

# scientific data



OPEN

DATA DESCRIPTOR

## A dataset for benchmarking Neotropical anuran calls identification in passive acoustic monitoring

Juan Sebastián Cañas<sup>1</sup>✉, María Paula Toro-Gómez<sup>1</sup>, Larissa Sayuri Moreira Sugai<sup>2</sup>, Hernán Darío Benítez Restrepo<sup>3</sup>, Jorge Rudas<sup>1</sup>, Breyner Posso Bautista<sup>1</sup>, Luís Felipe Toledo<sup>1,4</sup>, Simone Dena<sup>5</sup>, Adão Henrique Rosa Domingos<sup>6</sup>, Franco Leandro de Souza<sup>7</sup>, Selvino Neckel-Oliveira<sup>8</sup>, Anderson da Rosa<sup>8</sup>, Vítor Carvalho-Rocha<sup>1,8</sup>, José Vinícius Bernardy<sup>9</sup>, José Luiz Massao Moreira Sugai<sup>9</sup>, Carolina Emilia dos Santos<sup>9</sup>, Rogério Pereira Bastos<sup>9</sup>, Diego Llusia<sup>10,11,12</sup> & Juan Sebastián Ulloa<sup>1</sup>

Global change is predicted to induce shifts in anuran acoustic behavior, which can be studied through passive acoustic monitoring (PAM). Understanding changes in calling behavior requires automatic identification of anuran species, which is challenging due to the particular characteristics of neotropical soundscapes. In this paper, we introduce a large-scale multi-species dataset of anuran amphibians calls recorded by PAM, that comprises 27 hours of expert annotations for 42 different species from two Brazilian biomes. We provide open access to the dataset, including the raw recordings, experimental setup code, and a benchmark with a baseline model of the fine-grained categorization problem. Additionally, we highlight the challenges of the dataset to encourage machine learning researchers to solve the problem of anuran call identification towards conservation policy. All our experiments and resources have been made available at <https://soundclim.github.io/anuraweb/>.

### Background & Summary

Global anthropogenic biodiversity loss is a major challenge of contemporary society<sup>1</sup>. With severe wildlife population declines and extinctions over the planet, monitoring and predicting species responses to global changes became an urgent task for conservation. Novel technologies now offer remote, non-invasive, and automated methods to survey and monitor biodiversity at unprecedented spatial and temporal scales<sup>2</sup>. For instance, passive acoustic monitoring (PAM) has been largely adopted in ecological research and is increasingly used in conservation applications<sup>3</sup>. Based on acoustic sensor networks, PAM enables us to remotely and automatically record the vocal activity of wild animals, increasing our ability to study biological communities. However, a critical

<sup>1</sup>Instituto de Investigación de Recursos Biológicos Alexander von Humboldt, Avenida Paseo Bolívar 16-20, Bogotá, Colombia. <sup>2</sup>K Lisa Yang Center for Conservation Bioacoustics, Cornell Lab of Ornithology, Cornell University, 159 Sapsucker Woods road, 14850, Ithaca, New York, USA. <sup>3</sup>Pontifícia Universidad Javeriana Seccional Cali, Calle 18 No 118-250, Cali, Valle del Cauca, Colombia. <sup>4</sup>Laboratório de História Natural de Anfíbios Brasileiros (LaHNAB), Universidade Estadual de Campinas, Campinas, SP, Brazil. <sup>5</sup>Museu de Diversidade Biológica (MDBio), Universidade Estadual de Campinas, Campinas, SP, Brazil. <sup>6</sup>Instituto de Pesquisa da Biodiversidade (IPBio), Reserva Betary, Iporanga, São Paulo, Brazil. <sup>7</sup>Universidade Federal de Mato Grosso do Sul, Instituto de Biociências, Campo Grande, MS, Brazil. <sup>8</sup>Departamento de Ecología e Zoología, Universidad Federal de Santa Catarina, Florianópolis, SC, Brazil. <sup>9</sup>Universidade Federal de Goiás, Goiânia, GO, Brazil. <sup>10</sup>Terrestrial Ecology Group, Departamento de Ecología, Universidad Autónoma de Madrid, C/Darwin, 2, Ciudad Universitaria de Cantoblanco, Facultad de Ciencias, Edificio de Biología, 28049, Madrid, Spain. <sup>11</sup>Centro de Investigación en Biodiversidad y Cambio Global (CIBC), Universidad Autónoma de Madrid. C/Darwin 2, 28049, Madrid, Spain. <sup>12</sup>Laboratório de Herpetologia e Comportamento Animal, Departamento de Ecología, Instituto de Ciências Biológicas, Universidade Federal de Goiás, Goiás, Brazil. ✉e-mail: [jcanas@humboldt.org.co](mailto:jcanas@humboldt.org.co)

bottleneck for the widespread use of this method is the need for automated techniques to retrieve biologically meaningful information in the huge time-series audio datasets collected by PAM. Manual inspection of these recordings is unattainable due to the human specialist workload when audio data collected reach the big data scale<sup>4</sup>.

In the last decade, the three fundamental reasons for the success of machine learning (ML) techniques have been the advancement in high-computing hardware, novel algorithms, and the curation of high-quality datasets for standardized benchmark<sup>5</sup>. As a consequence, ML has emerged as a key solution and a general accelerator for multiple domains in which biodiversity monitoring programs, animal ecology, and global change research are not an exception<sup>6,7</sup>. Particularly, the growth of ML for ecological applications now depends on the variety, quality, and availability of public datasets that define ML tasks for determined contexts and problems<sup>7,8</sup>. Despite recent efforts to curate datasets for ecological research, available data remains taxonomically and geographically biased<sup>9</sup>. ML has opened up exciting possibilities for research in this area<sup>10–14</sup>, but limitations in the diversity of existing datasets must be acknowledged. In the field of bioacoustics and PAM, datasets aimed at supporting acoustic identification have been developed for a limited number of taxonomic groups, mainly birds<sup>15,16</sup>, mosquitoes<sup>17</sup>, and mammals<sup>8,18,19</sup>. These datasets have also served as general benchmarks for the detection and classification of the recorded individuals into species<sup>20</sup>. Altogether, the increasing number of curated datasets coming from bioacoustics research generates a unique opportunity to foster the culture of open data, open models, and benchmarks in conservation research<sup>21</sup>. PAM has special importance in applied conservation, where datasets may impact the robustness of biodiversity monitoring programs that support ecological<sup>22</sup> and policy-related<sup>23</sup> decision-making.

