Overview of BirdCLEF 2021: Bird call identification in soundscape recordings

Stefan Kahl, Tom Denton, Holger Klinck, Hervé Glotin, Hervé Goeau

Overview of BirdCLEF 2021: Bird call identification in soundscape recordings

Stefan Kahl[1], Tom Denton[2], Holger Klinck[1], Hervé Glotin[3], Hervé Goëau[4], Willem-Pier Vellinga[5], Robert Planqué[5] and Alexis Joly[6]

[1]K. Lisa Yang Center for Conservation Bioacoustics, Cornell Lab of Ornithology, Cornell University, Ithaca, USA
[2]Google LLC, San Francisco, USA
[3]University of Toulon, AMU, CNRS, LIS, Marseille, France
[4]CIRAD, UMR AMAP, Montpellier, France
[5]Xeno-canto Foundation, Groningen, Netherlands
[6]Inria, LIRMM, University of Montpellier, CNRS, Montpellier, France

Abstract

Conservation of bird species requires detailed knowledge of their spatiotemporal occurrence and distribution patterns. Over the past decade, passive acoustic monitoring (PAM) has become an essential tool to collect data on birds on ecologically relevant scales. However, these PAM efforts generate extensive datasets, and their comprehensive analysis remains challenging. Improved and fully automated acoustic analysis frameworks are needed to advance the field of avian conservation. The 2021 BirdCLEF challenge focused on developing and assessing automated analysis frameworks for avian vocalizations in continuous soundscape data. The primary task of the challenge was to detect and identify all bird calls within the hidden test dataset. This paper describes how the various algorithms were evaluated and synthesizes the results and lessons learned.

Keywords

LifeCLEF, bird, song, call, species, retrieval, audio, collection, identification, fine-grained classification, evaluation, benchmark, bioacoustics, passive acoustic monitoring

1. Introduction

Birds are widely used to monitor ecosystem health because they live in most environments and occupy almost every niche within those environments. Traditionally, human observers monitor bird populations by conducting point count surveys in an area of interest. At sampling locations along a transect, the domain expert will visually and aurally count every bird in a given time window (e.g., 3 or 5 minutes). However, conducting these surveys is time-consuming and requires expert knowledge in the identification of birds. Because the number of observers is typically limited, the spatiotemporal resolution of the surveys is limited as well.

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

sk2487@cornell.edu (S. Kahl); tmd@google.com (T. Denton); Holger.Klinck@cornell.edu (H. Klinck); herve.glotin@univ-tln.fr (H. Glotin); herve.goeau@cirad.fr (H. Goëau); wp@xeno-canto.org (W. Vellinga); bob@xeno-canto.org (R. Planqué); alexis.joly@inria.fr (A. Joly)
0000-0002-2411-8877 (S. Kahl); 0000-0003-1078-7268 (H. Klinck); 0000-0001-7338-8518 (H. Glotin); 0000-0003-3296-3795 (H. Goëau); 0000-0003-3886-5088 (W. Vellinga); 0000-0002-0489-5425 (R. Planqué); 0000-0002-2161-9940 (A. Joly)

In contrast, passive acoustic monitoring (PAM) uses autonomous recording units (ARUs) to monitor the acoustic environment, often continuously, in the vicinity of the deployment location over extended periods (weeks to months). These data sets complement traditional bird surveys and help to improve our ability to accurately monitor the status and trends of bird populations and avian diversity more broadly.

While PAM surveys are very cost-effective to conduct, the handling and analysis of vast amounts of collected data (often tens or even hundreds of Terabytes) remains challenging. In the past, researchers frequently subsampled the collected data or focused on specific call types to circumnavigate this challenge. However, as a consequence, large amounts of data remain untouched, and new analysis frameworks are required to mine these datasets thoroughly. Effective analysis frameworks coming out of BirdCLEF and other competitions have the potential to revolutionize how we monitor and conserve birds and biodiversity in the future.

The *LifeCLEF Bird Recognition Challenge* (BirdCLEF) focuses on the development of reliable analysis frameworks to detect and identify avian vocalizations in continuous soundscape data. Launched in 2014, it has become one of the largest bird sound recognition competition in terms of dataset size and species diversity, with multiple tens of thousands of recordings covering up to 1,500 species [1, 2].

2. BirdCLEF 2021 Competition Overview

Recent advances in the development of machine listening approaches to identify animal vocalizations have improved our ability to comprehensively analyze long-term acoustic datasets [3, 4]. However, it remains difficult to generate analysis outputs with high precision and recall, especially when targeting a high number of species simultaneously. Bridging the domain gap between high-quality training samples (focal recordings) and noisy test samples (soundscape recordings) is one of the most challenging tasks in the area of acoustic event detection and identification. The 2021 BirdCLEF competition tackled this complex task and was held on Kaggle. This year's edition was a so-called '*code competition*' which encourages participants to publish their code for the benefit of the community.

