visual appearance when dried and placed on a herbarium sheet as it can be seen in Fig. 1.

2.2 Dataset and Evaluation Protocol

The challenge is based on a dataset of 997 species mainly focused on the South America's Guiana Shield (Fig. 2), an area known to have one of the greatest diversity of plants in the world. The challenge was evaluated as a cross-domain classification task where the training set consist of 321,270 herbarium sheets and 6,316 photos in the field to enable learning a mapping between the two domains. A valuable asset of this training set is that a set of 354 plant observations are provided with both herbarium sheets and field photos to potentially allow a more precise mapping between the two domains.

The test set relied on two highly trusted experts and was composed of 3,186 photos in the field related to 638 plant observations.

Participants were allowed to use complementary training data (e.g. for pre-training purposes) but on the condition that (i) the experiment is entirely reproducible, i.e. that the used external resource is clearly referenced and accessible

[5] http://portal.idigbio.org/portal/search.
[6] https://explore.recolnat.org/search/botanique/type=index.

Fig. 1. Field photos and herbarium sheets of the same specimen (*Tapirira guianensis* Aubl.). Despite the very different visual appearances between the two types of images, similar structures and shapes of flowers, fruits and leaves can be observed.

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 training data provided this year.

The main evaluation measure for the challenge was the Mean Reciprocal Rank (MRR), which is defined as

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

where Q is the number of plant observations and rank_q is the predicted rank of the true label for the qth observation.

A second metric was again the MRR but computed on a subset of observations of species that are rarely photographed in the field. The species were chosen based on the most comprehensive estimates possible from different data sources (IdigBio, GBIF, Encyclopedia of Life, Bing and Google Image search engines, previous datasets related to PlantCLEF and ExpertCLEF challenges). It is therefore a more challenging metric because it focuses on the species which impose a mapping between herbarium and field photos.

2.3 Participants and Results

68 participants registered for the PlantCLEF challenge 2020 (PC20) and downloaded the data set, and 7 research groups succeeded in submitting runs, *i.e.* files containing the predictions of the system(s) they ran. Details of the methods and systems used in the runs are synthesized in the overview working note paper of the task [21] and further developed in the individual working notes of most of the participants (Holmes [7], ITCR PlantNet [54], SSN [46], LU [58]). The remaining teams did not provide an extended description of their systems but sometimes a few informal descriptions were provided in the metadata associated with the submissions and partially contributed to the comments below. We report in Fig. 3 the performance achieved by the 49 collected runs.

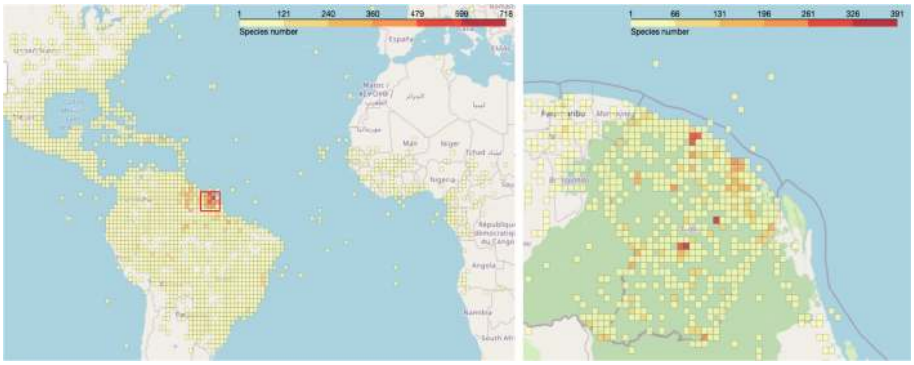


Fig. 2. Density grid maps of the number of species of geolocated plants in PlantCLEF2020. Many species have also been collected to a lesser extent in other regions outside French Guiana, such as the Americas and Africa.

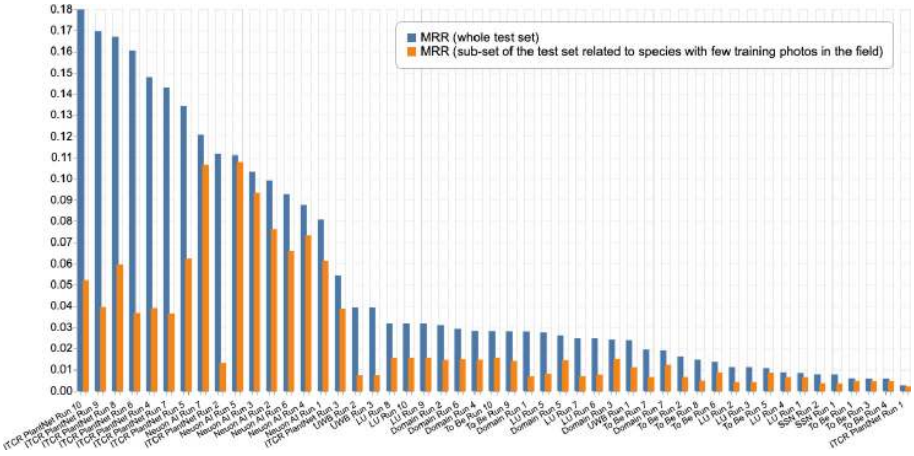


Fig. 3. PlantCLEF 2020 results

The Most Difficult Plant Challenge Ever. This year's challenge is confirmed to be the most difficult of all previous editions, with at best a quite low MRR value of 0.18. As already noticed last year, tropical flora is inherently more difficult than the generalist flora explored during the previous eight years, even for experts [20]. The asymmetry between training data based on herbarium sheets and test data based on field photos did not make the task any easier.