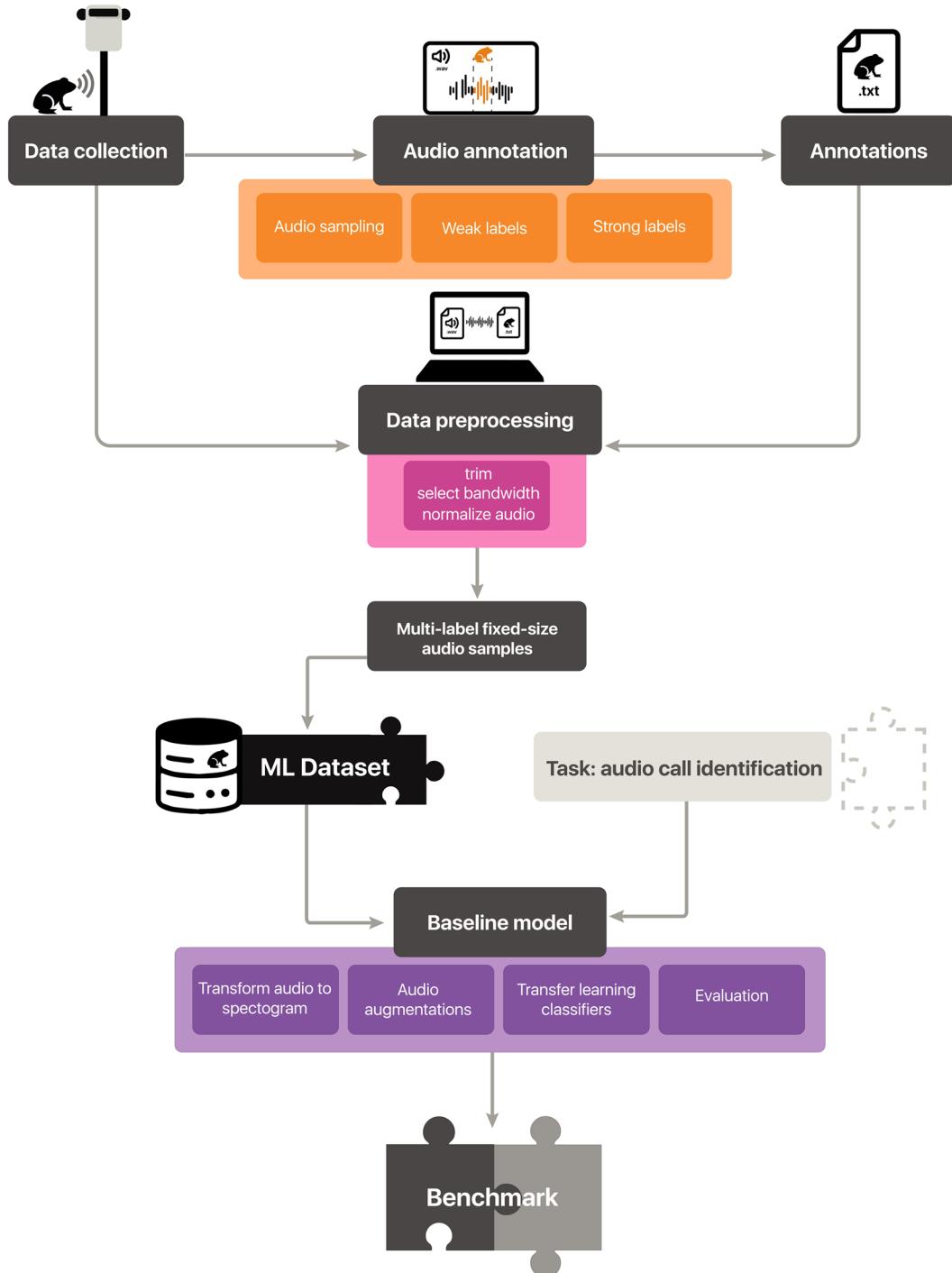
Amphibians are one of the most endangered vertebrate groups in the world, with more than 40% of the species endangered to extinction<sup>24</sup>. In the tropics, amphibian communities exhibit high diversity<sup>25</sup> and are more prone to extinction<sup>26</sup> compared to other regions. To monitor these communities, researchers can take advantage of PAM techniques which are a non-invasive data collection that allows incorporating information from both rare and cryptic species, as well as from common and abundant ones. Acoustic communication has a central role in the reproductive behavior of anurans<sup>27</sup>. During the breeding season, males call, for example, to attract females, defend territories, and deter competitors<sup>28</sup>. Thus, a wide range of research relies on the identification and quantification of these sounds, with an increasing number of applications. However, there is a lack of open datasets for this highly vocal group that can support the development of ML models for PAM research.

This study introduces a large-scale annotated dataset of Neotropical anuran calls: *AnuraSet*. This dataset was compiled through a country-wide collaborative PAM program across Brazil between 2019 and 2021, and it is composed of 1612 1-minute annotated audio recordings, equivalent to 26.87 hours of audio. We collected data from four strategically selected sites in the Neotropics and generated precise annotations on the recordings. Subsequently, we preprocessed the data to train deep learning models, enabling us to conduct a baseline experiment and launch a benchmarking initiative for the automated identification of anuran calls (Fig. 1). The preprocessing and baseline code is released under the MIT License and all the data is under the CC0 license to support reproducible research. *AnuraSet* will potentially provide a common and realistic-scale evaluation task for species identification in Neotropical soundscapes. In addition, *AnuraSet* is a solid starting point for a comprehensive and accessible dataset of anuran calls and choruses. Since tropical acoustic environments are highly complex and manually annotated datasets are scarce, *AnuraSet* has the capacity to accelerate the development of robust machine listening models for wildlife monitoring in biodiversity hotspots. Furthermore, we summarize the main challenges and propose a roadmap to foster a culture of collaboration, experimentation, research, and exploration in ML for applied ecology. In our viewpoint, this culture is essential for advancing ML techniques and ecological inferences for conservation policies. In addition, the challenges posed by biodiversity acoustic monitoring provide a unique opportunity for exploring new avenues in the field of ML.

In summary, our contributions are (i) a collection of manually annotated PAM recordings of Neotropical anurans calling activity, with information on species composition (presence-absence data) and audio quality of the recordings; (ii) a curated, preprocessed, and in the wild acoustic dataset, with a detailed description of the data challenges; and (iii) baseline models for benchmarking the problem of species identification towards the creation of robust classifiers and the fast development of new models. Overall, our goal is to support a community of ML researchers and conservationists who can work together to develop innovative solutions for biodiversity monitoring. All our experiments and resources have been made available at <https://soundclim.github.io/anuraweb/>. By providing open-access resources and encouraging the exploration of new techniques, we aim to contribute to developing powerful tools for conservation and ecological research.

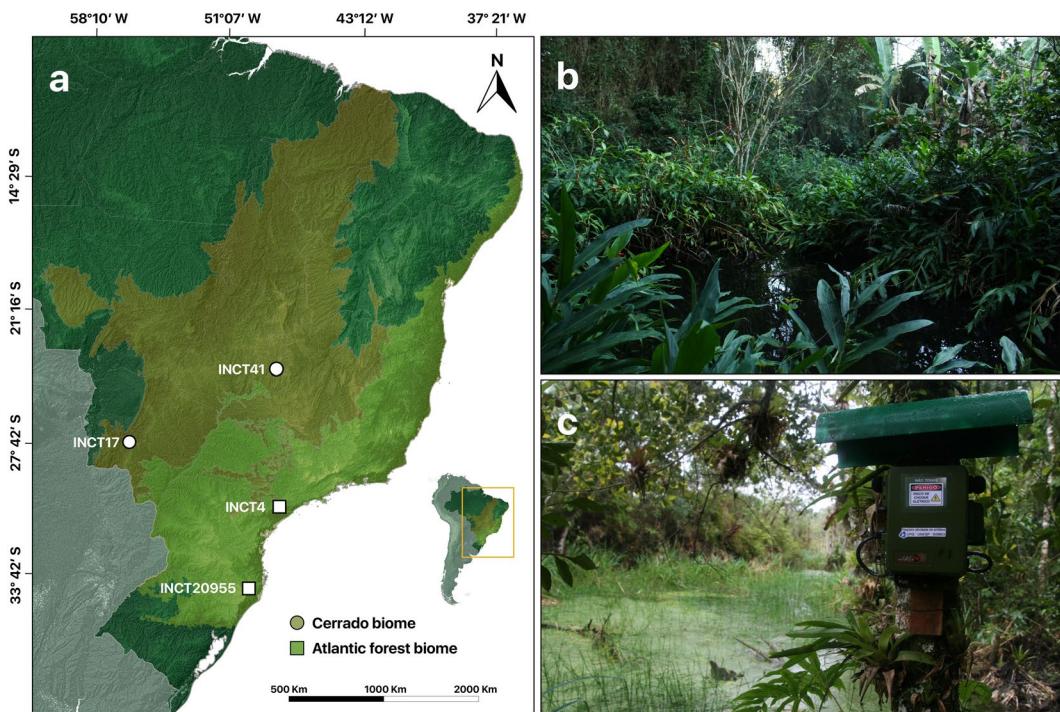
## Methods

**Data Collection.** Calling activity of Neotropical anuran communities was monitored from 2019 to 2021 in four sites located at the Cerrado (INCT17, INCT41) and Atlantic Forest (INCT20955, INCT4) biomes, known for their critical role as global biodiversity hotspots (Fig. 2). INCT refers to *Institutos Nacionais de Ciência, Tecnologia e Inovação* (National Institutes of Science and Technology). At the edge of the water bodies of each site, we installed an acoustic sensor equipped with omnidirectional microphones (SM4, Wildlife Acoustics, Inc., Concord, MA, USA) that were fixed on trees or wooden bases, at about 1.5 m above the ground. Each recorder was configured to register one min every 15 min over 24 h a day (a total of 1.6 hours per day), with a sampling rate of 22050 Hz and 16-bit depth resolution. Audios were recorded in stereo mode, with 10 dB and 16 dB gain on each channel. We considered aspects of anuran calling behavior to choose this recording schedule: a) detectability of pond-breeding anurans is often high, as individuals engage in calling activity from aggregations on the margins of the ponds where the recorders are installed, and b) 1 minute of recording every 15 minutes was the best compromise between obtaining data at a high daily temporal resolution while enabling the sampling over longer periods (e.g. 3 to 4 months)<sup>29</sup>.



