Overview of BirdCLEF 2023: Automated Bird Species Identification in Eastern Africa

Stefan Kahl[1,2,*], Tom Denton[3], Holger Klinck[1], Hendrik Reers[4], Francis Cherutich[5], Hervé Glotin[6], Hervé Goëau[7], Willem-Pier Vellinga[8], Robert Planqué[8] and Alexis Joly[9]

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

Abstract

The BirdCLEF 2023 challenge focused on bird species classification in a dataset of Kenyan soundscape recordings. Kenya is home to over 1,000 species of birds, covering a wide range of ecosystems, from the savannahs of the Maasai Mara to the Kakamega rainforest, and even alpine regions on Kilimanjaro and Mount Kenya. Tracking this vast number of species with ML can be challenging, especially with minimal training data available for many species. This year the competition switched back to threshold-free evaluation metric, and introduced a two-hour time limit on inference to ensure the practical usability of models.

Keywords

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

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
*Corresponding author.
stefan.kahl@cornell.edu (S. Kahl); tmd@google.com (T. Denton); holger.klinck@cornell.edu (H. Klinck); reers@oekofor.de (H. Reers); 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)

1. Introduction

Passive acoustic monitoring (PAM), utilizing autonomous sound recorders to observe animals and their habitats at ecologically relevant scales, has emerged as an indispensable survey method in conservation [1]. The accessibility of inexpensive commercial off-the-shelf sound recorders has made data collection a straightforward process for the community. Arrays of sound recorders are frequently deployed over extended periods (weeks to months), generating large amounts of data that offer valuable insights into the abundance and distribution of vocalizing animals with high spatiotemporal resolution [2]. Nevertheless, several challenges persist in PAM. It is not uncommon for data collection endeavors to produce tens of Terabytes of acoustic data that must be efficiently managed, stored, and analyzed [3]. Particularly, the analysis task, which involves reliably extracting relevant signals from often intricate soundscapes, remains an active area of research. Furthermore, while common species typically have ample representation in existing training datasets, data for rare, listed, or endangered species is often limited, necessitating the development of novel and innovative algorithmic approaches to monitor these species in need.

Eastern African species play a crucial role in both ecological systems and evolutionary processes, making them of utmost importance for scientific research and conservation efforts [4, 5]. This region is renowned for its exceptional biodiversity, particularly in terms of avian species, which exhibit a remarkable diversity of vocalizations [6, 7]. The vocal signals produced by Eastern African species are integral to their communication and social interactions and serve as valuable indicators of their presence and behavior. However, despite their significance, Eastern African bird species are often underrepresented in sound collections like Xeno-canto[1] or Macaulay Library[2]. This lack of comprehensive acoustic data poses challenges for developing and applying machine learning algorithms aimed at monitoring and studying these species. Insufficient training data for these algorithms hinders their ability to detect and classify the vocalizations of Eastern African species accurately, impeding advancements in automated monitoring techniques. Consequently, addressing this issue and raising awareness is crucial for facilitating the development of robust machine-learning algorithms and enabling effective monitoring and conservation strategies for these important and underrepresented species.

The *Bird Recognition Challenge* (BirdCLEF) is part of LifeCLEF 2023 [8] and 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 competitions in terms of dataset size and species diversity, with several tens of thousands of recordings covering up to 1,500 species [9, 10]. The BirdCLEF 2023 competition challenged participants to develop reliable analysis frameworks to detect and identify the vocalizations of bird species in continuous Eastern African soundscapes, coping with limited training data for many species.

