Overview of LifeCLEF 2018: a Large-scale Evaluation of Species Identification and Recommendation Algorithms in the Era of AI

Alexis Joly[1], Hervé Goëau[2], Christophe Botella[1,3], Hervé Glotin[4], Pierre Bonnet[2], Willem-Pier Vellinga[5], Robert Planqué[5], Henning Müller[6]

[1] Inria, LIRMM, Montpellier, France
[2] CIRAD, UMR AMAP, France
[3] INRA, UMR AMAP, France
[4] AMU, Univ. Toulon, CNRS, ENSAM, LSIS UMR 7296, IUF, France
[5] Xeno-canto foundation, The Netherlands
[6] HES-SO, Sierre, Switzerland

Abstract. Building accurate knowledge of the identity, the geographic distribution and the evolution of living species is essential for a sustainable development of humanity, as well as for biodiversity conservation. Unfortunately, such basic information is often only partially available for professional stakeholders, teachers, scientists and citizens, and often incomplete for ecosystems that possess the highest diversity. In this context, an ultimate ambition is to set up innovative information systems relying on the automated identification and understanding of living organisms as a means to engage massive crowds of observers and boost the production of biodiversity and agro-biodiversity data. The LifeCLEF 2018 initiative proposes three data-oriented challenges related to this vision, in the continuity of the previous editions, but with several consistent novelties intended to push the boundaries of the state-of-the-art in several research directions. This paper describes the methodology of the conducted evaluations as well as the synthesis of the main results and lessons learned.

1 LifeCLEF Lab Overview

Identifying organisms is a key for accessing information related to the uses and ecology of species. This is an essential step in recording any specimen on earth to be used in ecological studies. Unfortunately, this is difficult to achieve due to the level of expertise necessary to correctly record and identify living organisms (for instance flowering plants are one of the most difficult groups to identify with an estimated number of 400,000 species). This *taxonomic gap* has been recognized since the Rio Conference of 1992, as one of the major obstacles to the global implementation of the Convention on Biological Diversity. Among the diversity of methods used for species identification, Gaston and O'Neill [10] discussed in 2004 the potential of automated approaches typically based on machine learning and multimedia data analysis methods. They suggested that, if the scientific

community is able to (i) overcome the production of large training datasets, (ii) more precisely identify and evaluate the error rates, (iii) scale up automated approaches, and (iv) detect novel species, it will then be possible to initiate the development of a generic automated species identification system that could open up vistas of new opportunities for theoretical and applied work in biological and related fields.

Since the question raised in Gaston and O'Neill [10], *automated species identification: why not?*, a lot of work was done on the topic (*e.g.* [30,7,46,45,47,23]) and it is still attracting much research today, in particular using deep learning techniques. In parallel to the emergence of automated identification tools, large social networks dedicated to the production, sharing and identification of multimedia biodiversity records have increased in recent years. Some of the most active ones like eBird[7] [43], iNaturalist[8], iSpot [39], Xeno-Canto[9] or Tela Botanica[10] (respectively initiated in the US for the two first ones and in Europe for the three last ones), federate tens of thousands of active members, producing hundreds of thousands of observations each year. Noticeably, the Pl@ntNet initiative was the first one attempting to combine the force of social networks with that of automated identification tools [23] through the release of a mobile application and collaborative validation tools. As a proof of their increasing reliability, most of these networks have started to contribute to global initiatives on biodiversity, such as the Global Biodiversity Information Facility (GBIF[11]) which is the largest and most recognized one. Nevertheless, this explicitly shared and validated data is only the tip of the iceberg. The real potential lies in the automatic analysis of the millions of raw observations collected every year through a growing number of devices but for which there is no human validation at all. However, this is still a challenging task: state-of-the-art multimedia analysis and machine learning techniques are actually still far from reaching the requirements of an accurate biodiversity monitoring system working. In particular, we need to progress on the number of species recognized by these systems. Indeed, the total number of living species on earth is estimated to be around 10K for birds, 30K for fishes, more than 400K for flowering plants (cf. State of the World's Plants 2017[12]) and more than 1.2M for invertebrates [2]. To bridge this gap, it is required to boost research on large-scale datasets and real-world scenarios.

To evaluate the performance of automated identification technologies in a sustainable, repeatable and scalable way, the LifeCLEF[13] research platform was created in 2014 as a continuation of the plant identification task [24] that was run within the ImageCLEF lab [14] the three years before [14,15,13,33]. LifeCLEF enlarged the evaluated challenge by considering birds and marine animals in ad-

[7] http://ebird.org/content/ebird/
[8] http://www.inaturalist.org/
[9] http://www.xeno-canto.org/
[10] http://www.tela-botanica.org/
[11] http://www.gbif.org/
[12] https://stateoftheworldsplants.com/
[13] http://www.lifeclef.org
[14] http://www.imageclef.org/

dition to plants, and audio and video content in addition to images. In this way, it aims at pushing the boundaries of the state-of-the-art in several research directions at the frontier of information retrieval, machine learning and knowledge engineering including (i) large scale classification, (ii) scene understanding, (iii) weakly-supervised and open-set classification, (iv) transfer learning and fine-grained classification and (v), humanly-assisted or crowdsourcing-based classification. As described in more detail in the following sections, each task is based on big and real-world data and the measured challenges are defined in collaboration with biologists and environmental stakeholders so as to reflect realistic usage scenarios. The main novelties of the 2018 edition of LifeCLEF compared to the previous years are the following:

1. **Expert vs. Machines plant identification challenge**: As the image-based identification of plants has improved considerably in the last few years (in particular through the PlantCLEF challenge), the next big question is how far such automated systems are from the human expertise. To answer this question, following the study of [4], we launched a new challenge, ExpertLifeCLEF, which involved 9 of the best expert botanists of the French flora who accepted to compete with AI algorithms.
2. **Location-based species recommendation challenge**: Automatically predicting the list of species that are the most likely to be observed at a given location is useful for many scenarios in biodiversity informatics. To boost the research on this topic, we also launched a new challenge called GeoLifeCLEF.