**Fig. 1** Overview of the AnuraSet methodological workflow that encompasses the process of dataset creation and benchmarking. It begins with the collection of passive acoustic monitoring data from four sites in the Neotropics. Subsequently, we annotated the recordings with both weak and strong labels. Leveraging these annotations, we undertook a preprocessing of the data to construct a machine learning-compatible dataset. For solving the problem of anuran call identification, we frame the problem as a multilabel classification challenge, and to establish a baseline model, we adopted a transfer learning approach. Furthermore, we merged a specific task with the dataset, culminating in the creation of a benchmark.

**Audio Annotation.** We developed an annotation protocol in order to build automated tools for determining the species recorded with PAM. We combined weak labels (temporal precision was limited to the 60 s duration) with strong labels (providing exact temporal segments of the audio recording where the anuran call was active). The weak labels were annotated by local herpetologists and bioacoustics experts and the strong labels were annotated by a herpetologist over a selected subsample of all raw recordings to obtain a presence-absence dataset at the



**Fig. 2** Data collection of calling activity of Neotropical anuran communities. (a) Geographic location of the four sites where the passive acoustic monitoring data was collected. Sites at the Cerrado biome, INCT17 and INCT41 (dots), and at the Atlantic Forest biome, INCT20955 and INCT4 (squares). (b) Photograph of INCT4 monitoring site at the Atlantic Forest biome. (c) Details of the acoustic sensor used to record anuran calls at the edge of a water body.

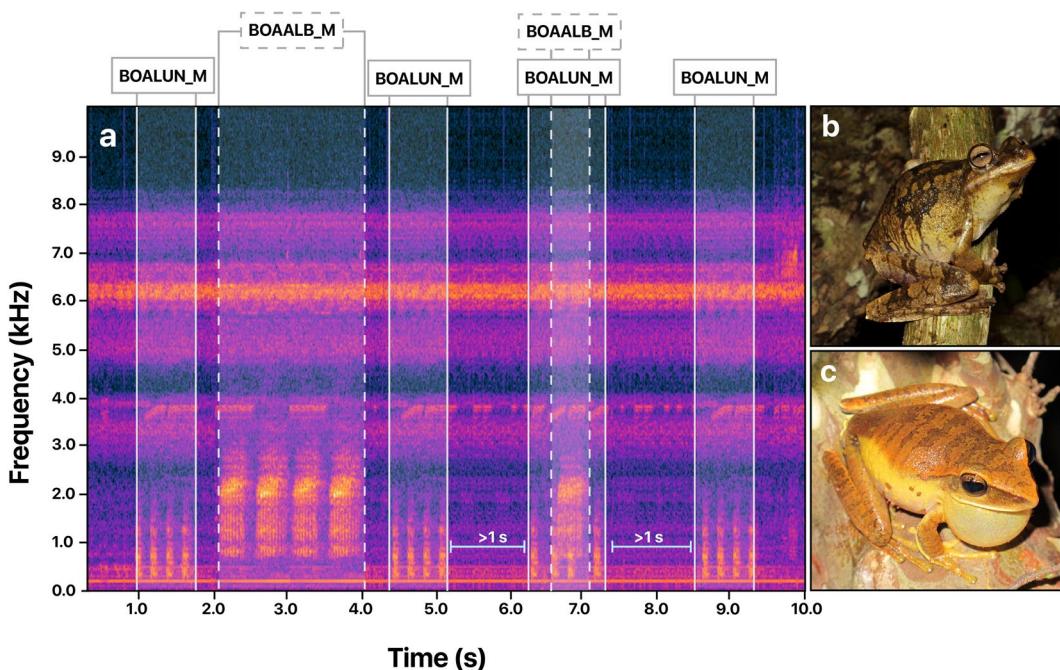
scale of the audio recording. All annotators had previous experience detecting anurans calls in recordings. Since the list of species at each study site was initially unknown, we first searched each 1-min recording of the species using local expert knowledge in the form of weak labels. After that, we used strong labels as they are better to solving the audio event classification problem<sup>30</sup>. The protocol that we developed consists of three steps specifically tailored for the identification of anuran calls. However, it can be easily adapted and customized for any taxon.

**Step 1. Audio sampling.** To annotate audio files, train, and validate ML models, we first obtained a stratified sample of audio recordings from each site that was representative of both seasonal periods and daily periods of highest calling activity. Samples were drawn from months considering the extent of the breeding season, as informed by the principal investigators at each site (3–6 months), at night time (from 1 h before sunset until 1 h before sunrise). From these strata, we randomly selected a total of 300 to 600 files, depending on the amount of months informed by the researchers. These files were processed using two sequential steps, with first, inspection to generate weak labels (see step 2) followed by strong label annotations (step 3). In total, we selected 1612 1-min audio files (26.87 hours) over the four study sites.

**Step 2. Weak labeling.** To identify anuran species recorded in the selected samples (420, 354, 472 and 366 files for INCT04, INCT17, INCT20955, INCT41, respectively), local herpetologists and bioacoustics experts (JVB, SD, JLMMS, AdaR) performed a visual and auditory analysis of spectrograms using Audacity® 3.2.5 software (<https://audacityteam.org/>). Local annotators were asked to report the level of calling activity of each recorded anuran species based on the Amphibian Calling Index<sup>31</sup> (Table 1), according to the species-specific calling activity level in each 1-minute audio file (weak labeling).

**Step 3. Strong labeling.** To provide precise annotations within the 1-min files, we identified bouts of advertisement calls and generated strong labels (step 1). Using Audacity 3.2, we conducted a detailed visual and aural inspection of the spectrogram to identify temporal limits (beginning and end) containing species-specific calls with an inter-call interval of less than 1 second. These annotations ensured fine-scale specificity (Fig. 3). For longer inter-call intervals, we boxed calls separately and labeled them independently. Detailed labels assigned to time boxes were composed of (i) the species ID, tagged with a unique 6-letter code built from the scientific name of each identified species (Supplementary Table 1), and (ii) the perceived quality of the recorded signal, included as a single letter indicating a Low (L), Medium (M), or High (H) quality (Fig. 4). To ensure consistency among the perceptual quality labels, we set up the following criteria: A high-quality call has a high signal-to-noise ratio, no overlap with other sounds, has a well-identifiable structure on the spectrogram, and can be easily visualized on the oscilloscope. A medium-quality call can be visually identified on the spectrogram but may overlap with other sounds that can be difficult to identify in the oscilloscope. A low-quality call shows a low signal-to-noise ratio, is partially masked by other sounds, appears with low intensity on the spectrogram, and cannot be easily

Calling activity level	Score	Description of chorus activity
Absence	0	No anuran calls recorded
Low	1	Individual anurans can be counted and calls do not overlap
Moderate	2	Few anuran individuals that cannot be accurately counted, and with both separate and overlapping calls
High	3	Intense chorus, with continuous and overlapping calls from different individuals

