Overview of BirdCLEF 2022: Endangered bird species recognition in soundscape recordings

Stefan Kahl[1], Amanda Navine[2], Tom Denton[3], Holger Klinck[1], Patrick Hart[2], Hervé Glotin[4], Hervé Goëau[5], Willem-Pier Vellinga[6], Robert Planqué[6] and Alexis Joly[7]

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

Abstract

As the "extinction capital of the world", Hawai'i has lost 68% of its native bird species, the consequences of which can harm entire ecosystems. With physical monitoring difficult, scientists have turned to sound recordings, as this approach could provide a passive, low labor, and cost-effective strategy for monitoring endangered bird populations. Current methods for processing large bioacoustic datasets involve manual review of each recording. This requires specialized training and prohibitively large amounts of time. Recent advances in machine learning have made it possible to automatically identify bird songs for common species with ample training data. However, it remains challenging to develop such tools for rare and endangered species. The main goal of the 2022 edition of BirdCLEF was to advance automated detection of rare and endangered bird species that lack large amounts of training data. The competition challenged participants to develop reliable analysis frameworks to detect and identify the vocalizations of rare bird species in continuous Hawaiian soundscapes utilizing limited training data.

Keywords

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

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

stefan.kahl@cornell.edu (S. Kahl); navine@hawaii.edu (A. Navine); tmd@google.com (T. Denton); Holger.Klinck@cornell.edu (H. Klinck); pjhart@hawaii.edu (P. Hart); 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); 0000-0003-4827-461X (A. Navine); 0000-0003-1078-7268 (H. Klinck); 0000-0003-4781-8785 (P. Hart); 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)

CEUR Workshop Proceedings (CEUR-WS.org)

1. Introduction

Passive acoustic monitoring (PAM), using autonomous sound recorders to monitor animals and their habitats at ecologically relevant scales, has become a critical survey tool in conservation [1]. Inexpensive commercial off-the-shelf sound recorders are readily available to the community, making the data collection straightforward. Arrays of sound recorders are often deployed for extended periods (weeks to months) and collect vast amounts of data, providing valuable information on the abundance and distribution of vocalizing animals with high spatio-temporal resolution [2]. However, several challenges with PAM remain. It is not uncommon that data collection efforts result in tens of Terabytes of acoustics data that need to be efficiently handled, stored, and analyzed [3]. Especially the analysis, reliably extracting signals of interest in often complex soundscapes, is an active area of research. In addition, while common species are typically well represented in available training datasets, data for rare, listed, or endangered species is often sparse, requiring the development of new and innovative algorithmic approaches to monitor these species in need.

The Hawaiian Islands were once home to an estimated 152 species of terrestrial bird species, but today, habitat loss, habitat degradation by introduced plants and ungulates, and introduced diseases and predators have left only 45 endemic species remaining, 33 of which are listed as endangered [4, 5, 6]. Despite their plight, island birds often receive less conservation attention and funding than continental species [4]. Further, many of the endangered birds of Hawai‘i only persist at low population densities in remote, high-elevation, and/or densely forested habitats, which are difficult to access and monitor. In light of these obstacles, Hawaiian birds are prime examples of species that could benefit greatly from PAM techniques that can reduce labor demands and cost. However, these species remain challenging to incorporate into automated detection frameworks because they are severely underrepresented in online acoustic data repositories such as Xeno-canto[1].

The BirdCLEF 2022 competition challenged participants to develop reliable analysis frameworks to detect and identify the vocalizations of rare bird species in continuous Hawaiian soundscapes utilizing limited training data.

2. BirdCLEF 2022 Competition Overview

In recent years, research in the domain of bioacoustics shifted towards deep neural networks for sound event recognition [7, 8]. In past editions, we have seen many attempts to utilize convolutional neural network (CNN) classifiers to identify bird calls based on visual representations of these sounds (i.e., spectrograms) [9, 10, 11]. Despite their success for bird sound recognition in focal recordings, the classification performance of CNNs on continuous and omnidirectional soundscape recordings remained low. Passive acoustic monitoring can be a valuable sampling tool for habitat assessments and observations of environmental niches, which often are threatened. However, manual processing of large collections of soundscape data is not desirable, and automated attempts can help to advance this process [12]. Yet, the lack of suitable validation and test data prevented the development of reliable techniques to solve this task.

