



## Learning to detect an animal sound from five examples

Ines Nolasco<sup>c,1</sup>, Shubhr Singh<sup>c,1</sup>, Veronica Morfi<sup>c,1</sup>, Vincent Lostanlen<sup>d,e</sup>, Ariana Strandburg-Peshkin<sup>f,h,g</sup>, Ester Vidaña-Vila<sup>l</sup>, Lisa Gill<sup>j,i</sup>, Hanna Pamuła<sup>k</sup>, Helen Whitehead<sup>m</sup>, Ivan Kiskin<sup>n</sup>, Frants H. Jensen<sup>p,q,r</sup>, Joe Morford<sup>o</sup>, Michael G. Emmerson<sup>c</sup>, Elisabetta Versace<sup>c</sup>, Emily Grout<sup>h,f,g</sup>, Haohe Liu<sup>n</sup>, Burooj Ghani<sup>a</sup>, Dan Stowell<sup>a,b,\*</sup>

<sup>a</sup> Naturalis Biodiversity Center, Leiden, the Netherlands

<sup>b</sup> Tilburg University, Tilburg, the Netherlands

<sup>c</sup> Queen Mary University of London, London, UK

<sup>d</sup> Centre National de la Recherche Scientifique (CNRS), France

<sup>e</sup> Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

<sup>f</sup> Biology Department, University of Konstanz, Universitätsstrasse 10, Konstanz, Germany

<sup>g</sup> Centre for the Advanced Study of Collective Behaviour, University of Konstanz, Universitätsstrasse 10, Konstanz, Germany

<sup>h</sup> Department for the Ecology of Animal Societies, Max Planck Institute of Animal Behavior, Büchelstrasse 5, Konstanz, Germany

<sup>i</sup> Landesbund für Vogel- und Naturschutz, Hilpoltstein, Germany

<sup>j</sup> Naturkundemuseum Bayern/BIOTOPIA Lab, Munich, Germany

<sup>k</sup> AGH University of Science and Technology, al. Adama Mickiewicza 30, Krakow, Poland

<sup>l</sup> La Salle Campus Barcelona, Universitat Ramon Llull, Sant Joan de La Salle, 42, Barcelona, Spain

<sup>m</sup> School of Science, Engineering and Environment, University of Salford, Manchester, UK

<sup>n</sup> Centre for Vision, Speech and Signal Processing, FEPS, University of Surrey, Surrey, UK

<sup>o</sup> The Oxford Navigation group, Dept. of Zoology, University of Oxford, Oxford, UK

<sup>p</sup> Aarhus University, Department of Ecoscience, Frederiksborgvej 399, 4000 Roskilde, Denmark

<sup>q</sup> Syracuse University, 107 College Place, Syracuse, NY 13244, USA

<sup>r</sup> Woods Hole Oceanographic Institution, 266 Woods Hole Rd, Woods Hole, MA 02543, USA

### ARTICLE INFO

#### Keywords:

Bioacoustics

Deep learning

Event detection

Few-shot learning

### ABSTRACT

Automatic detection and classification of animal sounds has many applications in biodiversity monitoring and animal behavior. In the past twenty years, the volume of digitised wildlife sound available has massively increased, and automatic classification through deep learning now shows strong results. However, bioacoustics is not a single task but a vast range of small-scale tasks (such as individual ID, call type, emotional indication) with wide variety in data characteristics, and most bioacoustic tasks do not come with strongly-labelled training data. The standard paradigm of supervised learning, focussed on a single large-scale dataset and/or a generic pre-trained algorithm, is insufficient. In this work we recast bioacoustic sound event detection within the AI framework of *few-shot learning*. We adapt this framework to sound event detection, such that a system can be given the annotated start/end times of as few as 5 events, and can then detect events in long-duration audio—even when the sound category was *not known* at the time of algorithm training. We introduce a collection of open datasets designed to strongly test a system's ability to perform few-shot sound event detections, and we present the results of a public contest to address the task. Our analysis shows that prototypical networks are a very common used strategy and they perform well when enhanced with adaptations for general characteristics of animal sounds. However, systems with high time resolution capabilities perform the best in this challenge. We demonstrate that widely-varying sound event durations are an important factor in performance, as well as non-stationarity, i.e. gradual changes in conditions throughout the duration of a recording. For fine-grained bioacoustic recognition tasks without massive annotated training data, our analysis demonstrate that few-shot sound event detection is a powerful new method, strongly outperforming traditional signal-processing detection methods in the fully automated scenario.

\* Corresponding author at: Tilburg University, Tilburg, the Netherlands.

E-mail address: [d.stowell@tilburguniversity.edu](mailto:d.stowell@tilburguniversity.edu) (D. Stowell).

<sup>1</sup> Equal first authors.

<https://doi.org/10.1016/j.ecoinf.2023.102258>

Received 22 May 2023; Received in revised form 9 August 2023; Accepted 9 August 2023

Available online 10 August 2023

1574-9541/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Machine listening, defined as the application of machine learning to audio content analysis, has untapped potential in the life sciences, applied to animal vocalisations. Because animal vocalisations vary systematically across species, across social/environmental/emotional contexts, and across individuals (Brown and Riede, 2017; Marler and Slabbekoorn, 2004), machine listening has the potential to provide crucial information on animal populations and communities as well as on individuals and their behavioral states. Hence, automated detection and analysis of animal vocalisations is not only valuable for our understanding of sound production but also for diverse research fields including animal behavior, animal welfare, neuroscience and ecology (Caiger et al., 2020; Gillespie et al., 2009; Gillings and Scott, 2021; Riede, 2018). Recent advances in consumer electronics have considerably lowered the cost and weight of digital audio acquisition, thus allowing deployment of autonomous recording units at large spatio-temporal scales (Hill et al., 2018; Roe et al., 2021; Sethi et al., 2020). However, massively distributed bioacoustic surveys have resulted in a “data deluge”, where data collection outgrows information management. This issue is not limited to scientific research, where audio corpora serve to conduct statistical hypothesis testing. Difficulties in handling, analysing and interpreting large amounts of data also extend to applied fields in which animals can be monitored using sound: farming, conservation, and wind energy, to name a few.

Since the beginning of the 21st century, the need for large-scale analyses of animal sounds has spurred the emergence of “computational bioacoustics” approaches, complementary to human surveys (Stowell, 2018). Methods have often been inspired by developments in neighbouring subfields of machine listening—music information retrieval and speech technology—as well as by computer vision. In this regard, the breakthrough of deep learning in automatic speech recognition, around the year 2012, has profoundly influenced the orientation of computational bioacoustics research (Hinton et al., 2012). In particular, most deep learning systems for bioacoustics are trained as sound event *classifiers*: given a short audio excerpt, usually of constant duration, they return an element within a predefined class. This approach is derived from the “phone classification task” used in speech analysis, with animal vocalisations in lieu of human utterances, and a species-specific catalogue in lieu of a phonetic alphabet (Ganchev, 2017).

However, the paradigm of supervised sound event classification based on speech is reaching its limits in computational bioacoustics. Indeed, the extrapolation between speech to other animal sounds is difficult and limited, due to differences in sound duration and units of interest, differences in context and taxonomy, as well as differences in recording conditions, among others. First, detecting the start and end time of animal sounds has a key role in community ecology, since so much of the structure lies in call-and-response and other patterns of influence (Logue and Krupp, 2016; Stowell et al., 2016). Secondly, bioacoustic practitioners operate at many different levels of granularity, from coarse (e.g., species classification) to fine (e.g., distinguishing call types or syllables from one individual); whereas speech science relies on limited levels of granularity where human phonemes or words are the fundamental units. Thirdly, non-human animal sounds are acquired with a plethora of diverse equipment, including far-field, on-body, and underwater, whereas speech sounds are typically acquired with an individual device, that is usually controlled by the person speaking.

A main limitation in bioacoustics is the lack of a unified framework that can be applied to different vocalisations. Today, the literature on computational bioacoustics is fragmented into subdomains: marine versus terrestrial, individual versus species identification, handheld versus fixed equipment, and so forth (Frazao et al., 2020; Kahl et al., 2020; Linhart et al., 2022). Overall, all these subdomains share a common definition of what constitutes a “sound event”: i.e., a recognisable auditory perception with an onset and offset. However, the spectrotemporal characteristics underlying these events vary dramatically

across species and domains. Thus, bioacoustic event detection does not appear as a single “big-data” problem; but instead, as a juxtaposition of many small-data problems, each currently addressed by specialised systems. The field benefits from the common coarse-scale task of species classification, which has provided a clear and useful focus to drive computational bioacoustics into the deep learning era (Joly et al., 2019; Kahl et al., 2021). Yet, systems trained for coarse-scale tasks, even with massive data, do not automatically acquire the ability to make fine-grained or local distinctions, and must be further trained or customised (Lauha et al., 2022; Van Horn et al., 2021). Thus, much recent work (re)trains deep learning systems anew for each specific new task.

Such fragmentation hinders the practical usability of deep learning in bioacoustics, and thus in the life sciences at large. Indeed to date, the success of deep neural networks in the supervised regime depends on the availability of a massive corpus of audio examples for the sound events of interest, paired with human annotations. Yet, temporally-precise and fine-grained annotation of audio demands expertise, and is thus costly and time-consuming. In many cases, the obstacle is not only to acquire annotations, but also the audio examples themselves: e.g. for rare species, remote locations, or costly equipment. Furthermore, these numerous small-data scenarios remain outside the scope of digital bioacoustic archives, such as Xeno-Canto and the Macaulay Library.

In this article, we aim to guide the development of a unified method that works across the many subdomains of computational bioacoustic sound event detection (SED). The benefit of doing so resides in the development of a robust and versatile system that could serve the scientific community at large. Hence, we assembled a collection of 14 small-scale datasets, between 10 min and 10 h in duration. Each of them reflects a genuine but slightly different application setting and are obtained from completely different sources. The main originality of our work is in the proposal that, instead of training 14 individual machine listening systems (one per dataset), we train a single system to detect sound events on many different datasets, in which each dataset has a different category of sound event to be detected—that category only defined at “query time”. Furthermore, when being evaluated on an audio file, the system is prompted with the first five occurrences of the sound event of interest. This paradigm of machine learning is known as “few-shot learning” (FSL) (Snell et al., 2017; Wang et al., 2020a).

Stated otherwise, our hypothesis is that bioacoustic event detectors can take advantage of whichever bioacoustic datasets are available at training time, and then generalise from a few (five) examples of the new target at deployment time. This is difficult under a standard supervised paradigm because the training set may not reflect real-world deployment conditions, nor cover all sound categories of possible interest. For these reasons, we place the concept of domain adaptation at the heart of the few-shot learning paradigm in bioacoustics: our goal is not only to learn a detector from limited labelled data but also to learn domain-agnostic representations of animal sounds which can readily adapt to unforeseen recording conditions (cf. Beery et al. (2018) in computer vision).

In order to diversify methods and accelerate progress, we have organized an open-science challenge for a community of researchers named DCASE: “Detection and Classification of Acoustic Scenes and Events”. The challenge was open to everyone and consisted of public datasets, evaluation metrics, documentation, and baseline systems.

In this paper we formulate bioacoustic sound event detection (SED) as a few-shot learning task. We describe our approach in the form of two ML systems customised to this new task (published openly as baseline methods), and we report on a public data challenge conducted over three years to generate and evaluate novel algorithmic solutions.<sup>2</sup> We evaluate various dimensions of the ML paradigms that have been put

<sup>2</sup> Development data: <https://zenodo.org/record/6482837> Evaluation data: <https://zenodo.org/record/6517414> Code: <https://github.com/c4dm/dcase-few-shot-bioacoustic/>

forward for this task, and explore their ability to adapt to aspects of bioacoustic data presented in our datasets. Our study demonstrates that few-shot SED is a feasible way forward in bioacoustics.

### 1.1. Related work

Few-shot *classification* has been investigated generally, and also for audio (acoustic) data (Naranjo-Alcazar et al., 2022; Pons et al., 2019; Shi et al., 2020; Snell et al., 2017). However, SED has different requirements from classification: typically, the desired output includes the onset and offset times for each detected event (Mesaros et al., 2016), roughly similar to the “object detection” task in computer vision.

One important insight behind few-shot learning is that of *meta-learning* (“learning to learn”), or the idea of leveraging past experience to speed-up new learning by improving the performance of the learner (Schaul and Schmidhuber, 2010). One approach to meta-learning is training a system across many loosely-similar tasks/datasets, such that the system is then well-configured to generalise from a few examples of a novel class (Ravi and Larochelle, 2017; Wang et al., 2020a). This depends on a system learning something of the implicit commonalities and analogies across the tasks, which might then influence an algorithm’s learnt feature extraction, or its measure of similarity between data points, for example. Related work in computer vision explores the challenge of fine-grained classification and object detection in images from camera traps in novel conditions (Beery et al., 2018).

Van Horn et al. (2021) introduced a wildlife image dataset with multiple subsets each defining a different binary question. This has a similar meta-learning spirit as our work, with the aim that a sophisticated first stage of “representation learning” across multiple tasks can make future tasks simple. However, unlike the few-shot setting that we use, a lightweight (shallow) classifier must be trained for each new question from a non-trivial number of positive/negative examples.

A previous data-driven challenge on animal vocalisation audio detection was focused on birds (Stowell et al., 2019). That challenge also aimed to generalise robustly to conditions not seen in the training data but was simpler than ours, in that it did not require systems to predict event onset or offset times, only presence/absence; it stayed within the framework of supervised classification rather than generalising from examples of new categories; and it didn’t include as broad a range of animal taxa.

### 1.2. Novel approaches

Deep learning models for few-shot learning problems can be broadly categorized into two approaches: meta-learning and transfer learning. Prototypical networks and matching networks (Vinyals et al., 2016), are good examples of meta-learning that have performed well in few-shot learning tasks across both image and audio domain.

Meta-learning based methods rely on the assumption that the tasks belong to a single distribution, for example metric learning based methods require the tasks all coming from a similar domain such that there exists a uniform metric that could work across tasks (Wang et al., 2019a). However, in real world scenarios this assumption does not always hold such as in case of our task where the datasets vary in terms of species, recording conditions and microphones, essentially rendering the problem as a cross-domain few-shot learning. In such cases, a hybrid meta-learning approach towards the task may be required, which moves beyond the assumption that future tasks are well-represented by the set of training tasks. A few hybrid methods are as follows:

- **Cross-domain few-shot learning** - Very few methods specifically designed to account for cross-domain scenarios have been previously explored. Feature-wise transformation layers were introduced in Tseng et al. (2020) for augmenting the features using affine transforms, in order to adapt to domain shift across tasks. In Dong and

Xing (2019), an adversarial network based model is used for one-shot domain adaption from source to target domain.

- **Transductive few-shot learning** - Meta learning methods aim to learn on scarce data in order to generalise to unseen tasks, which makes the problem fundamentally difficult. In order to mitigate the difficulty, transductive based methods utilise *the information present in the unlabeled examples from the query set* to adapt the model and improve its predictions. In Liu et al. (2018), the samples in support and query set are jointly modelled as nodes of a graph and the prediction on query set is conducted by label-propagation algorithm. In Hou et al. (2019), a cross-attention based map is learnt between support set and query set in order to make predictions on individual query examples.