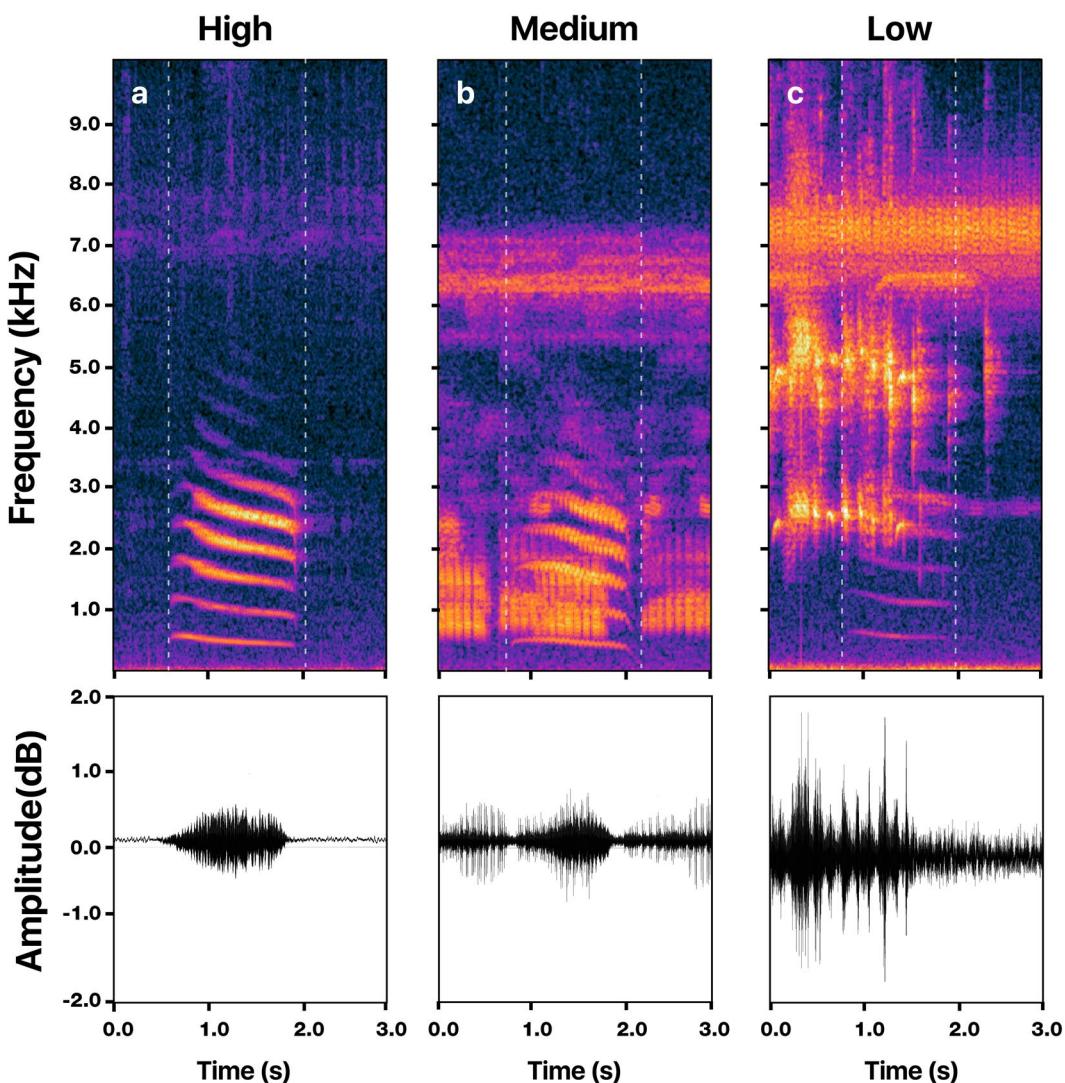
**Table 1.** Levels of anuran calling activity for the weak labeling.

**Fig. 3** Strong labeling process of raw data using audio and spectrogram. (a) Example of strong labeling. For each 1-min raw audio file sampled, the herpetologist annotator identified and selected the temporal limits of the advertisement call. We annotated calls using separate time selections when spaced more than 1 second apart. (b) Image of an individual of *Boana lundii* its advertisement call coded as BOALUN in the annotation process (c) A calling male of *Boana albopunctata* (BOAALB).

identified on the oscillogram. This information was used to promote the usability of the data and improve the error analysis of the learning model.

We followed a consistent annotation procedure for all the data, performed by a single trained herpetologist (MPTG). We used Audacity® 3.2.5 software to visualize the spectrograms and create the labels in steps 2 and 3. We optimized the visualization of the acoustic signals by setting the spectrogram configuration parameters as follows: linear scale for frequency, the maximum frequency of 10 kHz, gain of 20 dB, range 80 dB, FFT algorithm with a window size of 1024, and Hann type, and standard color range to represent sound energy.

**Data Preprocessing.** We framed the species identification problem as a multi-label classification task considering the common occurrence of call overlap in PAM. We applied a set of transformations over the raw audio files and annotations to obtain a dataset suitable for use with ML algorithms. First, reading the metadata of the 1-minute raw audio files, we obtained samples of a 3-second fixed-length window applying a 1-second sliding window. This produced a two-third overlap between samples<sup>10,32</sup>. Second, we assigned a multilabel species label to each sample whenever a portion of a species call appeared within one of these windows. This procedure was applied to all calls, regardless of their quality. Third, we preprocessed each 1-minute annotated audio file using the scikit-maad python package<sup>33</sup> and applied the sliding window approach described above. After trimming the 3-second in time and the frequency limits between 1 Hz and 10000 Hz, we applied a bandwidth filter which uses a bandpass filter to process a 1D signal with an infinite impulse response (IIR) Butterworth filter of order 5. After that, we normalized the audio signal to a maximum amplitude of 0.7 decibel full-scale value (dBFS) and saved it as an uncompressed WAV format. Finally, we selected each 1-minute recording containing weak labels to split the dataset between training and test. We summed the occurrences of all species and applied an iterative stratification for the multi-label setting<sup>34,35</sup> to the unbalanced proportions in the different subsets, with 70% in training and 30% in the test. In this step, we used the 1-minute recording level to avoid data leakage (same 1-minute audio with samples in train and testing subsets).

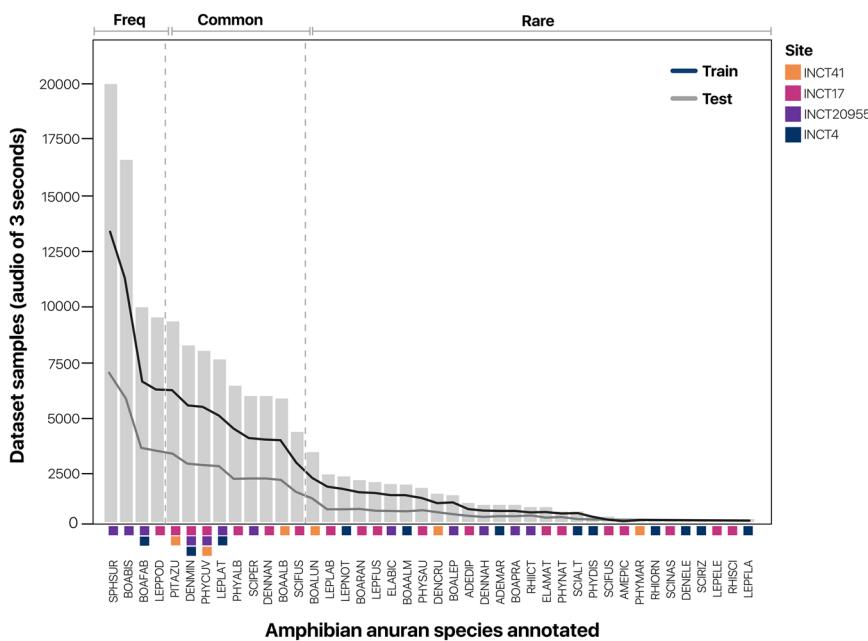


**Fig. 4** An illustrative example of the advertisement call of *Physalaemus albonotatus* for the three audio quality categories. (a) High-quality call (H) shows a high signal-to-noise ratio, no overlap with other sounds, a well-identifiable structure on the spectrogram, and can be easily visualized on the oscillosogram. (b) Medium-quality call (M) can be visually identified on the spectrogram but may overlap with other sounds that can be difficult to identify in the oscillosogram. (c) Low-quality call (L) shows a low signal-to-noise ratio, is partially masked by other sounds, appears with low intensity on the spectrogram, and cannot be easily identified on the oscillosogram.