Traditional CNNs Performed Poorly. Figure 3 shows a great disparity between the performance obtained by the different submissions. To explain that we have first to distinguish between approaches based on CNNs alone (typically pretrained on ImageNet and finetuned with the provided training data) and approaches that additionally incorporate an explicit and formal Domain Adaptation (DA) technique between the herbarium and field domains. As expected regarding the low number of field photos in the training set for numerous species, directly finetuned CNNs with the PC20 data obtained the lowest scores (ITCR Run 1, SSN Run 1&2, UWB Run 1).

External Training Data on Traditional CNNs Did Not Really Improve Performances. CNNs can be improved by the use of external data, involving more field photos, as it is demonstrated with the UWB runs 2 & 3 and ITCR Run 2. All these runs extended the training data with the previous year's PC19 training data and the GBIF training data provided by [49]). ITCR Run 2 made a greater improvement on the overall MRR probably by using a two stage training strategy: they first finetuned an ImageNet-pretrained ResNet50 with all the herbarium sheets from PC20, and then finetuned it again with all the field photos extracted from PC20 and the external training data. This two stages strategy can be seen as a naive DA technique because the second stage shifts the learned features in an initial herbarium feature space to a field photo feature space. However, regarding the second MRR metric focusing on the most difficult species with few field photos in the training set, performance for all these runs is still quite low. This means that the performance of a traditional CNN approach (without a more formal adaptation technique) is too dependent from the number of field photos available in the training data, and is not able to efficiently transfer visual knowledge from herbarium domain to field photos domain.

Adversarial DA Techniques Performed the Best. Among other submissions, two participants stood out from the crowd with two quite different DA techniques. ITCR PlantNet team based all its remaining runs on a Few Shot Adversarial Domain Adaptation approach [45] (FSADA), directly applied in the run 3. FSADA approach uses a discriminator that helps the initial encoder trained on herbarium sheets to shift the learned feature representations to a domain agnostic feature space where the discriminator is no longer able to distinguish if a picture comes from the herbarium or the photo domain, while maintaining the discriminative power regarding the final species classification task. The basic FSADA approach (ITCR Run 3) clearly outperformed the traditional CNN approach (run 1), while both approaches are based on the same initial finetuned ResNet50 model on the PC20 training herbarium data. It should be noted that the LU team also used an adversarial approach but with less success.

Mapping DA Technique Reached an Impressive Genericity on Difficult Species. While the adversarial DA technique used by the ITCR PlantNet team obtained the best result on the main MRR metric, the Neuon AI team obtained the best results on the second MRR metric focusing on the most difficult species in the test set. This last team used two encoders, one trained on

the herbarium sheets in PC20 and a second one trained on the photos from the PC17 dataset. Then they learned a distance function based on a triplet loss to maximize the embedding distance of different species and at the same time minimize the distance of the same species. Performances measured from the Neuon AI Run 5, which is an ensemble of 3 instances of their initial approach, gave especially impressive results with quite high MRRs and above all similar values between the two MRR metrics. It means that Neuon AI's approach is very robust to the lack of training field photos and able to generalize on rare difficult species in the test set. In other words, their approach is able to transfer knowledge to rare species which was the underlying objective of the challenge.

External Data Improved DA Approaches. ICTR Run 4 shows a significant impact on the main MRR metric from using external training data compared to the same adversarial DA approach (run 3), while maintaining the same level of genericity on rare species with similar MRRs value on the second metric. Unfortunately it is not possible to measure this impact on the Neuon AI method because they did not provide a run using only this year's training data.

Auxiliary tasks have impact, notably by the use of upper taxon level information in a multi classification task way integrated to the FSADA approach (ITCR Run 6 is better than run 4 with a single species classification task). This is the first time over all the years of PlantCLEF challenges that we clearly observe an important impact of the use of genus and family information to improve the species identification. Many species with few training data have apparently been able to benefit indirectly from a "sibling" species with many data related to a same genus or family. The impact is probably enhanced this year because of the lack of visual data on many species. To a lesser extent, self supervision auxiliary task such as jigsaw solving prediction task (ITCR Run 5 improved a little the baseline of this team (run 4), and the best submission over all this year challenge is an ensemble of all FSADA approaches, combining self supervision or not, upper taxons or not.

3 BirdCLEF Challenge: Bird Sound Recognition in Complex Acoustic Environments

A detailed description of the task and a more complete discussion of the results can be found in the dedicated overview paper [39].

3.1 Objective

The *LifeCLEF Bird Recognition Challenge* (BirdCLEF) 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], [40]. 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[7] launched in 2005 and hosts bird sounds from all continents and daily receives new recordings from some of the remotest places on Earth. The Xeno-canto archive currently consists of more than 550,000 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.

The diversity of this data made BirdCLEF a demanding competition and required participating research groups to develop efficient processing and classification pipelines. The large number of recordings often forced participants to reduce the training data and the number of features—strongly implying the deficiencies of low-level audio feature classification for extremely large datasets. In 2016, Sprengel et al. applied the classical scheme of image classification with deep neural networks to the domain of acoustic event recognition and introduced a convolutional neural network (CNN) classifier trained on extracted spectrograms that instantly outperformed all previous systems by a significant margin [12]. The success of deep neural networks in the domain of sound identification led to the disappearance of MFCCs, SVMs and decision trees which dominated previous editions.

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. Yet, the lack of suitable validation and test data prevented the development of reliable techniques to solve this task. This changed in 2019 when 350 h of fully annotated soundscapes were introduced as test data. Participants were asked to design a detection system that was trained on focal recordings (provided by the Xeno-canto community) and applied to hour-long soundscapes. 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.

3.2 Dataset and Evaluation Protocol

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 from around the world 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

[7] https://www.xeno-canto.org/.

