BEANS: THE BENCHMARK OF ANIMAL SOUNDS

Masato Hagiwara, Benjamin Hoffman, Jen-Yu Liu, Maddie Cusimano, Felix Effenberger, Katie Zacarian

Earth Species Project

ABSTRACT

The use of machine learning (ML) based techniques has become increasingly popular in the field of bioacoustics over the last years. Fundamental requirement for the successful application of ML based techniques are curated, agreed upon, high-quality datasets and benchmark tasks to be learned on a given dataset. However, the field of bioacoustics so far lacks such public benchmarks which cover multiple tasks and species to measure the performance of ML techniques in a controlled and standardized way and that allows for benchmarking newly proposed techniques to existing ones. Here, we propose BEANS (the BEnchmark of ANimal Sounds), a collection of bioacoustics tasks and public datasets, specifically designed to measure the performance of machine learning algorithms in the field of bioacoustics. The benchmark consists of two common tasks in bioacoustics: classification and detection. It includes 12 datasets covering various species, including birds, land and marine mammals, anurans, and insects. In addition to the datasets, we also present the performance of a set of standard ML methods as the baseline for task performance. The benchmark and baseline code is made publicly available\(^1\) in the hope of establishing a new standard dataset for ML-based bioacoustic research.

Index Terms— bioacoustics, benchmark

1. INTRODUCTION

Due to their increasing affordability, recording and storage devices are now widely used to collect bioacoustic data. These devices enable animal welfare and wildlife conservation applications, such as passive acoustic monitoring (PAM), which offers tools for wildlife population assessment and conservation research in a non-invasive and unbiased manner [1]. However, they produce a large amount of bioacoustic data, and the manual processing and analysis of these data has become a bottleneck [2, 3].

For this reason, machine learning (ML) has increasingly been used to automate the processing and analysis of bioacoustics tasks. It has been successfully applied to a variety of tasks such as classification of species, individuals, and various other characteristics of calls [4], detection and recognition of vocalizations in passive recordings [5], and automatic discovery of vocalization units [6]. Furthermore, the so-called “deep learning” (DL) models based on deep artificial neural networks have drastically reduced error rates and are increasingly being used for these tasks in recent years.

However, typical research studies in bioacoustics focus only on a small number of species and/or specific types of methodology [7, 8]. This narrow focus has led to a proliferation of ML models and algorithms that perform well on the tasks in question on a given data set, but not necessarily outside of their scope. Moreover, the lack of publicly available, agreed-upon standard datasets in bioacoustics [9, 10] makes it difficult to reproduce and compare different approaches in a standardized way.

If we turn our attention to other fields of ML, much of the recent progress has been driven by standard “benchmarks”. A benchmark is a collection of datasets along with tasks to be performed on the data, specifically designed to measure the performance of ML algorithms in a standardized way. For example, standardized datasets such as MNIST, CIFAR-10/100, and ImageNet [11] have been used to measure the performance of image classification algorithms in computer vision for decades. Other examples include GLUE [12] and SuperGLUE [13] for natural language processing, and SUPERB [14] and HEAR [15] for human speech/audio processing. These standardized benchmarks allow for an objective and quantitative comparison of different approaches, the identification of strengths and weaknesses of different methodologies, and additionally for an assessment of the overall progress made in the field of research.

In this paper, we propose BEANS (the BEnchmark of ANimal Sounds), a collection of publicly available bioacoustics public datasets along with tasks to be performed on those, specifically designed to measure the performance of ML algorithms in the bioacoustics domain in a standardized manner. The benchmark includes two common bioacoustics tasks, classification and detection, and consists of twelve datasets covering diverse species, including birds, land and marine mammals, anurans, and insects. We run various non-DL and DL algorithms as the baseline on BEANS and show that there is considerable room for improvement, especially for the detection task. We release the entire code and the baseline implementations as open source to encourage the further development of generic bioacoustic methods.

---

\(^1\)https://github.com/earthspecies/beans
2. BENCHMARK DESIGN

The goal of BEANS is to accurately measure and compare ML models through a collection of bioacoustic datasets covering diverse species. We wish to encourage the development of ML models that work well not only on specific species and dataset, but also on a diverse set of species with as little species-specific modification or training data as possible. Such models are of interest to the bioacoustic community due to better generalizability and lower development cost.