Alternatively, transfer learning based methods rely on adapting to a new task through the transfer of knowledge from a related task that has already been learned (Parnami and Lee, 2022). First, a deep learning model is trained on large training set of base class and then fine-tuned on a few examples of the novel class. Fine-tuning on a few examples of the novel class can often lead to poor generalisation, hence techniques have to be adopted in order to avoid overfitting. For example, in Wang et al. (2021), a dynamic few-shot learning approach is adopted where an auxiliary model is used as a few-shot classification “weight generator” which uses an attention map between the existing classification weight vector of the base classes and the few-shot examples of the novel classes. SimpleShot (Wang et al., 2019b) uses a pretrained deep network to get feature embeddings for the input and query set and performs L2 normalization on the obtained features, subsequently, an Euclidean distance based nearest neighbour classification is performed. A similar approach with cosine-distance was proposed in Chen et al. (2020).

Through the outcomes of the public challenge, we evaluate some combinations of these novel approaches for the particular domain of bioacoustic SED.

## 2. Method

### 2.1. Task formulation

We formulate few-shot bioacoustic sound event detection (FSED) as follows:

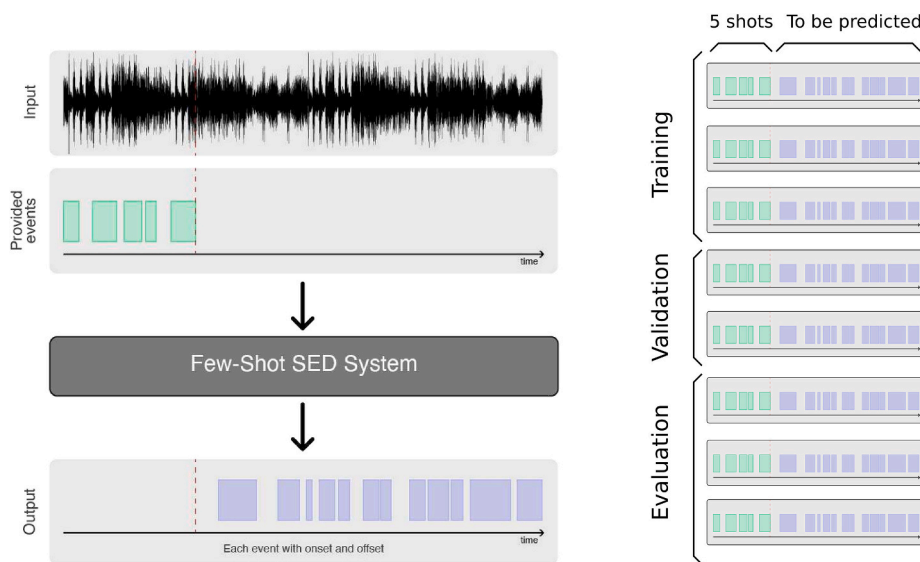
*Given one long audio recording (or multiple audio recordings), as well as annotations on the onset and offset time for each of the first five sound events of interest, identify the onset and offset times for all other such sound events in the recording(s) (Fig. 1a).*

To train a system for this using meta-learning, we make use of multiple bioacoustic datasets (Fig. 1b) representing a range of taxa and recording conditions, each annotated with a different target sound category (see next section).

Note that we do not consider multiple classes in one dataset (Mesaros et al., 2019; Naranjo-Alcazar et al., 2022): each dataset represents a single-class problem. Other sounds are undoubtedly present in almost all audio recordings, but these are considered to be background noise (clutter/distractor events). Our formulation is easily extended to multiple classes in a scene by applying inference separately for each category of interest.

We choose few-shot rather than one-shot learning because animal sounds of interest often cover a range of variability: for example, there may be multiple call types in the set of sounds of interest, or calls from multiple distinct individuals within a group. Five as the number of examples is a conventional choice in few-shot learning, but could vary (Pons et al., 2019; Ravi and Larochelle, 2017; Snell et al., 2017).

Note also that we choose to use the *first* occurring events as examples, rather than a randomly-selected set. This reflects typical practice in bioacoustics, in that acoustic data are typically labelled in contiguous time segments which may or may not be fully representative of the entire data set, and should be tractable for future users of few-shot



**Fig. 1.** (a) Few-shot sound event detection: the first 5 sound events are given as examples—in standard supervised learning they would be considered the training set—and the remainder must then be detected. (b) Few-shot sound event detection as a *meta-learning* problem. Each of our datasets represents a different but related few-shot task. The overall goal is to use the training and validation datasets collectively to train or otherwise develop a system that, when presented with 5 sound events from any of the evaluation datasets, can perform well at detecting the remaining events.

acoustic systems. It offers one benefit, that an algorithm may make use of the strong assumption that the periods between the first five examples fall into the negative class. It also aligns with common scenarios, such as manually labelling data during a pilot phase and then deploying a recognition system to automatically label incoming data. However, it also presents a risk: the first sequence of examples may be similar to each other in some way which is not representative of the whole set of events, for example if the acoustic environment or animal behaviours exhibit non stationary characteristics but change over time.

## 2.2. Datasets

A conventional machine learning experiment uses a single dataset partitioned into three subsets, used for training, validation, and evaluation (test). In the few-shot formulation, we also divide the data into these three partitions (training/validation/evaluation), but each in fact contains *multiple datasets*, and each dataset represents one example of a few-shot task. Within each dataset, there are one or more audio files, each accompanied by a CSV text file giving the start time, end time and label of the targeted audio events. The label can be POS for a positive example of the target class, NEG for a negative example (background or non-target sound event), or UNK for unknown cases, where the human annotator(s) were unsure whether a sound event should be considered in the positive class. Such UNK cases may often occur in complex wildlife sound scenes; our chosen strategy was to explicitly label these cases, allowing algorithm designers to make their own decisions on how to handle them, but to exclude UNK time-regions from the evaluation measures (described later) since their correct label is ambiguous. For each dataset, the first five POS events are the “few shots” from which the rest should be inferred.

A *development set* was provided for the task when the challenge was launched, consisting of the predefined training and validation sets to be used for system development. The development set consists of datasets from multiple sources with audio recordings and associated reference annotations in our specified format. More specifically, for the training set multi-class temporal annotations were provided for each recording (with multiple POS/NEG/UNK columns in the data, one per class), while for the validation set single-class temporal annotations (POS/UNK) were provided for each recording.

A separate *evaluation set* was kept for evaluating the performance of the systems. During the task, only the five POS event annotations were provided for each of the recordings for the class of interest. The developed systems had to use those five annotated events and then learn to

detect the same type of events throughout the rest of the recording. The true annotations for the rest of the recording were kept private for evaluation.

**Table 1:** Information on each dataset used in the 2022 challenge. “Density” is calculated as in signal processing: (total duration of events) / (total duration of audio), thus values close to 0 are sparse, and close to 1 are dense.

**BirdVox-DCASE-10 h (BV-Training set):** The BV dataset was produced as part of the BirdVox project Website of the BirdVox project: WebsiteoftheBirdVoxproject:\url{https://wp.nyu.edu/birdvox}, whose goal is to monitor bird migration with autonomous recording units (Lostanlen et al., 2018). The recordings were obtained in four locations of Tompkins County, NY, US, during the 2015 fall migration season. An expert ornithologist, Andrew Farnsworth, has annotated 2662 flight calls from 11 species of passerines, e.g., Swainson’s thrush (*Catharus ustulatus*) and White-throated sparrow (*Zonotrichia albicollis*). These flight calls have a duration in the range 50–150 milliseconds and a fundamental frequency in the range 2–10 kHz.

**Hyenas (HT-Training set):** The HT dataset contains five recordings from hyenas. Spotted hyena vocalisation data were recorded on custom-developed audio tags (DTAG) designed by Mark Johnson and integrated into combined GPS/acoustic collars (Followit Sweden AB) by Frants Jensen and Mark Johnson, Johnson and Tyack (2003b). Collars were deployed on female hyenas of the Talek West hyena clan at the MSU-Mara Hyena Project (directed by Kay Holekamp) in the Masai Mara, Kenya as part of a multi-species study on communication and collective behavior. Spotted hyenas are a highly social species that live in “fission-fusion” groups where group members range alone or in smaller subgroups that split and merge over time. Hyenas use a variety of types of vocalisations to coordinate with one another over both short and long distances (Lehmann, 2020). Recordings used as part of this task contain a variety of different vocalisations which were identified and classified into types based on the established hyena vocal repertoire (Leblond et al., 2021). Fieldwork was carried out from November 2016–February 2017 by Kay Holekamp, Andrew Gersick, Frants Jensen, Ariana Strandburg-Peshkin, Benson Pion, Morgan Lucot, and Rebecca LeFleur; labelling was done by Kenna Lehmann and colleagues.

**Meerkats (MT-Training set, ME-Validation set):** The MT and ME datasets contains two recordings each from meerkats. Recordings used in this task were acquired at the Kalahari Meerkat Project (Kuruman River Reserve, South Africa; directed by Marta Manser and Tim Clutton-Brock), as part of a multi-species study on communication and collective behavior. Recordings of the development set (MT) were recorded on



**Table 1**

Overview of all the datasets in the development and evaluation sets used in the 2022 edition of the challenge.

| 2022       | Dataset | Taxon            | Mic type | # Files | Total duration | # Labels | # Events | Density | Mean event length (sec) |
|------------|---------|------------------|----------|---------|----------------|----------|----------|---------|-------------------------|
| Training   | BV      | Birds            | fixed    | 5       | 10 h           | 11       | 9026     | 0.038   | 0.15                    |
|            | HT      | Mammals          | on-body  | 5       | 5 h            | 5        | 611      | 0.047   | 1.42                    |
|            | MT      | Mammals          | on-body  | 2       | 70 mins        | 4        | 1294     | 0.042   | 0.15                    |
|            | JD      | Birds            | on-body  | 1       | 10 mins        | 1        | 357      | 0.062   | 0.11                    |
| Val.       | WMW     | Birds            | various  | 161     | 5 h            | 26       | 2941     | 0.25    | 1.54                    |
|            | HB      | Insects          | handheld | 10      | 2.4 h          | 1        | 712      | 0.67    | 11.67                   |
|            | PB      | Birds            | fixed    | 6       | 3 h            | 2        | 292      | 0.003   | 0.11                    |
|            | ME      | Mammals          | handheld | 2       | 20 mins        | 2        | 73       | 0.011   | 0.19                    |
| Evaluation | CHE     | Birds            | fixed    | 18      | 3 h            | 3        | 2550     | 0.263   | 1.94                    |
|            | DC      | Birds            | fixed    | 10      | 95 mins        | 3        | 967      | 0.350   | 1.66                    |
|            | CT      | Mammals          | on-body  | 3       | 48 mins        | 3        | 365      | 0.017   | 0.16                    |
|            | MS      | Birds            | fixed    | 4       | 40 mins        | 1        | 1087     | 0.084   | 0.18                    |
|            | QU      | Mammals (marine) | on-body  | 8       | 74 mins        | 1        | 3441     | 0.045   | 0.06                    |
|            | MGE     | Birds            | fixed    | 3       | 32 mins        | 2        | 1195     | 0.194   | 0.27                    |

small audio devices (TS Market, Edic Mini Tiny+ A77, 8 kHz) integrated into combined GPS/audio collars which were deployed on multiple members of meerkat groups to monitor their movements and vocalisations. Recordings of the evaluation set (ME) were recorded by an observer following a focal meerkat with a Sennheiser ME66 directional microphone from a distance of typically less than 1 m. Meerkats are a highly social mongoose species that live in stable social groups and use a variety of distinct vocalisations to communicate and coordinate with one another. Recordings were carried out during daytime hours while meerkats were primarily foraging and include several different call types. The meerkat vocal repertoire has been well characterised based on previous research, allowing calls to be reliably classified by human labellers (Manser, 1998; Manser et al., 2014). Fieldwork was carried out by Ariana Strandburg-Peshkin, Baptiste Averly, Vlad Demartsev, Gabriella Gall, Rebecca Schaefer and Marta Manser; and the recordings were labelled by Baptiste Averly, Vlad Demartsev, Ariana Strandburg-Peshkin, and colleagues.

**Jackdaws (JD-Training set):** The JD dataset contains a 10-min on-bird sound recording of one male jackdaw during the breeding season in 2015. This individual was recorded in a larger multi-year field study (Max-Planck-Institute for Ornithology, Seewiesen, Germany) in which wild jackdaws were equipped with small backpacks containing miniature voice recorders (Edic Mini Tiny A31, TS-Market Ltd., Russia) to investigate the vocal behavior of individuals interacting with their group and behaving freely in their natural environment. Jackdaws are highly vocal corvid songbirds that usually breed, forage and sleep in large groups, but form a pair bond with the same partner for life. The sound recordings contain loud recordings of the focal bird, as well as background sounds from non-focal birds and other sound sources. Fieldwork was conducted by Lisa Gill, Magdalena Pelayo van Buuren and Magdalena Maier. Sound files were manually annotated by Lisa Gill, using Audacity software, following a previously established video-validation in a captive setting (Stowell et al., 2017).

**Western Mediterranean Wetlands Bird Dataset (WMW-Training set):** The WMW dataset contains 161 files with bird sounds from 20 endemic species that are typically inhabitants of the “Aiguamolls de l’Empordà” natural park in Girona, Spain. The audio files that compose this dataset were originally retrieved from the Xeno-Canto portal (Velting and Planqué, 2015). Xeno-Canto is a portal in which citizens can upload wildlife sounds. As the audio files are collected by a wide community of people, the recording devices used for gathering data can be different in every audio file. Depending on the species, audios contain vocalisations such as bird calls or songs; or sounds such as bill clapping (*Ciconia ciconia* species) or drumming (*Dendrocopos minor* species). For the WMW dataset, Juan Gómez-Gómez, Ester Vidaña-Vila and Xavier Sevillano manually cleaned and labelled downloaded audio files using the Audacity software (Gómez-Gómez et al., 2023). The cleaning and labelling process consisted in listening to every audio file in the dataset and annotating the specific parts of the file in which the bird is

vocalizing, thus separating the bird vocalisations from background noise.

**HumBug (HB-Validation set):** The HB dataset contains sounds of lab-cultured *Culex quinquefasciatus* mosquitoes from Oxford, UK, and various species captured in the wild in Thailand, placed into plastic cups (Li et al., 2018). Mosquitoes produce sound both as a by-product of their flight and as a means for communication and mating. Fundamental frequencies vary in the range of 150 to 750 Hz (Kiskin et al., 2020). As part of the HumBug project, acoustic data was recorded with a high specification field microphone (Telinga EM-23) coupled with an Olympus LS-14. The recordings used in this challenge are a subset of the datasets marked as ‘OxZoology’ and ‘Thailand’ from HumBugDB (Kiskin et al., 2021).<sup>3</sup>

**Polish Baltic Sea bird flight calls (PB-Validation set):** The PB dataset consists of six 30-min recordings of bird flight calls recorded along the Polish Baltic Sea coast (Daąbrowice near Darłowo). Three autonomous recording units were used with the same hardware settings (Song Meters SM2, Wildlife Acoustics, Inc). They were deployed close to each other (<100m) - near the lake, on the dune, and on the forest clearing - to provide diverse acoustic background. The recordings were acquired during the 2016, 2017 and 2018 fall migration seasons. The recordings are the excerpt from Hanna Pamula’s project, focused on the acoustic monitoring of birds migrating at night along the Polish Baltic Sea coast (Pamula, 2022; Pamula, 2022). The passerines night flight calls were annotated by Hanna Pamula using the Audacity software. In each recording, only one bird species is the target class: song thrush, *Turdus philomelos* (3 recordings); blackbird, *Turdus merula* (3 recordings). The usual fundamental frequency range for calls of the chosen species is 5–9 kHz, with standard call duration in the range of 10–250 milliseconds.