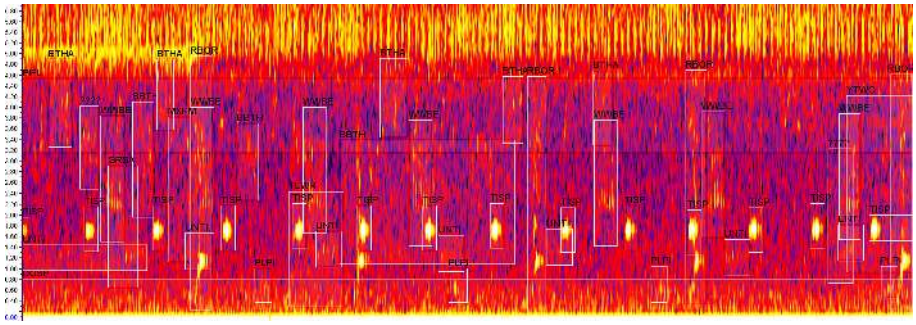


Fig. 4. South American soundscapes often have an extremely high call density. The 2020 BirdCLEF test data contains 48 fully annotated soundscapes recorded in Peru.

editions, training data was provided by the Xeno-canto community and consisted of more than 70,000 recordings covering 960 species from three continents (South and North America and Europe). Participants were allowed to use this and other (meta) data to develop their systems. A representative validation dataset with two hours of soundscape data was also provided, but participants were not allowed to use this data for training—detection systems had to be trained on focal recordings only.

In addition to the 2019 test data, soundscapes from three other recording sites were added in the 2020 edition of BirdCLEF. All audio data were collected with passive acoustic recorders from deployments in Germany (GER), Peru (PER), the High Sierra Nevada (HSN) of California, USA and the Sapsucker Woods area (SSW) in New York, USA. In an attempt to lower the entry level of this challenge, the total amount of soundscape data was reduced to 153 recordings with a duration of ten minutes each. Expert ornithologists provided annotations for often extremely dense acoustic scenes with up to eight species vocalizing at the same time (1.9 on average, see Fig. 4).

The goal of the task was to localize and identify all audible birds within the provided soundscape test set. Each soundscape was divided into segments of 5 seconds, and a list of species associated to probability scores had to be returned for each segment. The used evaluation metric was the classification mean Average Precision ($cmAP$), considering each class c of the ground truth as a query. This means that for each class c, all predictions with $ClassId = c$ are extracted from the run file and ranked by decreasing probability in order to compute the average precision for that class. The mean across all classes is computed as the main evaluation metric. More formally:

$$cmAP = \frac{\sum_{c=1}^{C} AveP(c)}{C}$$

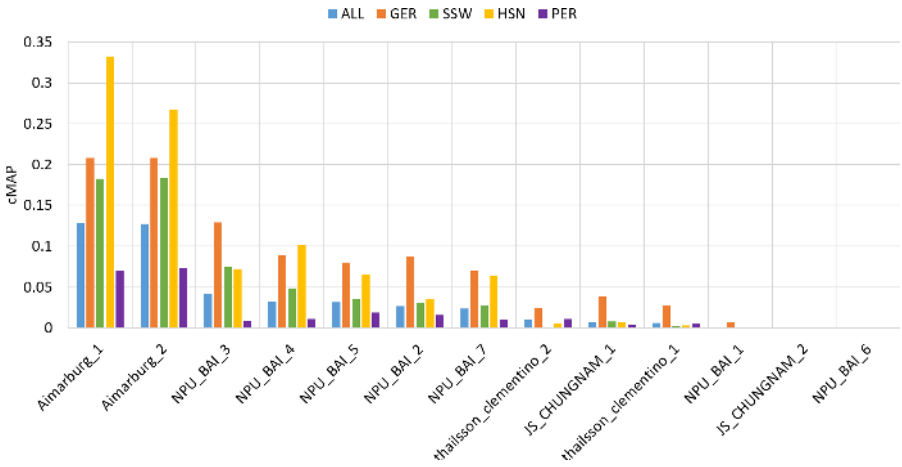


Fig. 5. Scores achieved by all systems evaluated within the bird identification task of LifeCLEF 2020.

where C is the number of classes (species) in the ground truth and $AveP(c)$ is the average precision for a given species c computed as:

$$AveP(c) = \frac{\sum_{k=1}^{n_c} P(k) \times rel(k)}{n_{rel}(c)}.$$

where k is the rank of an item in the list of the predicted segments containing c, n_c is the total number of predicted segments containing c, $P(k)$ is the precision at cut-off k in the list, $rel(k)$ is an indicator function equaling 1 if the segment at rank k is a relevant one (*i.e.* is labeled as containing c in the ground truth) and $n_{rel}(c)$ is the total number of relevant segments for class c.

3.3 Participants and Results

69 participants registered for the BirdCLEF 2020 challenge and downloaded the dataset. Four teams succeeded in submitting runs. Details of the methods and systems used in the runs are synthesized in the overview working notes paper of the task [39] and further developed in the individual working notes of the participants ([1,8]). In Fig. 5 we report the performance achieved by the 13 collected runs.

All submitted runs featured a CNN classifier trained on extracted audio features and all approaches employ current best practices from past editions. Established neural network architectures like VGG, Inception v3, EfficientNet, Xception, or the baseline repository [41] were used in the majority of the submitted runs. Most attempts used log-scale spectrograms as input, only one team used a custom Gabor wavelet layer in their network design. All participants used pre-processed data and distinguished between salient audio chunks and noise (i.e. non-events) to improve the performance of their classifier. Data augmentation is

key for generalization and all participating research groups used a set of domain-specific augmentation methods. The results reflect the slight imbalance of the test data in terms of number of soundscapes per recording site and individual vocalization density. The highest scoring team achieved a class-wise mean average precision of 0.128 across all four recording sites (0.148 on validation data). Some of the participating groups did not manage to score above a cmAP of 0.01 which highlights the demanding nature of this task despite the versatility of deep neural networks. This becomes even more apparent when investigating the classification performance for the South American split of the test data. The highest scoring system achieved a cmAP of only 0.07, on average, the cmAP across all submission was 0.017 for this portion of the test set. Participants scored best for soundscapes recorded in North America with a maximum score of 0.333 for the High Sierra Nevada data. Species composition and recording characteristics play a significant role and the detection quality highly depends on avian call density. Additionally, significant improvements of current classifiers are needed to develop a reliable bird sound recognition system for highly endangered habitats in South America. Current training regimes and neural network architectures might not be suited for this task.