Besides these two main novelties, we decided to continue running the BirdCLEF challenge without major changes over the 2017 edition. The previous results actually showed that there was still a large margin of progress in terms of performance, in particular on the *soundscapes* data (long audio recordings). More generally, it is important to remind that an evaluation campaign such as LifeCLEF has to encourage long-term research efforts so as to (i) encourage non-incremental contributions, (ii) measure consistent performance gaps, and (iii), enable the emergence of a strong community.

Overall, 57 research groups from 22 countries registered to at least one of the three challenges of the lab. 12 of them finally crossed the finish line by participating in the collaborative evaluation and by writing technical reports describing in details their evaluated system. In the following sections, we provide a synthesis of the methodology and main results of each of the three challenges of LifeCLEF2018. More details can be found in the overview reports of each challenge and the individual reports of the participants (references provided below).

2 Task1: ExpertLifeCLEF

Automated identification of plants has improved considerably in the last few years. In the scope of LifeCLEF 2017 in particular, we measured impressive identification performance achieved thanks to recent convolutional neural network models. This raised the question of how far automated systems are from

the human expertise and of whether there is a upper bound that can not be exceeded. A picture actually contains only a partial information about the observed plant and it is often not sufficient to determine the right species with certainty. For instance, a decisive organ such as the flower or the fruit, might not be visible at the time a plant was observed. Some of the discriminant patterns might be very hard or unlikely to be observed in a picture such as the presence of pills or latex, or the morphology of the root. As a consequence, even the best experts can be confused and/or disagree between each other when attempting to identify a plant from a set of pictures. Similar challenges arise for most living organisms including fishes, birds, insects, etc. Quantifying this intrinsic data uncertainty and comparing it to the performance of the best automated systems is of high interest for both computer scientists and expert naturalists.

2.1 Dataset and Evaluation Protocol

Test set: to conduct a valuable experts vs. machines experiment, image-based identifications from the best of the best experts in the plant domain in France were collected according to the following procedure. 125 plants were photographed between May and June 2017, in a botanical garden called the *Parc floral de Paris* and in a natural area located in the north of Montpellier city (southern part of France, close to the Mediterranean sea). The photos were produced with two best-selling smartphones by a botanist and an amateur under his supervision. The species were selected by several criteria including (i) their membership to a difficult plant group (*i.e.* a group known as being the source of many confusions), (ii) the availability of well developed specimens with visible organs on the spot and (iii), the diversity of the selected set of species in terms of taxonomy and morphology. About fifteen pictures of each specimen were acquired to cover all the informative parts of the plant. However, only 1 to 5 pictures were randomly selected for all specimen to intentionally hide a part of the information and increase the difficulty of the identification. In the end, the set contains 75 plants illustrated by a total of 216 images and is related to 33 families and 58 genera. The species labels were cross-validated by other experts in order to have a near-perfect gold standard. Finally, the set was mixed into a larger one containing about 2000 observations (and about 7000 associated images) coming from the data flow of the mobile application Pl@ntNet[15],[16]. The added observations are necessarily related to species belonging to the list of the 10,000 species of the training set and are mainly wild plant species coming from the Western European flora and the North American flora but also plant species used all around the world as cultivated or ornamental plants including some endangered species.

Training set(s): As training data, all the datasets of the previous PlantCLEF challenges were made available to the participants. It can be divided into 3 subsets: first a **"Trusted"** training set contains 256,287 pictures related to the 10,000 most populated species in the online collaborative Encyclopedia Of

[15] `https://itunes.apple.com/fr/app/plantnet/id600547573?mt=8`
[16] `https://play.google.com/store/apps/details?id=org.plantnet`

Life (EoL) after a curation pipeline made by the organizers of the PlantCLEF 2017 task (taxonomic alignment, duplicates removal, herbaria sheets removal, no plant pcitures removal). A second ***Noisy*** training set is an extension of the ***Trusted*** training set adding about 900,000 images collected through the Bing image search engine during Autumn 2016 (and to a lesser extent with the Google image search engine). Lastly, a ***PlantCLEFPrevious*** training set is the concatenation of images collected through the Pl@ntNet project and shared during the challenges PlantCLEF 2011 to 2017, related to more than 100,000 images and 1100 species. In the end, the whole training set contains more than 1.2 million pictures and has the specificity to be strongly unbalanced with for instance a minimum of 4 pictures for the *Plectranthus sanguineus* species while the a maximum is 1732 pictures for *Fagus grandifolia*.

Task and evaluation: the goal of the task was to return the most likely species list by decreasing probability for each observation of the test set, and the main evaluation metric was the top-1 accuracy.

2.2 Participants and Results

28 research groups registered for the ExpertCLEF challenge 2018 and downloaded the dataset. Only 4 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 notes paper of the task [12] and further developed in the individual working notes of the participants (CMP [42], MfN [29], Sabanci [1] and TUC MI [21]. We report in Figure 1 the performance achieved by the 19 collected runs and the 9 participating human experts, while Figure 2 reports the results on the whole test dataset.

The main outcomes we derived from the results of the evaluation are the following ones:

A difficult task, even for experts: as a first noticeable outcome, none of the botanist correctly identified all observations. The top-1 accuracy of the experts is in the range $0.613-0.96$. with a median value of 0.8. This illustrates the difficulty of the task, especially when reminding that the experts were authorized to use any external resource to complete the task, Flora books in particular. It shows that a large part of the observations in the test set do not contain enough information to be identified with confidence when using classical identification keys. Only the four experts with an exceptional field expertise were able to correctly identify more than 80% of the observations.

Deep learning algorithms were defeated by the best experts but the margin of progression is becoming tighter and tighter. The top-1 accuracy of the evaluated systems is in the range $0.32 - 0.84$ with a median value of 0.64. This is globally lower than the experts but it is noticeable that the best systems were able to perform better than 5 of the highly skilled participating experts.

We give hereafter more details of the 2 systems that performed the best.