**Transfer-Exposure-Effects dataset (CHE-Evaluation set):** The CHE dataset contains bird vocalisations from the Chornobyl Exclusion Zone (CEZ). Data were collected using unattended acoustic recorders (Songmeter 3) to capture the Chornobyl soundscape and investigate the longterm effects of the nuclear power plant accident on the local ecology. This dataset comes from the Transfer Exposure-Effects (TREE) research project.<sup>4</sup> To date, the study has captured approximately 10,000 h of audio from the CEZ. The fieldwork was designed and undertaken by Mike Wood (University of Salford), Nick Beresford (UK Centre for Ecology & Hydrology) and Sergey Gashchak (Chornobyl Center). Common Chiffchaff (*Phylloscopus collybita*) and Common Cuckoo (*Cuculus canorus*) vocalisations were manually annotated and labelled from these recordings by Helen Whitehead.

**BIOTOPIA Dawn Chorus (DC-Evaluation set):** The DC dataset used as part of the evaluation set stems from dawn chorus recordings, made

<sup>3</sup> <https://github.com/HumBug-Mosquito/HumBugDB/>

<sup>4</sup> <https://tree.ceh.ac.uk/>

using Zoom H2 recorders at three different locations in Southern Germany (Haspelmoor, Munich's Nymphenburg Schlosspark, and Natesbuch). Many bird species produce vocalisations during the entire day, but their vocally most active period by far usually occurs around dawn. This natural phenomenon of *dawn chorus* has received a lot of attention in biological studies, and also appears to be the perfect time window for species detection, as it provides the largest likelihood of most individuals of the same and of different species signaling. Yet the sheer complexity of undirected dawn chorus recordings have made automatic species classification extremely difficult, leaving this potentially rich source of acoustically determined species data largely untapped. The Dawn Chorus project is a worldwide citizen science and arts project bringing together amateurs and experts to experience and record the dawn chorus at their doorstep, to draw a global picture of bird biodiversity through sound. The three recordings used in the present study were made and donated by two participants (Moritz Hertel and Rudi Schleich). The vocalisations of three target species (Common cuckoo, *Cuculus canorus*; European robin, *Erithacus rubecula*; Eurasian wren, *Troglodytes troglodytes*) were manually annotated by Lisa Gill, using Audacity.

**Coati (CT-Evaluation set):** The CT dataset contains audio recordings collected from two adult females from the same group on Barro Colorado Island, Panama in March 2020. These data are part of the Communication and Coordination Across Scales project. The two coatis wore collars which recorded high-resolution GPS data with an external attachment of a small audio recording device (TS Market, Edic Mini Tiny+ A77, 22.05 KHz). Audio data were recorded during their active foraging period in daytime hours when a variety of social and aggressive calls are commonly emitted. Coatis are omnivorous diurnal mammals that live in stable social groups consisting of females and related juvenile and subadult males. Coatis produce a number of call types that are used across a variety of different behavioral contexts. The documentation of their complete vocal repertoire is currently under development. The target calls used in this dataset are growls, chitters and chirp-grunts. Growls and chitters are used in aggressive contexts, whereas chirp-grunts are contact calls emitted when foraging and moving with the group. Several other call types that might be confused with the targets were captured in the recordings which present the main challenging aspect of this data. Fieldwork was carried out by Emily Grout, Josué Ortega and Ben Hirsch. Calls were labelled by Emily Grout using Adobe Audition.

**Manx Shearwater (MS-Evaluation set):** The MS dataset contains vocalisations from Manx Shearwater individuals, which are procellariiform seabirds that breed in dense island colonies in the North Atlantic, mostly in the British Isles, and winter in the South Atlantic off the South American coast. In a multi-year study, Audiomoth recorders were placed in burrows on Skomer Island to record the vocalisations of both adult Manx shearwaters and chicks during the breeding season. Adult Manx shearwaters make loud, distinctive vocalisations while present at their breeding colony in various contexts: in duets with their partner in their nesting burrow, to broadcast from their burrow, and during flight. Pairs of Manx shearwaters raise single chicks in underground burrows, regularly visiting the breeding colony at night to feed their chick. During these visits, the chick vocalises to 'beg' for food from the parent shearwater; these vocalisations typically comprise bouts of short high-pitched 'peeps'. Fieldwork was undertaken by various members of the Oxford Navigation Group (OxNav), associated with the Oxford University Department of Biology and led by Professor Tim Guilford. Annotation of individual chick begging vocalisations was carried out by Joe Morford using Sonic Visualiser; these vocalisations, therefore, represent the target class in this dataset.

**Dolphin Quacks (QU-Evaluation set):** The QU dataset contains recordings from bottlenose dolphin sounds from Sarasota, FL, obtained using sound-and-movement recording DTAGs Johnson and Tyack (2003a), attached with suction cups by Frants Jensen in collaboration with Drs. Peter Tyack, Vincent Janik, Laela Sayigh, Randall Wells and

the Sarasota Dolphin Research Program. All tags were deployed during routine health assessments conducted by the Sarasota dolphin research project and under a National Marine Fisheries Service research permit to Dr. Randall Wells of Chicago Zoological Society. Bottlenose dolphins are highly acoustic animals with an expansive repertoire of acoustic signals used for social interactions. Male bottlenose dolphins (*Tursiops truncatus*) in Sarasota form close pair bonds with other males that help them consort with females during the mating season. The target class is Quacks, which are short, low-frequency narrowband signals (around 100 ms duration and main energy below a few kHz) Simard et al. (2011), and emitted at relatively high rates by one or both males in the alliance. These calls are produced in bouts often with hundreds of quacks in a single short vocal bout. Bouts of quacks were extracted from 4 bottlenose dolphins tagged in 2013, 2014, and 2015. Quacks were labelled by Austin Dziki and validated by Frants Jensen using DTAG auditing tools in Matlab.

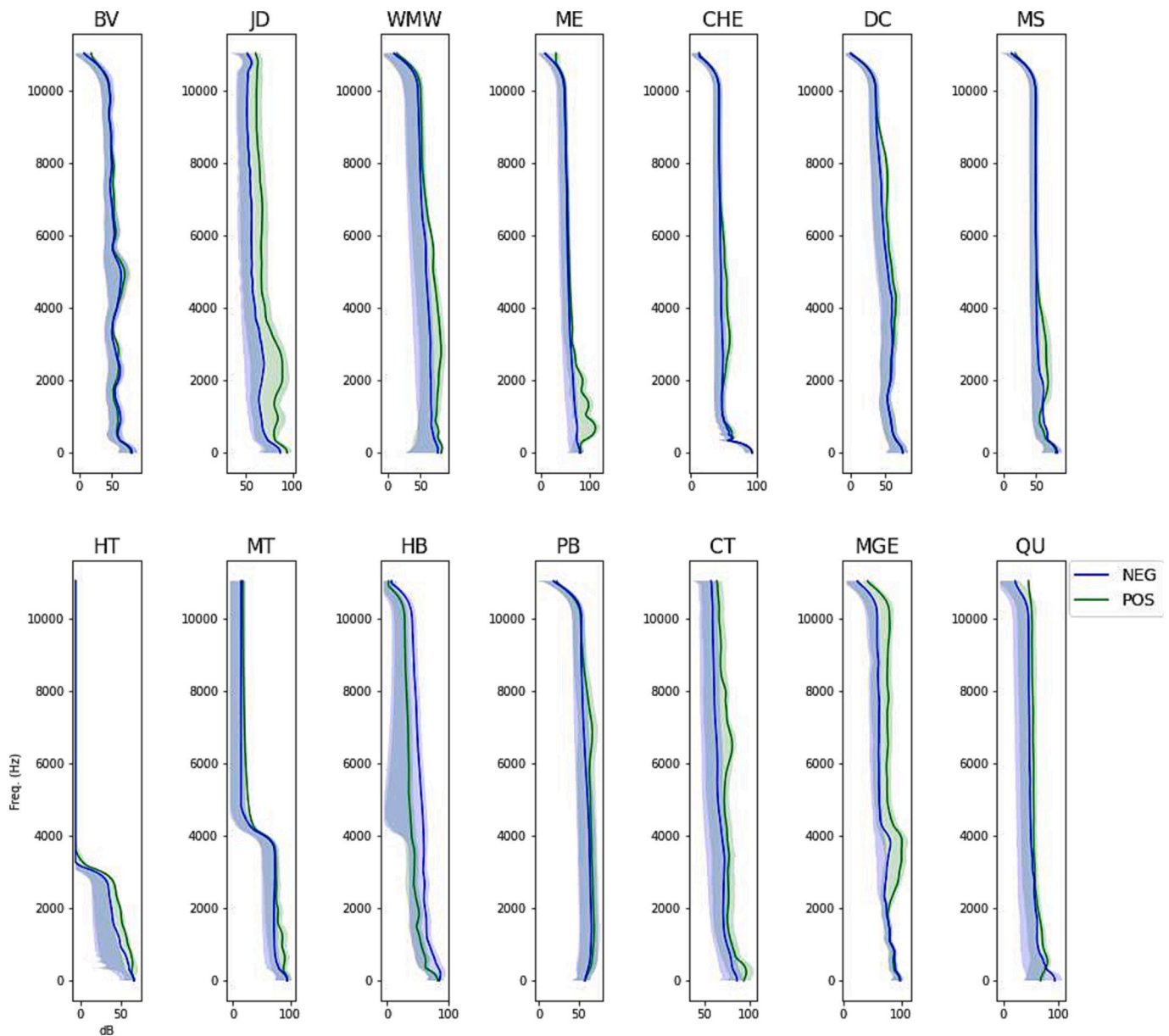
**Chick calls (MGE-Evaluation set):** The MGE dataset contains three 10-min recordings from three 1-day old domestic chicks (*Gallus gallus*). Vocalisations have been recorded in 2019 and annotated using Sonic Visualiser in the Prepared Minds Lab (Queen Mary University of London<sup>5</sup>) by Dr. Versace's staff (Shuge Wang, Michael Emmerson, Laura Freeland, Elisabetta Versace). Individual chicks had just been removed from the hatchery, and were free to explore the experimental arena. Chicks have been recorded in the controlled environment of the laboratory, a 24–48 h after hatching. Chickens are a precocial social bird species and upon hatching they establish a strong attachment to their social companions, via a process called imprinting, where acoustic information strengthens affiliative responses Versace et al. (2017). During and after the imprinting process, chicks vocalise signaling that they are in close proximity to their social partners (i.e. pleasure calls) or that they are distant or separated from them (i.e. contact calls). The data gathered in the dataset present uneven time distribution. Calls typically have a short duration (100–400 milliseconds). In the dataset, only pleasure calls were annotated in recordings from chicks one and two, only contact calls were annotated in recordings from chick three. We defined calls based on previous literature (Marx et al., 2001).

To summarise, these datasets together represent some of the wide variety of bioacoustic SED tasks, and were selected to give broad coverage of some of the key axes of variation, such as rate of occurrence of the target sound, length of calls, background noise (SNR), taxa, etc. Some of these quantitative characteristics are summarized in Table 1, and a visual representation of each dataset is presented in fig. A.8. Descriptive analysis of the datasets further illustrates the variation in temporal and spectral characteristics, for the target sounds as well as the background soundscapes. The spectral profile of each dataset is presented in Fig. 2, this shows the energy distribution across frequency bands for the POS and background in separate. A similar representation is used to create the temporal profiles shown in Fig. 3.

The datasets represent diverse challenges for the few-shot SED systems that are trained and evaluated on them. For each dataset, the provided 5 events are used to specify the class of target sounds. The extent to which a small set of calls can be representative depends on various factors including stereotypy - the degree of how stereotyped are the calls, and vocabulary size.

To approximately quantify stereotypy, for each class in the evaluation set, we calculate similarity between sound events. We do this between the selected five events and the remaining events, as well as for the annotated calls more generally (Fig. 4). Together with the SNR and the sparsity/density of call events, this stereotypy aspect is expected to be one of the axes of variation among our datasets. (details in Appendix A.2).

<sup>5</sup> <https://www.preparedminds.org/home>



**Fig. 2.** Spectral summary profiles of each dataset. For each frequency, we show mean and 90% confidence intervals of the energy distribution, for the foreground (POS events) and negative regions (background and non target sounds) separately.

### 2.3. Baseline methods

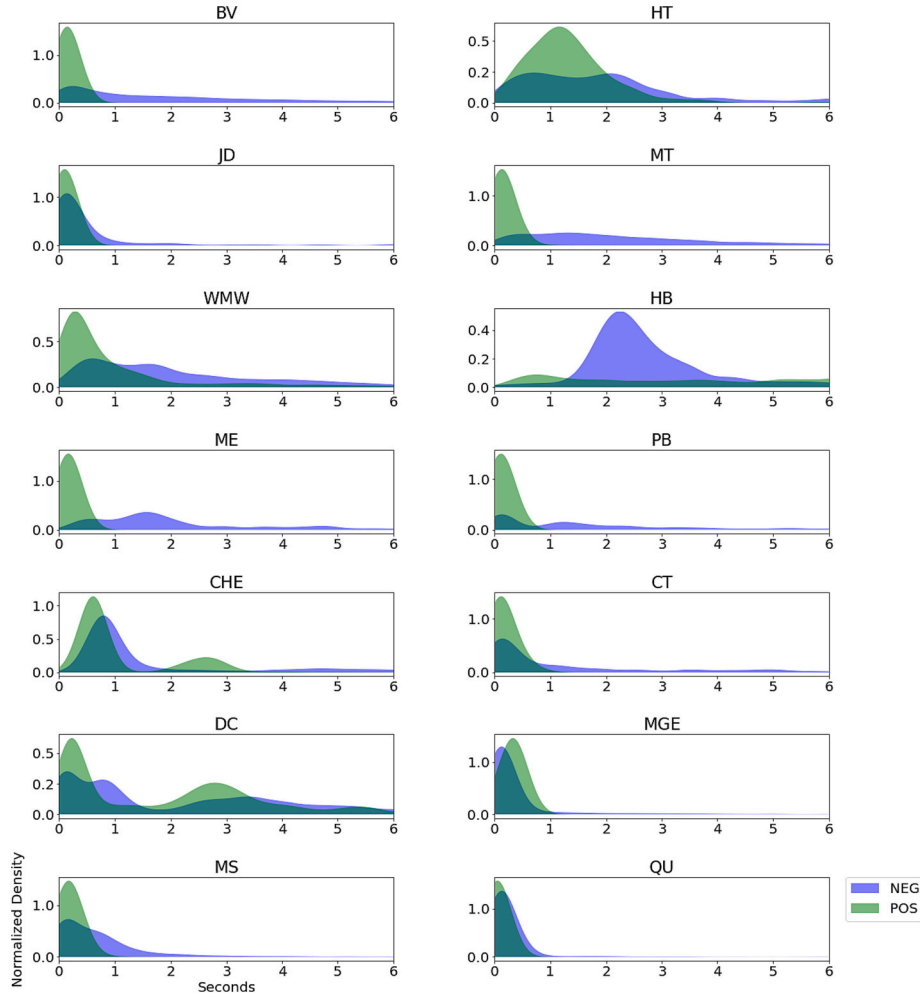
We propose two systems as baselines, representative of standard good-quality methods that can be applied to the task, and against which to measure the performance of novel submitted methods. One is an approach commonly used in bioacoustics based on spectrogram cross-correlation and the other is a deep learning approach based on prototypical networks, which have been used in other FSL work.