4 GeoLifeCLEF Challenge: Species Distribution Prediction Based on Occurrence Data, Environmental Data and Remote Sensing Data

A detailed description of the task and a more complete discussion of the results can be found in the dedicated working note [10].

4.1 Objective

Automatic prediction of the list of species most likely to be observed at a given location is useful for many scenarios related to biodiversity management and conservation. First, it could 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. More generally, it could facilitate biodiversity inventories through the development of location-based recommendation services (*e.g.* on mobile phones), encourage the involvement of citizen scientist observers, and accelerate the annotation and validation of species observations to produce large, high-quality data sets. Last but not least, this could be used for educational purposes through biodiversity discovery applications with features such as contextualized educational pathways.

4.2 Data Set and Evaluation Protocol

Data Collection: A detailed description of the GeoLifeCLEF 2020 dataset is provided in [9]. In a nutshell, it consists of over 1.9 million observations in US

and France covering 31,435 plant and animal species (as illustrated in Figure 7). Each species observation is paired with high-resolution covariates (RGB-IR imagery, land cover and altitude) as illustrated in Fig. 6. These high-resolution covariates are resampled to a spatial resolution of 1 m per pixel and provided as 256×256 images covering a 256 m $\times$ 256 m square centered on each observation. RGB-IR imagery come from the 2009–2011 cycle of the National Agriculture Imagery Program (NAIP) for the U.S.[8], and from the BD-ORTHO® 2.0 and ORTHO-HR® 1.0 databases from the IGN for France[9]. Land cover data originates from the National Land Cover Database (NLCD) [31] for the U.S. and from CESBIO[10] for France. All elevation data comes from the NASA Shuttle Radar Topography Mission (SRTM)[11]. In addition, the dataset also includes traditional coarser resolution covariates: bio-climatic rasters (1 km^2/pixel, from WorldClim [30]) and pedologic rasters (250 m^2/pixel, from SoilGrids [29]).

Train-Test Split: The full set of occurrences was split in a training and testing set using a spatial block holdout procedure (see Fig. 7). This limits the effect of *spatial auto-correlation* in the data as explained in [50]. This means that a model cannot achieve a high performance by simply interpolating between training samples. The split was based on a global grid of 5 km $\times$ 5 km quadrats. 2.5% of the quadrats were randomly sampled for the test set, and the remaining quadrats 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 with associated confidence scores. The main evaluation criterion is an adaptive variant of the top-K accuracy. Contrary to a classical top-K accuracy, this metric assumes that the number of species K may not be the same at each location. It is computed by thresholding the confidence score of the predictions and keeping only the species above that threshold. The threshold is determined automatically so as to have $K = 30$ results per occurrence on average. See [9] for full details and justification.

4.3 Participants and Results

40 participants registered for the GeoLifeCLEF 2020 challenge and downloaded the dataset. Only two of them succeeded in submitting runs: **Stanford** and **LIRMM**. A major hindrance to participation was the volume of data as well as the computing power needed to train the models (e.g. almost two weeks to train a convolutional neural network on 8 GPUs). Details of the methods and systems used in the runs of both participants are synthesized in the overview working note paper for this task [10]. Runs of the LIRMM team are further developed in

[8] National Agriculture Image Program, https://www.fsa.usda.gov.
[9] https://geoservices.ign.fr.
[10] http://osr-cesbio.ups-tlse.fr/~oso/posts/2017-03-30-carte-s2-2016/.
[11] https://lpdaac.usgs.gov/products/srtmgl1v003/.

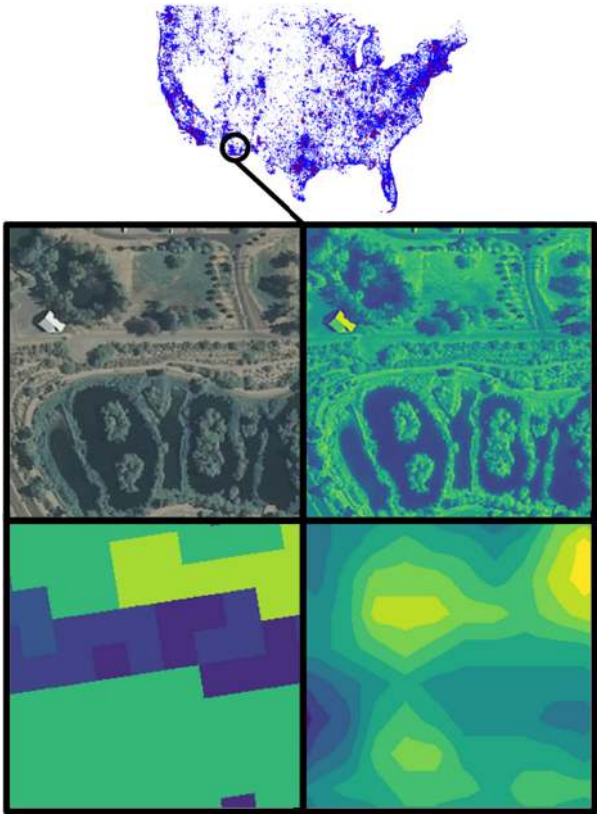
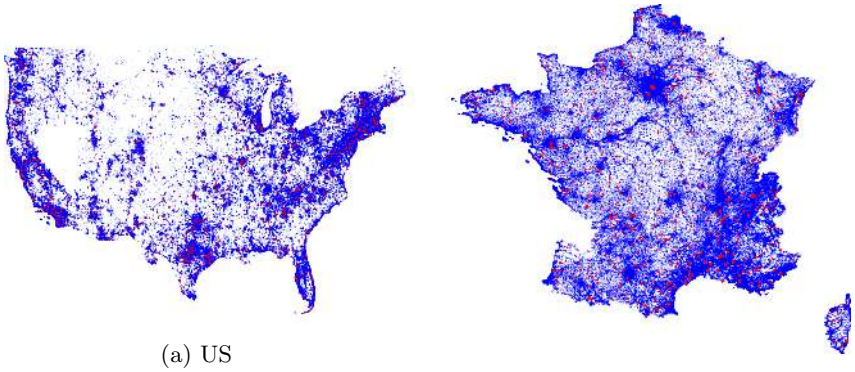