In this benchmark, we focus on classification and detection tasks, which are the two tasks most commonly considered in the bioacoustics literature [2]. Classification is a task in which each sample is assigned one or more labels from a set of predefined classes, such as a set of individuals or a set of species. Here, we use a regular single-label, multi-class setting (Figure 1 left).

Detection is a task in which one uses ML algorithms to identify subsections of interest and their properties from long recordings (often obtained from passive acoustic monitoring). We adopt a sliding window approach, as commonly done for detection tasks [5, 16], where long recordings are broken up into short (potentially overlapping) segments and the ML algorithm makes a prediction per segment (Figure 1 right). In order to address multiple overlapping vocalizations, more than one label can be assigned to a segment, making it a multi-label, multi-class classification setting.

Due to the design of these tasks, from the perspective of the ML algorithm, the two tasks can be both solved by the same classification model. This homogeneity and simplicity in terms of the structure of the tasks allows the benchmark users to use almost the same algorithm with little modification and encourages the use of generic models.

As metrics for evaluating task performance, we use accuracy $A = \frac{\sum_{c}(tp_c + tn_c)}{C \cdot N}$, for classification tasks and mean average precision mAP $= \frac{1}{N} \sum_{c \in \{0,1,\ldots,10\}} P_c(r)$ for detection tasks, where $N$ is the total number of samples in the dataset, $C$ is the number of classes, and $tp_c, fn_c, P_c(r)$ are the number of true positives, false positives, the interpolated precision at recall $r$ for class $c$, respectively. The benchmark comes with predefined train, validation, and test splits, as well as a baseline implementation, to encourage consistent comparison and reproducibility.

### Table 1. Datasets included in the benchmark. * The numbers of samples for classification and the number of 1-minute “chunks” for detection (see Section 4 for more details).

| Dataset       | # Insts. | # Labels (type) | SR (Hz) |
|---------------|----------|-----------------|---------|
| watkins       | 1,695    | 31 (species)    | 44.1k   |
| bats          | 10,000   | 10 (individual) | 250k    |
| ebi           | 21,375   | 264 (species)   | 44.1k   |
| dogs          | 693      | 10 (individual) | 44.1k   |
| humbugdb      | 13,011   | 14 (species)    | 44.1k   |
| dcase         | 1,166    | 20 (species)    | var.    |
| enabirds      | 382      | 34 (species)    | 32k     |
| hiceas        | 674      | 1 (species)     | 500k    |
| rfcx          | 4,724    | 24 (species)    | 48k     |
| gibbons       | 2,077    | 3 (call type)   | 9.6k    |
| esc-50        | 2,000    | 50 (sound type) | 16k     |
| sc: human     | 105,830  | 35 (word)       | 44.1k   |

In this section, we describe the datasets that we included in the benchmark. To make this choice, we surveyed many bioacoustics datasets from the literature and chose 5 datasets for classification and another 5 for detection tasks (Table 1) based on the following criteria:

- **Availability**: Is the dataset publicly and freely available for research purposes?
- **Difficulty**: Is the dataset moderately difficult for ML algorithms to solve?
- **Size**: Is the dataset large enough for ML algorithms to learn meaningful patterns from it? Is the dataset small enough so that the training is within the reach of an average compute budget of typical users?
- **Diversity**: Does the benchmark represent a diverse collection of sound-making animal species?

We also included two “auxiliary” datasets commonly used to evaluate environmental sound detection and speech classification systems, two domains closely related to bioacoustics, in order to encourage the development of ML models that generalize beyond bioacoustics. These two datasets are not officially part of the benchmark, but benchmark users can choose to report performance numbers on them as a reference. Below, we describe the five classification datasets:

**watkins** [17]: The Watkins Marine Mammal Sound Database is a database of marine mammal sounds. We used the preprocessed dataset hosted on the Internet Archive 2 which contains the recordings of 32 species from the ‘Best of cuts’ section (except for weddell seal recordings, which had only a few samples). We randomly split the dataset into 6:2:2 train:valid:test portions with stratification. All recordings are resampled to 44.1kHz.

---

2 [https://archive.org/details/watkins_202104](https://archive.org/details/watkins_202104)
**bats** [18]: The original dataset contains annotated recordings of Egyptian fruit bats (*Rousettus aegyptiacus*) vocalizations recorded at a sampling rate of 250kHz. We used the preprocessed dataset of individual calls of up to a few seconds long. The target label is the emitter ID (individual).