#### 2.3.1. Template matching (cross-correlation)

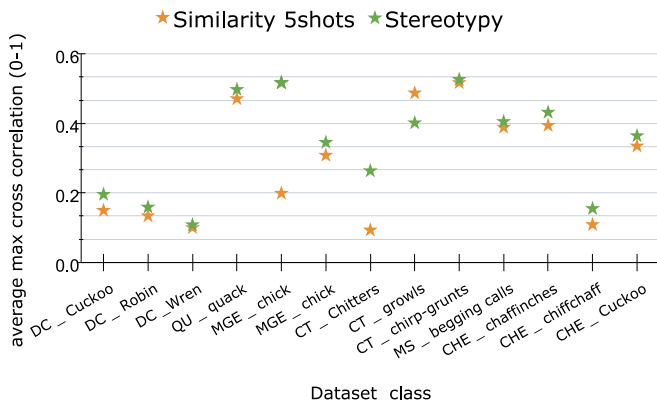
Signal-processing methods have been used for decades to detect events of possible interest in audio data (Gillespie et al., 2009; Towsey et al., 2012). Common approaches include energy thresholding, which can work in low-noise scenarios only, and template matching, usually based on cross-correlation (matched filtering) of waveforms or spectrograms. Template matching can work well in noisy audio, providing the target signal is acoustically (a) distinct from the background sounds and (b) stereotyped, i.e. not strongly varying in character. We thus expect template matching to work well in some of the scenarios we

study, but to perform very poorly in others.

Our baseline cross-correlation method is based on scikit-image's `match_template` function applied to spectrograms: it uses fast, normalised cross-correlation to find instances of a template in an image, returning values ranged between  $-1.0$  and  $1.0$ , with higher values corresponding to higher correlation. Our few-shot template matching method computes cross-correlation across the time axis between each of the events (shots) provided for a file and the rest of the recording. A different detection threshold is set for each audio file based on the max value of the cross-correlation results between the shots provided. Peak picking is performed on the results of the template matching algorithm, with any peak above the threshold corresponding to the center of a detected event in that recording. Borders of the predicted event are assumed to align with the beginning and end of the template when it matches. Each of the 5 templates is used separately for matching, and the resulting event predictions are collapsed into a single binary prediction vector which will produce the final events predicted for the class of interest.



**Fig. 3.** Temporal profiles of each dataset. We show the empirical distributions (kde smoothed) of durations of marked regions, for the foreground (POS events) and negative regions (all non-POS regions) separately.



**Fig. 4.** Values of similarity between the annotated calls and the first 5 events (shots), and stereotypy for each class in the evaluation set. Classes are indicated in the horizontal axis by DatasetName\_ClassName. Both factors are computed using a similarity metric based on the average maximum cross correlation between events. It ranges between 0 and 1, where values closer to 1 represent higher similarity. (the details on how these values are computed are presented in Appendix A.2).

### 2.3.2. Prototypical networks

Our second baseline is based on prototypical networks, a deep learning technique whose training procedure is designed especially for few-shot learning (Snell et al., 2017). The networks are trained using *episodic training*: each “episode” is configured as an “ $N$ -way- $k$ -shot” classification task, where  $N$  denotes the number of classes and  $k$  the number of known samples per class. In the present work  $k = 5$ , and  $N = 2$  when there is only one sound event type to consider, in which case the two classes then represent active and inactive. Prototypical networks have previously been evaluated as highly promising for few-shot audio classification tasks (Pons et al., 2019).

A *prototype* in this method is a coordinate in some vector representation, which is calculated as a simple centroid (mean) of the coordinates for each of the  $k$  examples. The training data consist of a *support set*  $S$  consisting of  $k$  labelled samples from each class, with the remaining samples comprising the *query set*  $Q$ . Prototypical networks compute a class prototype  $c_n$  through an embedding function  $f_\phi : \mathbb{R}^D \rightarrow \mathbb{R}^M$  with learnable parameters  $\phi$ . In our baseline system  $D = 128$  and  $M = 64$ , and  $f_\phi$  is a neural network. The prototype for class  $n$  is computed as the mean of the embedded support points belonging to that class:

$$c_n = \frac{1}{k} \sum_{(x_i) \in S_n} f_\phi(x_i) \quad (1)$$

where  $S_n$  represents the subset of  $S$  from class  $n$ .



Then, for each sample  $x_q$  from the query set, a distance function is used to calculate the Euclidean distance of  $x_q$  from each prototype, following which a softmax function over the distances produces a distribution over the classes. This directly implies that training the neural network to optimise these distances should move prototypes and their corresponding query points closer together in the embedding space created by  $f_\phi$ , and further away from non-matching points. In other words, the training procedure creates a general representation in which similar sounds are close to each other. Nearest-neighbour algorithms such as k-means can then be used to label future data points—even those from novel categories, after a simple procedure of calculating the prototype of a novel category as the centroid of its  $k$  shots.

During evaluation, we adopt a binary classification strategy inspired by Wang et al. (2020b). The first 5 positive (POS) annotations are used for calculation of positive class prototype and the rest of the audio file is treated as the negative class, based on the assumption that the positive class is relatively sparse in the recording. We randomly sample time regions from the negative class to calculate the negative prototype. Each query sample is predicted to have the target sound active, if its embedding coordinate is closer to the positive prototype than the negative. The prediction process for each file is repeated 5 times, with the negative prototype created by random sampling each time. The final prediction probability for each query frame is the average of predictions across all iterations. Finally, post-processing is applied to the outputs in order to remove possible false positives. For each audio file, predicted events with shorter duration than 60% of the duration of the shortest shot provided for that file are removed.

## 2.4. Evaluation and public challenge

For the evaluation of this task, we employ an event-based F-measure with macro-averaged metric, to evaluate the match between true and predicted events. The main complexity is related to the detection of a match between ground truth events and predicted events. Traditional approaches use onset detection based metrics and fixed-size evaluation windows (Mesaros et al., 2019). Given the great variation between datasets and characteristics of the events we want to detect in this task, these approaches are not suitable. Instead, we use the Intersection over Union (IoU), with 30% minimum overlap to produce a list of possible matches of the predictions. Applied to temporal events we get a list of predicted events that overlap at least 30% with the ground truth events and thus are candidate matches. For each ground truth event, a single best match is selected by applying the Hopcroft-Karp-Karzanov algorithm for bipartite graph matching, a similar procedure as used in the sed eval toolbox.<sup>6</sup>

In a SED task we can define True Positives (TP) as predicted events that match ground truth events, False Positives (FP) as predicted events that do not match any ground truth events, and False Negatives (FN) as ground truth events that are not predicted. In this task, ground truth events consist of POS events of the class and UNK events that have some uncertainty associated to the assigned class. The UNK label is typically assigned if the annotator is not sure if the event belongs to the class of interest, in other situations the annotator knows it is not the class of interest however the event looks like it could be. This decision is thus very subjective and dependent on each dataset. For evaluation purposes, matches to UNK events do not count towards calculating TP, FP or FN. In doing so, we ensure that the subjectivity associated with assigning the UNK label does not impact the performance score of the systems.

The procedure we employ is:

1. Apply IoU and bipartite graph matching between predicted events and ground truth POS events only, resulting in TP.
2. Apply IoU and bipartite graph matching between remaining predicted events, that did not match with any POS event, and ground truth UNK events only.
3. Compute FP as the number of predicted events that were not matched to either POS or UNK events.
4. Compute FN as the number of POS ground truth events that were not matched by any predicted event.

This is applied to each dataset in the evaluation set where we compute the F-score metric. The reported results are the harmonic mean over all the datasets, which is appropriate for combining percentage results, and ensures that a system should perform well across all datasets to achieve a strong score.

We thus use an averaged F-score as our main summary statistic for each submitted system. To explore system performance in more detail, we also inspect the F-scores per dataset, and per class in each dataset, in particular to examine whether differences in acoustic characteristics correlate with differences in performance.

The F-score metric is designed to summarise how well a system's outputs correspond to the desired outputs. However, there are many factors that affect the usefulness of such outputs, meaning that it is difficult to estimate a technology readiness level from only numerical scores. Hence, in addition to our quantitative analysis, we conduct a qualitative user-oriented analysis of selected system outputs, gathering feedback from expert users (annotators of the datasets).

## 3. Results

We report here the results of our public challenge. We have conducted three editions to date (2021, 2022, 2023), and each year the evaluation dataset has been extended to cover a wider range of bio-acoustic sources. After the first edition, it was agreed that evaluation datasets should be refined and expanded to give a more robust estimation of system performance. We thus report here on the systems submitted to the second and third editions, evaluated on the datasets of the second edition for comparability. For completeness, a summary of the 2021 outcomes is given in the Supplementary Information.

In the 2022 edition, 15 teams participated submitting a total of 46 systems and in 2023 there were 6 teams with a total of 22 systems. We present in Table 2 the overall scores of the best system submitted by each team in these two editions of the challenge. The challenge can be seen to be a difficult one: the baseline systems, and many teams, obtained F-score averages below 25%. On the other hand, methods could be designed which reach well over 40% F-score average, and up to 60% (Table 2). Such performances were much stronger than expected based on the task difficulty and 2021 results.

Several systems adopted a prototypical network approach, perhaps influenced by the baseline code and/or the outcomes of the 2021 edition. Simple improvements over the baselines were achieved by applying data augmentation techniques and intelligent post-processing. Better ways to construct the negative prototype were also explored by some teams who reported improved results (Liu\_Surrey, XuQianHu\_NUDT\_BIT, Jung\_KT, Wu\_SHNU, Jung\_KT, Willbo\_RISE). *Transductive inference*—adapting the learnt feature space at test-time based on the newly-presented positive and negative events—was also applied by some participants (Liu\_Surrey, XuQianHu\_NUDT\_BIT, Li\_QMUL, Tan\_WHU, Zou\_PKU).

The top 2 scoring systems, (Du\_NERCSLIP\_23 and Du\_NERCSLIP) belong to the same team, who was able to achieve the best score at both editions. Their implementation is based on the idea of learning frame-level embeddings, instead of an embedding for a whole segment. This confers to the system a high time resolution capability, which is important to perform particularly well on classes of very short duration such as QU, (Fig. 5). For the third edition of the task, they have

<sup>6</sup> [http://tut-arg.github.io/sed\\_eval/generated/sed\\_eval\\_util.event\\_matching.bipartite\\_match.html](http://tut-arg.github.io/sed_eval/generated/sed_eval_util.event_matching.bipartite_match.html)

**Table 2**

F-score results (in %) per team (best scoring submission) on 2022 evaluation and validation sets. Systems are ordered by higher scoring rank on the evaluation set. These results and technical reports for the submitted systems can be found on task 5 results page (DCASE, 2022) and (DCASE, 2023).

| Team   | # Best submission | Evaluation (95% CI) | Validation |
|--|-------------------|---------------------|------------|
| Du_NERCSLIP_23 (Yan et al., 2023)                              | 2                 | 61.83 (61.23–62.32) | 75.6       |
| Du_NERCSLIP (Tang et al., 2022)                                | 2                 | 60.22 (59.66–60.70) | 74.4       |
| Liu_Surrey (Liu et al., 2022a)                                 | 2                 | 48.52 (48.18–48.85) | 50.03      |
| Martinsson_RISE (Martinsson et al., 2022)                      | 1                 | 47.97 (47.48–48.40) | 60         |
| Hertkorn_ZF (Hertkorn, 2022)                                   | 2                 | 44.98 (44.44–45.42) | 61.76      |
| Liu_BIT-SRCB (Liu et al., 2022b)                               | 4                 | 44.26 (43.85–44.62) | 64.77      |
| Wu_SHNU (Wu and Long, 2022)                                    | 1                 | 40.93 (40.48–41.30) | 53.88      |
| XuQianHu_NUDT_BIT (Liu et al., 2023)                           | 3                 | 37.71 (36.98–38.23) | 63.94      |
| Moummad_JMT (Moummad et al., 2023)                             | 2                 | 37.32 (36.82–37.74) | 63.46      |
| Zgorzynski_SRPOL (Zgorzynski and Matuszewski, 2022)            | 4                 | 33.24 (32.69–33.69) | 57.2       |
| Gelderblom_SINTEF (Gelderblom et al., 2023)                    | 2                 | 26.79 (26.13–27.29) | 36.6       |
| Mariajohn_DSPC (Mariajohn, 2022)                               | 1                 | 25.66 (25.40–25.91) | 43.89      |
| Jung_KT (Lee et al., 2023)                                     | 3                 | 23.74 (23.14–24.17) | 81.52      |
| Willbo_RISE (Willbo et al., 2022)                              | 4                 | 21.67 (21.32–21.97) | 47.94      |
| Zou_PKU (Yang et al., 2022)                                    | 1                 | 19.20 (18.88–19.51) | 51.99      |
| Huang_SCUT (Huang et al., 2022)                                | 1                 | 18.29 (18.01–18.56) | 54.63      |
| Tan_WHU (Tan et al., 2022)                                     | 4                 | 17.22 (16.82–17.55) | 54.53      |
| Li_QMUL (Li et al., 2022)                                      | 1                 | 15.49 (15.16–15.77) | 47.88      |
| Wilkinghoff_FKIE (Wilkinghoff and Cornaggia-Urrigshardt, 2023) | 4                 | 13.31 (12.83–13.67) | 62.64      |
| baseline-TempMatch (Morfi et al., 2021)                        | –                 | 12.35 (11.52–12.75) | 3.37       |
| baseline-ProtoNet (Morfi et al., 2021)                         | –                 | 5.3 (5.1–5.2)       | 28.45      |
| Zhang_CQU (Zhang et al., 2022)                                 | 4                 | 4.34 (3.74–4.56)    | 44.17      |
| Kang_ET (Kang, 2022)   | 2                 | 2.82 (2.76–2.87)    | –          |

incorporated this frame-level embedding idea into a multi-task learning framework, that also includes a speaker voice activity detection branch. These modifications are responsible for a score improvement of almost 2% points.

The next system in rank, Liu\_Surrey, implements a novel approach designed to optimise the contrast between positive events and negative prototypes. This, together with an adaptive segment length dependent on each target class, works well across all the evaluation sets.

The problem of very different lengths of events across target classes was also directly addressed by other submissions. Both Martinsson\_RISE and Zgorzynski\_SRPOL implemented an ensemble approach where each individual model focuses on a different input size range. In Liu\_BIT-SRCB this is explored through a multi-scale ResNet, and in Willbo\_RISE with a wide ResNet containing many channels. Also in XuQianHu\_NUDT\_BIT, they implement a novel adaptive mechanism - squeeze/excitation block - designed to assign different weights to different channels of the feature map.