Fig. 6. Each species observation is paired with high-resolution covariates (clockwise from top left: RGB imagery, IR imagery, altitude, land cover).



(a) US

(b) France

Fig. 7. Occurrences distribution over the US and France. Blue dots represent training data, red dots represent test data. (Color figure online)

the individual working note [11]. Due to convergence issues for runs of Stanford team, after discussion with the authors, it was mutually agreed that they would not provide additional working notes for their runs.

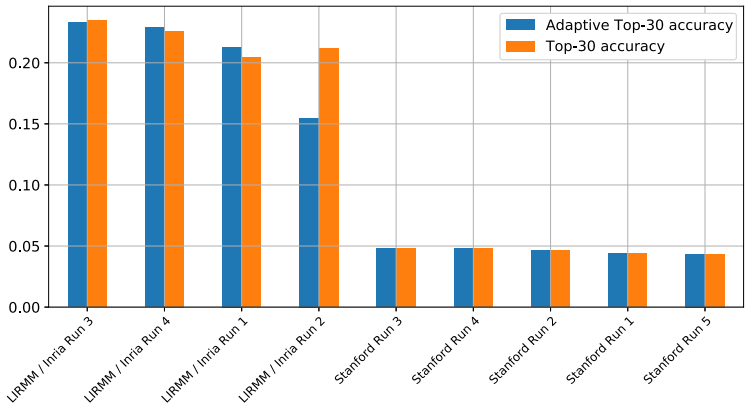


Fig. 8. Adaptive top-30 accuracy and top-30 accuracy per run and participant on GeoLifeCLEF 2020 task.

In Fig. 8 we report the performance achieved by the 9 collected runs[12]. The main outcome of the challenge was that the method achieving the best results (LIRMM/Inria Run 3) was based solely on a convolutional neural network (CNN) trained on the high-resolution covariates (RGB-IR imagery, land cover, and altitude). It did not make use of any bioclimatic variable or soil type variable whereas these variables are often considered as the most informative in the ecological literature. On the contrary, the method used in LIRMM/Inria Run 1 was based solely on the punctual environmental variables using a machine learning method classically used for species distribution models (Random Forest, [13]). This shows two things: (i) important information explaining the species composition is contained in the high-resolution covariates and (ii), convolutional neural networks are able to capture this information. An important following question would be to know whether the information captured by the high-resolution CNN is complementary to the one captured from the bioclimatic and soil variables. This was the purpose of LIRMM/Inria Run 4 that merged the prediction of both models by averaging their outputs. Unfortunately, this was not really conclusive. Either the high-resolution CNN already captured most of the information contained in the bioclimatic variables, or the fusion method was not able to take the best of each model.

[12] Most of the Stanford team's methods were based on deep neural networks, but the authors informed us that they encounter convergence issues resulting in performance poorer than expected.

5 SnakeCLEF Challenge: Automated Snake Species Identification Based on Images and Two-Level Geographic Location Data (Continent and Country)

A detailed description of the task and a more complete discussion of the results can be found in the dedicated overview paper [48].

5.1 Objective

To create an automatic and robust 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, having a system that is capable to recognize or differentiate various snake species from images could significantly improve eco-epidemiological data and treatment outcomes (e.g. based on specific use of antivenoms) [3,6].

Rhombic Night Adder | African Egg-eating Snake

Variable Coralsnake | Variegated False Coralsnake

Fig. 9. Medically important snake species (left) and similar-looking non-venomous species (right). © Peter Vos, iNaturalist, CC-BY-NC and © Alex Rebelo, iNaturalist, CC-BY-NC and © Peter Vos, iNaturalist, CC-BY-NC and © Iris Melgar, iNaturalist, CC-BY-NC.

Since snake species identification is a fine-grained visual categorization task, the main difficulty of this challenge is the high intra-class and low inter-class

Fig. 10. Two observations of the same snake species (Boomslang, *Dispholidus typus*) with high visual dissimilarity related to sex (female left, male right). © Mark Heystek, iNaturalist, CC-BY-NC and © Daniel Rautenbach, iNaturalist, CC-BY-NC.

variances. In other words, certain classes could be highly variable in appearance depending on geographic location, sex, or age (Fig. 9) and at the same time could be visually similar to other species (e.g. mimicry) (Fig. 10). The goals and usage of image-based snake identification are complementary with those of other challenges: classifying snake species in images and predicting the list of species that are the most likely to be observed at a given location.

5.2 Dataset and Evaluation Protocol

Dataset Overview: For this challenge we have prepared a dataset with 259,214 images belonging to 783 snake species from 145 countries. The dataset has a heavy long-tailed class distribution, where the most frequent species (*Thamnophis sirtalis*) is represented by 12,201 images and the least frequent by just 17 (*Naja pallida*). Such a distribution with small inter-class variance and high intra-class variance creates a challenging task.