[1] https://xeno-canto.org

Bridging the acoustic gap between high-quality training recordings and complex soundscapes with varying ambient noise levels is one of the most challenging tasks in the domain of audio event recognition. This is especially true when the amount of training data is insufficient, as is the case for many rare and endangered bird species around the globe. Despite the vast amounts of data collected on Xeno-canto and other online sound libraries, audio data for endangered birds is still sparse. However, those endangered species are most relevant for conservation, rendering acoustic monitoring of endangered birds particularly difficult.

2.1. Goal and Evaluation Protocol

The main goal of the 2022 edition of BirdCLEF was to advance automated detection of rare and endangered bird species that lack large amounts of training data. The competition was hosted on Kaggle[2] to attract machine learning experts from around the world to participate in the challenge. The overall task design was consistent with previous editions, but the focus was shifted towards species with very few training samples. Participants were asked to detect and identify 21 target bird species within the provided soundscape test set. Each soundscape was divided into segments of 5 seconds, and a list of audible species within each segment had to be reported by the participants.

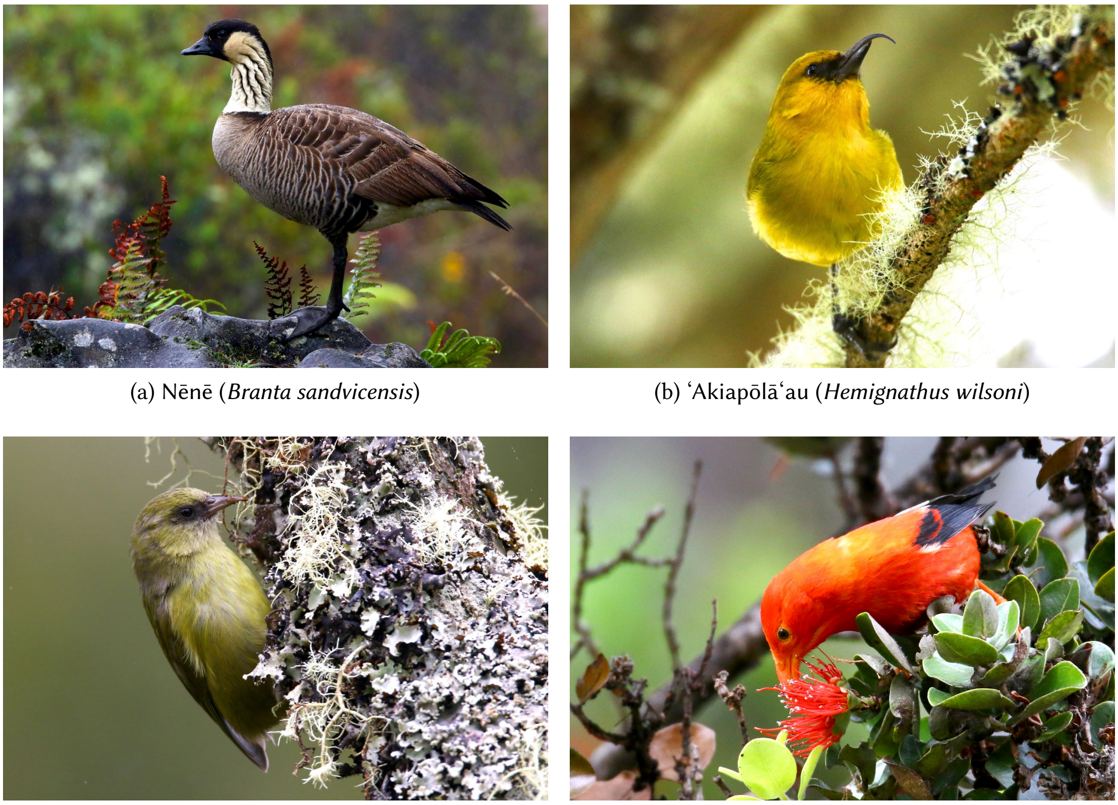
The competition's evaluation metric was a weighted variant of the macro-averaged F1-score, which weighted all target classes equally, thus emphasizing rare acoustic events. In earlier editions, ranking metrics were used to assess the overall classification performance. However, when applying bird call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

To obtain a weighted F1 score, we compute a separate weight for each classification output - i.e., a separate weight for every possible label for every 5-second segment. These weights are chosen such that a) the total weight of positive and negative samples for each species is equal, and b) the total weight for each species is equal.