### Data Records

The dataset and the raw data are provided under the Public Domain Dedication license (CC0) and are deposited in Zenodo<sup>36</sup>. We collected data for 42 neotropical anuran amphibian species from 12 genera and 5 families (Supplementary Table 1). Taxonomic nomenclature followed Frost<sup>37</sup>. A total of 16,000 time boxes equivalent to approximately 31 hours of cumulative duration and 27 hours of human-generated annotations was created, considering all individual or series of calls from these species. It is important to note that due to significant overlap in time boxes among different species, the cumulative duration exceeded the sum of the recording time. Among the collected data, approximately 20% of the 1-minute raw audio files did not contain anuran calls but contained soundscapes with geophonic sources like rain and wind, as well as biophonic sources such as other vocalizing species like insects and birds. The strong labeled annotated data was unevenly distributed across the sites INCT17, INCT20955, INCT41, and INCT4 at 42.5%, 33%, 13.5%, and 11%, respectively. The distribution of samples per species in the final dataset exhibits a long-tailed pattern, which coincides with the typical species diversity pattern in tropical environments (Fig. 5). This reflects the local number of registered species and their vocal activity levels, which depends on the regional pool of species and other contexts regarding the ecology of species. Additionally, we observed a high degree of variability in species composition between sites; specifically, only five species were detected in more than one site.

Here we provide two main data resources: (i) the raw audio files with an associated table containing annotations, and (ii) the preprocessed input dataset for ML with 93378 3-second audio samples, both sharing a similar folder structure. The raw data was divided into separate folders per site. Inside each folder, there is a collection of



**Fig. 5** Frequency distribution of 3-second samples per anuran species. The long-tailed distribution is a typical distribution of a real-world species diversity dataset. We split species into the classes of ‘common’, ‘frequent’, and ‘rare’ to determine the effect of sample size on the performance of the species identification problem. Additionally, the occurrence of the same species in different sites is represented by different colored squares at the bottom of the histogram. The training and test set distributions obtained by using the split strategy are depicted with black and blue lines, respectively.

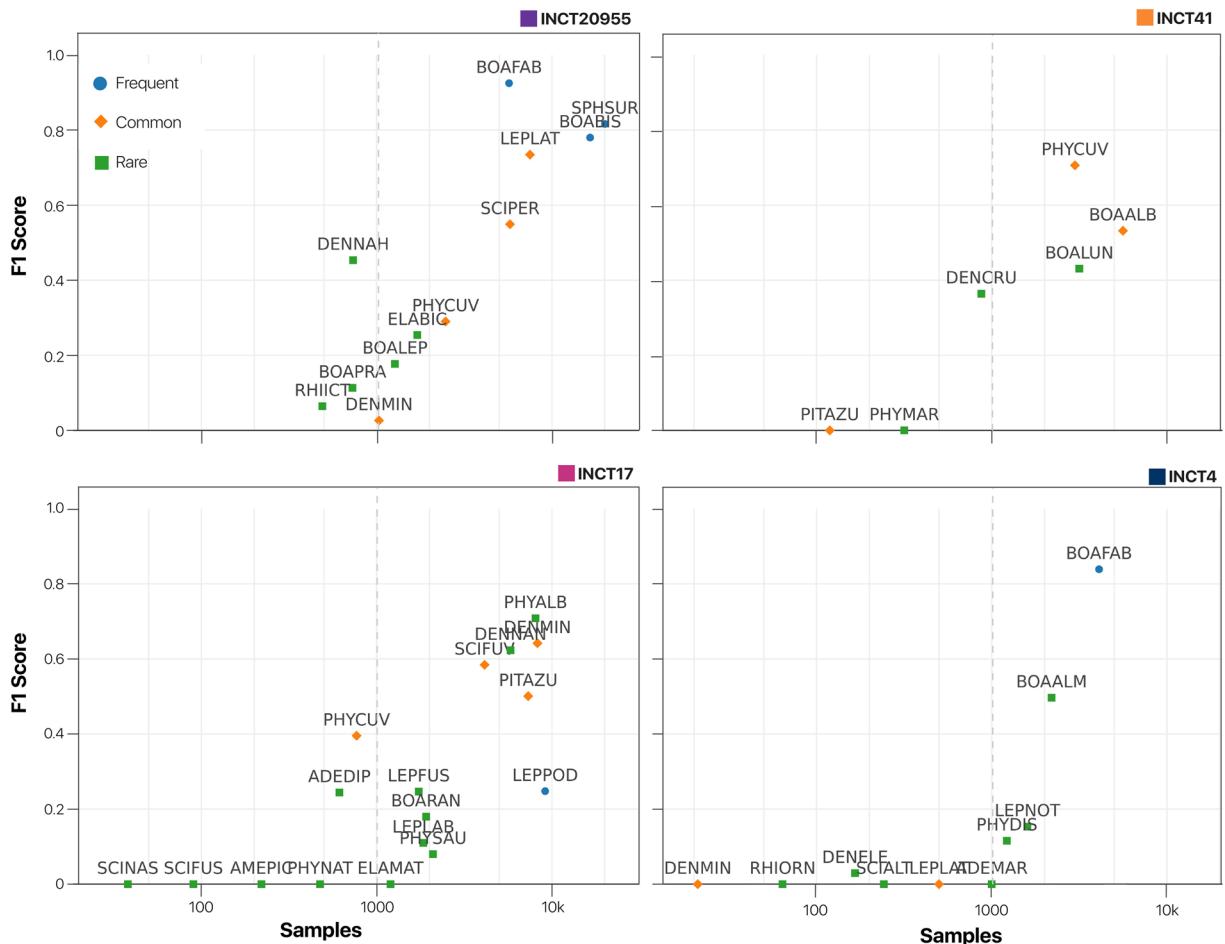
1-minute recordings in WAV format with self-explanatory filenames that include the site name, the date, and the time as follows: {site}—{date}—{time}.wav. For example, the file INCT20955\_20190830\_231500.wav is located in the folder of site INCT20955 and was obtained on 30 August 2019 at 23:15 (BRT time zone). In the same way, the preprocessed dataset follows the same folder and naming structure but also includes the start and final second of the audio segment: {site}—{date}—{time}—{start second}—{final second}.wav. Following the previous example, INCT20955\_20190830\_231500\_30\_33.wav means that the sample starts in the second 30 and ends at the second 33. The dataset folder contains 2 files and one folder containing separate folders per site. The samples are WAV audio files with fixed 3-second lengths, obtained with 22.05 kHz sampling frequency and 16-bit depth. The two other files are a README file describing the structure and construction of the dataset and a metadata CSV file containing the labels for each sample as follows:

- sample\_name: the unique identifier of each sample that corresponds to a unique audio file in the audio folder and follows the structure {site}—{date}—{hour}—{start second}—{final second}.wav. The next 5 columns were constructed based on this column.
- fname: raw audio filename extracted from a site and used by annotators to create weak labels.
- min\_t: second where the annotation starts in a fixed window length.
- max\_t: second where the annotation ends in a fixed window length.
- site: identifier of the recording site.
- date: datetime of the recording.
- subset: training or test subset.
- species\_number: total number of species in each sample. The sum of the next 42 columns per row.
- {species} × 42 Binary columns of each species where 1 if some portion of the call is in the sample, 0 else. The 42 species column names are the codes shown in Supplementary Table 1.

### Technical Validation

**Experimental setup.** The main goal for creating the AnuraSet is to provide a solution for the species identification problem and boost ecological inferences in PAM-based anuran monitoring programs. We frame the species identification problem as a multi-label classification problem using the data from all 4 sites without temporal or site distinction. Following the ecological conditions of the large-scale analysis bioacoustics project<sup>32</sup>, we choose the F1-score as the performance classification metric using the usual 0.5 threshold. For the case of multi-label classification, we selected the Macro version of the F1-score to give the same importance to all species. To better understand the dependency between the number of samples and performance, we grouped species into “Common”, “frequent”, and “rare” categories using the samples frequency similar to the Auto Arborist Dataset<sup>12</sup>. The grouping reflected the label frequency within each anuran assemblage with 2 breakdowns, where species