2.1. Goal and Evaluation Protocol

The 2021 BirdCLEF challenge focused on developing and assessing automated analysis frameworks for avian vocalizations in continuous soundscape data. The primary task of the challenge was to detect and identify all bird calls within the hidden test dataset. Each soundscape was divided into 5 second segments, and participants were tasked to return a list of audible species for each segment. The row-wise micro-averaged F1 score was used for evaluation. In previous editions, ranking metrics were used to assess the overall classification performance. However, when applying bird call identification frameworks to real-world data, a suitable confidence threshold must be set to balance precision and recall. The F1 score reflects this circumstance best. However, the selected threshold can significantly impact the overall performance, especially when applied to the hidden test dataset.

Precision and recall were determined based on the total number of true positives (TP), false positives (FP), and false negatives (FN) for each segment (i.e., row of the submission). More formally:

$$\textit{Micro-Precision} = \frac{TP}{TP + FP}, \qquad \textit{Micro-Recall} = \frac{TP}{TP + FN}$$

The micro-F1 score as harmonic mean of the micro-precision and micro-recall for each segment was defined as:

$$\textit{Micro-F1} = 2 \times \frac{\textit{Micro-Precision} \times \textit{Micro-Recall}}{\textit{Micro-Precision} + \textit{Micro-Recall}}$$

The average across all (segment-wise) micro-F1 scores was used as the final metric. Segments that did not contain a bird vocalization had to be marked with the '*nocall*' label, which acted as an additional class label for non-events. The micro-averaged F1 score reduced the impact of rare events, which only contributed slightly to the overall metric if misidentified. The classification performance on common classes (i.e., species with high vocal presence) was well reflected in the metric.

2.2. Dataset

The 2021 BirdCLEF challenge featured one of the largest, fully annotated collections of soundscape recordings from four different recording locations in North and South America. Concerning real-world use cases, labels and metrics were chosen to reflect the vast diversity of bird vocalizations and variable ambient noise levels in omnidirectional recordings.

2.2.1. Training Data

As in previous editions, training data were provided by the Xeno-canto community and consisted of more than 60,000 (high-quality, focal) recordings covering 397 species from two continents (North and South America). The maximum number of recordings for one species was limited to 500, which only affected a dozen species and resulted in a highly unbalanced dataset. Nine species contained less than 25 recordings, making it difficult to train reliable classifiers without an appropriate few-shot approach. Participants were allowed to use various metadata to develop their frameworks. Most notably, we provided detailed location information on recording sites of focal and soundscape recordings, allowing participants to account for migration and spatial distribution of bird species. Other metadata as secondary labels, call type, and recording quality were also provided, allowing participants to apply pre- and post-processing schemes which were not only based on audio inputs.

2.2.2. Test Data

In this edition, test data were hidden and only accessible to participants during the inference process. This required participants to fine-tune their systems without knowing the value

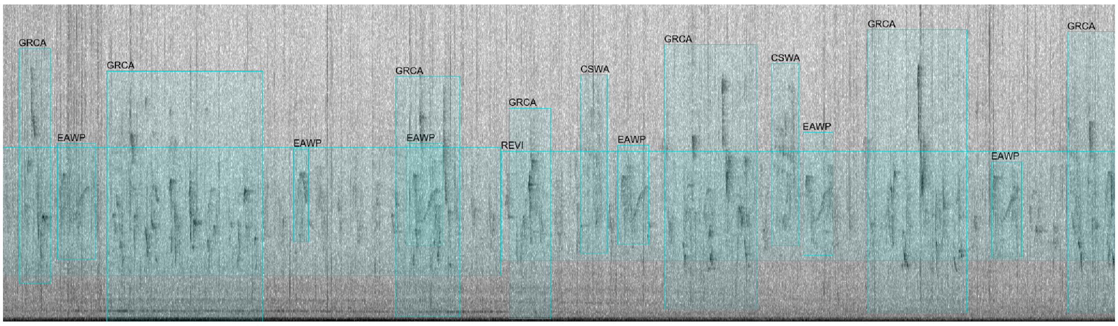


Figure 1: Dawn chorus soundscapes (like this example from the SNE dataset) often have an extremely high call density. The 2021 BirdCLEF dataset contained 100 fully annotated 10-minute soundscapes recorded in North and South America. Expert ornithologists provided bounding box labels for all soundscape recordings.

distribution of the test data. This approach more closely resembles real-world use cases where vast majorities of the recorded audio data have an unknown species composition. The hidden test data contained 80 soundscape recordings of 10-minute duration covering four distinct recording locations. All audio data were collected with passive acoustic recorders (SWIFT recorders, K. Lisa Yang Center for Conservation Bioacoustics, Cornell Lab of Ornithology[1]) deployed in Colombia (COL), Costa Rica (COR), the Sierra Nevada (SNE) of California, USA and the Sapsucker Woods Sanctuary (SSW) in Ithaca, New York, USA. Expert ornithologists provided annotations for a variety of quiet and extremely dense acoustic scenes (see Figure 1). In addition, a validation dataset with 200 minutes (20x 10-minute recordings) of soundscape data were also provided to allow participants to get a better understanding of the acoustic target domain. Participants were allowed to use these data for validation or during training. Soundscapes from the validation data only covered two (COR, SSW) of the four recording locations.

