FEW-SHOT BIOACOUSTIC EVENT DETECTION AT THE DCASE 2022 CHALLENGE

I. Nolasco[1], *S. Singh*[1], *E. Vidaña-Vila*[2], *E. Grout*[3,4], *J. Morford*[5], *M.G. Emmerson*[6] *F. H. Jensen*[7], *I. Kiskin*[8] *H. Whitehead*[9], *A. Strandburg-Peshkin*[3,4], *L. Gill*[10], *H. Pamuła*[11], *V. Lostanlen*[12], *V. Morfi*[1,13], *D. Stowell*[14]

[1] Centre for Digital Music (C4DM), Queen Mary University of London, London, UK
[2] La Salle, University Ramon Llull, Barcelona, ES
[3] Dept. of Biology & Centre for the Advanced Study of Collective Behaviour, University of Konstanz, DE
[4] Dept. for the Ecology of Animal Societies, Max Planck Institute of Animal Behavior, DE
[5] The Oxford Navigation group, Dept. of Zoology, Oxford University, Oxford, UK
[6] School of Biological and Behavioural Sciences, Queen Mary University of London, London, UK
[7] Biology Dept, Syracuse University, NY, USA
[8] Institute for People-Centred AI, FHMS, University of Surrey, Surrey, UK
[9] School of Science, Engineering and Environment, University of Salford, Manchester, UK
[10] BIOTOPIA Naturkundemuseum Bayern, Munich, DE
[11] AGH University of Science and Technology, Kraków, PL
[12] Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, FR
[13] Sonantic Limited, London, UK
[14] Tilburg University, Tilburg, The Netherlands; Naturalis Biodiversity Centre, Leiden, NL

ABSTRACT

Few-shot sound event detection is the task of detecting sound events, despite having only a few labelled examples of the class of interest. This framework is particularly useful in bioacoustics, where often there is a need to annotate very long recordings but the expert annotator time is limited. This paper presents an overview of the second edition of the few-shot bioacoustic sound event detection task included in the DCASE 2022 challenge. A detailed description of the task objectives, dataset, and baselines is presented, together with the main results obtained and characteristics of the submitted systems. This task received submissions from 15 different teams from which 13 scored higher than the baselines. The highest F-score was of 60.2% on the evaluation set, which leads to a huge improvement over last year's edition. Highly-performing methods made use of prototypical networks, transductive learning, and addressed the variable length of events from all target classes. Furthermore, by analysing results on each of the subsets we can identify the main difficulties that the systems face, and conclude that few-show bioacoustic sound event detection remains an open challenge.

Index Terms— Few-shot learning, bioacoustics, sound event detection, DCASE challenge

1. INTRODUCTION

The task of bioacoustic sound event detection refers to the retrieval of animal vocalisations from audio recordings in terms of onset and offset times. It shares a common methodology with other sound event detection (SED) contexts, yet, the application domain of bioacoustics is particularly challenging for SED. Deep learning contributed to overcome some of these difficulties in bioacoustic SED, however it also established strong requirements regarding the amount of annotated data needed [1]. Collecting and annotating a large dataset of animal vocalisations is often not feasible given that species are unequally abundant [2] and may be rarely observed; and audio annotation is costly and time-consuming [3]. In contrast to traditional deep learning approaches that use a large amount of data to train models, few-shot learning tries to build accurate models with very few training data [4]. Few-shot learning is usually studied using N-way-k-shot classification, where N denotes the number of classes and k the number of known examples for each class.

This problem was first evaluated as a task on the DCASE 2021 challenge [5]. This year, the setup and goal remain the same: Given the first 5 events of a target class, can systems detect the subsequent events of the same class in the remaining of the audio recording? Diverse approaches have been used to address the few-shot learning problem for classification. Some use prior knowledge about similarity between sounds by computing embeddings (learnt representation spaces) designed to help discriminate between unseen classes [4], while others exploit prior knowledge about the structure of the data by using augmentation to synthesise new data [6]. Finally, some approaches can learn models with parameters that can be fine-tuned to smaller datasets [7]. More recent works use meta-learning and/or prototypical networks for acoustic few-shot learning [8], [9]. All of the above approaches deal with classification tasks rather than detection. Indeed, SED in a few shot setup is commonly approximated as an audio tagging task and few works have addressed the actual detection of onsets and offsets of events [10]. At last year's task edition, the best ranked system improved over the baseline prototypical approach by applying a transductive inference method and a mutual learning framework designed to make the feature extraction network more task dependent [11]. The overall best results were just below 40% F-score which indicates the difficulty of this task. This year, we added more and diverse datasets, and

increased the task difficulty (dataset diversity); yet the task doubled the amount of participants and the best overall F-score in the evaluation set reached the 60% level. This paper is structured as follows. Section 2 presents the bioacoustic datasets used for developing and evaluating submitted systems. Section 3 presents the two baseline methods proposed for the task, followed by the evaluation procedure. Finally, section 4 presents the results of the submitted systems and a discussion about the overall task and future steps in the field of few-shot bioacoustic event detection.