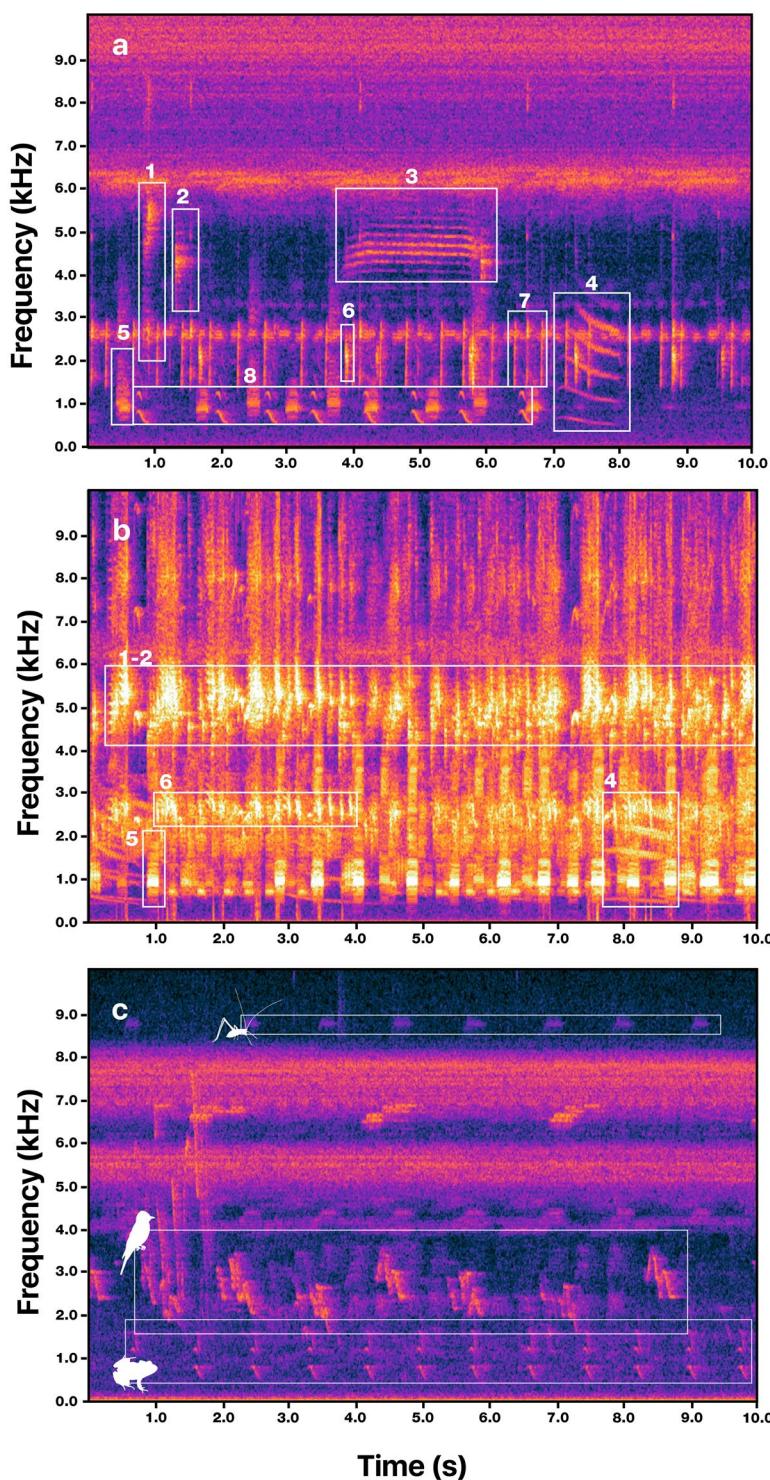
	Macro F1-score (%) ↑		
	ResNet18	ResNet50	ResNet152
Frequent	61.6	62.3	<b>68.4</b>
Common	52.0	53.9	<b>56.8</b>
Rare	14.8	9.9	<b>15.7</b>
All	34.9	33.2	<b>37.8</b>

**Table 2.** Performance of ResNet family models in F1-score percentage using all sites and species.**Fig. 6** Performance for benchmarking the species identification problem. Using the ResNet152 model, we evaluated the species identification problem (see section ‘Experimental Setup’) in each site. The x-axis is the number of samples in the logarithmic scale and the y-axis is the F1-score. Across sites, we found a positive relationship between samples and performance.

with more than 10.000 samples were classified as common species, less than 5.000 samples were classified as rare species, and those between 5.000 and 10.000 samples were classified as frequent species.

**Baseline Models.** Following the pipelines of previous studies<sup>10,32</sup>, we applied a Mel Spectrogram transformation on audio recordings using a window size of 512, a hop length of 28, and the number of mel filter banks of 128. Then we applied SpecAugmentation in time and frequency<sup>38</sup> as spectrogram augmentation strategies and resize. The transformations and augmentations described above generated the inputs in ResNet<sup>39</sup> family models. Specifically, we tested the ResNet18, ResNet50, and ResNet152. All our baseline experiments were implemented using the PyTorch<sup>40</sup> framework and the torchaudio<sup>41</sup> library which are publicly available in the repository <https://github.com/soundclim/anuraset>.

**Benchmark Results.** After testing the ResNet family models, we grouped the performance by species according to their classes of sample frequency (Table 2). The best model in all cases was the ResNet152, with a percentage (%) F1-score of 68.4 for the Frequent group, 56.8 for the Common, and 15.7 for the Rare classes. The total Macro F1-score was 37.8. This result suggested that the number of samples strongly influences the general



**Fig. 7** Analytical challenges of the AnuraSet. **(a)** spectrogram showing eight different species that were recorded calling in less than eight seconds, highlighting the degree of co-occurrences, call overlap, and fine-grained identification; **(b)** spectrogram showing a dense chorus with a low signal-to-noise ratio and high sound masking. **(c)** spectrogram showing the richness and co-occurrence of sounds from different taxa (silhouettes from bottom to top depicting frogs, birds, and orthopterans, respectively).

performance of the models. The F1-score performance of each species in each site is reported in Fig. 6. In this Figure, we confirmed the challenge of *learning from small samples*, which is related to the problem of creating machine learning models using just a few samples for training, for example in Fig. 6 we can see that in less than 1000 samples the algorithms perform percentage F1-score less than 20% in all cases. This problem is still an open research area in deep learning for computational bioacoustics<sup>42</sup>.

## Usage Notes

**Data Challenges and Open Problems.** During the annotation and dataset-building process, we faced challenges inherent to Neotropical, real-world datasets in PAM. We encourage researchers to experiment with the AnuraSet, from heuristics to understand the optimal parameters in preprocessing steps, including augmentation strategies to novel techniques for advancing the anuran call identification problem and other tasks yet to be discovered. With the goal of paving the way for new directions and advancements in ML research for bioacoustics and ecoacoustics, we summarize these challenges in the following topics.

**The devil is in the tails.** As expected, the number of audio samples per species is highly imbalanced (Fig. 5), forming a long-tailed distribution<sup>43</sup>. The characteristics of a large number of categories and small training examples pose a challenge for obtaining good classifiers in all species. As we see in the benchmark results, there was a dependency between the number of samples and performance. This situation is especially relevant when rare species are of interest for ecological and conservation applications. AnuraSet is a suitable dataset to test different methods such as algorithmic solutions<sup>44–46</sup> or augmentation strategies that have been proposed to overcome the long-tailed problem. Furthermore, this problem can be formulated as a *Learning from small samples* problem to explore state-of-the-art approaches<sup>42</sup> like few-shot learners<sup>47,48</sup> or self-supervised learning<sup>49,50</sup>.

**Human Intensive Annotation.** Another manifestation of the *Learning from small samples* challenges happens in the early beginning of the annotation process. As we showed in the annotation protocol section, this is a human labor-demanding process. To scale in rich and large datasets it is necessary to use new ways to annotate data points as have been shown in clever approaches like Auto Arborist Dataset<sup>12</sup>. For example, one possible path is to explore a hybrid approach for human-machine collaboration labeling using an active labeling and learning scheme where each step of the learning procedure is actively assisted by a learning algorithm<sup>51</sup>. Recent work<sup>52</sup> shows that weak labels combined with unsupervised learning approaches can improve the performance of classifiers. The evaluation of such methods on the AnuraSet dataset can facilitate advancements in efficient and scalable annotation techniques.

**Fine-grained audio in natural environments.** Despite a classic dataset such as ImageNet<sup>53</sup>, where the classes can be easily identified for a human, the classes annotated in the AnuraSet rely on the expert knowledge of local herpetologists on sound-based species identification. Additionally, the recordings were collected in complex environments, generating variability in the signal-to-noise ratio of the data due to neotropical soundscape diversity in the different biomes (Fig. 7). We confirm that the presence of calls in noisy conditions is a typical situation encountered in tropical environments investigated by PAM. This kind of problem, which involves distinguishing between subtle differences may imply other approaches<sup>54,55</sup> compared with generic object recognition.