CMP system[42]: used an ensemble of a dozen Convolutional Neural Networks (CNNs) based on 2 state-of-the-art architectures (Inception-ResNet-v2

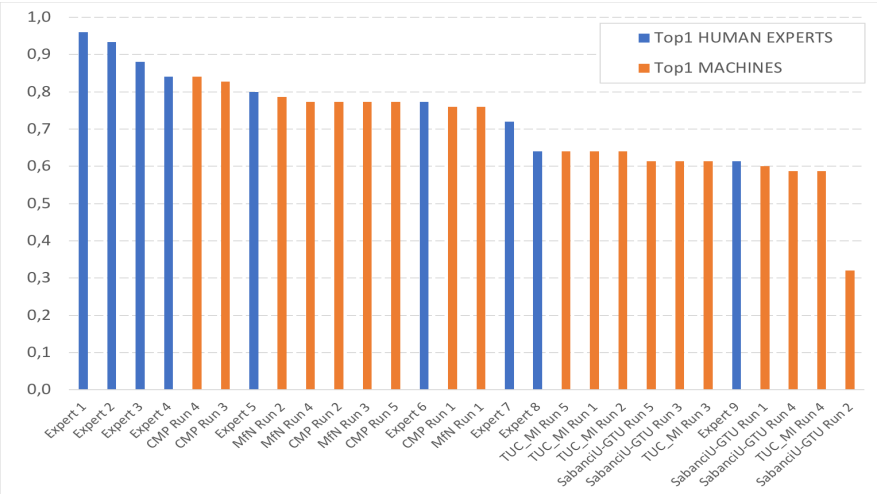

Fig. 1. ExpertLifeCLEF 2018 results: Identification performance achieved by the evaluated systems and the participating human experts

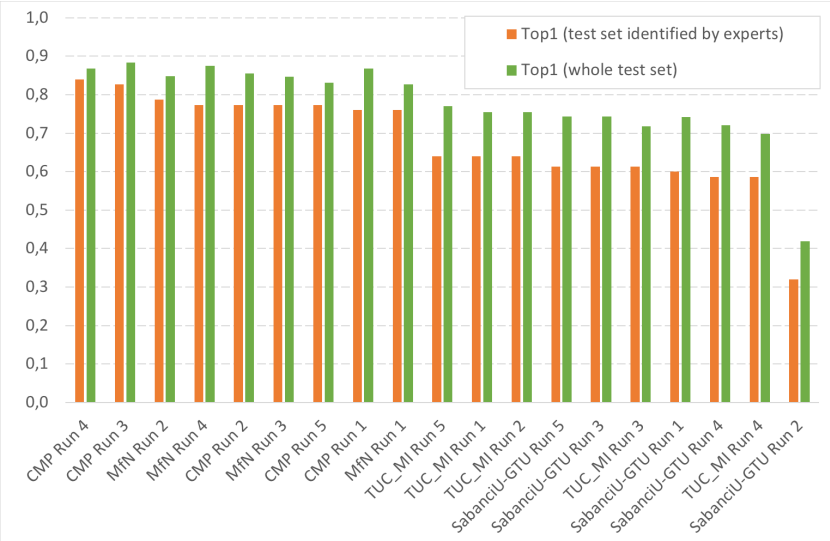

Fig. 2. Identification performance achieved by machines: top-1 accuracy on the whole test dataset and on the subpart also identified by the human experts.

and Inception-v4). The CNNs were initialized with weights pre-trained on ImageNet, then fine-tuned with different hyper-parameters and with the use of data augmentation (random horizontal flip, color distortions and random crops for some models). Each single test image is also augmented with 14 transformations (central/corner crops, horizontal flips, none) to combine and improve the pre-

dictions. Still at test time, the predictions are computed using the *Exponential Moving Average* feature of TensorFlow, *i.e.* by averaging the predictions of the set of models trained during the last iterations of the training phase (with an exponential decay). This popular procedure is inspired from Polyak averaging method [36] and is known to sometimes produce significantly better results than using the last trained model solely. As a last step in their system, assuming that there is a strong unbalanced distribution of the classes between the test and the training sets, the outputs of the CNNs are adjusted according to an estimation of the class prior probabilities in the test set based on an Expectation Maximization algorithm. The best score of 88.4% top-1 accuracy during the challenge was obtained by this team with the largest ensemble (CMP Run 3). With half less combined models, the CMP Run 4 reached a close top-1 accuracy and even obtained a slightly better accuracy on the smaller test subset identified by human experts. It can be explained by the strategy during the training of using the trusted and noisy sets: a comparison between CMP Run 1 and 4 clearly illustrates that refining further a model with only the trusted training set after learning it on the whole noisy training set is not relevant. CMP Run 3 which combines all the models seems to have its performances degraded by the inclusion of the models refined on the trusted training set when we compare it with CMP Run 4 on the test subset identified by human experts.

MfN system[29]: followed quite similar approaches used last year during the PlantCLEF2017 challenge [27]. This participant used an ensemble of fine-tuned CNNs pretrained on ImageNet, based on 4 architectures (GoogLeNet, ResNet-152, ResNeXT, DualPathNet92), each trained with bagging techniques. Data augmentation was used systematically for each training, in particular random cropping, horizontal flipping, variations of saturation, lightness and rotation. For the three last transformations, the intensity of the transformation is correlated to the diminution of the learning rate during training to let the CNNs see patches progressively closer to the original image at the end of the training. Test images followed similar transformations for combining and boosting the accuracy of the predictions. MfN Run 1 used basically the best and winning approach during PlantCLEF2017 by averaging the prediction of 11 models based on 3 architectures (GoogLeNet, ResNet-152, ResNeXT). However, surprisingly, the runs MfN Run 2 and 3, which are based on only one architecture (respectively ResNet152 and DualPathNet92), performed both better than the Run 1 combining several architectures and models. The combination of all the approaches in MfN Run 4 seems even to be penalized by the winning approach during PlantCLEF2017.

3 Task2: BirdCLEF

The general public as well as professionals like park rangers, ecological consultants and of course ornithologists are potential users of an automated bird song identifying system. A typical professional use would be in the context of wider

initiatives related to ecological surveillance or biodiversity conservation. Using audio records rather than bird pictures is justified [7,46,45,6] since birds are in fact not that easy to photograph and calls and songs have proven to be easier to collect and have been found to be species specific.

The 2018 edition of the task shares similar objectives and scenarios with the previous edition: (i) the identification of a particular bird species from a recording of one of its sounds, and (ii) the recognition of all species vocalising in so-called *soundscapes* that can contain up to several tens of birds vocalising. The first scenario is aimed at developing new automatic and interactive identification tools, to help users and experts to assess species and populations from field recordings obtained with directional microphones. The soundscapes, on the other side, correspond to a much more passive monitoring scenario in which any multi-directional audio recording device could be used without or with very light user's involvement. These (possibly crowdsourced) passive acoustic monitoring scenarios could scale the amount of annotated acoustic biodiversity records by several orders of magnitude.