[1] https://xeno-canto.org
[2] https://www.macaulaylibrary.org

2. BirdCLEF 2023 Competition Overview

Recent advancements in the development of machine listening techniques for the identification of animal vocalizations have enhanced our ability to analyze long-term acoustic datasets comprehensively [11, 12]. However, generating analysis outputs with high precision and recall, particularly when targeting a large number of species simultaneously, remains challenging. Bridging the gap between high-quality training samples (focal recordings) and noisy test samples (soundscape recordings) is a difficult task in the field of acoustic event detection and classification. The 2023 BirdCLEF competition addressed this intricate challenge and was hosted on Kaggle[3]. This year's edition focused on identifying which birds are calling in long recordings made in Kenya. The competition was held in a "code competition" format, encouraging participants to share their code for the benefit of the community, especially scientists and practitioners who monitor bird populations for conservation purposes in Kenya. In addition, submissions were required to complete inference in less than 2 hours. We implemented this time constraint to ensure that the developed models can run efficiently on modest compute resources available to conservationists.

2.1. Goal and Evaluation Protocol

This year's competition featured two major changes compared to the previous few years: A new metric was used for evaluation (padded class-averaged mean-average precision, or pcmAP) and a time-limit of two CPU hours was placed on inference.

2.2. Metric

The cmAP metric was used in BirdCLEF competitions before the move to Kaggle, which previous to this year could not support cmAP. As a result, the competition in 2020, 2021, 2022, have used variants of F1 score. The downside to F1 is that it requires choosing a binary label for species in each inference window. There are numerous ways to reduce a model's output probabilities to a binary decision, which has led to substantial effort in recent years to find clever threshold selection techniques. Unfortunately, this makes it hard to evaluate the base model quality. In practice, correct selection of thresholds depends on the goals of the end-user (eg, according to preference in trading off precision and recall), so there is some preference for threshold-free evaluation of model quality.

This year, Kaggle added support for custom metrics, which allowed the competition to use the cmAP score, which is defined as:

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

where C is the number of target classes, and $AveP(c)$ is the average precision for the cth species, computed as:

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

[3] https://www.kaggle.com/c/birdclef-2023

where k is the rank of an item in the list of examples containing class c, $P(k)$ is the precision at cut-off k in the list, $rel(k)$ is an indicator function indicating whether class c is present in the kth example, and $n_{rel}(c)$ is the total number of examples containing class c.

cmAP computes the per-class mean average precision (treating each model output as an independent binary classifier), and then averages over all classes. An advantage of cmAP is that all species are weighted equally, regardless of the number of examples in the inference set. A disadvantage is that for species with very few examples in the dataset, the individual species MAP can be quite noisy, which in turn is reflected in the overall cmAP score.

To deal with this, we proposed a modification of $cmAP$ metric called *padded cmAP*, or *pcmAP*, in which p 'free' examples are added to the top of each class. This limits the dynamic range of the per-species MAP scores, reducing the impact of noise in species with very few labels. For the competition, we used $p = 5$.

2.3. Time Limits

Competitors were limited to two hours of inference time on CPU. This ensures that models are cost-effective for real-world usage. A side effect is reducing the impact of *ensembling*, a common Kaggle tactic which also obscures underlying model quality.