2.2. Dataset

Current methods for processing large bioacoustic datasets involve manual annotation of each recording. This requires specialized training and prohibitively large amounts of time. Thankfully, recent advances in machine learning have made it possible to automatically identify bird songs for common species with ample training data. However, it remains challenging to develop such tools for rare and endangered species, such as those in Hawaiʻi (Figure 1). Deploying a bird sound recognition system to a new recording and observation site requires classifiers that generalize well across different acoustic domains. Focal recordings of bird species form an excellent base to develop such a detection system. However, the lack of annotated soundscape data for a new deployment site poses a significant challenge.

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



(a) Nēnē (*Branta sandvicensis*)

(b) ʻAkiapōlāʻau (*Hemignathus wilsoni*)

(c) ʻAlawī (*Loxops mana*)

(d) ʻIʻiwi (*Drepanis coccinea*)

Figure 1: Four of the threatened (a, d) and endangered (b, c) Hawaiʻi endemic species featured in BirdCLEF 2022. Photo credit: Ann Tanimoto-Johnson.

2.2.1. Training Data

As in previous editions, training data was provided by the Xeno-canto community and consisted of more than 14,800 recordings covering 152 species. Participants were allowed to use metadata to develop their systems. Most notably, we provided detailed location information on recording sites of focal and soundscape recordings, allowing participants to account for spatio-temporal occurrence patterns of bird species.

2.2.2. Test Data

In this edition of the BirdCLEF competition, the test data consisted of 5,356 one-minute soundscapes (amounting to approximately 90 hours of recordings). This dataset was hidden and only accessible to participants during the inference process. These soundscapes were collected for various research projects by the Listening Observatory for Hawaiian Ecosystems (LOHE) at the University of Hawaiʻi at Hilo at 7 sites across the islands of Hawaiʻi, Maui, and Kauaʻi. Because these data were not collected specifically to be used in the BirdCLEF 2022 contest, different

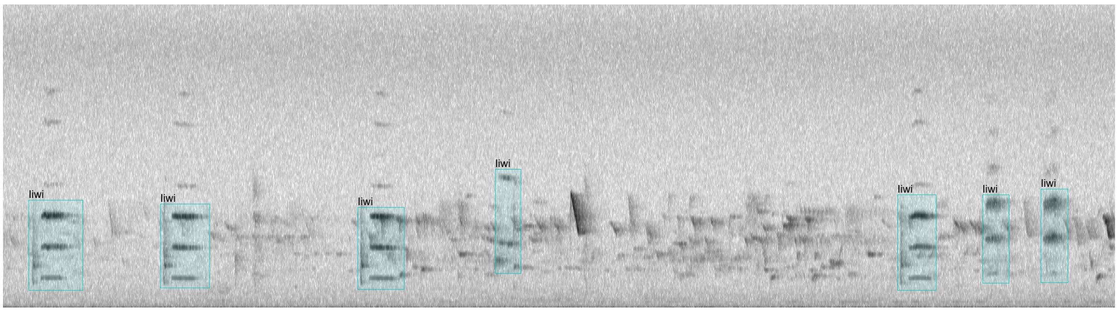


Figure 2: Expert ornithologists provided bounding box labels for all soundscape recordings indicating calling of 21 target species. In this example, all ʻIʻiwi calls were annotated, while vocalizations of other species were not labeled. This labeling scheme was applied to all test data soundscapes.

acoustic recording equipment (Song Meter SM2 or SM4[3]) and recording settings (sampling rate, mono versus stereo microphones, audio gain, etc.) were used to collect the data. The recordings also varied in duration, time of year, and time of day. All soundscapes received some level of manual bird vocalization annotation by trained members of the LOHE lab using Raven Pro 1.5 software[4], however some recordings had a select few target species annotated, while others were annotated for every detectable species (Figure 2). In light of these annotation strategies, only the subset of species for which every vocalization was annotated were scored for any given file, which resulted in a total of 21 scored bird species, 15 species endemic to the Hawaiian Islands and 6 introduced species. This allowed us to include a much larger test dataset in the contest than would have otherwise been possible and exactly replicates real-world use cases where often whatever data is easily accessible is used to answer biological questions.