2.2.3. Colombia (COL) & Costa Rica (COR)

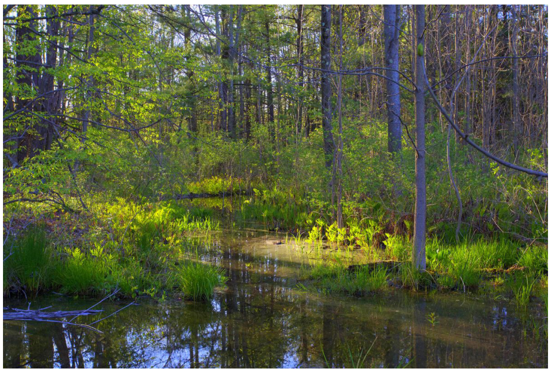
The Nespresso Biodiversity Index aims at quantifying the impact of eco-friendly coffee farms on the avian diversity in surrounding areas. In collaboration with the Cornell Lab of Ornithology, passive acoustic recorders were deployed on coffee farms in Colombia and Costa Rica to measure the transformative effects of sustainable farming by analyzing large amounts of acoustic data. Surveys are carried out twice a year to be able to capture how bird species are using the coffee landscape when both Neotropical resident and migratory birds are present (Nov-Mar) and around the peak of the breeding season for resident birds (April-Jun) [5]. Developing automated detection and identification frameworks can help to provide reliable results over relevant spatiotemporal scales and help researchers and decision-makers meet their conservation goals. Expert annotators provided annotations for 40 soundscape recordings of 10-minute duration collected at various recordings sites in Sep-Nov 2019. Additionally, ten fully annotated recordings from Costa Rica were provided to participants as training or validation data. Soundscapes from Colombian recording locations were exclusively part of the hidden test data. In contrast to many

[1] https://www.birds.cornell.edu/ccb/swiftone/

(a) COL recording habitat

(b) COR recording habitat

(c) SNE recording habitat



(d) SSW recording habitat

Figure 2: Test data recording locations. ARU were used to collect audio data of targeted ecosystems at large spatial scales. Photos: Fernando Cediel, Alejandro Quesada, Connor Wood, Brian Maley.

other tropical recordings sites, these soundscapes did not contain a very high vocal diversity (due to the proximity to farmland). However, some rare species, for which only very few training examples were available, were present in the data. Therefore, the data from these two recording sites could be considered the most challenging of the competition.

2.2.4. Sierra Nevada (SNE)

Measuring the effects of landscape management activities in the Sierra Nevada, California, USA can reveal a potential correlation with avian population density and diversity. Passive acoustic monitoring can help to reduce the costs of observational studies and expand the scale at which these studies can be conducted, provided there are robust bird call recognition systems [6]. For this dataset, passive acoustic surveys were conducted in the Lassen and Plumas National Forests in May-August 2018. Survey grid cells (4 km^2) were randomly selected from a 6,000-km^2 area, and recording units were deployed at acoustically advantageous locations (e.g., ridges rather than gullies) within those cells. The recordings were made from 04:00 to 08:00 for 5–7 days between May 9 and June 10 (sunrise was roughly 05:35–05:50 during that time) [7].

Because of this, call density was particularly high in this dataset - most soundscapes reflected the species diversity during the dawn chorus. We randomly selected 20 expertly annotated 10-minute soundscape recordings, which were exclusively part of the hidden test data. Although sufficient amounts of training data were available for most annotated species, the high number of overlapping sounds posed a significant challenge.

2.2.5. Sapsucker Woods (SSW)

As part of the Sapsucker Woods Acoustic Monitoring Project (SWAMP), the K. Lisa Yang Center for Conservation Bioacoustics at the Cornell Lab of Ornithology deployed 30 SWIFT recorders in the surrounding bird sanctuary area in Ithaca, NY, USA. This ongoing study aims to investigate the vocal activity patterns and diversity of local bird species. The data are also used to assess the impact of noise pollution on the behavior of birds. In 2018, expert birders annotated 20 full days of audio data recorded between January and June 2017 and provided almost 80,000 labels across randomly selected recordings. The 2019 edition of BirdCLEF [8] used twelve of these days as test data and three as validation data. This year, the amount of test data were limited to twenty 10-minute recordings, including previously unreleased data from this deployment. This reduction became necessary to balance the test data and to reduce the bias towards a specific dataset. Additionally, ten randomly selected recordings were provided as validation data to allow participants to fine-tune their frameworks.