**Multi-label dataset.** Tropical anuran assemblages recorded via PAM exhibit a distinctive feature of dense choruses with high call overlap, comprising different call types. This characteristic often leads to sound masking and makes the identification of individual calls challenging. Species calls in PAM recordings from AnuraSet are highly overlapped, therefore, calls often overlap not only between conspecifics but also between heterospecifics. As Fig. 7a shows, 8 different anuran calls were recorded in less than 8 seconds. This characteristic is unique to PAM data and poses a challenge that is different from other wildlife monitoring sensors like camera trap images. These overlaps are related to the classic problem of the *cocktail party*, in which we try to search for an audio signal of interest like the anuran call, while other species, geophony, and biophony sounds co-occur or overlap with the signal of interest. Recent studies<sup>56–58</sup> show promising progress in the context of bioacoustics.

**Towards abundance and behavior classification.** In the weak labeling process, we go beyond binary presence-absence annotation and use four categories to capture calling activity, similar to the Amphibian Calling Index<sup>31</sup> (Table 1). By mixing this assignment of weak labels with strong labels in the AnuraSet it is possible to work towards call activity classifiers that can measure anuran abundance and behavior in an ecologically meaningful way. These classifiers could help us understand species co-occurrence, temporal patterns of vocal activity, and chorus formation. Measuring abundance in bioacoustics is not straightforward, as it depends on factors such as variability of animal vocalization behavior, overlap, and interference of sounds from different sources. However, the AnuraSet provides a dataset with these properties in a natural and complex environment that will allow the development of new classification techniques that consider sources of error and bias.

## Code availability

The dataset and the raw data are hosted in Zenodo <https://doi.org/10.5281/zenodo.8342596> under the CC0 license<sup>36</sup>. All the code for reproducing the experimental protocol, the building and preprocessing of the dataset, and the use of the baseline model are available in the repository <https://github.com/soundclim/anuraset> under the MIT license. We open the Python code to fast development of new deep learning models and experiments in Pytorch.

Received: 20 June 2023; Accepted: 19 October 2023;

Published online: 06 November 2023

## References

- Urban, M. C. *et al.* Improving the forecast for biodiversity under climate change. *Science* **353**, aad8466 (2016).
- Sugai, L. S. M., Silva, T. S. F., Ribeiro, J. W. Jr & Llusia, D. Terrestrial passive acoustic monitoring: review and perspectives. *BioScience* **69**, 15–25 (2019).
- Gibb, R., Browning, E., Glover-Kapfer, P. & Jones, K. E. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods Ecol. Evol.* **10**, 169–185 (2019).

4. Beery, S. Scaling Biodiversity Monitoring for the Data Age. *XRDS Crossroads ACM Mag. Stud.* **27**, 14–18 (2021).
5. Hardt, M. & Recht, B. *Patterns, predictions, and actions: Foundations of machine learning*. (Princeton University Press, 2022).
6. Rolnick, D. *et al.* Tackling climate change with machine learning. *ACM Comput. Surv.* **55**, 1–96 (2022).
7. Tuia, D. *et al.* Perspectives in machine learning for wildlife conservation. *Nat. Commun.* **13**, 792 (2022).
8. Dufourq, E. *et al.* Automated detection of Hainan gibbon calls for passive acoustic monitoring. *Remote Sens. Ecol. Conserv.* **7**, 475–487 (2021).
9. Luccioni, A. S. & Rolnick, D. Bugs in the Data: How ImageNet Misrepresents Biodiversity. *Proc. AAAI Conference on Artificial Intelligence.* **37**, 14382–14390, <https://doi.org/10.1609/aaai.v37i12.26682> (2023).
10. Van Horn, G. *et al.* Exploring Fine-Grained Audiovisual Categorization with the SSW60 Dataset. in *Proc. ECCV European Conference on Computer Vision.* 271–289, [https://doi.org/10.1007/978-3-031-20074-8\\_16](https://doi.org/10.1007/978-3-031-20074-8_16) (2022).
11. Van Horn, G. *et al.* The inaturalist species classification and detection dataset. in *Proc. IEEE conference on computer vision and pattern recognition.* 8769–8778, <https://doi.org/10.1109/CVPR.2018.00914> (2018).
12. Beery, S. *et al.* The Auto Arborist Dataset: A Large-Scale Benchmark for Multiview Urban Forest Monitoring Under Domain Shift. *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 21294–21307, <https://doi.org/10.1109/CVPR52688.2022.02061> (2022).
13. Kay, J. *et al.* The Caltech Fish Counting Dataset: A Benchmark for Multiple-Object Tracking and Counting. in *Proc. ECCV European Conference on Computer Vision.* 290–311, [https://doi.org/10.1007/978-3-031-20074-8\\_17](https://doi.org/10.1007/978-3-031-20074-8_17) (2022).
14. Beery, S., Van Horn, G. & Perona, P. Recognition in terra incognita. in *Proc. ECCV European conference on computer vision.* 456–473, [https://doi.org/10.1007/978-3-030-01270-0\\_28](https://doi.org/10.1007/978-3-030-01270-0_28) (2018).
15. Lostanlen, V., Salamon, J., Farnsworth, A., Kelling, S. & Bello, J. P. Birdvox-full-night: A dataset and benchmark for avian flight call detection. in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 266–270, <https://doi.org/10.1109/ICASSP2018.8461410> (2018).
16. Chronister, L. M., Rhinehart, T. A., Place, A. & Kitzes, J. An annotated set of audio recordings of Eastern North American birds containing frequency, time, and species information. *Ecology.* **102**, e03329 (2021).
17. Kiskin, I. *et al.* HumBugDB: a large-scale acoustic mosquito dataset. in *Conference on Neural Information Processing Systems 35<sup>th</sup> (NeurIPS) Track on Datasets and Benchmarks.* <https://doi.org/10.48550/arXiv.2110.07607> (2021).
18. Aodha, O. M. *et al.* Towards a General Approach for Bat Echolocation Detection and Classification. Preprint at *bioRxiv*, <https://doi.org/10.1101/2022.12.14.520490> (2022).
19. Prat, Y., Taub, M., Pratt, E. & Yovel, Y. An annotated dataset of Egyptian fruit bat vocalizations across varying contexts and during vocal ontogeny. *Sci. Data.* **4**, 170143 (2017).
20. Hagiwara, M. *et al.* BEANS: The Benchmark of Animal Sounds. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 1–5, <https://doi.org/10.1109/ICASSP49357.2023.10096686> (2022).
21. Baker, E. & Vincent, S. A deafening silence: a lack of data and reproducibility in published bioacoustics research? *Biodivers. Data J.* **7**, e36783 (2019).
22. Ross, S. R.-J. *et al.* Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions. *Funct. Ecol.* **37**, 959–975 (2023).
23. August, T. *et al.* Realising the potential for acoustic monitoring to address environmental policy needs. *JNCC Rep.* N° 707 (2022).
24. Stuart S.N. *et al.* *Threatened amphibians of the world* (Lynx Edicions, 2008).
25. Pyron, R. A. & Wiens, J. J. Large-scale phylogenetic analyses reveal the causes of high tropical amphibian diversity. *Proc. R. Soc. B Biol. Sci.* **280**, 20131622 (2013).
26. Duarte, H. *et al.* Can amphibians take the heat? Vulnerability to climate warming in subtropical and temperate larval amphibian communities. *Glob. Change Biol.* **18**, 412–421 (2012).
27. Narins, P. M. & Feng, A. S. *Hearing and sound communication in amphibians: prologue and prognostication*. (Springer, 2006).
28. Köhler, J. *et al.* The use of bioacoustics in anuran taxonomy: theory, terminology, methods and recommendations for best practice. *Zootaxa.* **4251**, 1–124 (2017).
29. Sugai, L. S. M., Desjonquères, C., Silva, T. S. F. & Llusia, D. A roadmap for survey designs in terrestrial acoustic monitoring. *Remote Sens Ecol Conserv* **6**, 220–235 (2020).
30. Hershey, S. *et al.* The Benefit of Temporally-Strong Labels in Audio Event Classification. in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 366–370, <https://doi.org/10.1109/ICASSP39728.2021.9414579> (2021).
31. Mossman, M. J. & Weir, L. A. North American amphibian monitoring program (NAAMP). in *Amphibian declines.* 307–313, (University of California Press, 2005).
32. Kahl, M. S. S. Identifying birds by sound: large-scale acoustic event recognition for avian activity monitoring. *Wissenschaftliche Schriftenreihe Dissertationen der Medieninformatik Chemnitz University of Technology.* **10**, 2195–2574 (2019).
33. Ulloa, J. S., Haupert, S., Latorre, J. F., Aubin, T. & Sueur, J. scikit-maad: An open-source and modular toolbox for quantitative soundscape analysis in Python. *Methods Ecol. Evol.* **12**, 2334–2340 (2021).
34. Szymański, P. & Kajdanowicz, T. A Network Perspective on Stratification of Multi-Label Data. *Proc. PMLR Machine learning Research.* **74**, 22–35 (2017).
35. Sechidis, K., Tsoumakas, G. & Vlahavas, I. On the Stratification of Multi-label Data. *Machine Learning and Knowledge Discovery in Databases ECML.* **6913**, 145–158, [https://doi.org/10.1007/978-3-642-23808-6\\_10](https://doi.org/10.1007/978-3-642-23808-6_10) (2011).
36. Cañas, J. S. *et al.* AnuraSet: A dataset for benchmarking neotropical anuran calls identification in passive acoustic monitoring. *Zenodo*, <https://doi.org/10.5281/zenodo.8342596> (2023).
37. Frost, D. R. Amphibian Species of the World: an Online Reference. Version 6.2. <https://amphibiansoftheworld.amnh.org/index.php>, <https://doi.org/10.5531/db.vz.0001> (2023).
38. Park, D. S. *et al.* Specaugment: A simple data augmentation method for automatic speech recognition. *Proc. Interspeech* 2613–2617, <https://doi.org/10.21437/Interspeech.2019-2680> (2019).
39. Targ, S., Almeida, D. & Lyman, K. Resnet in resnet: Generalizing residual architectures. Preprint at ArXiv, <https://doi.org/10.48550/arXiv.1603.08029> (2016).
40. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, 8024–8035 (2019).
41. Yang, Y. Y. *et al.* Torchaudio: Building blocks for audio and speech processing. in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 6982–6986, <https://doi.org/10.1109/ICASSP43922.2022.9747236> (2022).
42. Stowell, D. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ.* **10**, e13152 (2022).
43. Van Horn, G. & Perona, P. The devil is in the tails: Fine-grained classification in the wild. Preprint at ArXiv, <https://doi.org/10.48550/arXiv.1709.01450> (2017).
44. Menon, A. K. *et al.* Long-tail learning via logit adjustment. in *Proc. ICLR International Conference on Learning Representations.* <https://doi.org/10.48550/arXiv.2007.07314> (2021).
45. Cui, Y., Jia, M., Lin, T. Y., Song, Y. & Belongie, S. Class-Balanced Loss Based on Effective Number of Samples. in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 9260–9269, <https://doi.org/10.1109/CVPR.2019.00949> (2019).
46. Cao, K., Wei, C., Gaidon, A., Arechiga, N. & Ma, T. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. in. *Adv. Neural Inf. Process. Syst.* **32**, 1567–1578 (2019).