**cbi** [19]: This is the dataset from the Cornell Bird Identification competition hosted on Kaggle. The training set consists of bird recordings uploaded to xeno-canto by volunteer users. Since the test set labels are hidden, we split the train set into 6:2:2 train:valid:test portions in such a way that there is no overlap in recordists between splits.

**dogs** [20]: This dataset consists of barks recorded from 10 individual domestic dogs in different situations (disturbance, isolation, and play) originally at 48 kHz and resampled to 44.1kHz. Each recording is annotated with the individual and the situation, but we used the individual as the target label. We randomly split the dataset into 6:2:2 train:valid:test portions with stratification.

**humbugdb** [21]: HumBugDB is a collection of wild and cultured mosquito wingbeat sounds recorded in various settings (including sounds when the animals were located in cups and under bednets). The purpose is to detect and classify species that can be vectors of diseases such as malaria. We took their species classification dataset, while collapsing any species that can be vectors of diseases such as malaria. We used the 33 most frequent species labels and treated all infrequent labels as “OTHERS.” We randomly split the dataset into 6:2:2 train:valid:test portions with stratification.

The following are the details of the five detection datasets:

**dcase** [22]: This is the dataset used for DCASE 2021 Task 5: Few-shot Bioacoustic Event Detection. It contains mammal and bird multi-species recordings annotated with species, onset, and offset times. We repurposed their few-shot development dataset as a (regular) detection dataset by partitioning long recordings into 1-minute chunks, and used the first 60% for training, the next 20% for validation, and the final 20% for testing. We only retained positive (POS) labels. The dataset contains files recorded at various sample rates, but we up/down sampled them all at 16kHz.

**enabirds** [23]: The dataset contains recordings of bird dawn chorus, annotated with the onset/offset time, the frequency range, and the species. We used the 33 most frequent species labels and treated all infrequent labels as “OTHERS.” We partitioned the data set into training, validation, and testing portions as described in dcase.

**hiceas** [24]: The dataset consists of a subset of passive acoustic data collected using a multi-channel towed hydrophone during the Hawaiian Islands Cetacean and Ecosystem Assessment Survey (HICEAS) in 2017. We used the human-audible Minke whale “boing” vocalization annotations of the data set, rendering this a single-class (binary) detection task. We sampled 1/20th of all the files in the dataset, downsampled them from 500kHz to 22.5kHz, and assigned them to train, valid and test splits randomly with a 6:2:2 ratio. We redistribute the preprocessed dataset as part of the BEANS repository.

**rfcx** [5]: This is a dataset of continuous soundscape recordings of 24 species of frogs and birds collected by Rainforest Connection (RFCx). The data were annotated with the onset / offset time, as well as the frequency range. We randomly assigned the files into train, valid, and test splits with a 6:2:2 ratio. We did not use the false positive annotation.

**hainan-gibbons** [16]: The dataset contains continuous recordings of Hainan gibbon calls. The data were annotated with onset/offset times and call types (one pulse, multiple pulse, duet). There are a total of 14 files, each corresponding to ~8 hours of recording for a particular day. Due to its large size, we sub-sampled 1/3 of the dataset after splitting into chunks. We used the first 6 files for training, the next 3 files for validation, and the remaining ones for testing.

Finally, we included two auxiliary datasets:

**esc50** [25]: A dataset of environmental audio recordings including animal, nature, human (non-speech), interior/domestic, and exterior/urban sounds. The dataset is commonly used as a benchmark for environmental sound classification and comes with predefined splits. We used split 4 for validation, split 5 for testing, and the rest for training.

**speech-commands** [26]: A dataset of single-word utterances covering 35 English words including digits, commands, and others, spoken by multiple speakers. The dataset is commonly used to benchmark speech systems. The dataset comes with predefined train, validation, and test splits. We used version .02 of the data set.

### 4. EXPERIMENTS

We ran a wide range of traditional non-DL and DL algorithms to establish a baseline for BEANS. All the waveforms were mixed to a single channel (mono) with 16 bit depth. For classification, each waveform was padded with silence at the end if it was shorter than the minimum duration threshold set for the dataset, or truncated if it was longer. For detection, a sliding window algorithm was applied to partition the chunks into instances. The length of sliding windows was 2 seconds for dcase and enabirds, 10 seconds for hiceas and rfcx, and 4 seconds for hainan-gibbons. An instance is marked positive if the amount of overlap with any annotation is larger than 20%.