3. Results

1,004 participants from 70 countries on 816 teams entered the BirdCLEF 2021 competition and submitted a total of 9,307 runs. Figure 3 illustrates the performance achieved by the top 50 collected runs. The private leaderboard score is the primary metric and was computed on roughly 65% of the test data (based on a random split). It was revealed to participants after the submission deadline to avoid probing the hidden test data. Public leaderboard scores were visible to participants over the course of the entire challenge and were determined on 35% of the entire test data.

The baseline F1 score in this year's edition was 0.4799 (public 0.5467), with all segments marked as non-events (i.e., *nocall*), and 686 teams managed to score above this threshold. The best submission achieved an F1 score of 0.6932 (public 0.7736), and the top 10 best-performing systems were within only 2% difference in score. Top-scoring participants were required to publish their code and associated write-up, lower-ranked participants opted to do so as well, which resulted in a vast collection of publicly available online resources. It also allowed organizers to inspect frameworks and approaches to assess the current state-of-the-art in this domain. Unsurprisingly, deep convolutional neural networks were the go-to tool in this competition, similar to previous editions. In many cases, participants chose to use off-the-shelve architectures pre-trained on ImageNet (like EfficientNet [9], DenseNet [10], or ResNet [11]). The vast majority of systems used mel scale spectrograms as input data, applied mixup [12] and specaugment [13] to diversify the training data. Provided metadata like time and location of training recordings were used to estimate the occurrence probability of individual bird species to post-filter predictions in many submissions.

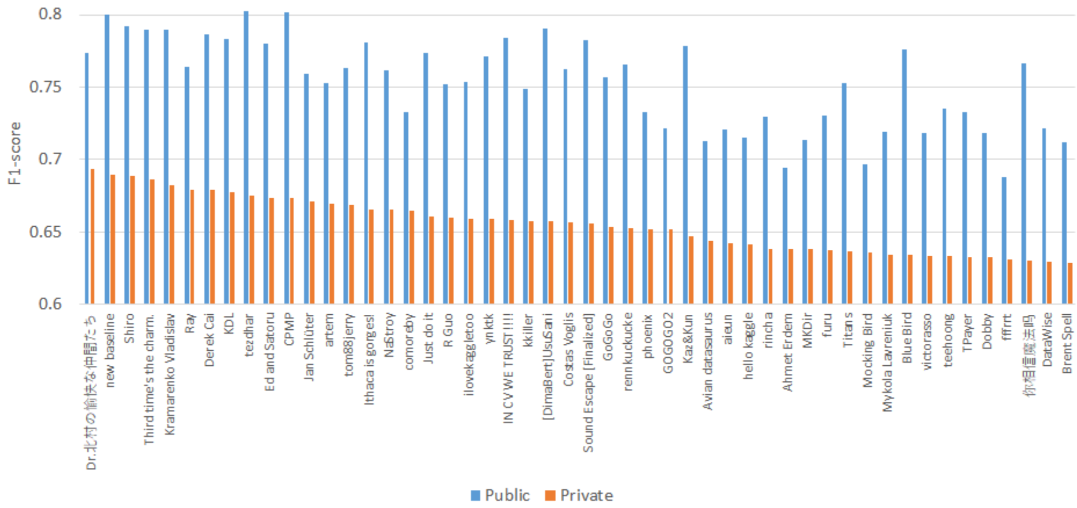


Figure 3: Scores achieved by the best systems evaluated within the bird identification task of LifeCLEF 2021. Public and private test data were split randomly, private scores remained hidden until the submission deadline. Participants were able to optimize the recognition performance of their systems based on public scores, which might explain some differences in scores.

In addition to code repositories and online write-ups, eight teams also submitted full working notes, which are summarized below:

Murakami, Tanaka & Nishimori [14], Team Dr.北村の愉快な仲間たち: This team factorized the problem into three tasks: *Nocall* detection, bird call classification, and post-processing based on provided metadata. The detection and classification backbone was a ResNet50, which used pre-computed mel scale spectrograms as inputs. This team employed a sophisticated scheme of post-processing, using gradient boosting decision trees to eliminate false detections. The overall approach is computationally very efficient and required only few resources for training and inference. The final submission achieved an F1 score of 0.6932 (public 0.7736).

Henkel, Pfeiffer & Singer [15], Team new baseline: Off-the-shelve model architectures pre-trained on ImageNet worked well in this competition. This team used an ensemble of 9 different pre-trained CNN architectures which used 30-second mel scale spectrograms as input. Most notably, this team used a sample mixup scheme which diversified the training data within and across samples by using non-event samples from previous editions and the freefield1010 dataset [16]. Context windows were split into 5-second segments for inference. The final submission achieved an F1 score of 0.6893 (public 0.7998).

Conde, et al. [17], Team Ed and Satoru: This team's solution also relied on the performance of pre-trained models. In this case, the backbone consisted of a ResNeSt which used spectrograms as input. During post-processing, the rolling mean of model confidence scores and clip wise confidence scores were used to eliminate false positives. The final submission achieved an F1 score of 0.6738 (public 0.7801) and consisted of 13 different models, including best performing models from the 2020 Kaggle bird call recognition competition.