3. Results

A total of 1,019 participants (801 teams) from 62 countries participated in the BirdCLEF 2022 competition and submitted 23,352 runs. In Figure 3 we report the performance achieved by the top 50 runs. The private leaderboard score is the primary metric and was revealed to participants after the submission deadline to avoid probing the hidden test data. Public leaderboard scores were visible to participants throughout the challenge.

The baseline weighted F1-score in this year's edition was 0.5112 (public 0.4849) with all scored birds marked as silent (False) for all segments, and 665 teams managed to score above this threshold. The best submission achieved a weighted F1-score of 0.8527 (public 0.9128) and the top 10 best performing submissions only differed by 7% in score. Most approaches were based on convolutional neural network ensembles and mainly differed in pre- and post-processing strategies and neural network backbone architectures. Interestingly, few-shot learning techniques were vastly underrepresented despite the fact that some target species only had a handful of training samples. Participants employed various sophisticated post-processing schemes, most

[3] Wildlife Acoustics Inc., Maynard, Massachusetts
[4] The Cornell Lab of Ornithology, Ithaca, New York

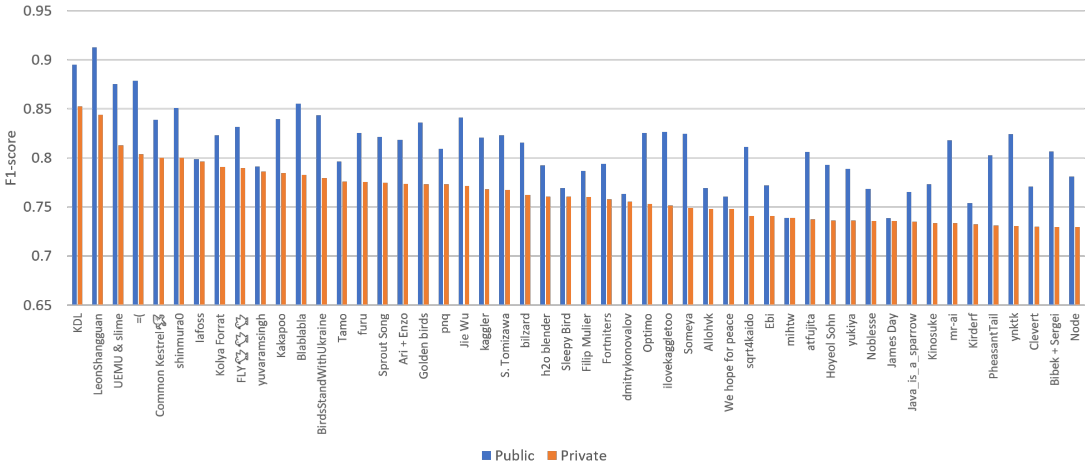


Figure 3: Scores achieved by the best systems evaluated within the primary bird identification task of LifeCLEF 2022. Public and private test data were split randomly, private scores remained hidden until the submission deadline. Participants were able to optimize the recognition performance of their systems based on public scores, which likely explains some differences in scores.

notably a percentile-based thresholding approach established during the 2021 edition [13]. Some participants experimented with different loss functions, focal loss being the most notable.Some teams used audio transformers, but the results were inconsistent and led to discussions about whether these methods were appropriate for the task of bird call identification.

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

Conde & Choi [14]: This team explored a number of CNN backbones and evaluated their performance for this task, finding that EfficientNet architectures work best (i.e., EfficientNet-B0). The overall training process was adapted from previous attempts, predominantly last year's second-ranked solution by Henkel et al. [13]. In addition, the authors experimented with various post-processing steps to improve results. In particular, a penalization which reduces output probabilities proportional to amount of training data was introduced. On top of that, quantile-based class-wise thresholds, tuned by grid search and leaderboard probing, were employed.