2.4. Dataset

2.4.1. Training Data

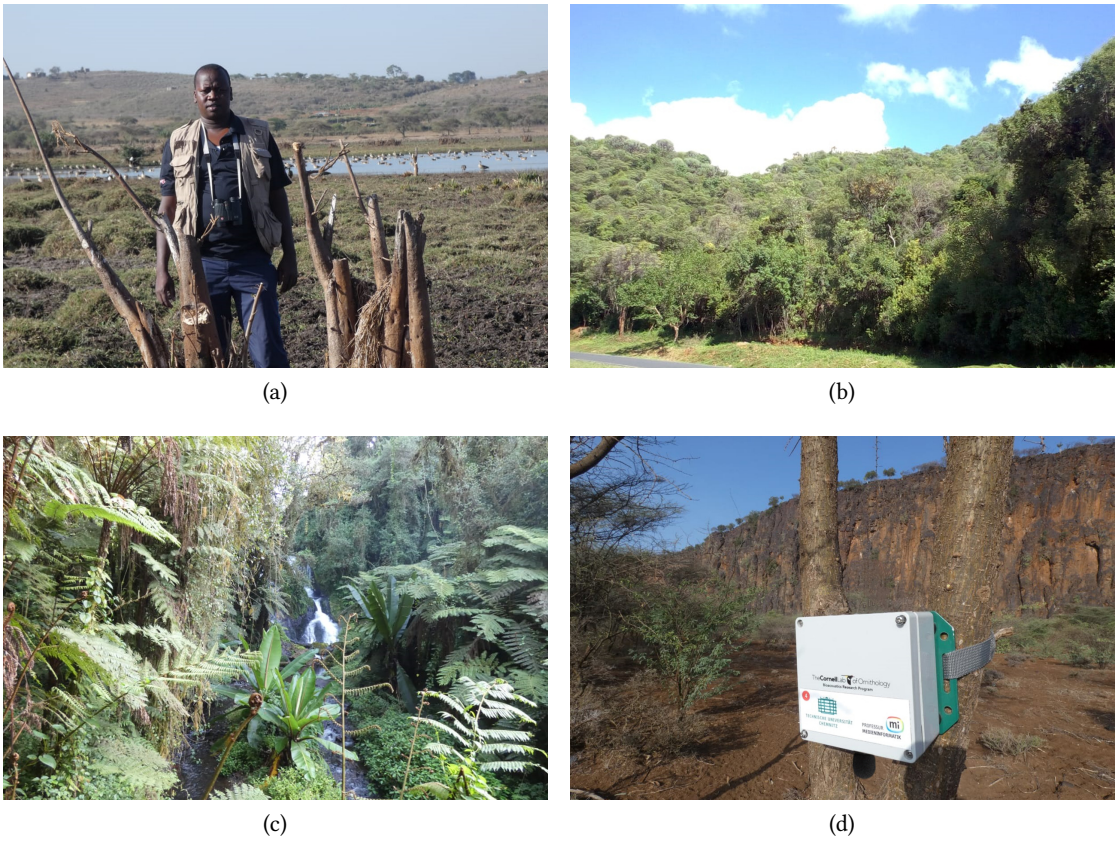
As in previous editions, training data was provided by the Xeno-Canto community and consisted of more than 16,900 recordings covering 264 species. Participants were allowed to use metadata to develop their systems, and were also allowed to gather more recordings from Xeno-Canto. Most notably, we provided detailed information on where and when focal and soundscape recordings were made, allowing participants to account for spatiotemporal occurrence patterns of bird species.

2.4.2. Test Data

As in previous years on Kaggle, the test data was completely hidden from participants. Hidden test data consisted of 191 soundscapes of 10-minute duration and were recorded at multiple locations west and southwest of Lake Baringo in Baringo County, Kenya (see figure 1). Soundscapes were expertly annotated by local expert Francis Cherutich who provided 10,294 labels for 176 species.

3. Results

This year, we had 1,189 teams and nearly 1,400 competitors, and 21,519 total submissions. As in other recent years, two-thirds of the test data was used for the private leaderboard, and one third for the public leaderboard. The padded cmAP metric yielded a fairly high baseline score. A baseline using BirdNET 2.2 with no modifications gave a score of 0.771 on the public board and 0.664 on the private leaderboard. The overall winner achieved a public score of 0.844 and a



(a) (b)

(c) (d)

Figure 1: Pictures illustrating the different habitat types in Kenya where the soundscape recordings were collected. Photos Francis Cherutich.

private score of 0.764. There was relatively little 'shake-up' in the results — the top public entry dropped to third, and otherwise the top five maintained their order on the private leaderboard. This stability may be due to better cross-validation practices in recent competitions.

A few common themes emerged in the top solutions.

First, the top competitors went out of their way to obtain more data and ensure that the data they were working with was clean. The pre-packaged Xeno-Canto data was limited to 500 samples per species to keep file sizes manageable; multiple top competitors went back to XC and downloaded the additional data. The second-place competitor also noticed some inconsistencies in converting the XC scientific names to the ebird codes used in previous-year test data; this may be due to subsequent changes in both the ebird and IOC taxonomy used by Xeno-Canto, but in any case reinforces the need for careful data handling.