Puget [18], Team CPMP: Transformers are the go-to model architecture for text processing. Only recently, vision transformers achieved state-of-the-art results on ImageNet [19] and even for acoustic event recognition [20]. This team tried to adapt vision transformers to the task of bird call recognition and achieved very strong results without the need for large CNN model ensembles. Again, mel scale spectrograms were used as input data, however, patch extraction which accounted for the sequential nature of acoustic data allowed the use of pre-trained transformer models despite visually distorted input data. The final input consisted of a 256x576 pixel spectrogram in which each of the 576 time steps contains 16x16 pixels. This way the entire spectrogram can be reshaped to 24x24 patches of size 16x16 - the input size of pre-trained vision transformers - while still exploiting the sequential structure of an audio signal. The best performing submission achieved an F1 score of 0.6736 (public 0.8015) with the best performing single transformer model achieving an F1 score of 0.6667 (public 0.7569).

Schlüter [21], Team Jan Schlüter: This team used random 30-second crops from each training recording with binary labels for primary and secondary species to train an 18 model ensemble of CNNs. Notably, among mixup as a basic augmentation method, other strategies such as magnitude warping and linear fading of high frequencies were used to emulate variations seen in soundscape recordings. Predictions were made per file by pooling scores over consecutive windows. These per-file predictions were then used to post-filter predictions for 5-second segments. Using additional metadata such as location and time did not help to improve the results. The best submission achieved an F1 score of 0.6715 (public 0.7595).

Shugaev, et al. [22], Team Just do it: Strong *nocall* detection performance appeared to have significant impact on the overall score in this year's competition. This team manually labeled non-event segments in the training data and used these segments to train a binary bird/no bird detection system. Additionally, the *nocall* probability is used to weight weakly labeled training samples during the main model training. This team explored different combinations of spectrogram window length and threshold tuning to improve scores. The best submission achieved an F1 score of 0.6605 (public 0.7736).

Das & Aggarwal [23], Team Error_404: This team designed custom convolutional model architectures which used raw audio samples as 1D inputs. In addition, an elaborated scheme of different attention mechanisms was employed. The two-step recognition process consisted of a binary bird/no bird detector and a species classification model. Specaugment and Mixup were used to diversify the training data, and the final submission achieved an F1 score of 0.6179 (public 0.6878) through the combination of 1D and 2D convolutional classifiers.

Sampathkumar & Kowerko [24], Team Arunodhayan: Domain-specific augmentation appeared to be key to improve the overall recognition performance. This team focused on data augmentation, exploring the impact of different methods on the classification accuracy. During local evaluation, this team was able to improve their baseline F1 score of 0.58 to 0.64 by adding background samples comprised of a variety of non-events from different data sources. This team also explored different schemes of weighting ensemble predictions, the best ensemble consisted of 9 models with ResNet and DenseNet backbones, and achieved an F1 score of 0.5890 (public 0.6799).

3.1. Per-Species Analysis

Because the test-sets and labels were hidden from competitors, their approaches were mostly blind to the test set composition. While this was by design (to encourage the creation of strong general-purpose classifiers, avoiding overfitting to dataset-specific priors), it does inhibit the kind of iterative problem-solving which is usually available in a research context.

A correct identification requires correctly classifying whether a segment contains a bird, assigning a logit to the presence of each particular species, and then determining from the set of logits which birds are actually present. For the micro-averaged F1 score, the number of "*nocall*" segments in the soundscapes dominates the number of segments containing any particular species, so any solution's performance on the *nocall* label has a very large impact on the model's success in the competition.

Taking the top submissions, per-species F1 scores can be computed for each submission. All species with less than five observations in the test were discarded, to reduce variance; this leaves 92 species. These per-species F1 scores over the top 15 submissions were then aggregated. A histogram of the max and mean per-species F1 scores is given in Figure 4.

Had the metric been computed only on segments with birds (removing the no-bird classification problem), the top submissions would have ranked very differently: The eight-place submission (Team KDL[2]) had the best average F1 over species. We acknowledge, though, that changing the metric would have also changed team's tuning strategies, limiting the usefulness of this contra-positive scenario.