2. DATASETS

A *development dataset* consists of predefined training and validation sets to be used for system development[1]. The training set contains multi-class temporal annotations, provided for each recording as: positive (POS), negative (NEG) and unknown (UNK). For the validation set only single-class temporal annotations (POS/UNK) were provided for each recording. A separate *evaluation set* was kept for evaluating the performance of the systems[2]. During the challenge, only the first five POS events of the class of interest were provided for each of the recordings. Table 1 presents an overview of all the datasets in the development and evaluation sets.

BirdVox-DCASE-10h (BV): This dataset contains audio files recorded in 2015, during the fall migration season by four different autonomous recording units located in Tompkins County, NY, USA. An expert ornithologist, Andrew Farnsworth, has annotated flight calls from four families of passerines, namely: American sparrows, cardinals, thrushes, and New World warblers. These flight calls have a duration in the range 50–150 ms and a fundamental frequency in the range 2–10 kHz.

Hyenas (HT): Hyenas use a variety of types of vocalizations to coordinate with one another over both short and long distances. Spotted hyenas were recorded on custom-developed audio tags designed by Mark Johnson and integrated into combined GPS/acoustic collars (Followit Sweden AB) by Frants Jensen and Mark Johnson. 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 behaviour. The recordings contain 5 different vocalisations which were identified and classified into types based on the established hyena vocal repertoire [12]. Field work was carried out by Kay Holekamp, Andrew Gersick, Frants Jensen, Ariana Strandburg-Peshkin, and Benson Pion; labelling was done by Kenna Lehmann and colleagues.

Meerkats (MT, ME): 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 of meerkats 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 behaviour. Recordings of the training set (MT) were recorded on 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. Recordings of the validation set (ME) were recorded by an observer following a focal meerkat with a Sennheiser ME66 directional microphone (44.1 kHz) from a distance of less than 1m. Recordings were carried out during daytime hours while meerkats were primarily foraging and include several different call types. Field work was carried out by Ariana Strandburg-Peshkin, Baptiste Averly, Vlad Demartsev, Gabriella Gall, Rebecca Schaefer and Marta Manser. Audio recordings were labelled by Baptiste Averly, Vlad Demartsev, Ariana Strandburg-Peshkin, and colleagues.

Jackdaws (JD): Jackdaws are corvid songbirds that usually breed, forage and sleep in large groups.In a multi-year field study (Max-Planck-Institute for Ornithology, Seewiesen, Germany), wild jackdaws were equipped with small backpacks containing miniature voice recorders (Edic Mini Tiny A31, TS-Market Ltd., Russia) to investigate the vocal behaviour of individuals interacting with their group and behaving freely in their natural environment. Field work was conducted by Lisa Gill, Magdalena Pelayo van Buuren and Magdalena Maier. Sound files were annotated by Lisa Gill.

Western Mediterranean Wetlands Bird Dataset (WMW): Contains 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 portaland were manually cleaned and labelled by Juan Gómez-Gómez, Ester Vidaña-Vila and Xavier Sevillano using the Audacity software [13].

HumBug (HB): 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 [14]. 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 [15][3]. The recordings contain the sound of lab-cultured *Culex quinquefasciatus* mosquitoes from Oxford, UK, and various species captured in the wild in Thailand, placed into plastic cups.[16].

Polish Baltic Sea bird flight calls (PB): The PB dataset consists of bird flight calls recordings from Hanna Pamuła's project focused on the acoustic monitoring of birds migrating at night along the Polish Baltic Sea coast. Three autonomous recording units (Song Meters SM2, Wildlife Acoustics, Inc) were deployed close to each other (<100m) near a lake, on the dune, and on a forest clearing. The passerines night flight calls were annotated by Hanna Pamuła. Target classes belong to: song thrush, *Turdus philomelos* and blackbird, *Turdus merula*. Event lengths vary between 8 to 400 milliseconds and the usual fundamental frequency range for calls is 5 to 9 kHz.