Training-Validation Split: To allow participants to easily validate their intermediate results, we have split the full dataset into a training subset with 245,185 images, and validation subset with 14,029 images. Both subsets have similar class distribution, while the minimum number of validation images per class is one.

Testing Dataset: Apart from other LifeCLEF challenges, the final testing set remains undisclosed as it is a composition of private images from individual reporters and natural history museums who have not put those images online in any form. A brief description of this closure method is as follows - twice as big as the validation set, contains all 973 classes, and observations from almost all the countries presented in training and validation sets.

Geographical Information: For approximately 80% of the images we provided a two levels of geographical information - country and continent. We have collected observations across 145 countries and all continents. Such information could be crucial for the AI based recognition as it is useful for human experts.

Fig. 11. Randomly selected images from the SnakeCLEF 2020 training set. © stewartb, iNaturalist, CC-BY-NC and © Jennifer Linde, iNaturalist, CC-BY-NC and © Gilberto Ponce Tejeda, iNaturalist, CC-BY-NC and © Ryan van Huyssteen, iNaturalist, CC-BY-NC and © Jessica Newbern, iNaturalist, CC-BY-NC.

Evaluation: The main goal of this challenge was to build a system that is autonomously able of recognizing 973 snake species based on the given image and geographical location input. Every participant had to submit their whole solution into the GitLab based evaluation system that performed evaluation over the secret testing set. Since data were secret each participated team could submit up to 5 submissions per day. The main evaluation metric for this challenge was the Dice Similarity Coefficient (DSC), also known as F1 score.

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

This score represents the harmonic mean of the Precision and the Recall.

$$Precision = \frac{TP}{TP + FN}; \qquad Recall = \frac{TP}{TP + FN}$$

The secondary metric was calculated as Multi-class Classification Logarithmic Loss e.g. Cross Entropy Loss.

$$LogLoss = -\sum_{c=1}^{M} y_{o,c} \cdot \log(p_{o,c})$$

This metric considers the uncertainty of a given prediction based on how much it differs from the actual label. This gives us a more subtle evaluation of the performance.

5.3 Participants and Results

Out of 8 registered teams in the SnakeCLEF 2020 challenge, only 2 teams managed to submit a working version of their recognition system. Even though participants were able to evaluate their system 5 times a day, we have registered only 27 submissions. Details of the methods and systems used in the runs are synthesized in the overview working note paper of the task [48] and further developed in the individual working notes (FHDO_BCSG [2]], Gokuleloop [44]). In a nutshell, both participants featured deep convolutional neural network architectures (ResNet50 and EfficientNet). They completely avoided CNN ensembles and used geological locations in a test time. The Gokuleloop team approaches were focused on the domain specific fine-tuning where this team tried different

pre-trained weights. With the Imagenet-21k weights, ResNet50 architecture, and naive probability weighting approach, Gokuleloop team achieved top F1 score of 0.625 while having a Log Loss of 0.83. The FHDO_BCSG team approaches combined two stages. Firstly, they used a Mask R-CNN instance detection method for snake detection. Secondly, different EfficientNet models were used to classify regions detected by the previous stage. Their best submitted model was an EfficientNet-B4 fine-tuned from the ImageNet pre-trained checkpoint. This model achieves F1 score of 0.404 and a Log-Loss of 6.650. The high Log-Loss was achieved due to the application of softmax normalization after the multiplication of the location data which leads to small differences in the predictions. All submission and their achieved scores are reported in the Fig. 12.

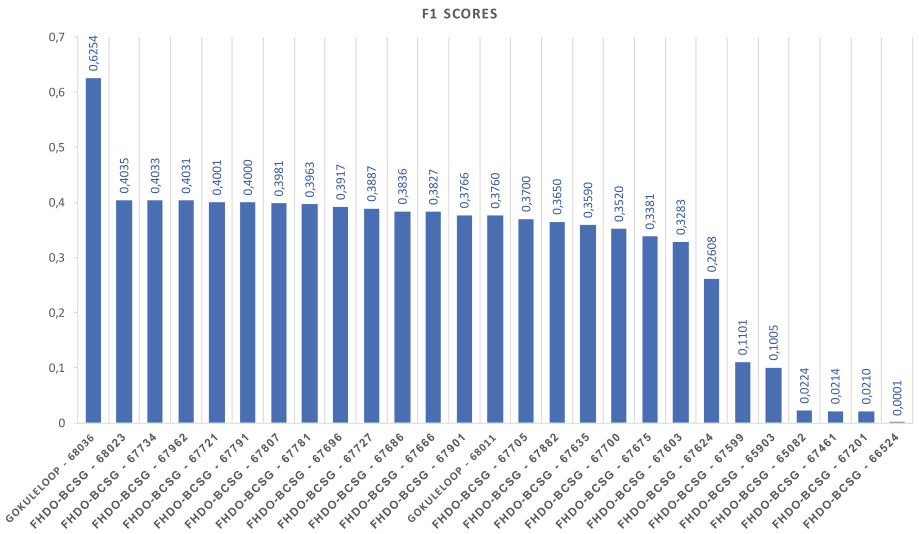


Fig. 12. F1 Scores achieved within the SnakeCLEF 2020.

6 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. The results of the PlantCLEF challenge, in particular, revealed that the last advances in domain adaptation enable the use of herbarium data to facilitate the identification of rare tropical species for which no or very few other training images are available. The results of the GeoLifeCLEF challenge were also highly relevant, revealing that deep convolutional neural networks trained on high-resolution geographic images are able to effectively predict species distribution even without using bioclimatic or soil variables. Furthermore, the results of the SnakeCLEF challenge showed that both traditional approaches and deep convolutional neural networks can benefit from geographical information.