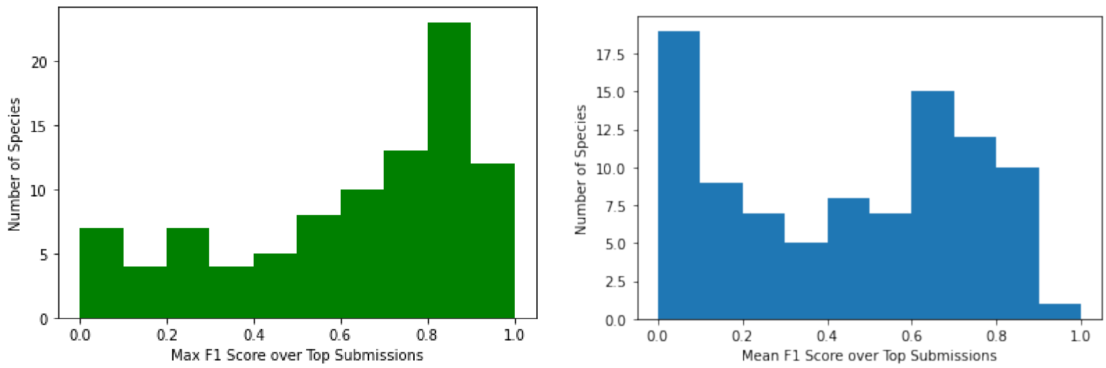
The per-species F1 scores was highly dependent on a submission's (hidden) choice of thresholds. As a result, it was hard to compare performance of particular models on a given species: A small change of threshold could have a large impact on the F1 score. However, we examined the species for which the top submissions were uniformly poor.

The Fox Sparrow (*Passerella iliaca*, *foxspa*) and Green-Tailed Towhee (*Pipilo chlorurus*, *gnttow*) are an interesting pair. Both had very low mean and max F1 scores across all submissions. Their complex songs are well-known to be easily confused by human listeners. The Fox Sparrow has a mean F1 score of 0.01; all but one competitor scored 0. Meanwhile, for *gnttow* the scores are a bit better, with mean F1 0.17 and max F1 0.48. In addition to the complex song, the *gnttow* has a diagnostic call which is easily identifiable.

This indicates that models could improve significantly if they find some way to better distinguish easily-confused species. There are a range of reasons to believe that this is possible. Some easily-confused species can be distinguished by human experts. Secondly, the birds themselves are likely able to distinguish their own species from other species [25]. Thirdly, birds have superior temporal integration for consecutive tones of different frequency [26]. This means that there could be fine structure in these songs which a high-resolution ML algorithm may be able to distinguish. Finally, in some cases, hard-to-distinguish species have non-overlapping geographical distribution ranges. Inclusion of species-specific metadata in decision-making can help in these cases.

The second class of common failure was on birds with very few training resources, especially in the tropical regions, where coverage in Xeno-canto is less comprehensive. The Steely-Vented Hummingbird (*Amazilia saucerottei*, *stvhum2*) is a good example: The max (and mean) F1 score

[2] https://www.kaggle.com/hidehisaarai1213/birdclef2021-infer-between-chunk



(a) (b)

Figure 4: Aggregated per-species max (a) and mean (b) F1 scores over the best submissions from the top 15 competitors.

for *stvhum2* was 0.0 over the 15 submissions examined. This species has just 8 recordings in the training data, most of which have low user ratings and a lot of background noise. The few high quality recordings also demonstrate a fair amount of variability, though the recorded vocalizations do have a recognizable 'hummingbird' timbre.

This indicates that improved few-shot learning may help with identification tasks, especially for identification in geo-regions systemically lacking in training data. Improved few-shot learning models can help with species where there are few training examples, but could also help in cases where a species has an extremely variable song structure.

3.2. Per-Location Analysis

Recording equipment and annotation scheme were identical for all four recording sites. Because of this, variation in recognition performance based on location-specific differences could be explored. Figure 5 shows average scores achieved by the top-15 participants across all recording locations. Despite the uniform recording and label quality of all four datasets, significant differences in achieved scores were observed. When considering all ground truth annotations (incl. *nocall*), submitted systems performed best on the SSW data. The SSW test data had the highest number of *nocall* annotations (i.e., the highest number of 5-second segments without bird vocalization) and by far the largest amount of focal training data across all audible bird species. The performance of an automated distinction between bird calls and non-events (i.e., *nocall*-detector) can be considered one of the primary reason for the drop of 0.182 in F1 score when only segments with a bird call were considered.

This drop in scores can be observed for all locations. However, call density does not seem to affect recognition performance as strongly as other factors. The SNE test dataset almost entirely consisted of dawn chorus recordings with the highest call density of all four sites (1.19 calls per 5-second segment compared to 0.66 for COL, 0.5 for COR, and 0.52 for SSW). Yet, performance across all segments that contained a bird was still very strong, with almost no drop in precision. It appears that other dataset properties had significantly more impact on