3.1 Data and tasks description

SubTask1: monospecies (monophone) recordings The dataset was the same as the one used for BirdCLEF 2017 [17], mostly based on the contributions of the Xeno-Canto network. The training dataset contains 36,496 recordings covering 1500 species of south America (more precisely species observed in Brazil, Colombia, Venezuela, Guyana, Suriname, French Guiana, Bolivia, Ecuador and Peru) and it is the largest bioacoustic dataset in the literature to our knowledge. It has a massive class imbalance with a minimum of four recordings for *Laniocera rufescens* and a maximum of 160 recordings for *Henicorhina leucophrys*. Recordings are associated to various metadata such as the type of sound (call, song, alarm, flight, etc.), the date, the location, textual comments of the authors, multilingual common names and collaborative quality ratings. The test set for the monophone sub-task contains 12,347 recordings of the same type (mono-phone recordings). More details about that data can be found in the overview working note of BirdCLEF 2017 [17].

The goal of the task is to identify the species of the most audible bird (*i.e.* the one that was intended to be recorded) in each of the provided test recordings. Therefore, the evaluated systems have to return a ranked list of possible species for each of the 12,347 test recordings. The used evaluation metric is the Mean Reciprocal Rank (MRR), 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:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{Q} \frac{1}{rank_i}$$

where $|Q|$ is the total number of query occurrences in the test set.

SubTask2: soundscape recordings As the soundscapes appeared to be very challenging during the 2015 and 2016 (with an accuracy below 15%), new soundscape recordings containing time-coded bird species annotations were integrated in 2017 in the test set (so as to better understand what makes state-of-the-art methods fail on such contents). This new data was specifically created for BirdCLEF thanks to the work of Paula Caycedo Rosales (ornithologist from the Biodiversa Foundation of Colombia and Instituto Alexander von Humboldt, Xeno-Canto member), Hervé Glotin (bio-accoustician, co-author of this paper) and Lucio Pando (field guide and ornithologist in Peru). In total, about 6,5 hours of audio recordings were collected and annotated in the form of time-coded segments with associated species name. A baseline and validation package developed by Chemnitz University of Technology was shared with the participants[17]. The validation package contains 20 minutes of annotated soundscapes split into 5 recordings took of the last year test dataset. The baseline package offers a tools and a workflow to assist the participants in the development of their system: spectrograms extraction, deep neural network training, audio classification task, local validation (more details can be found in [26]).

Task Description Participants were asked to run their system so as to identify all the actively vocalising birds species in each test recording (or in each test segment of 5 seconds for the soundscapes). The submission *run files* had to contain as many lines as the total number of identifications, with a maximum of 100 identifications per test segment). Each prediction had to be composed of a species name belonging to the training set and a normalized score in the range $[0, 1]$ reflecting the likelihood that this species is singing in the 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. Then, the mean across all classes is computed as the main evaluation metric. More formally:

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

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

[17] https://github.com/kahst/BirdCLEF-Baseline

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.

3.2 Participants and results

29 research groups registered for the BirdCLEF 2018 challenge and downloaded the data. Six of them finally submitted run files and technical reports. Details of the systems and the methods used in the runs are synthesized in the overview working note of the task [16] and further developed in the individual working notes of the participants ([20,28,37,25,34]). Below we give more details about the 2 systems that performed the best:

MFN system [28]: this participant trained an ensemble of fine-tuned Inception-V3 models [44] feeded by mel spectrograms and using various data augmentation techniques in the temporal and frequency domains. According to some preliminary experiments they conducted [28], Inception-V3 is likely to outperform more recent and/or larger architectures (such as ResNet152, DualPathNet92, InceptionV4, DensNet, InceptionResNetV2, Xception, NasNet), presumably because of its auxiliary branch that acts as an effective regularizer. Among all the data augmentation techniques they experimented [28], the most contributing one is the addition of background noise or sounds from other files belonging to the same bird species with random intensity, in order to simulate artificially numerous contexts where a given species can be recorded. The other data augmentation types, all together, also improve the prediction but none of them is prevalent. Among them, we can mention a low-quality degradation based on a MP3 encoding-decoding, jitter on duration (+/- 0.5 sec), random factor to signal amplitude, random cyclic shift, random time interval dropouts, global and local pitch shift and frequency stretch, color jitter (brightness, contrast, saturation, hue). MfN Run 1 selected for each subtask the best single model learned during preliminary evaluations. The two models mainly differ in the pre-processing of audio files and choice of FFT parameters. MfN Run 2 combines both models, MfN Run 3 added a third declination of the model with other FFT parameters, but combined the predictions of the two best snapshots per model (regarding performance on the validation set) for averaging 3x2 predictions per species. MfN Run 4 added 4 more models and snapshots, reaching a total combination of 18 predictions per species.

OFAI system [37]: this participant used a quite different approach than MFN, without massive data augmentation and without relying on very deep image-oriented CNN architectures. OFAI rather used an ensemble of more shallow and compact CNN architectures (4 networks in total in OFAI Run 1). The first one, called *Sparrow*, was initially built for detecting the presence of bird calls in audio recordings [18]. *Sparrow* has a total of 10 layers (7 convolution, 2 pooling, 1 dense+softmax), taking as input rectangular gray mel spectrograms pictures. The second model is a variant of *Sparrow* where two pairs of convolution layers were replaced by two residual network blocks. During the training,

the first model focused on the foreground species as targets, while the second one used also the background species. Additional models were based on the same architectures but were learned as Born-Again Networks (BANs), a distillation technique where student models are not designed for compacting teacher models but where they are parameterized identically to them, surpassing finally the performance of the teachers [9]. For the species prediction a temporal pooling with log-mean-exp is applied for combining the outputs given by the *Sparrow* model for all chunks of 5 seconds from a single audio recording, while a temporal attention is used for the second model *Sparrow-resnet*. The predictions are combined after temporal pooling, but before the softmax. In addition to the four convolutional neural networks, eight Multi-Layer Perceptrons (MLPs) with two hidden leaky ReLU layers were learned on the meta-data vector associated to each audio recording (yearly circular date, longitude, latitude and elevation). A Gaussian blurring was applied to that data as a data augmentation technique to avoid overfitting. The 4 CNN and the 8 MLPs were finally combined into a single ensemble that was evaluated through the submission of OFAI Run 2. OFAI Run 3 is the same as Run 2 but exploited the information of the year of introduction of the test samples in the challenge as a mean to post-filter the predictions. OFAI Run 4 corresponds to the performance of a single *Sparrow* model.