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. Bai, J., Chen, C., Chen, J.: Xception based system for bird sound detection. In: CLEF Working Notes 2020, CLEF: Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 2020 (2020)
2. Bloch, L., et al.: Combination of image and location information for snake species identification using object detection and efficientnets. In: CLEF Working Notes 2020, CLEF: Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 2020 (2020)
3. Bolon, I., et al.: 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)
4. Bonnet, P., et al.: Plant identification: experts vs. machines in the era of deep learning. In: Joly, A., Vrochidis, S., Karatzas, K., Karppinen, A., Bonnet, P. (eds.) Multimedia Tools and Applications for Environmental & Biodiversity Informatics. MSA, pp. 131–149. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76445-0_8
5. Cai, J., Ee, D., Pham, B., Roe, P., Zhang, J.: Sensor network for the monitoring of ecosystem: bird species recognition. In: 2007 3rd International Conference on Intelligent Sensors, Sensor Networks and Information, ISSNIP 2007 (2007). https://doi.org/10.1109/ISSNIP.2007.4496859
6. de Castañeda, R.R., et al.: Snakebite and snake identification: empowering neglected communities and health-care providers with AI. Lancet Digit. Health **1**(5), e202–e203 (2019)
7. Chulif, S., Chang, Y.L.: Cross-domain plant identification on French Guyana Flora: neuon submission to LifeCLEF 2020 plant. In: CLEF Working Notes 2020, CLEF: Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 2020 (2020)
8. Clementino, T., Colonna, J.G.: Using triplet loss to bird species recognition on BirdCLEF 2020. In: CLEF Working Notes 2020, CLEF: Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 2020 (2020)
9. Cole, E., et al.: The GeoLifeCLEF 2020 dataset. arXiv preprint arXiv:2004.04192 (2020)
10. Deneu, B., et al.: Overview of LifeCLEF location-based species prediction task 2020 (GeoLifeCLEF). In: CLEF task overview 2020, CLEF: Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 2020 (2020)
11. Deneu, B., Servajean, M., Joly, A.: Participation of LIRMM/Inria to the GeoLifeCLEF 2020 challenge. In: CLEF Working Notes 2020, CLEF: Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 2020 (2020)
12. Sprengel, E., Jaggi, M., Kilcher, Y., Hofmann, T.: Audio based bird species identification using deep learning techniques. In: CLEF Working Notes 2016, CLEF: Conference and Labs of the Evaluation Forum, Évora, Portugal, September 2016 (2016)

13. Evans, J.S., Murphy, M.A., Holden, Z.A., Cushman, S.A.: Modeling species distribution and change using random forest. In: Drew, C., Wiersma, Y., Huettmann, F. (eds.) Predictive Species and Habitat Modeling in Landscape Ecology, pp. 139–159. Springer, New York (2011). https://doi.org/10.1007/978-1-4419-7390-0_8
14. Gaston, K.J., O'Neill, M.A.: Automated species identification: why not? Philos. Trans. Roy. Soc. Lond. B: Biol. Sci. **359**(1444), 655–667 (2004)
15. Ghazi, M.M., Yanikoglu, B., Aptoula, E.: Plant identification using deep neural networks via optimization of transfer learning parameters. Neurocomputing **235**, 228–235 (2017)
16. Glotin, H., Clark, C., LeCun, Y., Dugan, P., Halkias, X., Sueur, J.: Proceedings of 1st Workshop on Machine Learning for Bioacoustics - ICML4B. ICML, Atlanta (2013). http://sabiod.org/ICML4B2013_book.pdf
17. Goëau, H., Bonnet, P., Joly, A.: Plant identification in an open-world (LifeCLEF 2016). In: CLEF Task Overview 2016, CLEF: Conference and Labs of the Evaluation Forum, September 2016, Évora, Portugal (2016)
18. Goëau, H., Bonnet, P., Joly, A.: Plant identification based on noisy web data: the amazing performance of deep learning (LifeCLEF 2017). In: CLEF Task Overview 2017, CLEF: Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 2017 (2017)
19. Goëau, H., Bonnet, P., Joly, A.: Overview of ExpertLifeCLEF 2018: how far automated identification systems are from the best experts? In: CLEF Task Overview 2018, CLEF: Conference and Labs of the Evaluation Forum, Avignon, France, September 2018 (2018)
20. Goëau, H., Bonnet, P., Joly, A.: Overview of LifeCLEF plant identification task 2019: diving into data deficient tropical countries. In: CLEF Task Overview 2019, CLEF: Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 2019 (2019)
21. Goëau, H., Bonnet, P., Joly, A.: Overview of LifeCLEF plant identification task 2020. In: CLEF Task Overview 2020, CLEF: Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 2020 (2020)
22. Goëau, H., et al.: The ImageCLEF 2013 plant identification task. In: CLEF Task Overview 2013, CLEF: Conference and Labs of the Evaluation Forum, Valencia, Spain, September 2013 (2013)
23. Goëau, H., et al.: The ImageCLEF 2011 plant images classification task. In: CLEF Task Overview 2011, CLEF: Conference and Labs of the Evaluation Forum, Amsterdam, Netherlands, September 2011 (2011)
24. Goëau, H., et al.: ImageCLEF 2012 plant images identification task. In: CLEF Task Overview 2012, CLEF: Conference and Labs of the Evaluation Forum, Rome, Italy, September 2012 (2012)
25. Goëau, H., Glotin, H., Planqué, R., Vellinga, W.P., Stefan, Kahl, J.A.: Overview of BirdCLEF 2018: monophone vs. soundscape bird identification. In: CLEF Task Overview 2018, CLEF: Conference and Labs of the Evaluation Forum, Avignon, France, September 2018 (2018)
26. Goëau, H., Joly, A., Bonnet, P.: LifeCLEF plant identification task 2015. In: CLEF Task Overview 2015, CLEF: Conference and Labs of the Evaluation Forum, Toulouse, France, September 2015 (2015)
27. Goëau, H., et al.: The ImageCLEF plant identification task 2013. In: Proceedings of the 2nd ACM International Workshop on Multimedia Analysis for Ecological Data, pp. 23–28. ACM (2013)