Inspecting the characteristics of the methods performing most strongly in the challenge, broadly across all editions, we observe some general tendencies (Table 3). Firstly, there is relatively little variation in the acoustic features extracted, and the neural network architecture: most systems use Mel spectrograms with PCEN, and standard CNNs. The main innovation in this aspect comes from You et al. (2023), where the CNN is replaced by the audio spectrogram transformer (AST). However, there is considerable variation in the method of training the network, and performing inference. There is a roughly equal balance of the two main paradigms: meta-learning with prototypical networks, versus fine-tuning or otherwise adapting a network trained using cross-entropy.

Within both paradigms there are instances of transductive inference (Yang et al., 2021), Liu\_Surrey. The ‘dynamic few-shot learning’ (DFSL) method employed by Wu\_SHNU is an alternative approach to query-time adaptation: the feature extraction, and the representation for previously-known classes, is never altered, but at query time the new task is considered to be a new class, whose representation is a weighted sum of those for the previously-known classes. This has the appealing characteristics of combining stability with dynamic adaptation, unlike standard fine-tuning in which care must be taken not to overfit to the new examples. Despite these innovations, it is notable that multiple teams achieved strong performance without test-time adaptation of the learnt feature space.

Many teams innovated in the way time-regions are selected for

training an algorithm, both for computing the positive and negative regions (foreground and background). Multiple teams made use of *pseudo-labelling* as a way to bootstrap the amount of data presented to the system: this means using the system to make a first ‘draft’ identification of which regions are positive/negative for the events of interest, and then using that estimated labelling to further train the system (Yang et al. (2021), Wu\_SHNU and Du\_NERCSLIP(23)). Pseudo-labelling has been explored in many machine learning domains for data-poor scenarios.

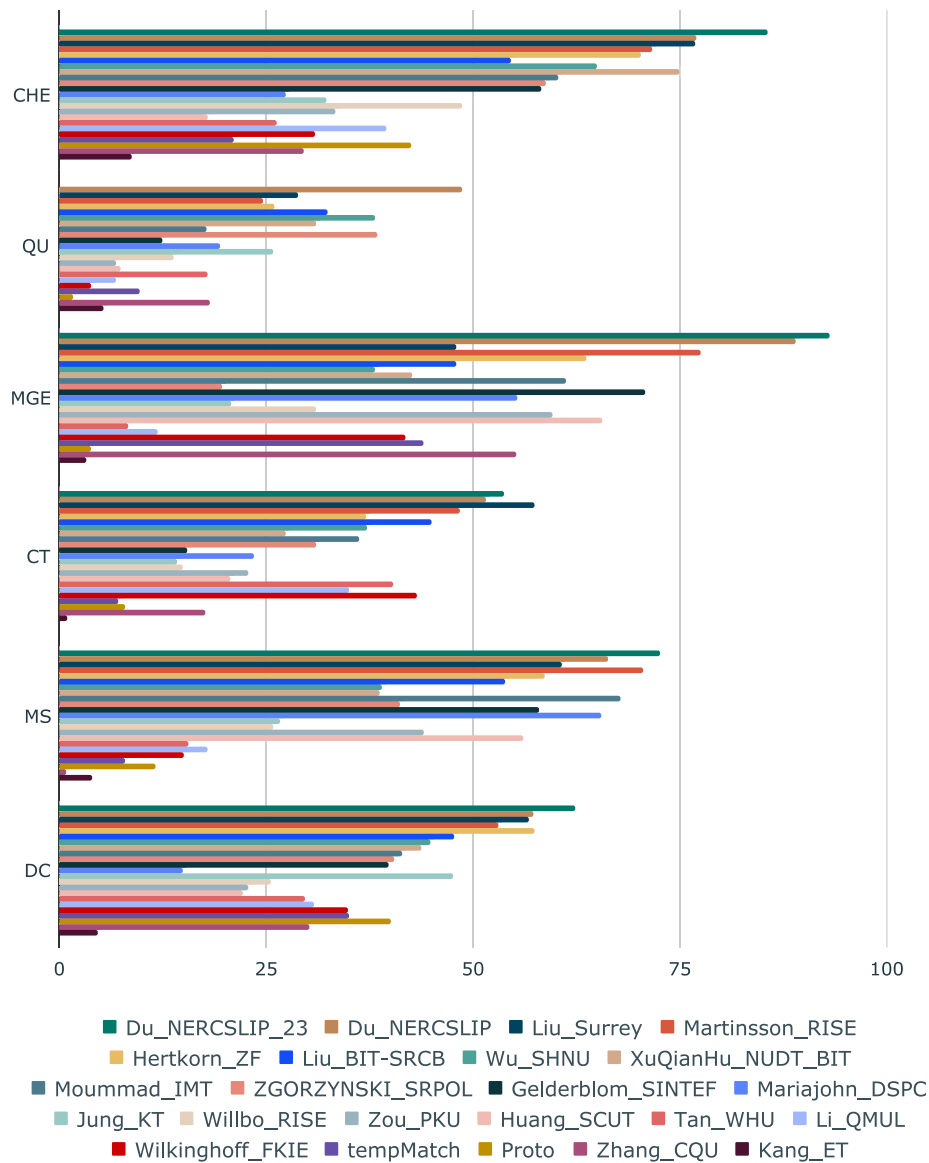
Successful systems also commonly used explicit methods to control the duration of the detected events. In many cases this consists of postprocessing predictions to delete/merge very short events, or estimating the typical duration from the examples. Du\_NERCSLIP(23) and Wolters et al. (2021) made use of neural network architectures specifically trained to infer and output region annotations.

Overall, the different approaches submitted illustrate the introduction of ideas to address challenges related to this task: how to deal with very different event lengths; how to construct a negative class when no explicit labels are given for this; and how to bridge the gap between classification and detection for few-shot sound event detection. These challenges derive from the combination of few-shot learning with sound event detection, and hence are not addressed in standard few-shot learning (Wang et al., 2020a).

### 3.1. Analysis of dataset dependencies

The submitted systems exhibit variations in their performance across our datasets (Fig. 5). The same is true even when we look at the target class level, Fig. 6 presents the Fscore results for the different target classes in each dataset. The easiest classes to be detected are CHE\_chaffinches, CT\_chirpgrunts and DC\_robins, where several systems reach above 75% F-score. On the other side, CT\_Chitters, DC\_Cuckoo and the QU\_Quacks seem to be the classes where systems struggled the most to make correct predictions. The disparity in score between systems is also evident. The performance on MGE\_Chick\_Pleasure\_calls is a good example where Du\_NERCSLIP’s systems show a significant advantage over the others.

To determine which data characteristics might be the strongest factors in these performance variations, we investigated five data attributes, three commonly considered in soundscape analysis: SNR, event sparsity, and event length, plus similarity between events and the 5 shots



**Fig. 5.** F-Score results of 2022 and 2023 systems on each dataset of the 2022 evaluation set. Systems are ordered by overall highest scoring rank on the evaluation set.

and stereotypy (both defined in Section 2.2 and Appendix A.2). We performed a multivariate regression with different combinations of these variables. By evaluating and selecting the best model, we can verify which would be the best attribute or combination of attributes that predicts the Fscore. The possible 31 combinations of these attributes were used as the predictors of the average F-score across all systems scoring above the baseline. The resulting regression models were then evaluated by inspecting the  $p$ -values, adjusted R-squared, Akaike information criterion (AIC) and the Bayesian information criterion (BIC). The results indicate that none of these factors translating bioacoustic considerations was a strong predictor of differences in performance. A similar conclusion can be reached by observing Fig. 7. Here, the average F-score across systems performing above the baselines is plotted against Stereotypy, Mean event duration, SNR and Event density. The absence of clear correlations indicates the difficulty in selecting which could be the most important factors impacting the f-score. Furthermore, this lack of strong visual relationships between Fscore and data characteristics remains even when only the scores of one individual system are used.

### 3.2. Ablation study

The developed systems are complex and most consist of various independent functional units coming together to solve the task.

Here, we present the results of the ablation study performed on Liu\_Surrey's system (Liu et al., 2022a). The choice to perform the ablation analysis on this system is due to it being the highest scoring system with open access code. An ablation study consists in removing different parts of the network and evaluating the impact these changes have on performance. This allows for some increased understanding of how a system works, while providing a way in which it is possible to measure the contributions of each individual unit for the overall level of performance achieved.

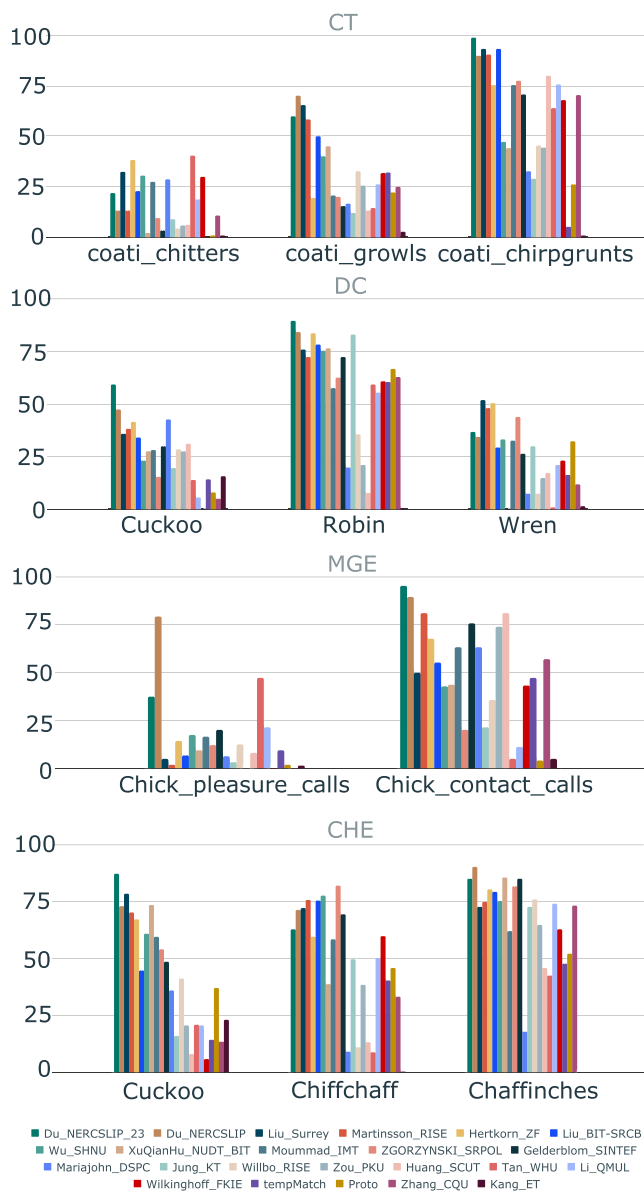
The experiments with variations to the system's architecture can be organized into different categories: 1) exploring different input features, 2) analysing the impact of Contrastive Learning and 3) impact of the number of "ways" used for episodic training (as in the Meta-learning setup  $N_{way}$ ,  $K_{shots}$ ): "ways" means the number of different sound categories considered at once). The F-score results are presented in Table 4. Because systems are developed based on the development and

**Table 3**

Methodological features of various systems of interest. Terms: Proto = prototypical network; x-ent = cross-entropy; Dist = distance-based; TI = transductive inference; DFSL = dynamic few shot learning; AST = Audio Spectrogram transformer; SCL = supervised constrastive learning; 5 = the original 5 shots are used; between-the-5 = the spaces between the 5 shots are used; pseudo-pos/pseudo-neg: pseudo-labelling is used to select additional examples; aug = data augmentation applied.

|   |  | Spectrogr.<br>features    | Neural net<br>arch.      | Training<br>objective  | New class<br>addition      | Inference                              | Feature space<br>changes? | Negatives<br>selection                     | Positives<br>selection | Segment<br>length<br>technique                      | Post-processing                                   |
|---|--|---------------------------|--------------------------|------------------------|----------------------------|--|---------------------------|--|------------------------|---|---|
| Baselines                                       | Prototypical                                       | Mel +PCEN                 | CNN                      | Proto                  | Proto                      | Dist:Proto                             | No                        | Whole audio                                | 5                      | Derived from<br>shots                               | Delete very short                                 |
|   | Template matching                                  | Lin                       | n/a                      | n/a                    | New<br>templates           | Cross-correl                           | No                        | n/a  | 5 + aug                | Template<br>length                                  | –   |
| Systems submitted<br>to the public<br>challenge | <a href="#">Yang et al. (2021)</a>                 | Mel                       | CNN                      | x-ent                  | Retrain (new<br>pos + neg) | Posterior                              | TI x-ent                  | Pseudo-neg                                 | Pseudo-pos             |   | Peak picking,<br>thresholding                     |
|   | <a href="#">Tang et al. (2021)</a>                 | Lin + PCEN                | CNN                      | Proto                  | Proto                      | Dist:Proto<br>(Attention-<br>weighted) | No                        | Whole audio                                | 5                      |   | Peak picking, median<br>filtering                 |
|   | Du_NERCSLIP  | Mel +PCEN                 | CNN<br>framework         | x-ent                  | Finetune last<br>layer     | Posterior                              | Finetune x-<br>ent        | Between-the-5                              | Pseudo-pos             | Adaptive<br>length fixed<br>shift                   | CRNN event filter                                 |
|   | Liu_Surrey   | Mel +PCEN &<br>delta-MFCC | CNN                      | Proto<br>(modified)    | Proto                      | Dist:Proto                             | TI, Retrain               | Between-the-5 +<br>Pseudo-neg<br>(SpecSim) | 5                      | Derived from<br>shots                               | Split-merge-filter;<br>delete very long/<br>short |
|   | Wu_SHNU (+Wu<br>2023 ICASSP) DFSL<br>Moummad_IMT   | Mel +PCEN                 | CNN (ResNet)             | x-ent                  | DFSL<br>attentive          | DFSL attentive                         | No                        | Pseudo-neg                                 | Pseudo-pos             | –   | –   |
| Other   | Wolters 2021 arxiv<br>Perceiver                    | Mel                       | CNN + CRNN<br>+Perceiver | Proto +RPN<br>(R-CRNN) | Proto                      | Dist:Proto                             | No                        | n/a  | 5                      | Adaptive<br>length<br>Region<br>proposal<br>network | –   |
|   | <a href="#">You et al. (2023)</a><br>(ICASSP 2023) | Mel                       | AST                      | Proto                  | Proto                      | Dist:Proto                             | Finetune, TI              | Between-the-5                              | 5 + aug                | Derived from<br>shots                               | thresholding, merge/<br>filter small events       |





**Fig. 6.** Fscore (%) results by class in the evaluation set. Note that QU and MS datasets only contain a single class and thus are not represented here. The systems are ordered by overall highest scoring rank on the evaluation set.

validation sets alone, the design decisions might not be what works best for the evaluation set, specifically in the case where these datasets vary substantially. This might explain why the original submitted architecture (first row of Table 4) did not result in the best performance across all the variations tested here. In fact, instead of applying PCEN and Delta MFCCs as input features, the system modified to use Log Mel spectrograms resulted in the top performance on F-score. Similarly, here we see that including Negative Contrastive learning does not work well on this evaluation set and indeed the performance of the system decreases. Experimenting with different number of ways, confirms the expected that as the number of ways in the support set increases from 10 to 30, so does the performance. Finally, ensembling all the variations of the original system leads to an improved F-score.