The main conclusions we can draw from the results of Figures 3 and 4 are the following:

The overall performance improved significantly over last year for the mono-species recordings but not for the soundscapes: The best evaluated system achieves an impressive MRR score of 0.83 this year whereas the best system evaluated on the same dataset last year [38] achieved a MRR of 0.71. On the other side, we do not measured any strong progress on the soundscapes. The best system of MfN this year actually reaches a c-mAP of 0.193 whereas the best system of last year on the same test dataset [38] achieved a c-mAP of 0.182.

Using dates and locations of the observations provides some improvements: Contrary to all previous editions of LifeCLEF, one participant succeeded this year in improving significantly the predictions of its system by using the date and location of the observations. More precisely, OFAI Run 2 combining CNNs and metadata-based MLPs achieves a mono-species MRR of 0.75 whereas OFAI Run 1, relying solely on the CNNs, achieves a MRR of 0.72.

Shallow and compact architectures can compete with state-of-the-art architectures: on one hand one, can say that network architecture plays a crucial role and taking an heavy and deep state-of-the-art architecture such as Inception-v3 (MfN) with massive data augmentation is the best performing approach. On the other hand systems with shallow and compact architectures such as the OFAI system can reach very competitive results, even with a minimal number of data augmentation techniques.

The use of ensembles of networks still improves the performance consistently: this can be seen for instance through OFAI Run 4 (single model) that is consistently outperformed by OFAI Run 1 (11 models), or through the MfN Run 1 vs MfN Run 4 (18 models).

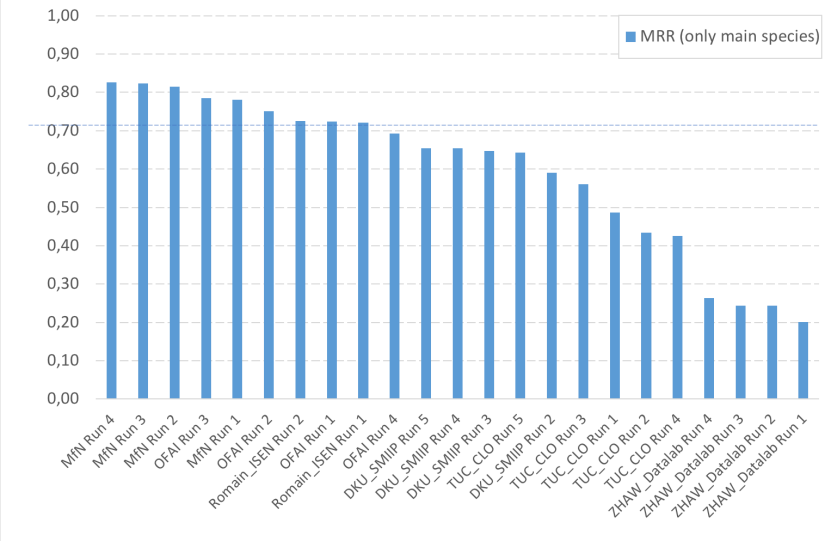

Fig. 3. BirdCLEF 2018 monophone identification results - Mean Reciprocal Rank. The blue dot line represents the last year's best system obtained by DYNI UTLN (Run 1) with a MRR of 0.714 [38]).

4 Task3: GeoLifeCLEF

The goal of the GeoLifeCLEF task is to automatically predict the list of plant species that are the most likely to be observed at a given location. This is useful for many scenarios in biodiversity informatics. First of all, it could improve species identification processes and tools by reducing the list of candidate species that are observable at a given location (be they automated, semi-automated or based on classical field guides or flora). More generally, it could facilitate biodiversity inventories through the development of location-based recommendation services (typically on mobile phones) as well as the involvement of non-expert nature observers. Last but not least, it might serve educational purposes thanks to biodiversity discovery applications providing innovative features such as contextualized educational pathways.

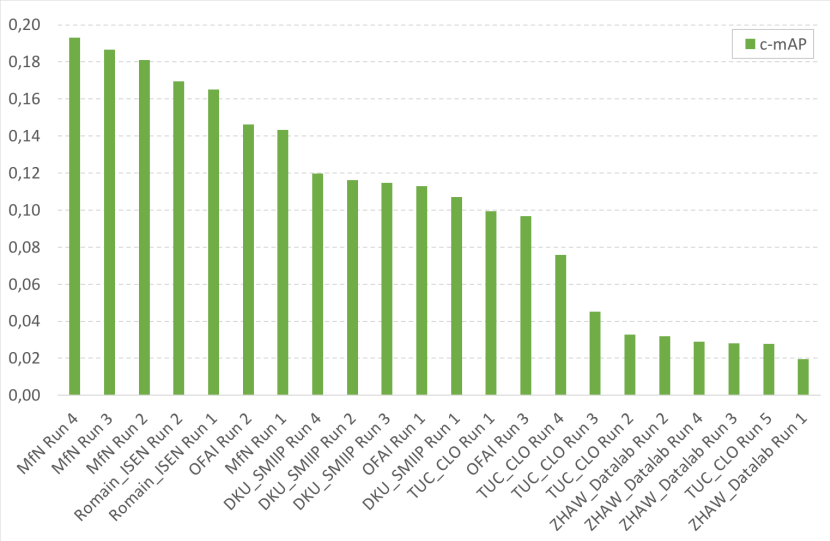

Fig. 4. BirdCLEF 2018 soundscape identification results - classification Mean Average Precision.