28. Goëau, H., et al.: The LifeCLEF 2014 plant images identification task. In: CLEF Task Overview 2014, CLEF: Conference and Labs of the Evaluation Forum, Sheffield, United Kingdom, September 2014 (2014)
29. Hengl, T., et al.: SoilGrids250m: global gridded soil information based on machine learning. PLoS One **12**(2), e0169748 (2017)
30. Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A.: Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol.: J. Roy. Meteorol. Soc. **25**(15), 1965–1978 (2005)
31. Homer, C., et al.: Completion of the 2011 national land cover database for the conterminous united states - representing a decade of land cover change information. Photogram. Eng. Rem. Sens. **81**(5), 345–354 (2015)
32. Joly, A., et al.: Interactive plant identification based on social image data. Ecol. Inf. **23**, 22–34 (2014)
33. Joly, A., et al.: Overview of LifeCLEF 2018: a large-scale evaluation of species identification and recommendation algorithms in the era of AI. In: Bellot, P., et al. (eds.) CLEF 2018. LNCS, vol. 11018, pp. 247–266. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98932-7_24
34. Joly, A., et al.: Overview of LifeCLEF 2019: identification of Amazonian plants, South & North American Birds, and Niche prediction. In: Crestani, F., et al. (eds.) CLEF 2019. Lecture Notes in Computer Science, vol. 11696, pp. 387–401. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28577-7_29. https://hal.umontpellier.fr/hal-02281455
35. Joly, A., et al.: LifeCLEF 2016: multimedia life species identification challenges. In: Fuhr, N., et al. (eds.) CLEF 2016. LNCS, vol. 9822, pp. 286–310. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44564-9_26. https://hal.archives-ouvertes.fr/hal-01373781
36. Joly, A., et al.: LifeCLEF 2017 lab overview: multimedia species identification challenges. In: Jones, G.J.F., et al. (eds.) CLEF 2017. LNCS, vol. 10456, pp. 255–274. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65813-1_24. https://hal.archives-ouvertes.fr/hal-01629191
37. Joly, A., et al.: LifeCLEF 2014: multimedia life species identification challenges. In: Kanoulas, E., et al. (eds.) CLEF 2014. LNCS, vol. 8685, pp. 229–249. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11382-1_20. https://hal.inria.fr/hal-01075770
38. Joly, A., et al.: LifeCLEF 2015: multimedia life species identification challenges. In: Mothe, J., et al. (eds.) CLEF 2015. Lecture Notes in Computer Science, vol. 9283, pp. 462–483. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24027-5_46
39. Kahl, S., et al.: Overview of BirdCLEF 2020: bird sound recognition in complex acoustic environments. In: CLEF Task Overview 2020, CLEF: Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 2020 (2020)
40. 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, Lugano, Switzerland, September 2019 (2019)
41. Kahl, S., Wilhelm-Stein, T., Klinck, H., Kowerko, D., Eibl, M.: Recognizing birds from sound - the 2018 BirdCLEF baseline system. arXiv preprint arXiv:1804.07177 (2018)
42. 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)

43. Lee, S.H., Chan, C.S., Remagnino, P.: Multi-organ plant classification based on convolutional and recurrent neural networks. IEEE Trans. Image Process. **27**(9), 4287–4301 (2018)
44. Moorthy, G.K.: Impact of pretrained networks for snake species classification. In: CLEF Working Notes 2020, CLEF: Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 2020 (2020)
45. Motiian, S., Jones, Q., Iranmanesh, S., Doretto, G.: Few-shot adversarial domain adaptation. In: Advances in Neural Information Processing Systems, pp. 6670–6680 (2017)
46. Krishna, N.H., Ram Kaushik, R., R.M.: Plant species identification using transfer learning - PlantCLEF 2020. In: CLEF Working Notes 2020, CLEF: Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 2020 (2020)
47. NIPS International Conference: Proceedings of Neural Information Processing Scaled for Bioacoustics, from Neurons to Big Data (2013). http://sabiod.org/nips4b
48. 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, Thessaloniki, Greece, September 2020 (2020)
49. Picek, L., Sulc, M., Matas, J.: Recognition of the Amazonian flora by inception networks with test-time class prior estimation. In: CLEF Working Notes 2019, CLEF: Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 2019 (2019)
50. Roberts, D.R., et al.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography **40**(8), 913–929 (2017)
51. Towsey, M., Planitz, B., Nantes, A., Wimmer, J., Roe, P.: A toolbox for animal call recognition. Bioacoustics **21**(2), 107–125 (2012)
52. 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. J. Acoust. Soc. Am. **123**, 2424 (2008)
53. Van Horn, G., et al.: The inaturalist species classification and detection dataset. In: CVPR (2018)
54. Villacis, J., Goëau, H., Bonnet, P., Mata-Montero, E., Joly, A.: Domain adaptation in the context of herbarium collections: a submission to PlantCLEF 2020. In: CLEF Working Notes 2020, CLEF: Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 2020 (2020)
55. Wäldchen, J., Mäder, P.: Machine learning for image based species identification. Methods Ecol. Evol. **9**(11), 2216–2225 (2018)
56. Wäldchen, J., Rzanny, M., Seeland, M., Mäder, P.: Automated plant species identification–trends and future directions. PLoS Comput. Biol. **14**(4), e1005993 (2018)
57. Yu, X., Wang, J., Kays, R., Jansen, P.A., Wang, T., Huang, T.: Automated identification of animal species in camera trap images. EURASIP J. Image Video Process. **2013**, 52 (2013)
58. Zhang, Y., Davison, B.D.: Adversarial consistent learning on partial domain adaptation of PlantCLEF 2020 challenge. In: CLEF Working Notes 2020, CLEF: Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 2020 (2020)