### 3.3. Expert use analysis

We are interested in understanding how far away the best scoring systems are from being incorporated into the annotation practice and

how helpful or misleading their predictions can be. The expert annotators of QU, CT and MS datasets were given the predictions resulting from the 3 top scoring systems of the 2022 edition of this task, (Du\_NERCSLIP, Liu\_Surrey and Martinsson\_RISE) and asked to analyse them in terms of a) Usability and b) Types of errors. Here are the main topics and highlights received. The full feedback can be read in Appendix A.3.

All consider that at least one of the systems results in useful predictions that can be used as a starting point for manual editing.

The best ranking system overall (Du\_NERCSLIP) is not always the one selected by the experts as the best predictor of events in the different datasets and it also changes for different classes within the same dataset. However the experts' selection almost always agree with the F-score results by class shown in Figs. 5 and 6.

As to the type of errors, the experts identified several instances of missed detections, misclassifications either on non-target calls or noise events, and in general imprecise detection of the duration of calls.

Another aspect highlighted for both QU and CT datasets were situations where the capability of the systems to produce correct predictions decreased over time, meaning that events happening further away from the beginning (where the 5 shots examples happen) were less well predicted.

The reason for some missed detections might be due to the selection of the 5 examples from which the systems need to learn the pattern of the target class. For both MS and CT datasets the experts commented that the range of variation of the target calls was not well captured within the 5 initial examples. This aspect is also expressed in Fig. 4.

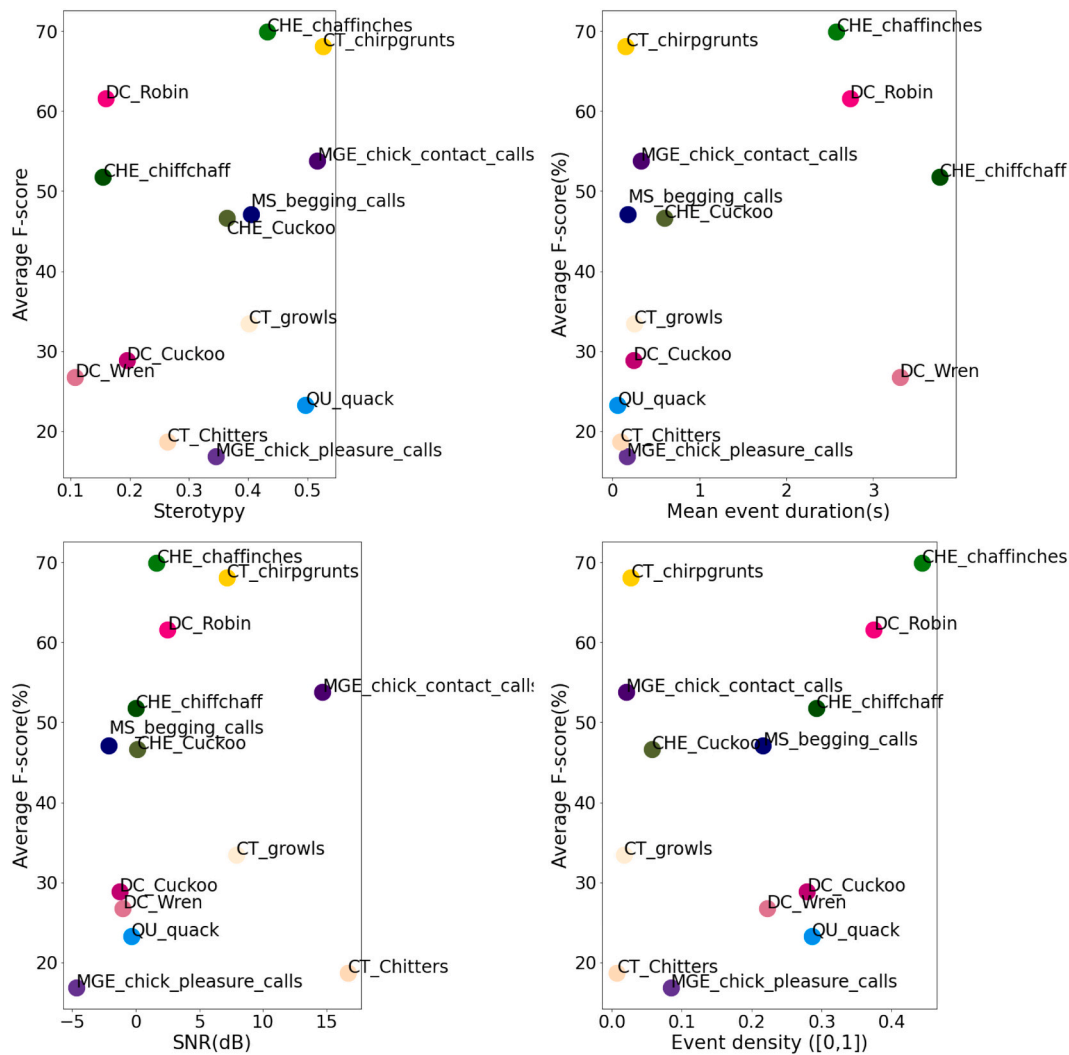
Finally the potential for using FSL to improve upon human manual annotations is illustrated in the feedback received for the CT dataset. The system ranked in second place overall, Liu\_Surrey, was able to predict 20 new Growls that the human annotator had not identified.

## 4. Discussion

In this work we have formulated few-shot bioacoustic event detection as a machine learning task. We have evaluated many approaches to the task, and demonstrated that both the meta-learning and the transfer learning methodologies can successfully generalise to novel sub-tasks in bioacoustic FSED—thus, transcribing animal sounds with a precision unobtainable with other automated methods, in the absence of huge training datasets. Our sub-tasks were chosen to be diverse and non-trivial: they differed in taxon, target sound characteristics, background noise, stereotypy, stationarity, duration, and more. We believe that we have shown that the many related recognition tasks in computational bioacoustics can be unified within a generalised approach to machine learning.

Leading systems achieve over 30% F-score on all 6 tasks. This is a dramatic improvement over classic template-matching, and also over a standard modern deep learning approach (both of which often achieved F-scores below 10%, in our baseline implementations). This reflects the fact that most bioacoustic sound event detection tasks have unique characteristics (such as noise, non-stereotypy, distractor sounds, non-stationarity) which make them distinct from each other and very hard to analyse with a conventional detection system. Although automatic detection has been in use for many years, it has often required manual tweaking of a system's parameters for each new situation. Based on this study we believe our formulation of FSED is a useful one. It is applicable across a wide selection of bioacoustics tasks, and provides a good target for machine learning development. It is not trivially solved by prior art in few-shot learning, nor by pretrained networks; yet we report very strong progress through the public challenges. We also consider that our chosen evaluation measure—an event-based F-score—has good external validity, since it aligns well with expert evaluations of automatic transcripts.

Our aim to generalise over a range of loosely-related datasets/tasks is of current interest in machine learning. There are some comparable initiatives in wildlife monitoring. The 'BEANS' project collects together



**Fig. 7.** Scatter plots illustrating the relationship between several data characteristics and the F-score for each class of the evaluation set. Note the vertical axis represents an averaged F-score, which is calculated from all the systems scoring above the baseline. The horizontal axis displays four factors, namely similarity to 5 shots, stereotypy, event length, event density and signal-to-noise ratio (SNR).

animal sound datasets and aims to provide a general evaluation benchmark (Hagiwara et al., 2022). Their work focuses on classification rather than temporal detection, and it does not consider few-shot learning or meta-learning—however it may be possible to re-use their data for such things. Similarly, in image recognition the NeWT benchmark provides a suite of tasks for wildlife images, using a classification framework to ask a very wide range of ecology questions from a single data representation (Van Horn et al., 2021). Key differences between these and our work are our few-shot setting, and the explicit inclusion of temporal structure in the input and output of our formulation.

Based on the results presented here, are few-shot bioacoustic SED systems good enough to use? Yes, as determined by feedback from our panel of experts. Although the outputs from such systems are far from perfect, they were judged to be of sufficient quality for active use, in place of fully-manual annotation. The quantitative results demonstrate that, when presented with a new dataset with no large training corpus, few-shot bioacoustic SED outperforms common methods such as template-matching. It is worth remembering, however, that our paradigm is designed for the case of detecting events for which no large training dataset is available. If large amounts of labelled data are available, or pre-trained networks whose training matches well with the intended use, then the more common machine learning method (i.e. supervised learning) would be expected to be the most reliable

approach.

#### 4.1. Aspects of bioacoustic datasets that affect performance

Many aspects of bioacoustic datasets make them complex to analyse: noise, highly sparse or dense events, varied levels of stereotypy, and non-stationarity (drift) in conditions. This is further illustrated when we show the variation in performance even across classes of the same

**Table 4**

F-Score results for the different system variations on the evaluation and validation sets. First row refers to the unchanged submitted system, all the other systems are simple modifications to this.

| System Variation                           | F-score-eval  | F-score-val   |
|--|---------------|---------------|
| PCEN+DeltaMFCC (original submitted system) | 35.019        | 38.350        |
| only PCEN                                  | 36.001        | 40.221        |
| <b>only LogMel</b>                         | <b>40.355</b> | 45.330        |
| LogMel+DeltaMFCC                           | 37.518        | 40.033        |
| w/o. Negative contrastive learning         | 39.637        | 46.614        |
| #ways 10                                   | 35.503        | 38.671        |
| #ways 20                                   | 35.866        | 43.634        |
| #ways 30                                   | 36.608        | 42.202        |
| #ways 40                                   | 36.021        | 42.015        |
| <b>Ensemble all</b>                        | <b>50.624</b> | <b>54.485</b> |

dataset (Fig. 6). We selected datasets which varied across many of these characteristics, and we sought to evaluate which of them were key factors influencing the difficulty of the task. A quantitative analysis (multivariate regression) was unable to identify any factors that consistently affected the F-score results across these datasets. However, qualitative feedback from expert users indicated that non-stationarity exerted an effect: this was shown by a reduction in performance for time-regions distant from the annotated examples.

A separate issue picked up by our annotators was that the support set was not always a good representation of the class to be detected, either because it did not include examples of every call type or due to the low stereotypy characteristics of certain classes. A trivial response is that we may include more than 5 examples, or curate the examples so they span the desired range of calls. In such cases we would expect stronger performance, at the slight cost of reintroducing some manual intervention. Our design decision to use the *first* few examples in a dataset is reflective of initialising a system before deploying it in a new recording situation. With offline analysis, and more flexibility in selecting examples, higher performance can be achieved.

Since bioacoustic targets may include multiple call types or non-stereotyped sounds, it is worth noting one aspect of prototypical networks. The formation of a single fixed prototype, by taking an average in a coordinate space, implies that the examples are in some sense all of one kind. This assumption is also challenged by non-stationarity, which we might think of as a prototype gradually drifting rather than remaining fixed. Transductive inference helps to reduce these issues by allowing the feature space to be updated at query time, and this was used by various strongly-performing systems. There remains much opportunity for multiplicity and drift to be included into designs for the *concept formation* of FSED systems.

#### 4.1.1. Duration of animal sound events

The annotated durations of animal sound events can range over multiple orders of magnitude, from milliseconds to minutes. They result from diverse physical processes, from the impulsive (dolphin clicks) to the continuous (mosquito flight). In retrospect it is clear that this needs to be handled carefully in the design of an SED system, because many computational methods have inbuilt assumptions or limitations in the durations they can process. When at first we formulated the task, we did not foresee that correct handling of event durations would be an important factor in evaluation performance, but this was indeed the case. The variable scale of event durations is not a limitation in itself—the difficulty comes when we try to solve all these different tasks with very different characteristics, together in one algorithm.

Many machine learning systems have pragmatic design constraints that limit the range of durations they can consider. Our template-matching method uses ranges directly inherited from the 5 annotated events, although there remain practical limits on very large templates, such as computer memory. In deep learning, long audio files are usually divided into shorter chunks (with fixed durations of e.g. 3 or 10 s), so that a whole batch can fit inside the limited memory of GPUs. To detect long events, detections that span these chunks are joined together in post-processing. This as well as other considerations meant that post-processing of outputs was an important aspect of all strongly-performing systems.

The (deep) feature extraction procedure itself can place limits on event durations. Firstly the resolution of spectrograms, with a typical granularity around 10 ms per ‘frame’, often predetermines the finest scale that can be resolved. Datasets QU and MGE contained very short sounds at around this scale; Du\_NERCSLIP(23)’s framewise CNN excelled on these datasets. Secondly, CNNs (used by all submitted systems) have relatively small “receptive fields” and do not consider the whole spectrogram but local feature patterns. Many of the strongest systems adapted themselves at query-time to the expected event length inferred from the 5 examples, in particular Martinsson\_RISE who trained a set of embedding functions, each designed for a different duration.

Some notable systems augmented their core CNN with architectures that are able to integrate information over long durations and directly infer onset and offset locations, such as a CRNN event filter (Du\_NERCSLIP), perceiver and/or region proposal network (Wolters et al., 2021). We envisage that future developments on these lines may be fruitful, perhaps using further techniques from object detection.

#### 4.2. A single method for bioacoustic SED?

Our baseline prototypical network is itself novel since FSL has been applied to sound event classification, but almost never (prior to our work) to SED. Through a public data challenge we have seen many different variations on this method, leading to strong results. Is it possible to recommend a single method to take forward for bioacoustic SED; and if so, does it use prototype-based meta-learning?

We find that prototype-based meta-learning works well when taking care about certain aspects of the method (namely the choice of negative examples, and duration filtering / postprocessing of events). However, many of the strongest performing systems avoided the prototypical net method entirely (Du\_NERCSLIP(23), Wu\_SHNU, Moummad\_IMT), showing that the paradigm is not a necessary component. Bioacoustic FSED can be addressed by either meta-learning or fine-tuning approaches.

Query-time adaptation (transductive inference) was shown in multiple cases to lead to very strong performance, within both the prototypical and fine-tuning paradigms. This comes at a cost of added complexity and added query-time computation, since typically a new run of statistical optimisation must be performed for a new query task. Thus, from the present results we can recommend that a system should include query-time adaptation for the best possible detections, but that a system without query-time adaptation should be a widespread default. Such fixed embeddings can easily be used off-the-shelf, in the same way that other pretrained networks are now commonly downloaded and used. The DFSL method employed by Wu\_SHNU is an alternative approach which combines an unchanging feature extraction with a query-time adaptive weighting. This combines stability with dynamic adaptation, and thus is worthy of further investigation.

#### 4.3. A single embedding for bioacoustic SED?

Contrary to query-time adaptation, in machine learning there is current interest in learning good feature representations (good embeddings) from data. If an embedding can be re-used unmodified, this has an appeal of providing a general, reusable, and potentially low-complexity analysis tool, a component to be used in many systems. For audio data, some of the most widely-used deep embeddings are those derived from pretraining with the large-scale *AudioSet* dataset, originally designed for classifying many different (human-centric) acoustic categories (Gemmeke et al., 2017). More recent work evaluates this and many more ways to create an embedding (Turian et al., 2022).

Our evaluation shows that improved prototypical network methods create powerful embeddings, useful even with no test-time adaptation. It is impressive that a single vector space could be used to represent our diverse bioacoustic tasks. The present work on few-shot learning thus offers a different perspective on representation learning for sound in general, and animal sound in particular.