4.1 Data and evaluation procedure

A detailed description of the protocol used to build the GeoLifeCLEF 2018 dataset is provided in [5]. In a nutshell, the dataset was built from occurrences data of the Global Biodiversity Information Facility (GBIF [18]), the world's largest open data infrastructure in this domain, funded by governments. It is composed of $291,392$ occurrences of $N = 3,336$ plant species observed on the French territory between 1835 and 2017. Each occurrence is characterized by 33 local environmental images of 64×64 pixels. These environmental images are windows cropped from wider environmental rasters and centered on the occurrence spatial location. They were constructed from various open datasets including Chelsea Climate, ESDB V2 soil pedology data, Corine Land Cover 2012 soil occupation data, CGIAR-CSI evapotranspiration data, USGS Elevation data (Data available from the U.S. Geological Survey.) and BD Carthage hydrologic data.
This dataset was split in 3/4 for training and 1/4 for testing with the constraints that: (i) for each species in the test set, there is at least one observation of it in the train set. and (ii), an observation of a species in the test set is distant of more than 100 meters from all observations of this species in the train set.
In the following, we usually denote as $x \in X$ a particular occurrence, each x being associated to a spatial position $p(x)$ in the spatial domain D, a species label $y(x)$ and an environmental tensor $\mathbf{g}(x)$ of size 64x64x33. We denote as P the set of all spatial positions p covered by X. It is important to note that a given spatial

[18] `https://www.gbif.org/`

position $p_0 \in P$ usually corresponds to several occurrences $x_j \in X, p(x_j) = p_0$ observed at that location (18 000 spatial locations over a total of 60 000, because of quantized GPS coordinates or Names-to-GPS transforms). In the training set, up to several hundreds of occurrences can be located at the same place (be they of the same species or not). The occurrences in the test set might also occur at identical locations but, by construction, the occurrence of a given species does never occur at a location closer than 100 meters from the occurrences of the same species in the training set.
The used evaluation metric is the Mean Reciprocal Rank (MRR). 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 correct answer. The MRR is the average of the reciprocal ranks for the whole test set:

$$MRR = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{rank_q}$$

where Q is the total number of query occurrences x_q in the test set and $rank_q$ is the rank of the correct species $y(x_q)$ in the ranked list of species predicted by the evaluated method for x_q.

4.2 Participants and results

22 research groups registered to the GeoLifeCLEF 2018 challenge and downloaded the dataset. Three research groups finally 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 of the task [5] and further developed in the individual working notes of the participants (FLO [3], ST [41] and SSN [35]). In a nutshell, the FLO team [3] developed four prediction models, (i) one convolutional neural network trained on environmental data (FLO_3), (ii) one neural network trained on co-occurrences data (FLO_2) and two other models only based on the spatial occurrences of species: (iii) a closest-location classifier (FLO_1) and (iv) a random forest fitted on the spatial coordinates (FLO_4). Other runs correspond to late fusions of that base models. The ST team [41] experimented two main types of models, convolutional neural networks on environmental data (ST_1, ST_3, ST_11, ST_14, ST_15, ST_18, ST_19) and Boosted Trees (XGBoost) on vectors of environmental variables concatenated with spatial positions (ST_6, ST_9, ST_10, ST_12, ST_13, ST_16, ST_17). For analysis purposes, ST_2 corresponds to a random predictor and ST_7 to a constant predictor returning always the 100 most frequent species (ranked by decreasing value of their frequency in the training set). The last team SSN [35], attempted to learn a CNN-LSTM hybrid model, based on a ResNext architecture [48] extended with an LSTM layer [11] aimed at predicting the plant categories at 5 different levels of the taxonomy (class, then order, then family,

then genus and finally species).

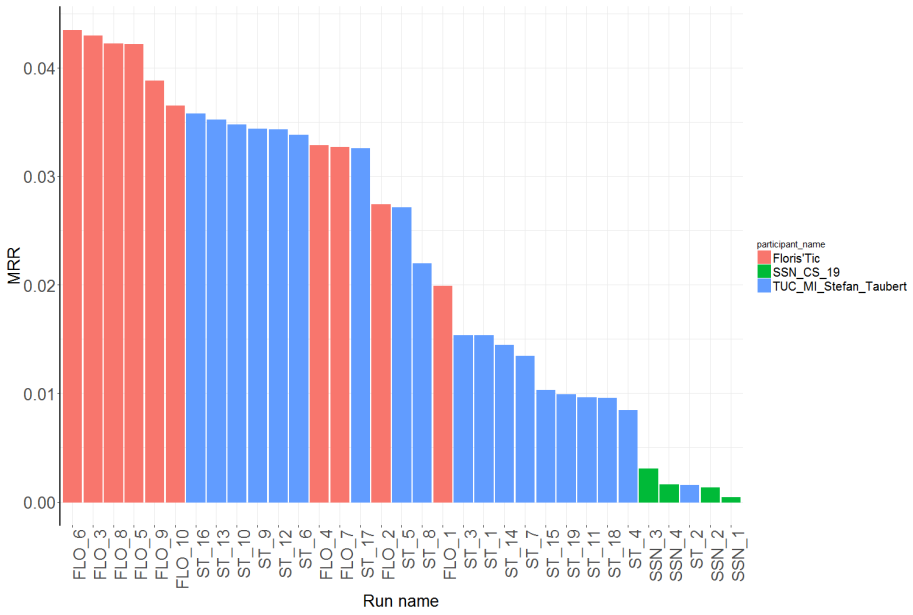

Fig. 5. GeoLifeCLEF 2018 results - Mean Reciprocal Rank of the evaluated systems

We report in Figure 5 the performance achieved by the 33 submitted runs. The main conclusions we can draw from the results are the following:

Convolutional Neural Networks outperformed boosted trees: Boosted trees are known to provide state-of-the-art performance for environmental modelling. They are actually used in a wide variety of ecological and studies [19,8,31,32]. Our evaluation, however, demonstrate that they can be consistently outperformed by convolutional neural networks trained on environmental data tensors. The best submitted run that does not result from a fusion of different models (**FLO_3**), is actually a convolutional neural network trained on the environmental patches. It achieved a MRR of 0.043 whereas the best boosted tree (**ST_16**) achieved a MRR of 0.035. As another evidence of the better performance of the CNN model, the six best runs of the challenge result from the combination of it with the other models of the Floris'Tic team. Now, it is important to notice that the CNN models trained by the ST team (ST_1, ST_3, ST_11, ST_14, ST_15, ST_18, ST_19) and SSN teams did not obtain good performance at all (often worse than the constant predictor based on the class prior distribution). This illustrates the difficulty of designing and fitting deep neural networks on new problems without former references in the literature. In particular, the approaches

trying to adapt existing complex CNN architectures that are popular in the image domain (such as VGG [40], DenseNet [22], ResNEXT [48] and LSTM [11]) were not successfull. High difference of performances in CNN learned with homemade architectures ($FLO\_6, FLO\_3, FLO\_8, FLO\_5, FLO\_9, FLO\_10$ compared to ST_3, ST_1) underlines the importance of architecture choices.