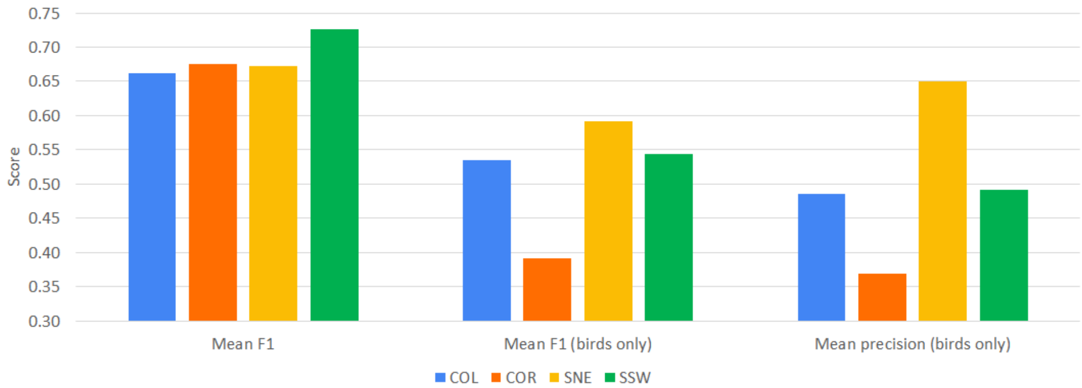


Figure 5: Per-location scores for top-15 participants. High species diversity and lack of focal training data (COR) appeared to have significantly more impact on the overall performance than call density (SNE).

the overall recognition performance. Differences in scores were largest for the COR dataset, which had the highest species diversity of all four sites. Additionally, the ground truth for this site contained 6 species with less than 25 training samples each. These 6 species accounted for 15% of all annotations, and we can assume that the combination of species diversity and lack of training data significantly impacts the overall performance. The availability of target domain (soundscape) training data does not seem to help to overcome the lack of focal training data. The COL test data had the least amount of focal training samples across all audible bird species, and the training data did not contain soundscape recordings from this site. However, performance was significantly higher (+0.144 in F1 score) compared to the COR data for which soundscape training data were available.

4. Conclusions and Lessons Learned

The 2021 BirdCLEF competition held on Kaggle featured a vast collection of training and test audio data. Participants were asked to develop robust bird call recognition frameworks to identify avian vocalizations within 5-second segments of soundscape recordings. The Xeno-canto community provided the primary training data. The test datasets were collected during passive acoustic surveys in North and South America. Species diversity, variability in call densities, and lack of training data for rare species posed significant challenges. Deep artificial neural networks which used spectrogram as input data were used across the board and provided remarkable results despite the domain gap between training and test data. Post-processing of detections and the use of additional metadata were key to achieve top results. However, the overall impact of metadata (e.g., location and time) was only incremental and significantly lower than expected. It appears that these data may be more useful in scenarios with significantly higher species diversity. Additionally, the competition setup encouraged the use of large model ensembles, which might not have real-world applicability.

Despite the high vocal activity in some test recordings, segments without audible bird vocalizations dominated the count. Because of this, threshold tuning (especially for '*nocall*' segments) had a significant impact, often masking the real performance of the algorithms. As a result, many participants relied on separate '*nocall*' detection systems to improve the overall score. Additionally, in this year's edition, off-the-shelve CNN backbones pre-trained on ImageNet provided strong results without the need to investigate the design of domain-specific architectures further. Hence, only very few participants explored alternative approaches like transformers or 1D convolutional networks. We will try to address this in upcoming editions.

Providing introductory code repositories and write-ups greatly improved participation and encouraged fast workflow development without the need for domain knowledge. We noticed that this year's participants quickly adapted to the core challenges of the competition and greatly appreciated the code notebooks provided by the organizers. In addition, prize money for highest scoring solutions, gamification elements on Kaggle, and the overall outreach of the platform had a significant impact on participation and helped attract a broader audience.

Acknowledgments

Compiling these extensive datasets was a major undertaking, and we are very thankful to the many domain experts who helped to collect and manually annotate the data for this competition. Specifically, we would like to thank (institutions and individual contributors in alphabetic order): Center for Avian Population Studies at the Cornell Lab of Ornithology (José Castaño, Fernando Cediel, Jean-Yves Duriaux, Viviana Ruiz-Gutiérrez, Álvaro Vega-Hidalgo, Ingrid Molina, and Alejandro Quesada), Google Bioacoustics Group (Julie Cattiau), K. Lisa Yang Center for Conservation Bioacoustics at the Cornell Lab of Ornithology (Russ Charif, Rob Koch, Jim Lowe, Ashik Rahaman, Yu Shiu, and Laurel Symes), Macaulay Library at the Cornell Lab of Ornithology (Jessie Barry, Sarah Dzielski, Cullen Hanks, Jay McGowan, and Matt Young), Nespresso AAA Sustainable Quality Program, Peery Lab at the University of Wisconsin, Madison (Phil Chaon, Michaela Gustafson, M. Zach Peery, and Connor Wood), and the outstanding Xeno-canto community.

We would also like to thank Kaggle for helping us host this competition and sponsoring the prize money. We are especially grateful for the incredible support and efforts of Addison Howard and Sohier Dane, who helped process the dataset and set up the competition website. Thanks to everyone who participated in this contest and shared their code base and write-ups with the Kaggle community.