Sampathkumar & Kowerko [15]: Data augmentation is an important processing step in bird sound recognition because of the domain shift between training and test recordings. In their work, this team focused on evaluating the best augmentation scheme for this task. Most transformations focus on adding different patterns of noise to the source recording, thus emulating noisy soundscape recordings. While the authors find that all augmentations methods improve the baseline experiment, Gaussian noise, loudness normalization and tanh distortion appear to be most impactful. This team also evaluated different CNN backbones, settling on a EfficientNet-B0 for their final submission.

Martynov & Uematsu [16]: This team also employed a sophisticated augmentation scheme to enhance the overall performance. Methods like mix-up, cut-mix and spec augment were chosen. Backbone architecture and training process are largely modeled after last year's submission by Henkel et al. [13] and adds species based thresholding as post-processing step. In order to overcome limitations of weakly-labeled training data, this team also decided to hand-label some source recordings. The final submission was an ensemble consisting of various CNN trained for different groups of birds.

Krishnan et al. [17]: Despite the dominance of CNN backbones for bird call detection, this team decided to exploit the sequential nature of audio data to train a prediction model based on Long Short-Term Memory (LSTM) units. Combined with a taxonomic sequence prediction task, this attempt consistently outperformed baseline setups based on multi-layer perceptrons. The authors conclude that the "expressive power of the hidden cells combined with the hierarchical set up for the task" can significantly improve performance. This might also apply, when added on top of other models (e.g., feeding logit outputs into LSTM units).

Miyaguchi et al. [18]: Unsupervised representation learning can be key when facing a training dataset with vast class imbalances and underrepresented classes. In this year's competition, some bird species only had a mere two training recordings, forcing participants to develop novel strategies to cope with these limitations. This team decided to explore unsupervised training approaches in combination with species classifiers. Motif mining [19] was used to extract features from training recordings, embeddings were generated by training a Tile2Vec model [20] with triplet loss and multi-layer perceptron as well as decision trees were used as classifiers. The authors note that the overall performance on downstream prediction tasks is still not competitive, but propose changes to the training scheme that need to be explored further.