Purely spatial models are not so bad: the random forest model of the FLO team, fitted on spatial coordinates solely (FLO_4), achieved a fair MRR of 0.0329, close to the performance of the boosted trees of the ST team (that were trained on environmental & spatial data). Purely spatial models are usually not used for species distribution modelling because of the heterogeneity of the observations density across different regions. Indeed, the spatial distribution of the observed specimens is often more correlated with the geographic preferences of the observers than with the abundance of the observed species. However the goal of GeoLifeClef is to predict the most likely species to observe given the real presence of a plant. Thus, the heterogeneity sampling effort should induce less bias than in ecological studies.

It is likely that the Convolutional Neural Network already captured the spatial information: The best run of the whole challenge (**FLO_6**) results from the combination of the best environmental model (CNN **FLO_3**) and the best spatial model (Random forest **FLO_4**). However, it is noticeable that the improvement of the fused run compared to the CNN alone is extremely tight (+ 0.0005), and actually not statistically significant. In other words, it seems that the information learned by the spatial model was already captured by the CNN. The CNN might actually have learned to recognize some particular locations thanks to specific shapes of the landscape in the environmental tensors.

A significant margin of progress but still very promising results: even if the best MRR scores appear to be very low at a first glance, it is important to relativize them with regard to the nature of the task. Many species (tens to hundred) are actually living at the same location so that achieving very high MRR scores is not possible. The MRR score is useful to compare the methods between each others but it should not be interpreted as for a classical information retrieval task. In the test set itself, several species are often observed at exactly the same location. So that there is a max bound on the achievable MRR equal to 0.56. The best run (**FLO_3**) is still far from this max bound (MRR=0.043) but it is much better than the random or the prior distribution based MRR. Concretely, it retrieves the right species in the top-10 results in 25% of the cases, or in the top-100 in 49% of the cases (over $3,336$ species in the training set), which means that it is not so bad at predicting the set of species that might be observed at that location.

5 Conclusions and Perspectives

The main outcome of this collaborative evaluation is a snapshot of the performance of state-of-the-art computer vision, bio-acoustic and machine learning techniques towards building real-world biodiversity monitoring systems. The re-

sults did show that very high identification rates can be reached by the evaluated systems, even on large number of species (up to 10,000 species). The most noticeable progress came from the deployment of new convolutional neural network architectures, confirming the fast growing progress of that techniques. Concerning the identification of plant images, our study did show that the performance of the best models is now very close from the expertise of highly skilled botanists. Concerning bird sounds identification, our study reports impressive performance when using monospecies recordings of good quality such as the one recorded by the Xeno-Canto community. Identifying birds in raw, multi-directional soundscapes, however, remains a very challenging task. We actually did not measure any progress compared to the previous year despite several participants are working hard on this problem. Last but not least, a new challenge was introduced this year for the evaluation of location-based species recommendation methods based on environmental and spatial data. Here again, CNNs trained on environmental tensors appeared to be the most promising models. They outperformed boosted trees which are usually known as the state-of-the-art in ecology. We believe this is the beginning of a new integrative approach to environmental modelling, involving multi-task deep learning models trained on very big multi-modal datasets.

Acknowledgements The organization of LifeCLEF 2018 was supported by the French project Floris'Tic (Tela Botanica, INRIA, CIRAD, INRA, IRD) funded in the context of the national investment program PIA. The organization of the BirdCLEF task was supported by the Xeno-Canto foundation for nature sounds as well as the French CNRS project SABIOD.ORG and EADM GDR CNRS MADICS, BRILAAM STIC-AmSud. The annotations of some soundscape were prepared with regreted wonderful Lucio Pando at Explorama Lodges, with the support of Pam Bucur, Marie Trone and H. Glotin.

References

1. Atito, S., Yanıkoglu, B., Aptoula, E., Ganiyusufoglu, I., Yildiz, A., Yildirir, K., Baris, S.: Plant identification with deep learning ensembles. In: Working Notes of CLEF 2018 (Cross Language Evaluation Forum) (2018)
2. Baillie, J., Hilton-Taylor, C., Stuart, S.N.: 2004 IUCN red list of threatened species: a global species assessment. Iucn (2004)
3. Benjamin Deneu, Maximilien Servajean, C.B., Joly, A.: Location-based species recommendation using co-occurrences and environment- geolifeclef 2018 challenge. In: CLEF working notes 2018 (2018)
4. Bonnet, P., Goëau, H., Thye Hang, S., Lasseck, M., Šulc, M., Malécot, V., Philippe, J., Melet, J.C., You, C., Joly, A.: Plant identification: Experts vs. machines in the era of deep learning. "Multimedia Technologies for Environmental & Biodiversity Informatics" A. Joly, P. Bonnet, S. Vrochidis, K. Karatzas and A. Karppinen, Springer Verlag (2018)
5. Botella, C., Bonnet, P., Joly, A.: Overview of geolifeclef 2018: location-based species recommendation. In: CLEF working notes 2018 (2018)
6. Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X.Z., Raich, R., Hadley, S.J., Hadley, A.S., Betts, M.G.: Acoustic classification of multiple simultaneous bird