### 5. Conclusion

In this study, we have considered how best to automate the task of sound event detection in bioacoustics, and highlighted its potential as a *few-shot learning* problem. Unifying a set of loosely-related detection tasks makes possible the training of systems that do not need to be designed afresh, or trained afresh, for each new bioacoustic dataset. We have curated a rich and diverse set of data representative of many sub-tasks in this scenario, varying across many of the aspects a

bioacoustician might consider (such as SNR or stereotypy).

We framed few shot bioacoustic event detection as a public challenge. Over the three editions of the challenge, the evaluation dataset has been extended with more bioacoustic sources, pushing participating teams to create systems with impressive generalisation capabilities. Our analysis indicates the validity of the few-shot formulation of the task. Submitted systems followed mainly two paradigms: meta-learning and transfer learning. Both methodologies can lead to good performance as long as certain aspects of the task are addressed, such as highly variable duration of events, high time resolution, and the selection of negative examples. Query-time adaptation is required for the best detection results, though even fixed embeddings can provide strong performance in situations where the computational expense of additional training is to be avoided. While we believe our initiative to have been successful, there is yet more scope to create generalisable bioacoustic detectors.

Finally, it is possible to start envisioning the implementation of such systems for practical use. Our paradigm is not at all limited to just 5 arbitrary examples per sound event, and can be used in any situation where training datasets are not large. Manually curating a small but high-quality set of examples falls outside the present study, but can easily be expected to boost performance beyond the fully hands-off results reported here. Our formulation, and the diverse strongly-performing systems analysed in our public evaluation, thus move us towards a post-template matching era for bioacoustic sound event detection.

#### CRedit authorship contribution statement

**Ines Nolasco:** Software, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Shubhr Singh:** Software, Validation, Investigation, Writing – original draft, Visualization. **Veronica Morfi:** Software, Validation, Investigation, Writing – original draft. **Vincent LOSTANLEN:** Conceptualization, Software, Data curation, Writing – original draft, Writing – review & editing. **Ariana Strandburg-Peshkin:** Conceptualization, Methodology, Data curation, Writing – review & editing, Writing – original draft. **Ester Vidana-Vila:** Data curation, Writing – original draft, Visualization. **Lisa Gill:** Methodology, Validation, Data curation, Writing – original draft, Writing – review & editing. **Hanna Pamula:** Data curation, Writing – original draft. **Helen Whitehead:** Data curation, Writing – original draft, Writing – review & editing. **Ivan Kiskin:** Data curation, Writing – original draft. **Frants H. Jensen:** Data curation, Writing – original draft. **Joe Morford:** Data curation, Writing – original draft. **Michael G. Emmerson:** Data curation. **Elisabetta Versace:** Data curation, Writing – original draft. **Emily Grout:** Data curation. **Burooj Ghani:** Investigation, Visualization. **Dan Stowell:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision, Project administration.

#### Declaration of Competing Interest

None.

#### Data availability

Development data: <https://zenodo.org/record/6482837>  
Evaluation data <https://zenodo.org/record/6517414>  
Code: <https://github.com/c4dm/dcse-few-shot-bioacoustic/>

#### Acknowledgments

The authors would like to thank the participants of the successive editions of the Few-shot Bioacoustic Event Detection task at the DCASE Challenge, without who this work would not have been possible.

IN is supported by the Engineering and Physical Sciences Research Council (grant number EP/R513106/1). SS is a research student at the

UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported jointly by UK Research and Innovation [grant number EP/S022694/1] and Queen Mary University of London. VL acknowledges funding from CNRS, MITI award CAPTEO. ASP acknowledges funding from Human Frontier Science Program award RGP0051/2019. The work was also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2117—422037984. ASP received additional support from the Gips-Schüle Stiftung and the Max Planck Institute of Animal Behavior.

The data for meerkats were collected at the Kalahari Meerkat Project in South Africa, currently supported by a European Research Council Advanced Grant (No. 742808 and No. 294494) to Tim H. Clutton-Brock, the MAVA Foundation, and the University of Zurich. We further thank the Trustees of the Kalahari Research Centre and the Directors of the Kalahari Meerkat Project.

Spotted hyena data were collected in collaboration with the MSU-Mara Hyena Project, and data collection was additionally supported by a grant from the Carlsberg Foundation to FHJ.

#### Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ecoinf.2023.102258>.

#### References

- Beery, Sara, Van Horn, Grant, Perona, Pietro, 2018. Recognition in terra incognita. In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473.
- Brown, C.H., Riede, T. (Eds.), 2017. *Comparative Bioacoustics: An Overview*. Bentham Science Publishers, Oak Park, IL, USA. ISBN 978-1-68108-317-9.
- Caiger, Paul E., Dean, Micah J., DeAngelis, Annamaria I., Hatch, Leila T., Rice, Aaron N., Stanley, Jenni A., Tholke, Chris, Zemeckis, Douglas R., Van Parijs, Sofie M., 2020. A decade of monitoring Atlantic cod *Gadus morhua* spawning aggregations in Massachusetts Bay using passive acoustics. *Mar. Ecol. Prog. Ser.* 635, 89–103. <https://doi.org/10.3354/meps13219>.
- Chen, Yinbo, Wang, Xiaolong, Liu, Zhuang, Huijuan, Xu, Darrell, Trevor, et al., 2020. A new meta-baseline for few-shot learning. *arXiv preprint arXiv:2003.04390*, 2 (3), 5.
- DCASE, 2022. DCASE challenge 2022 Few-Shot Bioacoustic Event Detection Task - Results Page. URL: <https://dcase.community/challenge2022/task-few-shot-bioacoustic-tic-event-detection-results>.
- DCASE, 2023. DCASE challenge 2023 Few-Shot Bioacoustic Event Detection Task - Results Page. URL: <https://dcase.community/challenge2023/task-few-shot-bioacoustic-tic-event-detection-results>.
- Dong, Nanqing, Xing, Eric P., 2019. Domain adaption in one-shot learning. In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*. Springer, pp. 573–588.
- Frazao, Fabio, Padovese, Bruno, Kirsebom, Oliver S., 2020. Workshop Report: Detection and Classification in Marine Bioacoustics with Deep Learning.
- Ganchev, Todor, 2017. *Computational Bioacoustics: Biodiversity Monitoring and Assessment*. Walter de Gruyter GmbH & Co KG.
- Gelderblom, Femke, Cretois, Benjamin, Johnsen, Pal, Remonato, Filippo, Reinen, Tor Arne, 2023. Few-Shot Bioacoustic Event Detection Using Beats. Technical report, DCASE2023 Challenge. June.
- Gemmeke, Jort F., Ellis, Daniel P.W., Freedman, Dylan, Jansen, Aren, Wade Lawrence, R., Moore, Channing, Plakal, Manoj, Ritter, Marvin, 2017. Audio set: An ontology and human-labeled dataset for audio events. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 776–780.
- Gillespie, Douglas, Mellinger, David K., Gordon, Jonathan, McLaren, David, Redmond, Paul, McHugh, Ronald, Trinder, Philip, Deng, Xiao-Yan, Thode, Aaron, 2009. PAMGUARD: semiautomated, open source software for real-time acoustic detection and localization of cetaceans. *Jasa* 125 (4), 2547 apr. <https://doi.org/10.1121/1.4808713>. URL: <https://doi.org/10.1121/1.4808713>.
- Gillings, Simon, Scott, Chris, 2021. Nocturnal flight calling behaviour of thrushes in relation to artificial light at night. *Ibis*. <https://doi.org/10.1111/ibi.12955>.
- Gómez-Gómez, Joan, Vidana-Vila, Ester, Sevillano, Xavier, 2023. Western mediterranean wetland birds dataset: a new annotated dataset for acoustic bird species classification. *Ecol. Inform.* 75, 102014. <https://doi.org/10.1016/j.ecoinf.2023.102014>. ISSN 1574-9541.
- Hagiwara, Masato, Hoffman, Benjamin, Liu, Jen-Yu, Cusimano, Maddie, Effenberger, Felix, Zacarian, Katie, 2022. BEANS: The Benchmark of Animal Sounds. *arXiv e-prints*, art. arXiv:2210.12300. October.
- Hertkorn, Michael, 2022. Few-shot bioacoustic event detection: Don't waste information. Technical report, June. URL: [https://dcase.community/documents/challenge2022/technical\\_reports/DCASE2022\\_Hertkorn\\_28\\_5.pdf](https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Hertkorn_28_5.pdf).
- Hill, Andrew P., Prince, Peter, Covarrubias, Evelyn Piña, Patrick Doncaster, C., Snaddon, Jake L., Rogers, Alex, 2018. *Audiomoth: evaluation of a smart open*