We used the following non-DL algorithms, often used for classification of tabular data: Logistic regression (LR), Support vector machine (SVM), Decision tree (DT), Gradient-boosted decision tree (GBDT), XGBoost (XGB, [27]). We first obtained 20 MFCCs features from power mel-spectrogram [28]. The length of the FFT window and the hop length were chosen as 50ms and 10ms, respectively. We computed four summary statistics, mean, standard deviation, min, and max of each MFCC dimension over time, resulting in a 80-dimension feature vector per sample.
Cornell Bird Identification (CBI) is the most challenging classification task in BEANS, due to the way the data is split (with no overlapping recordists) and the sheer number of target labels (264 species), which is close to how machine learning models are used in some real-world bioacoustic settings.

In terms of tasks, performance on some detection datasets (e.g., rfxc, gibbons) was lower than others. This might be due to the fact that a typical detection dataset is very sparse, meaning that only a minor portion of a typical recording is animal vocalization and has fewer training annotations per class than classification. Modern regularization techniques, such as mixup [34] may help improve the results.

For most of the datasets in BEANS, the precision of our baselines is well below 90%, meaning that there is large room for improvement for future, more specialized ML algorithms.

### 5. DISCUSSION

**Sample rate:** Bioacoustics data can have a wide range of sample rates, which can be challenging for ML models that are usually trained on data captured at a single and fixed sample rate. Even within BEANS, vocalizations range from human-audible calls that can be captured with a sampling rate around 10kHz to ultrasonic echolocation of bats that require sampling rates of 250kHz. Methods that can be adopted to different sampling frequencies (e.g., S4 [35]) could be a promising future direction.

**Overlapping calls:** Currently, we cannot address the cases where multiple individuals are vocalizing at the same time unless those individuals are explicitly and separately annotated. Sound separation models (e.g., BioCPPNet [36]) can be a promising preprocessing step to address this issue.

**Bias:** Some animal species and/or geographical locations can be over- or under-represented in a dataset. For example, North American birds are over-represented in BEANS, included in at least three datasets (cbi, dcase, and enabirds). We acknowledge that the curating process of datasets is inherently subjective [37, 38] and wish to address this bias in future iterations of this benchmark.

**Limitations:** We acknowledge that a benchmark is only a proxy for progress that we measure and can face construct validity issues [38, 39]. We also note the inherent fragility of the ML benchmarking process [40]. Therefore, we in particular do not recommend reporting a single metric averaged over datasets and tasks as a performance measure of an ML algorithm on this benchmark. Instead, BEANS is intended as a diagnostic tool to encourage the development of generic, well-balanced bioacoustics approaches.