species: A multi-instance multi-label approach. The Journal of the Acoustical Society of America 131, 4640 (2012)
7. 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)
8. De'Ath, G.: Boosted trees for ecological modeling and prediction. Ecology 88(1), 243–251 (2007)
9. Furlanello, T., Lipton, Z.C., Itti, L., Anandkumar, A.: Born again neural networks. In: Metalearn 2017 NIPS workshop. pp. 1–5 (Dec 2017)
10. Gaston, K.J., O'Neill, M.A.: Automated species identification: why not? Philosophical Transactions of the Royal Society of London B: Biological Sciences 359(1444), 655–667 (2004)
11. Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: Continual prediction with lstm (1999)
12. Goëau, H., Bonnet, P., Joly, A.: Overview of expertlifeclef 2018: how far automated identification systems are from the best experts? lifeclef experts vs. machine plant identification task 2018. In: CLEF 2018 (2018)
13. Goëau, H., Bonnet, P., Joly, A., Bakic, V., Barthélémy, D., Boujemaa, N., Molino, J.F.: The imageclef 2013 plant identification task. In: CLEF 2013. Valencia (2013)
14. 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 2011 (2011)
15. Goëau, H., Bonnet, P., Joly, A., Yahiaoui, I., Barthélémy, D., Boujemaa, N., Molino, J.F.: Imageclef2012 plant images identification task. In: CLEF 2012. Rome (2012)
16. Goëau, H., Glotin, H., Planqué, R., Vellinga, W.P., Stefan, Kahl, J.A.: Overview of birdclef 2018: monophone vs. soundscape bird identification. In: CLEF working notes 2018 (2018)
17. Goëau, H., Glotin, H., Vellinga, W., Planqué, B., Joly, A.: Lifeclef bird identification task 2017. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017. (2017)
18. 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)
19. Guisan, A., Thuiller, W., Zimmermann, N.E.: Habitat Suitability and Distribution Models: With Applications in R. Cambridge University Press (2017)
20. Haiwei, W., Ming, L.: Construction and improvements of bird songs' classification system. In: Working Notes of CLEF 2018 (Cross Language Evaluation Forum) (2018)
21. Haupt, J., Kahl, S., Kowerko, D., Eibl, M.: Large-scale plant classification using deep convolutional neural networks. In: Working Notes of CLEF 2018 (Cross Language Evaluation Forum) (2018)
22. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. vol. 1, p. 3 (2017)
23. 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)
24. Joly, A., Goëau, H., Bonnet, P., Bakic, V., Molino, J.F., Barthélémy, D., Boujemaa, N.: The imageclef plant identification task 2013. In: International workshop on Multimedia analysis for ecological data (2013)

25. Kahl, S., Wilhelm-Stein, T., Klinck, H., Kowerko, D., Eibl, M.: A baseline for large-scale bird species identification in field recordings. In: Working Notes of CLEF 2018 (Cross Language Evaluation Forum) (2018)
26. 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)
27. Lasseck, M.: Image-based plant species identification with deep convolutional neural networks. In: Working Notes of CLEF 2017 (Cross Language Evaluation Forum) (2017)
28. Lasseck, M.: Audio-based bird species identification with deep convolutional neural networks. In: Working Notes of CLEF 2018 (Cross Language Evaluation Forum) (2018)
29. Lasseck, M.: Machines vs. experts: Working note on the expertlifeclef 2018 plant identification task. In: Working Notes of CLEF 2018 (Cross Language Evaluation Forum) (2018)
30. 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)
31. Messina, J.P., Kraemer, M.U., Brady, O.J., Pigott, D.M., Shearer, F.M., Weiss, D.J., Golding, N., Ruktanonchai, C.W., Gething, P.W., Cohn, E., et al.: Mapping global environmental suitability for zika virus. Elife 5 (2016)
32. Moyes, C.L., Shearer, F.M., Huang, Z., Wiebe, A., Gibson, H.S., Nijman, V., Mohd-Azlan, J., Brodie, J.F., Malaivijitnond, S., Linkie, M., et al.: Predicting the geographical distributions of the macaque hosts and mosquito vectors of plasmodium knowlesi malaria in forested and non-forested areas. Parasites & vectors 9(1), 242 (2016)
33. Müller, H., Clough, P., Deselaers, T., Caputo, B. (eds.): ImageCLEF – Experimental Evaluation in Visual Information Retrieval, The Springer International Series On Information Retrieval, vol. 32. Springer, Berlin Heidelberg (2010)
34. Müller, L., Marti, M.: Two bachelor students' adventures in machine learning. In: Working Notes of CLEF 2018 (Cross Language Evaluation Forum) (2018)
35. Nithish B Moudhgalya, Sharan Sundar, S.D.M.P., Bose, C.A.: Hierarchically embedded taxonomy with clnn to predict species based on spatial features. In: CLEF working notes 2018 (2018)
36. Polyak, B.T., Juditsky, A.B.: Acceleration of stochastic approximation by averaging. SIAM Journal on Control and Optimization 30(4), 838–855 (1992)
37. Schlüter, J.: Bird identification from timestamped, geotagged audio recordings. In: Working Notes of CLEF 2018 (Cross Language Evaluation Forum) (2018)
38. Sevilla, A., Glotin, H.: Audio bird classification with inception v4 joint to an attention mechanism. In: Working Notes of CLEF 2017 (Cross Language Evaluation Forum) (2017)
39. Silvertown, J., Harvey, M., Greenwood, R., Dodd, M., Rosewell, J., Rebelo, T., Ansine, J., McConway, K.: Crowdsourcing the identification of organisms: A case-study of ispot. ZooKeys (480), 125 (2015)
40. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014)
41. Stefan Taubert, Max Mauermann, S.K.D.K., Eibl, M.: Species prediction based on environmental variables using machine learning techniques. In: CLEF working notes 2018 (2018)

42. Sulc, M., Picek, L., Matas, J.: Plant recognition by inception networks with test-time class prior estimation. In: Working Notes of CLEF 2018 (Cross Language Evaluation Forum) (2018)
43. Sullivan, B.L., Aycrigg, J.L., Barry, J.H., Bonney, R.E., Bruns, N., Cooper, C.B., Damoulas, T., Dhondt, A.A., Dietterich, T., Farnsworth, A., et al.: The ebird enterprise: an integrated approach to development and application of citizen science. Biological Conservation 169, 31–40 (2014)
44. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826 (2016)
45. Towsey, M., Planitz, B., Nantes, A., Wimmer, J., Roe, P.: A toolbox for animal call recognition. Bioacoustics 21(2), 107–125 (2012)
46. 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)
47. Wäldchen, J., Rzanny, M., Seeland, M., Mäder, P.: Automated plant species identification—trends and future directions. PLOS Computational Biology 14(4), 1–19 (04 2018), https://doi.org/10.1371/journal.pcbi.1005993
48. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. pp. 5987–5995. IEEE (2017)