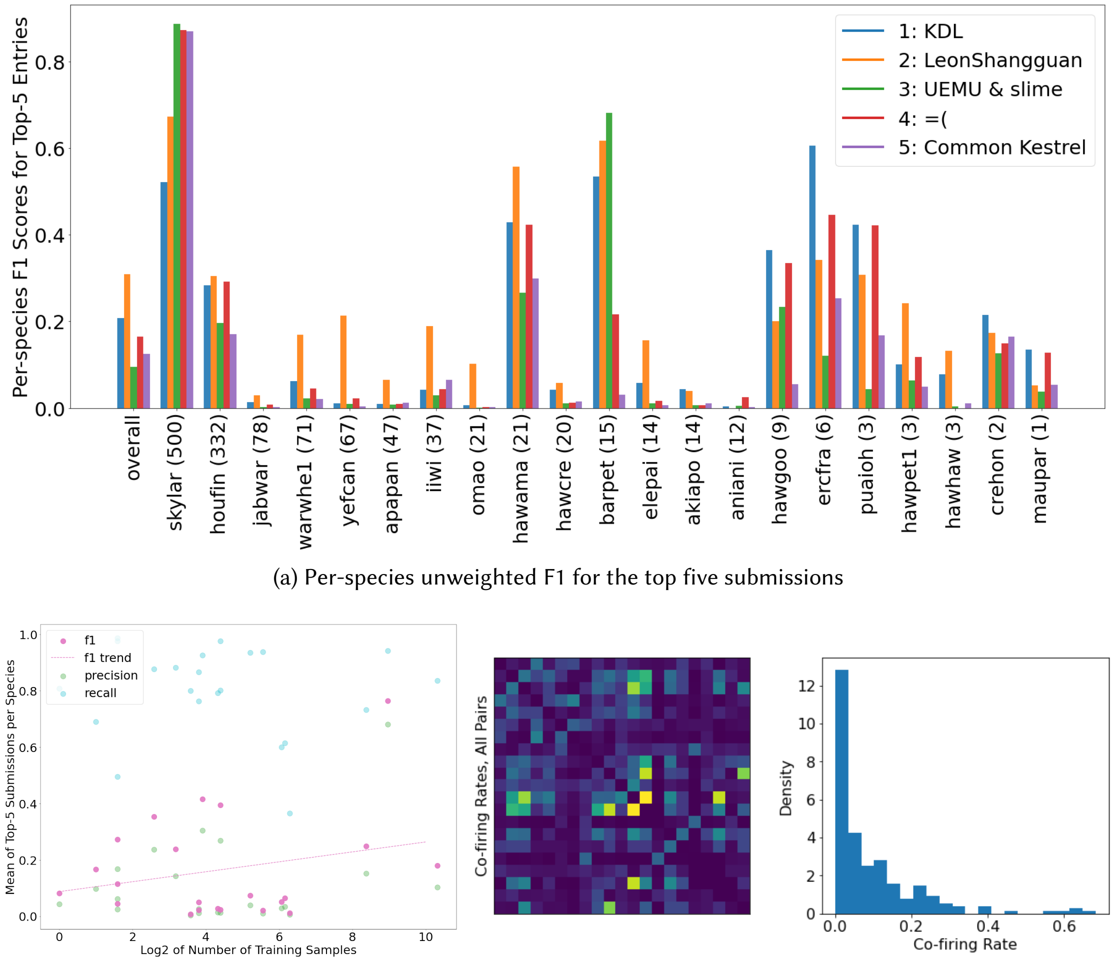
3.1. Per-Species Analysis

There was a wide variation in available Xeno-Canto training data for the various species. In Figure 4 a+b, we show the relationship between quantity of training data and unweighted model performance amongst the top competitors. Notice that quantity of training data does co-vary with model quality, but there are also significant per-species effects on model quality.

The competition metric was a weighted F1, chosen to equalize the importance of the 21 target species, and then within each species to equalize the importance of positive and negative labels. Because negative labels are far more common in the test data than positive labels, this weighting strongly emphasizes recall over precision.

A recent study of Hawaiian honeycreepers found 'cultural convergence' of birdsong in the face of population decline [21]. In this study, three species of honeycreeper were found to have decreased intra-species song variation over a 40-year period, but also decreased inter-species variation amongst the three species studied. This has implications for the acoustic identification task. While decreased intra-species variation should make identification easier, decreased inter-species variation will increase the difficulty of separating similar species.

For downstream users of high-recall models, we would expect conflation of easily confused species. Define the *co-firing rate* of two species as the probability that a binary classifier fires for both species when either species is present. For the top 5 submissions, the average co-firing



(a) Per-species unweighted F1 for the top five submissions

(b) Per-species precison, recall and weighted F1

(c) Co-firing rates

Figure 4: Per-species unweighted F1 for the top five submissions (a). Each species is listed by its eBird code, and the number of Xeno-Canto training recordings is given in parentheses. (b) Per-species precision, recall, and unweighted F1, averaged across the top five submissions. (c) Co-firing rates for all species pairs. The matrix of co-firing rates is on the left, and the density histogram of co-firing rates is on the right.

rate across all pairs of species was 9.3% (Figure 4 c). The species ‘I‘iwi and ‘Apapane are difficult even for humans to distinguish, and the co-firing rate for this pair was the highest amongst all species pairs at 68%. The 95th percentiles for the co-firing rate was 31.6%. The co-firing rates above this threshold all occurred for a cluster of nine native honeycreepers and thrushes.

The preference for high recall also led to over-firing when none of the target species were present. In the test data, about one third of segments contained at least one target species. However, the top five submissions had, on average, at least one label for 94% of segments. This suggests that these species classifiers should be paired with an upstream binary bird detector.

Using these high-recall species classifiers for the identification of Hawaiian species *could* help reduce the amount of data requiring attention, but only when trying to isolate a specific species. This would ultimately still require significant expert labor to separate vocalizations by species, given the high co-firing rates between species with similar vocalizations.

4. Conclusions and Lessons Learned

Despite being set up as a few-shot learning task, few teams decided to employ techniques other than CNNs. Pre-trained neural networks for image recognition still dominated the task, and participants tried to cope with the lack of training data through intensive data augmentation and transfer learning. Surprisingly, there was only a weak correlation between the number of training samples and overall per-species performance. This indicates that other factors - such as repertoire size and call patterns - might outweigh training data quantity. Automatic detection of endangered and rare species remains challenging. Still, this year's competition demonstrated that passive acoustic monitoring combined with machine learning could already be a powerful monitoring tool for some endangered species. BirdCLEF continues to engage a large number of data scientists from around the world to develop new and effective acoustic analysis solutions that aid avian conservation.