All results, code notebooks and forum posts are publicly available at:
https://www.kaggle.com/c/birdclef-2021

References

[1] A. Joly, H. Goëau, S. Kahl, L. Picek, T. Lorieul, E. Cole, B. Deneu, M. Servajean, R. Ruiz De Castañeda, I. Bolon, H. Glotin, R. Planqué, W.-P. Vellinga, A. Dorso, H. Klinck, T. Denton, I. Eggel, P. Bonnet, H. Müller, Overview of LifeCLEF 2021: a System-oriented Evaluation of Automated Species Identification and Species Distribution Prediction, in: Proceedings of the Twelfth International Conference of the CLEF Association (CLEF 2021), 2021.

[2] S. Kahl, M. Clapp, W. Hopping, H. Goëau, H. Glotin, R. Planqué, W.-P. Vellinga, A. Joly, Overview of BirdCLEF 2020: Bird sound recognition in complex acoustic environments, in: CLEF task overview 2020, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2020, Thessaloniki, Greece., 2020.

[3] S. Kahl, C. M. Wood, M. Eibl, H. Klinck, BirdNET: A deep learning solution for avian diversity monitoring, Ecological Informatics 61 (2021) 101236.

[4] Y. Shiu, K. Palmer, M. A. Roch, E. Fleishman, X. Liu, E.-M. Nosal, T. Helble, D. Cholewiak, D. Gillespie, H. Klinck, Deep neural networks for automated detection of marine mammal species, Scientific reports 10 (2020) 1–12.

[5] V. Ruíz, J. D. Román, J.-Y. Duriaux, The Sounds of Sustainability, 2021.

[6] C. M. Wood, V. D. Popescu, H. Klinck, J. J. Keane, R. Gutiérrez, S. C. Sawyer, M. Z. Peery, Detecting small changes in populations at landscape scales: A bioacoustic site-occupancy framework, Ecological Indicators 98 (2019) 492–507.

[7] C. M. Wood, S. Kahl, P. Chaon, M. Z. Peery, H. Klinck, Survey coverage, recording duration and community composition affect observed species richness in passive acoustic surveys, Methods in Ecology and Evolution 12 (2021) 885–896.

[8] S. Kahl, F.-R. Stöter, H. Glotin, R. Planqué, W.-P. Vellinga, A. Joly, Overview of birdclef 2019: Large-scale bird recognition in soundscapes, in: CLEF working notes 2019, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2019, Lugano, Switzerland., 2019.

[9] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.

[10] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[11] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[12] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, arXiv preprint arXiv:1710.09412 (2017).

[13] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le, Specaugment: A simple data augmentation method for automatic speech recognition, arXiv preprint arXiv:1904.08779 (2019).

[14] N. Murakami, H. Tanaka, M. Nishimori, Birdcall Identification using CNN and Gradient Boosting Decision Trees with Weak and Noisy Supervision, in: CLEF Working Notes 2021, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2021, Bucharest, Romania, 2021.

[15] C. Henkel, P. Pfeiffer, P. Singer, Recognizing bird species in diverse soundscapes under weak supervision, in: CLEF Working Notes 2021, CLEF: Conference and Labs of the

Evaluation Forum, Sep. 2021, Bucharest, Romania, 2021.
[16] D. Stowell, M. D. Plumbley, An open dataset for research on audio field recording archives: freefield1010, arXiv preprint arXiv:1309.5275 (2013).
[17] M. V. Conde, N. D. Movva, P. Agnihotri, S. Bessenyei, K. Shubham, Weakly-Supervised Classification and Detection of Bird Sounds in the Wild. A BirdCLEF 2021 Solution, in: CLEF Working Notes 2021, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2021, Bucharest, Romania, 2021.
[18] J.-F. Puget, STFT Transformers for Bird Song Recognition, in: CLEF Working Notes 2021, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2021, Bucharest, Romania, 2021.
[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
[20] Y. Gong, Y.-A. Chung, J. Glass, Ast: Audio spectrogram transformer, arXiv preprint arXiv:2104.01778 (2021).
[21] J. Schlüter, Learning to monitor birdcalls from weakly-labeled focused recordings, in: CLEF Working Notes 2021, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2021, Bucharest, Romania, 2021.
[22] M. Shugaev, N. Tanahashi, P. Dhingra, U. Patel, BirdCLEF 2021: Building a birdcall segmentation model based on weak labels, in: CLEF Working Notes 2021, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2021, Bucharest, Romania, 2021.
[23] G. Das, S. Aggarwal, Bird-Species Audio Identification, Ensembling 1D + 2D Signals, in: CLEF Working Notes 2021, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2021, Bucharest, Romania, 2021.
[24] A. Sampathkumar, D. Kowerko, TUC Media Computing at BirdCLEF 2021: Noise augmentation strategies in bird sound classification in combination with DenseNets and ResNets, in: CLEF Working Notes 2021, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2021, Bucharest, Romania, 2021.
[25] C. K. Catchpole, P. J. Slater, Bird song: biological themes and variations, Cambridge university press, 2003.
[26] R. J. Dooling, B. Lohr, M. L. Dent, Hearing in birds and reptiles, in: Comparative hearing: birds and reptiles, Springer, 2000, pp. 308–359.