Second, the top competitors took advantage of model optimization tools, managing to make ensembles of up to seven(!!) models run in the specified time limit. ONNX, a platform-independent model format, was particularly popular. A few competitors also saw large speed gains with OpenVINO, an Intel-specific model optimizer and inference library. Particularly creative: The third-place competitor (long-time BirdCLEF participant Mario Lasseck) built a large ensemble

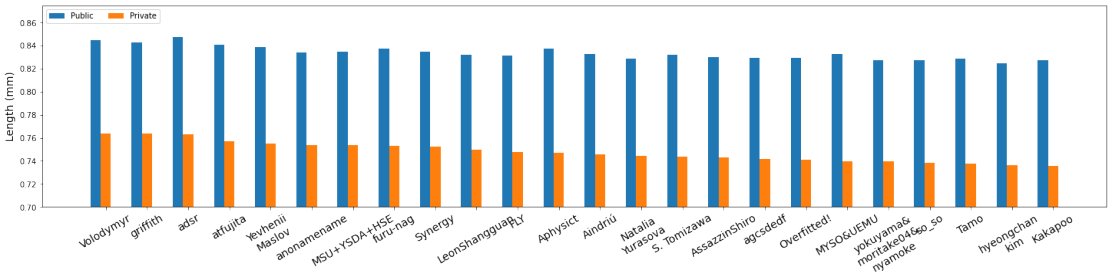


Figure 2: Top 25 private leaderboard scores achieved by the best systems evaluated within the primary bird identification task of LifeCLEF 2023. Public and private test data were split randomly, private scores remained hidden until the submission deadline. Participants were able to optimize the recognition performance of their systems based on public scores, which likely explains some differences in scores.

and used a timer to stop inference before time ran out. They also report that OpenVino only gave noticeable speedups over ONNX at small batch sizes. (This makes some sense: The primary tool for speeding up inference on CPU is SIMD computation. A kernel implementation which uses SIMD over the batch dimension will under-perform on small batch sizes.)

This year, many of the test-sets from previous years were released on Zenodo, though these datasets have minimal overlap with this year's African species list. Many top competitors used the Zenodo data, but most only used it to find no-call segments for augmentation and training.

The top competitors generally trained models from scratch from the Xeno-Canto data, first using a pre-training phase and then fine-tuning to the particular species in the competition list. Competitors mostly restricted pre-training to the 250 species appearing in previous competitions.

We accepted five working notes for the proceedings.

Miyaguchi, et al. [13]: This team experimented with using unsupervised source separation models for pseudo-labeling training data. They also included some UMAP plots of the model embedding space, which provide insight into the linear separability of the classes.

Mario Lasseck [14]: A clear write-up from a long-standing BirdCLEF competitor using a combination of methods (energy measurement and pseudo-labels) to combat the weak-label problem in the training data. Lasseck also found benefit from a new reverb augmentation and applying SED attention over time.

Lihang Hong [15]: The second-place entry used OpenVINO to get a full seven models running in an ensemble, while still meeting the two-hour inference time limit.

Paul Nussbaum [16]: Presents a method for measuring data loss as features develop in the classification network. By applying pseudo-inverse techniques to the activations at each layer, one can try to recover the original inputs and measure corruption.

Mihai Minut et al. [17]: An early effort from a new competitor, this write-up summarizes previous BirdCLEF results and presents experiments with data pre-processing and model architectures. They had more success with pre-trained ImageNet weights than with custom CNNs.

Notable excerpts from individual contributors who did not submit working notes:

- The fourth-place competitor, atfujita, used knowledge distillation, directly applying the techniques in (Knowledge distillation: A good teacher is patient and consistent) to distill the Google Bird Vocalizations model into a faster model.
- The fist-place competitor, Volodymyr, emphasized over-sampling low-data classes during training. They sampled each class according to the distribution which balances the training data distribution.
- Volodymyr also addressed around the 'weak-labels' problem in Xeno-Canto by training on longer 30-second segments and using an attention layer to aggregate the features. Since the attention-aggregated features are a weighted sum over the time dimension, the same model which is trained on 30 seconds of audio can then be applied to 5-second segments.