Acknowledgments

In particular, we want to thank the members of the Listening Observatory for Hawaiian Ecosystems (LOHE), led by Patrick Hart, who collected and annotated the soundscapes used in the test dataset. Special thanks to LOHE lab manager Ann Tanimoto-Johnson for gathering, organizing, annotating, and proof checking the soundscape data. Ann also provided photographs of the native birds shown on the competition website and in this paper.

We would also like to thank Kaggle for helping us host this competition. 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. Kaggle and Google jointly sponsored the prize money for this competition. 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-2022

References

[1] L. S. M. Sugai, T. S. F. Silva, J. W. Ribeiro Jr, D. Llusia, Terrestrial passive acoustic monitoring: review and perspectives, BioScience 69 (2019) 15–25.

[2] L. S. M. Sugai, C. Desjonqueres, T. S. F. Silva, D. Llusia, A roadmap for survey designs in terrestrial acoustic monitoring, Remote Sensing in Ecology and Conservation 6 (2020) 220–235.

[3] D. Tuia, B. Kellenberger, S. Beery, B. R. Costelloe, S. Zuffi, B. Risse, A. Mathis, M. W. Mathis, F. van Langevelde, T. Burghardt, et al., Perspectives in machine learning for wildlife conservation, Nature communications 13 (2022) 1–15.

[4] T. K. Pratt, C. T. Atkinson, P. C. Banko, B. L. Woodworth, J. D. Jacobi, Conservation Biology of Hawaiian Forest Birds: Implications for Island Avifauna, Yale University Press, 2009.

[5] M. Walther, Extinct Birds of Hawaii, first edition ed., Mutual Publishing, LLC, 2016.

[6] E. H. Paxton, M. Laut, J. P. Vetter, S. J. Kendall, Research and management priorities for hawaiian forest birds, The Condor 120 (2018) 557–565.

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

[8] 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.

[9] J. Schlüter, Bird identification from timestamped, geotagged audio recordings., in: CLEF working notes 2018, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2018, Avignon, France., 2018.

[10] M. Lasseck, Bird species identification in soundscapes., in: CLEF working notes 2019, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2019, Lugano, Switzerland., 2019.

[11] M. Mühling, J. Franz, N. Korfhage, B. Freisleben, Bird species recognition via neural architecture search, in: CLEF Working Notes 2020, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2020, Thessaloniki, Greece., 2020.

[12] 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.

[13] 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.

[14] M. V. Conde, U.-J. Choi, Few-shot Long-Tailed Bird Audio Recognition, in: CLEF Working Notes 2022, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2022, Bologna, Italy, 2022.

[15] A. Sampathkumar, D. Kowerko, TUC Media Computing at BirdCLEF 2022: Strategies in identifying bird sounds in a complex acoustic environment, in: CLEF Working Notes 2022, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2022, Bologna, Italy, 2022.

[16] E. Martynov, Y. Uematsu, Dealing with Class Imbalance in Bird Sound Classification, in: CLEF Working Notes 2022, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2022, Bologna, Italy, 2022.

[17] S. Krishnan, P. Khandelwal, R. Garg, Bird Species Classification: One Step at a Time, in: CLEF Working Notes 2022, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2022, Bologna, Italy, 2022.

[18] A. Miyaguchi, J. Yu, B. Cheungvivatpant, D. Dudley, A. Swain, Motif Mining and Unsupervised Representation Learning for BirdCLEF 2022, in: CLEF Working Notes 2022, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2022, Bologna, Italy, 2022.

[19] D. F. Silva, C.-C. M. Yeh, Y. Zhu, G. E. Batista, E. Keogh, Fast similarity matrix profile for music analysis and exploration, IEEE Transactions on Multimedia 21 (2018) 29–38.

[20] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, S. Ermon, Tile2vec: Unsupervised representation learning for spatially distributed data, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 3967–3974.

[21] K. L. Paxton, E. Sebastián-González, J. M. Hite, L. H. Crampton, D. Kuhn, P. J. Hart, Loss of cultural song diversity and the convergence of songs in a declining hawaiian forest bird community, Royal Society open science 6 (2019) 190719.