### 6. REFERENCES

[1] Larissa Sayuri Moreira Sugai et al., “Terrestrial passive acoustic monitoring: Review and perspectives,” BioScience, vol. 69, no. 1, 11 2018.
[2] Dan Stowell, “Computational bioacoustics with deep learning: a review and roadmap,” PeerJ, vol. 10, 2022.
[3] Devis Tuia et al., “Perspectives in machine learning for wildlife conservation,” Nature Communications, vol. 13, no. 1, Feb. 2022.
[4] Veronica Morfi et al., “Deep perceptual embeddings for unlabelled animal sound events,” The Journal of the Acoustical Society of America, vol. 150, no. 1, 2021.
[5] Jack LeBien et al., “A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network,” Ecological Informatics, vol. 59, 2020.
[6] Tim Sainburg et al., “Latent space visualization, characterization, and generation of diverse vocal communication signals,” bioRxiv, 2020.
[7] Kevin R. Coffey et al., “DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations,” Neuropharmacology, vol. 44, 2019.
[8] Yu Shiu et al., “Deep neural networks for automated detection of marine mammal species,” Scientific Reports, vol. 10, no. 607, 2020.
[9] Ed Baker and Sarah Vincent, “A deafening silence: a lack of data and reproducibility in published bioacoustics research?,” Biodiversity Data Journal, vol. 7, 2019.
[10] Michael P. McLoughlin et al., “Automated bioacoustics: methods in ecology and conservation and their potential for animal welfare monitoring,” Journal of The Royal Society Interface, vol. 16, no. 155, 2019.
[11] Jia Deng et al., “ImageNet: A large-scale hierarchical image database,” in CVPR, 2009.
[12] Alex Wang et al., “GLUE: A multi-task benchmark and analysis platform for natural language understanding,” in BlackboxNLP@EMNLP, 2018.
[13] Alex Wang et al., “SuperGLUE: A stickier benchmark for general-purpose language understanding systems,” in NeurIPS, 2019.
[14] Shu wen Yang et al., “SUPERB: Speech processing universal performance benchmark,” in Interspeech, 2021.
[15] Joseph Turian et al., “Hear: Holistic evaluation of audio representations,” in Proc. of the NeurIPS 2021 Competitions and Demonstrations Track, 2022.
[16] Emmanuel Dufourq et al., “Automated detection of Hainan gibbon calls for passive acoustic monitoring,” Remote Sensing in Ecology and Conservation, vol. 7, no. 3, 2021.
[17] Laela Sayigh et al., “The Watkins marine mammal sound database: An online, freely accessible resource,” Proc. of Meetings on Acoustics, vol. 27, no. 1, 2016.
[18] Yosef Prat et al., “An annotated dataset of Egyptian fruit bat vocalizations across varying contexts and during vocal ontogeny,” Scientific Data, vol. 4, 2017.
[19] Cornell Lab of Ornithology, “Cornell birdcall identification,” 2020.
[20] Sophia Yin and Brenda McCowan, “Barking in domestic dogs: context specificity and individual identification,” Animal Behaviour, vol. 68, no. 2, 2004.
[21] Ivan Kiskin et al., “HumBugDB: a large-scale acoustic mosquito dataset,” in Proc. of the NeurIPS: Track on Datasets and Benchmarks, 2021.
[22] Veronica Morfi et al., “Few-shot bioacoustic event detection: A new task at the dcase 2021 challenge,” in DCASE, 2021.
[23] Lauren M. Chronister et al., “An annotated set of audio recordings of Eastern North American birds containing frequency, time, and species information,” Ecology, vol. 102, no. 6, 2021.
[24] NOAA Pacific Islands Fisheries Science Center, “Hawaiian islands cetacean and ecosystem assessment survey (hicesas) towed array data,” DCLDE, 2022.
[25] Karol J. Piczak, “ESC: Dataset for Environmental Sound Classification,” in MM, 2015.
[26] P. Warden, “Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition,” ArXiv, 2018.
[27] Tianqi Chen and Carlos Guestrin, “XGBoost: A scalable tree boosting system,” in SIGKDD, 2016.
[28] Yao-Yuan Yang et al., “TorchAudio: Building blocks for audio and speech processing,” in ICASSP, 2022.
[29] Kaiming He et al., “Deep residual learning for image recognition,” CVPR, 2016.
[30] Shawn Hershey et al., “CNN architectures for large-scale audio classification,” in ICASSP, 2017.
[31] Zachary J. Ruff et al., “Automated identification of avian vocalizations with deep convolutional neural networks,” Remote Sensing in Ecology and Conservation, vol. 6, no. 1, 2020.
[32] Venkatesh Boddapati et al., “Classifying environmental sounds using image recognition networks,” KES, 2017.
[33] Sébastien Marcel and Yann Rodriguez, “Torchvision the machine-vision package of torch,” in MM, 2010.
[34] Hongyi Zhang et al., “mixup: Beyond empirical risk minimization,” ICLR, 2018.
[35] Albert Gu et al., “Efficiently modeling long sequences with structured state spaces,” in ICLR, 2022.
[36] Peter C. Bermant, “BioCPPNet: automatic bioacoustic source separation with deep neural networks,” Scientific Reports, vol. 11, 2021.
[37] Shreya Shankar et al., “No classification without representation: Assessing geodiversity issues in open data sets for the developing world,” in NIPS 2017 workshop: Machine Learning for the Developing World, 2017.
[38] Inioluwa Deborah Raji et al., “AI and the everything in the whole wide world benchmark,” ArXiv, 2021.
[39] Samuel R. Bowman and George Dahl, “What will it take to fix benchmarking in natural language understanding?,” in NAACL, 2021.
[40] Mostafa Dehghani et al., “The benchmark lottery,” ArXiv, 2021.