4. Conclusions and Lessons Learned

Thanks to the new metric and inference time limit, we saw a proliferation of new ideas this year, moving beyond the threshold tuning of the previous few years. The top entries included three approaches that we might call strong on Kaggle fundamentals, one entry which leveraged OpenVino to get a particularly large solution, and one entry which used model distillation to transfer strong pre-trained embeddings. We also saw more experimentation with modeling approaches overall.

Acknowledgments

We would also like to thank Kaggle for helping us host this competition. We are especially grateful for the incredible support and efforts of Maggie Demkin, Sohier Dane, and Addison Howard who helped process the dataset and set up the competition. Special thanks to Google for sponsoring the prize money for this competition. Last but not least, thanks to everyone who provided/annotated data, 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-2023

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] M. H. Neate-Clegg, M. A. Etterson, M. W. Tingley, W. D. Newmark, The combined effects of temperature and fragment area on the demographic rates of an afrotropical bird community over 34 years, Biological Conservation 282 (2023) 110051. URL: https://www.sciencedirect.com/science/article/pii/S0006320723001520. doi:https://doi.org/10.1016/j.biocon.2023.110051.

[5] T. Brooks, A. Balmford, N. Burgess, L. Hansen, J. Moore, C. Rahbek, P. Williams, L. Bennun, A. Byaruhanga, P. Kasoma, P. Njoroge, D. Pomeroy, M. Wondafrash, C. Williams, K. A, P. P, Conservation priorities for birds and biodiversity: Do east african important bird areas represent species diversity in other terrestrial vertebrate groups?, Ostrich Supplement (2001).

[6] T. Stevenson, J. Fanshawe, Field Guide to the Birds of East Africa: Kenya, Tanzania, Uganda, Rwanda, Burundi, Bloomsbury Publishing, 2020.

[7] R. B. Payne, Duetting and chorus singing in african birds, Ostrich 42 (1971) 125–146. doi:10.1080/00306525.1971.9633401.

[8] A. Joly, C. Botella, L. Picek, S. Kahl, H. Goëau, B. Deneu, D. Marcos, J. Estopinan, C. Leblanc, T. Larcher, R. Chamidullin, M. Šulc, M. Hrúz, M. Servajean, H. Glotin, R. Planqué, W.-P. Vellinga, H. Klinck, T. Denton, I. Eggel, P. Bonnet, H. Müller, Overview of LifeCLEF 2023: evaluation of AI models for the identification and prediction of birds, plants, snakes and fungi, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023.

[9] 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.

[10] 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.

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

[12] 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.

[13] A. Miyaguchi, N. Zhang, M. Gustineli, C. Hayduk, Transfer Learning with Semi-Supervised Dataset Annotation for Birdcall Classification, in: CLEF Working Notes 2023, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2023, Thessaloniki, Greece, 2023.
[14] M. Lasseck, Bird Species Recognition using Convolutional Neural Networks with Attention on Frequency Bands, in: CLEF Working Notes 2023, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2023, Thessaloniki, Greece, 2023.
[15] L. Hong, Acoustic Bird Species Recognition at BirdCLEF 2023: Training Strategies for Convolutional Neural Network and Inference Acceleration using OpenVINO, in: CLEF Working Notes 2023, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2023, Thessaloniki, Greece, 2023.
[16] P. Nussbaum, Reading the Robot Mind – Presenting Internal Data Flow of Deep Learning Neural Networks in a Format Familiar to Subject Matter Experts, in: CLEF Working Notes 2023, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2023, Thessaloniki, Greece, 2023.
[17] M. Minut, C. Simionescu, A. Iftene, Classic Approaches to Bird Song Classification, in: CLEF Working Notes 2023, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2023, Thessaloniki, Greece, 2023.