- acoustic device for monitoring biodiversity and the environment. *Methods Ecol. Evol.* 9 (5), 1199–1211.
- Hinton, Geoffrey, Deng, Li, Dong, Yu, Dahl, George E., Mohamed, Abdel-rahman, Jaitly, Navdeep, Senior, Andrew, Vanhoucke, Vincent, Nguyen, Patrick, Sainath, Tara N., et al., 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process. Mag.* 29 (6), 82–97.
- Hou, Ruibing, Chang, Hong, Ma, Bingpeng, Shan, Shiguang, Chen, Xilin, 2019. Cross attention network for few-shot classification. *Adv. Neural Inf. Process. Syst.* 32.
- Huang, Qisheng, Li, Yanxiong, Cao, Wenchang, Chen, Hao, 2022. Few-shot bio-acoustic event detection based on transductive learning and adapted central difference convolution. Technical report, June. URL: [https://dcase.community/documents/challenge2022/technical\\_reports/DCASE2022\\_Huang\\_59\\_5.pdf](https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Huang_59_5.pdf).
- Johnson, Mark P., Tyack, Peter L., 2003a. A digital acoustic recording tag for measuring the response of wild marine mammals to sound. *IEEE J. Ocean. Eng.* 28 (1), 3–12.
- Johnson, M.P., Tyack, P.L., 2003b. A digital acoustic recording tag for measuring the response of wild marine mammals to sound. *IEEE J. Ocean. Eng.* 28 (1), 3–12. <https://doi.org/10.1109/JOE.2002.808212>.
- Joly, Alexis, Goëau, Hervé, Glotin, Hervé, Spampinato, Concetto, Bonnet, Pierre, Vellinga, Willem-Pier, Lombardo, Jean-Christophe, Planqué, Robert, Palazzo, Simone, Müller, Henning, 2019. Biodiversity information retrieval through large scale content-based identification: A long-term evaluation. In: *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF*. Springer, pp. 389–413. [https://doi.org/10.1007/978-3-030-22948-1\\_16](https://doi.org/10.1007/978-3-030-22948-1_16).
- Kahl, Stefan, Mary Clapp, W., Hopping, Hervé Goëau, Glotin, Hervé, Planqué, Robert, Vellinga, Willem-Pier, Joly, Alexis, 2020. Overview of BirdCLEF 2020: Bird sound recognition in complex acoustic environments. In: *CLEF 2020*.
- Kahl, Stefan, Wood, Connor M., Eibl, Maximilian, Klinck, Holger, 2021. Birdnet: a deep learning solution for avian diversity monitoring. *Ecol. Inform.* 61, 101236.
- Kang, Taein, 2022. Few-shot bioacoustic event detection using good embedding model. Technical report, June. URL: [https://dcase.community/documents/challenge2022/technical\\_reports/DCASE2022\\_Kang\\_10\\_5.pdf](https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Kang_10_5.pdf).
- Kiskin, Ivan, Zilli, Davide, Li, Yunpeng, Sinka, Marianne, Willis, Kathy, Roberts, Stephen, 2020. Bioacoustic detection with wavelet-conditioned convolutional neural networks. *Neural Comput. & Applic.* 32 (4), 915–927.
- Kiskin, Ivan, Sinka, Marianne, Cobb, Adam D., Rafique, Waqas, Wang, Lawrence, Zilli, Davide, Gutteridge, Benjamin, Dam, Rinita, Marinos, Theodoros, Li, Yunpeng, et al., 2021. Humbugdb: A large-scale acoustic mosquito dataset. *arXiv preprint arXiv:2110.07607*.
- Lauha, Patrik, Somervuo, Panu, Lähiköinen, Petteri, Geres, Lisa, Richter, Tobias, Seibold, Sebastian, Ovaskainen, Otso, 2022. Domain-specific neural networks improve automated bird sound recognition already with small amount of local data. *Methods Ecol. Evol.* <https://doi.org/10.1111/2041-210X.14003>.
- Leblond, Camille, Grémillet, Camille, Meylan, Sandrine, Meylan, Sandrine, Whiting, Martin J., 2021. Group size and social complexity affect individual recognition in a social lizard. *Behav. Ecol. Sociobiol.* 75 (3), 1–11.
- Lee, Yuna, Chung, HaeChun, Jung, JaeHoon, 2023. Few-Shot Bioacoustic Detection Boosting with Fine Tuning Strategy Using Negative Based Prototypical Learning. Technical report., DCASE2023 Challenge. June.
- Lehmann, Kenna D.S., 2020. Communication and Cooperation in Silico and Nature. PhD thesis., Michigan State University.
- Li, Yunpeng, Kiskin, Ivan, Sinka, Marianne, Zilli, Davide, Chan, Henry, Herreros-Moya, Eva, Chareonviriyaphap, Theeraphap, Tisgratog, Rungarun, Willis, Kathy, Roberts, Stephen, 2018. Fast mosquito acoustic detection with field cup recordings: an initial investigation. In: *DCASE*, pp. 153–157.
- Li, Ren, Liang, Jinhua, Phan, Huy, 2022. Few-shot bioacoustic event detection using prototypical networks with resnet classifier technical report. Technical report, June. URL: [https://dcase.community/documents/challenge2022/technical\\_reports/DCASE2022\\_Li\\_90\\_5.pdf](https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Li_90_5.pdf).
- Linhart, Pavel, Mahamoud-Issa, Mathieu, Stowell, Dan, Blumstein, Daniel T., 2022. The potential for acoustic individual identification in mammals. *Mamm. Biol.* <https://doi.org/10.1007/s42991-021-00222-2> mar.
- Liu, Yanbin, Lee, Juh, Park, Minseop, Kim, Saehoon, Yang, Eunho, Hwang, Sung Ju, Yang, Yi, 2018. Learning to propagate labels: Transductive propagation network for few-shot learning. *arXiv preprint arXiv:1805.10002*.
- Liu, Haohe, Liu, Xubo, Mei, Xinhao, Kong, Qiuqiang, Wang, Wenwu, Plumbley, Mark D., 2022a. Surrey system for dcase 2022 task 5: Few-shot bioacoustic event detection with segment-level metric learning. Technical report, June. URL: [https://dcase.community/documents/challenge2022/technical\\_reports/DCASE2022\\_Haohe\\_85\\_5.pdf](https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Haohe_85_5.pdf).
- Liu, Miao, Zhang, Jianqian, Wang, Lizhong, Peng, Jiawei, Hu, Chenguang, Li, Kaige, Wang, Jing, Ma, Qiuyue, 2022b. Bit srcb team's submission for dcase2022 task5 - few-shot bioacoustic event detection. Technical report, June. URL: [https://dcase.community/documents/challenge2022/technical\\_reports/DCASE2022\\_Liu\\_43\\_5.pdf](https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Liu_43_5.pdf).
- Liu, Junyan, Zhou, Zikai, Sun, Mengkai, Kele, Xu, Qian, Kun, Bian, Hu., 2023. Se-Protonet: Prototypical Network with Squeeze-and-Excitation Blocks for Bioacoustic Event Detection. Technical report., DCASE2023 Challenge. June.
- Logue, David M., Krupp, Daniel Brian, 2016. Duetting as a collective behavior. *Front. Ecol. Evol.* 4 <https://doi.org/10.3389/fevo.2016.00007> feb.
- Lastanlen, Vincent, Salamon, Justin, Farnsworth, Andrew, Kelling, Steve, Bello, Juan Pablo, 2018. Birdvox-full-night: A dataset and benchmark for avian flight call detection. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 266–270.
- Manser, Martha B., 1998. The Evolution of Auditory Communication in Suricates, *Suricata suricatta*. PhD thesis., University of Cambridge.
- Manser, Marta B., Jansen, David A.W.A.M., Graw, Beke, Hollén, Linda I., Bousquet, Christophe A.H., Furrer, Roman D., le Roux, Aliza, 2014. Chapter six - vocal complexity in meerkats and other mongoose species. In: *Advances in the Study of Behavior*, vol. 46. Academic Press, pp. 281–310. <https://doi.org/10.1016/B978-0-12-800286-5.00006-7>. URL: <https://www.sciencedirect.com/science/article/pii/B9780128002865000067>.
- Mariajohn, Aaquila, 2022. Bioacoustic few shot learning with class augmentation technical report. Technical report, June. URL: [https://dcase.community/documents/challenge2022/technical\\_reports/DCASE2022\\_Mariajohn\\_104\\_5.pdf](https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Mariajohn_104_5.pdf).
- Marler, P.R., Slabbekoorn, H., 2004. *Nature's Music: The Science of Birdsong*. Academic Press, Massachusetts, USA.
- Martinsson, John, Willbo, Martin, Pirinen, Aleksis, Mogren, Olof, Sandsten, Maria, 2022. Few-shot bioacoustic event detection using a prototypical network ensemble with adaptive embedding functions. Technical report, June. URL: [https://dcase.community/documents/challenge2022/technical\\_reports/DCASE2022\\_Martinsson\\_78\\_5.pdf](https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Martinsson_78_5.pdf).
- Marx, G., Leppelt, J., Ellendorff, F., 2001. Vocalisation in chicks (*gallus gallus dom.*) during stepwise social isolation. *Appl. Anim. Behav. Sci.* 75 (1), 61–74.
- Mesáros, Annamaria, Heittola, Toni, Virtanen, Tuomas, 2016. Metrics for polyphonic sound event detection. *Appl. Sci.* 6 (6), 162. <https://doi.org/10.3390/app6060162>.
- Mesáros, Annamaria, Diment, Aleksandr, Elizalde, Benjamin, Heittola, Toni, Vincent, Emmanuel, Raj, Bhiksha, Virtanen, Tuomas, 2019. Sound event detection in the dcase 2017 challenge. *IEEE/ACM Trans. Audio Speech Lang. Process.* 27 (6), 992–1006.
- Morfi, Veronica, Nolasco, Inês, Lastanlen, Vincent, Singh, Shubhr, Strandburg-Peshkin, Ariana, Gill, Lisa F., Pamula, Hanna, Benvent, David, Stowell, Dan, 2021. Few-shot bioacoustic event detection: A new task at the dcase 2021 challenge. In: *DCASE*, pp. 145–149.
- Moummad, Ilyass, Serizel, Romain, Farrugia, Nicolas, 2023. Supervised Contrastive Learning for Pre-Training Bioacoustic Few Shot Systems. Technical report., DCASE2023 Challenge. June.
- Naranjo-Alcazar, Javier, Perez-Castanos, Sergi, Zuccarello, Pedro, Torres, Ana M., Lopez, Jose J., Ferri, Francesc J., Cobos, Maximo, 2022. An open-set recognition and few-shot learning dataset for audio event classification in domestic environments. *Pattern Recogn. Lett.* <https://doi.org/10.1016/j.patrec.2022.10.019>.
- Pamula, Hanna, 2022. Nocturnal Flight Calls Dataset: Long-Term Acoustic Monitoring of Birds Migrating at Night. May.
- Pamula, Hanna, 2022. Nowe Metody Akustycznej Identyfikacji ptaków Migrujących Noca / Novel Methods for Acoustic Identification of Birds Migrating at Night. PhD thesis., University of Science and Technology, Kraków, Poland.
- Parnami, Archit, Lee, Minwoo, 2022. Learning from few examples: A summary of approaches to few-shot learning. *arXiv preprint arXiv:2203.04291*.
- Pons, Jordi, Serrà, Joan, Serra, Xavier, 2019. Training neural audio classifiers with few data. In: *Proc ICASSP 2019*. URL: <https://arxiv.org/abs/1810.10274>.
- Ravi, Sachin, Larochelle, Hugo, 2017. Optimization as a model for few-shot learning. In: *Proc ICLR 2017*.
- Riede, Klaus, 2018. Acoustic profiling of Orthoptera: present state and future needs. *J. Orthop. Res.* 27 (2), 203–215. <https://doi.org/10.3897/jor.27.23700>.
- Roe, Paul, Eichinski, Philip, Fuller, Richard A., McDonald, Paul G., Schwarzkopf, Lin, Towsey, Michael, Trusking, Anthony, Tucker, David, Watson, David M., 2021. The Australian acoustic observatory. *Methods Ecol. Evol.* 12 (10), 1802–1808.
- Schaul, Tom, Schmidhuber, Jürgen, 2010. Metalearning. In: *Scholarpedia*, 5, p. 4650.
- Sethi, Sarab S., Ewers, Robert M., Jones, Nick S., Signorelli, Aaron, Picinali, Lorenzo, Christopher David, L., Orme., 2020. SAFE acoustics: an open-source, real-time eco-acoustic monitoring network in the tropical rainforests of Borneo. *Methods Ecol. Evol.* 11 (10), 1182–1185.
- Shi, Bowen, Sun, Ming, Puvvada, Krishna C., Kao, Chieh-Chi, Matsoukas, Spyros, Wang, Chao, 2020. Few-shot acoustic event detection via meta learning. In: *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 76–80. <https://doi.org/10.1109/ICASSP40776.2020.9053336>.
- Simard, Peter, Lace, Natalija, Gowans, Shannon, Quintana-Rizzo, Ester, Kuczaj, Stan A., Wells, Randall S., Mann, David A., 2011. Low frequency narrow-band calls in bottlenose dolphins (*tursiops truncatus*): signal properties, function, and conservation implications. *J. Acoust. Soc. Am.* 130 (5), 3068–3076.
- Snell, Jake, Swersky, Kevin, Zemel, Richard, 2017. Prototypical networks for few-shot learning. In: *Advances in Neural Information Processing Systems*, vol. 30.
- Stowell, Dan, 2018. Computational bioacoustic scene analysis. In: *Computational Analysis of Sound Scenes and Events*. Springer, pp. 303–333.
- Stowell, D., Gill, L.F., Clayton, D., 2016. Detailed temporal structure of communication networks in groups of songbirds. *J. R. Soc. Interface* 13 (119). <https://doi.org/10.1098/rsif.2016.0296>.
- Stowell, Dan, Benetos, Emmanouil, Gill, Lisa F., 2017. On-bird sound recordings: automatic acoustic recognition of activities and contexts. *IEEE/ACM Trans. Audio Speech Lang. Process.* 25 (6), 1193–1206.
- Stowell, Dan, Wood, Michael D., Pamula, Hanna, Stylianou, Yannis, Glotin, Hervé, 2019. Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. *Methods Ecol. Evol.* 10 (3), 368–380. April. <https://doi.org/10.1111/2041-210X.13103>.
- Tan, Yizhou, Xu, Lifan, Zhu, Chenyang, Li, Shengchen, Ai, Haojun, Shao, Xi, 2022. A new transductive framework for few-shot bioacoustic event detection task. Technical report, June. URL: [https://dcase.community/documents/challenge2022/technical\\_reports/DCASE2022\\_Tan\\_39\\_5.pdf](https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Tan_39_5.pdf).
- Tang, Tiantian, Liang, Yunhao, Long, Yanhua, 2021. Two Improved Architectures Based on Prototype Network for Few-Shot Bioacoustic Event Detection. Technical report., DCASE2021 Challenge. June.

- Tang, Jigang, Zhang, Xueyang, Gao, Tian, Liu, Diyuang, Fang, Jia Pan Xin, Wang, Qing, Jun, Du, Xu, Kele, Pan, Qinghua, 2022. Few-shot embedding learning and event filtering for bioacoustic event detection. Technical report, June. URL: [https://dcase.community/documents/challenge2022/technical\\_reports/DCASE2022\\_Du\\_122\\_5.pdf](https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Du_122_5.pdf).
- Towsey, M., Planitz, B., Nantes, A., Wimmer, J., Roe, P., 2012. A toolbox for animal call recognition. *Bioacoustics* 21 (2), 107–125. <https://doi.org/10.1080/09524622.2011.648753>.
- Tseng, Hung-Yu, Lee, Hsin-Ying, Huang, Jia-Bin, Yang, Ming-Hsuan, 2020. Cross-domain few-shot classification via learned feature-wise transformation. *arXiv preprint arXiv:2001.08735*.
- Turian, Joseph, Shier, Jordie, Khan, Humair Raj, Raj, Bhiksha, Schuller, Björn W., Steinmetz, Christian J., Malloy, Colin, Tzanetakis, George, Velarde, Gissel, McNally, Kirk, Henry, Max, Pinto, Nicolas, Noufi, Camille, Clough, Christian, Herremans, Dorien, Fonseca, Eduardo, Engel, Jesse, Salamon, Justin, Esling, Philippe, Manocha, Pranay, Watanabe, Shinji, Jin, Zeyu, Bisk, Yonatan, 2022. HEAR 2021: Holistic evaluation of audio representations. In: Kiela, Douwe, Ciccone, Marco, Caputo, Barbara (Eds.), *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176 of *Proceedings of Machine Learning Research*. PMLR, pp. 125–145. March. URL: <https://proceedings.mlr.press/v176/turian22a.html>.
- Van Horn, Grant, Cole, Elijah, Beery, Sara, Wilber, Kimberly, Belongie, Serge, Aodha, Oisín Mac, 2021. Benchmarking representation learning for natural world image collections. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12884–12893.
- Vellinga, Willem-Pier, Planqué, Robert, 2015. The xeno-canto collection and its relation to sound recognition and classification. In: *CLEF (Working Notes)*.
- Versace, Elisabetta, Spierings, Michelle J., Caffini, Matteo, ten Cate, Carel, Vallortigara, Giorgio, 2017. Spontaneous generalization of abstract multimodal patterns in young domestic chicks. *Anim. Cogn.* 20 (3), 521–529.
- Vinyals, Oriol, Blundell, Charles, Lillicrap, Timothy, Wierstra, Daan, et al., 2016. Matching networks for one shot learning. *Adv. Neural Inf. Proces. Syst.* 29.
- Wang, Duo, Yu, Cheng, Yu, Mo, Guo, Xiaoxiao, Zhang, Tao, 2019a. A hybrid approach with optimization and metric-based meta-learner for few-shot learning. *arXiv preprint arXiv:1904.03014*.
- Wang, Yan, Chao, Wei-Lun, Weinberger, Kilian Q., van der Maaten, Laurens, 2019b. SimpleShot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*.
- Wang, Yaqing, Yao, Quanming, Kwok, James T., Ni, Lionel M., 2020a. Generalizing from a few examples: a survey on few-shot learning. *ACM Comput. Surv.* 53 (3), 1–34. <https://doi.org/10.1145/3386252>.
- Wang, Yu, Salamon, Justin, Cartwright, Mark, Bryan, Nicholas J., Bello, Juan Pablo, 2020b. Few-shot drum transcription in polyphonic music. *CoRR*, abs/2008.02791. URL: <https://arxiv.org/abs/2008.02791>.
- Wang, Yu, Bryan, Nicholas J., Cartwright, Mark, Bello, Juan Pablo, Salamon, Justin, 2021. Few-shot continual learning for audio classification. In: *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 321–325.
- Wilkinghoff, Kevin, Cornaggia-Urrigshardt, Alessia, 2023. Few-Shot Bioacoustic Event Detection. Technical report, DCASE2023 Challenge. June.
- Willbo, Martin, Martinsson, John, Pirinen, Aleksis, Mogren, Olof, 2022. Wide resnet models for few-shot sound event detection. Technical report, June. URL: [https://dcase.community/documents/challenge2022/technical\\_reports/DCASE2022\\_Willbo\\_53\\_5.pdf](https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Willbo_53_5.pdf).
- Wolters, Piper, Daw, Chris, Hutchinson, Brian, Phillips, Lauren, 2021. Proposal-based few-shot sound event detection for speech and environmental sounds with perceivers. *arXiv preprint arXiv:2107.13616*.
- Wu, Xiaoxiao, Long, Yanhua, 2022. Few-shot continual learning for bioacoustic event detection. Technical report, June. URL: [https://dcase.community/documents/challenge2022/technical\\_reports/DCASE2022\\_Wu\\_4\\_5.pdf](https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Wu_4_5.pdf).
- Yan, Genwei, Wang, Ruoyu, Zou, Liang, Jun, Du, Wang, Qing, Gao, Tian, Fang, Xin, 2023. Multi-Task Frame Level System for Few-Shot Bioacoustic Event Detection. Technical report, DCASE2023 Challenge. June.
- Yang, Dongchao, Wang, Helin, Ye, Zhongjie, Zou, Yuexian, 2021. Few-Shot Bioacoustic Event Detection = a Good Transductive Inference Is all you Need. Technical report, DCASE2021 Challenge. June.
- Yang, Dongchao, Zou, Yuexian, Cui, Fan, Wang, Yujun, 2022. Improved prototypical network with data augmentation. Technical report, June. URL: [https://dcase.community/documents/challenge2022/technical\\_reports/DCASE2022\\_Zou\\_36\\_5.pdf](https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Zou_36_5.pdf).
- You, Liwen, Coyotl, Erika Pelaez, Gunturu, Suren, Van Segbroeck, Maarten, 2023. Transformer-based bioacoustic sound event detection on few-shot learning tasks. In: *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5.
- Zgorzynski, Bartłomiej, Matuszewski, Mateusz, 2022. Siamese network for few-shot bioacoustic event detection. Technical report, June. URL: [https://dcase.community/documents/challenge2022/technical\\_reports/DCASE2022\\_Zgorzynski\\_55\\_5.pdf](https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Zgorzynski_55_5.pdf).
- Zhang, Tianyang, Wang, Yuyang, Wang, Ying, 2022. A meta-learning framework for few-shot sound event detection. Technical report, June. URL: [https://dcase.community/documents/challenge2022/technical\\_reports/DCASE2022\\_Zhang\\_6\\_5.pdf](https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Zhang_6_5.pdf).

## Glossary

- FSED*: Few-shot bioacoustic sound event detection. 5, 28, 30
- FSL*: Few-shot learning. 4, 30
- ML*: Machine learning. 4
- PCEN*: Per-channel Energy Normalization. 26
- Query set*: The data for which predictions are to be generated in each sub-task (here, one or more long audio clips). 19
- SED*: Sound event detection. 5
- Support set*: The small set of data that helps define each new sub-task (here, 5 example sounds, and the background sound between). 19