47. Nolasco, I. *et al.* Learning to detect an animal sound from five examples. *Ecological Informatics*. **77**, 102258, <https://doi.org/10.1016/j.ecoinf.2023.102258> (2023).
48. Wang, Y., Bryan, N. J., Cartwright, M., Pablo Bello, J. & Salamon, J. Few-Shot Continual Learning for Audio Classification. in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 321–325, <https://doi.org/10.1109/ICASSP39728.2021.9413584> (2021).
49. Hagiwara, M. AVES: Animal Vocalization Encoder based on Self-Supervision. in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* 1–5, <https://doi.org/10.1109/ICASSP49357.2023.10095642> (2022).
50. Gontier, F. *et al.* Polyphonic training set synthesis improves self-supervised urban sound classification. *J. Acoust. Soc. Am.* **149**, 4309–4326 (2021).
51. Papadopoulos, D. P., Uijlings, J. R. R., Keller, F. & Ferrari, V. We Don't Need No Bounding-Boxes: Training Object Class Detectors Using Only Human Verification. in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 854–863, <https://doi.org/10.1109/CVPR.2016.99> (2016).
52. Michaud, F., Sueur, J., Le Cesse, M. & Haupert, S. Unsupervised classification to improve the quality of a bird song recording dataset. *Ecol. Inform.* **74**, 101952 (2023).
53. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 248–255, <https://doi.org/10.1109/CVPR.2009.5206848> (2009).
54. Cui, Y., Song, Y., Sun, C., Howard, A. & Belongie, S. Large Scale Fine-Grained Categorization and Domain-Specific Transfer Learning. in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4109–4118, <https://doi.org/10.1109/CVPR.2018.00432> (2018).
55. Yang, Z. *et al.* Learning to navigate for fine-grained classification. in *Proc. ECCV European Conference on Computer Vision*. 420–435, [https://doi.org/10.1007/978-3-030-01264-9\\_26](https://doi.org/10.1007/978-3-030-01264-9_26) (2018).
56. Berman, P. C. BioCPPNet: automatic bioacoustic source separation with deep neural networks. *Sci. Rep.* **11**, 23502 (2021).
57. Denton, T., Wisdom, S. & Hershey, J. R. Improving bird classification with unsupervised sound separation. in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*. 636–640, <https://doi.org/10.1109/ICASSP43922.2022.9747202> (2022).
58. Wisdom, S. *et al.* Unsupervised Sound Separation Using Mixture Invariant Training. *Adv. Neural Inf. Process. Syst.* **33**, 3846–3857 (2020).

## Acknowledgements

The authors acknowledge financial support from the intergovernmental Group on Earth Observations (GEO) and Microsoft, under the GEO-Microsoft Planetary Computer Programme (October 2021); São Paulo Research Foundation (FAPESP #2016/25358–3; #2019/18335–5); the National Council for Scientific and Technological Development (CNPq #302834/2020–6; #312338/2021–0, #307599/2021–3); National Institutes for Science and Technology (INCT) in Ecology, Evolution, and Biodiversity Conservation, supported by MCTIC/CNPq (proc. 465610/2014–5), FAPEG (proc. 201810267000023); CNPQ/MCTI/CONFAP-FAPS/PELD No 21/2020 (FAPESC 2021TR386); Comunidad de Madrid (2020-T1/AMB-20636, Atracción de Talento Investigador, Spain) and research projects funded by the European Commission (EAVESTROP–661408, Global Marie S. Curie fellowship, program H2020, EU); and the Ministerio de Economía, Industria y Competitividad (CGL2017–88764-R, MINECO/AEI/FEDER, Spain). We also thank Tom Denton for machine learning evaluation suggestions, dataset revision, and comments on the manuscript.

## Author contributions

J.S.C. and J.S.U. conceived and designed the experiments; J.S.U., L.S.M.S., and D.L. directed the project; J.S.U., L.S.M.S., D.L., H.B., R.P.B., and L.F.T. obtained funding sources; R.P.B., J.V.B., L.F.T., S.D., F.L.S., S.N.O., A.R., V.C.R., C.E.S., and A.H.R.D. participated in the data collection; M.P.T., J.S.U., L.S.M.S., and D.L. designed the annotation protocol; J.V.B., S.D., J.L.M.M.S., and A.R. made the weak labels, M.P.T. made the strong labels; J.S.C., J.S.U., H.B., J.R., and B.P. participated in the benchmark and dataset design; J.S.C. developed the code for dataset building, preprocessing, benchmark, and analysis tools; J.S.C. wrote the original draft; J.S.C., M.P.T., J.S.U., L.S.M.S., D.L., and H.B. wrote and edited the draft; all authors reviewed the final draft.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41597-023-02666-2>.

**Correspondence** and requests for materials should be addressed to J.S.C.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023