Transfer-Exposure-Effects dataset (CHE): This dataset comes from the Transfer Exposure-Effects (TREE) research project[4], Data were collected using unattended acoustic recorders (Songmeter 3) in the Chornobyl Exclusion Zone (CEZ) to capture the Chornobyl soundscape and investigate the longterm effects of the nuclear plower plant accident on the local ecology. 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): 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.The DC dataset stems from dawn chorus recordings,

[1]Dev set: https://doi.org/10.5281/zenodo.6012309
[2]Eval set: https://doi.org/10.5281/zenodo.6517413
[3]https://github.com/HumBug-Mosquito/HumBugDB/
[4]https://tree.ceh.ac.uk/

	Dataset	mic type	# audio files	total duration	# labels (excl. UNK)	# events
Development Set: Training	BV	fixed	5	10 hours	11	9026
	HT	various	5	5 hours	5	611
	MT	animal mounted	2	70 mins	4	1294
	JD	mobile	1	10 mins	1	357
	WMW	various	161	5 hours	26	2941
Development Set: Validation	HB	handheld	10	2.38 hours	1	712
	PB	fixed	6	3 hours	2	292
	ME	animal mounted	2	20 mins	2	73
Evaluation Set	CHE	fixed	18	3 hours	3	2550
	DC	fixed	10	95 mins	3	967
	CT	handheld	3	48 mins	3	365
	MS	fixed	4	40 mins	1	1087
	QU	animal mounted	8	74 mins	1	3441
	MGE	fixed	3	32 mins	2	1195

Table 1: Information on each dataset.

made using Zoom H2 recorders at 44100 Hz, at three different locations in Southern Germany (Haspelmoor, Munich's Nymphenburg Schlosspark, and Nantesbuch), by Moritz Hertel and Rudi Schleich. The vocalisations of three target species were annotated by Lisa Gill (Common cuckoo, *Cuculus canorus*; European robin, *Erithacus rubecula*; Eurasian wren, *Troglodytes troglodytes*). A challenging aspect of this data is related to recordings being very busy with various other birds vocalising at the same time.

Coati (CT): Coatis are omnivorous diurnal mammals that live in stable social groups ranging from 5 to 30 individuals. The target calls used in this dataset are growls, chitters and chirp-grunts. Several other call types that might be confused with the targets were captured in the recordings which configures the main challenging aspect of this data. Audio recordings were collected from two adult females from the same group on Barro Colorado Island, Panama in March 2020. These individuals wore collars which recorded high resolution GPS data with an external attachment of a small audio recording device (TS Market, Edic Mini Tiny+ A77, 22050 Hz). Audio data were recorded during their active foraging period in daytime hours.Fieldwork was carried out by Emily Grout, Josué Ortega and Ben Hirsch.

Manx Shearwater (MS): Manx shearwaters are procellariform seabirds that breed in dense island colonies in the North Atlantic and winter in the South Atlantic off the South American coast. Adult Manx shearwaters make loud, distinctive vocalisations while present at their breeding colony in various contexts.The target class is Chick begging vocalisations which typically comprise bouts of short, high-pitched 'peeps'. In a multi-year study, Audiomoth recorders were placed in burrows during the breeding season. Fieldwork on Skomer Island was undertaken by various members of the Oxford Navigation Group (OxNav) and annotation was carried out by Joe Morford.

Dolphin Quacks (QU): Bottlenose dolphins are highly acoustic animals with an expansive repertoire of acoustic signals used for social interactions. The target class is Quacks which are short signals (around 100 ms), low-frequencyand emitted at relatively high rates, often with 100s of quacks in a single short vocal bout. The recordings were obtained using sound-and-movement recording DTAGs (Johnson and Tyack 2003) attached with suction cups to bottlenose dolphins by F.Jensen in collaboration with Profs. Peter Tyack, Vincent Janik, and Laela Sayigh. 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. Individual quacks were labelled by Austin Dziki and validated by F. Jensen.

Chick calls (MGE): Chickens are a precocial bird and upon hatching undergo a process of 'filial imprinting' whereby they establish a strong attachment to their mother. Chicks are active participants in this filial imprinting process and use their calls to signal they are in close proximity to their mother and other family members (i.e. pleasure calls) and to signal distress during social separation in order to solicit maternal contact (i.e. contact calls). The dataset includes three chicks with each chick recorded for 10 minutes; pleasure calls were annotated in recordings from chicks one and two, contact calls were annotated in recordings from chick three. All data was collected by Elisabetta Versace, Shu Wang, and Laura Freeland as part of a project from the Prepared Minds Lab from Queen Mary University of London[5]. All data were annotated by Shu Wang, Laura Freeland, and Michael Emmerson.

3. BASELINE METHODS AND EVALUATION

The baseline systems proposed did not change considerably from last year's edition [5]. Template Matching is based on spectrogram cross-correlation and still commonly used in bioacoustics. This approach scored surprisingly well on last edition evaluation set and thus it remains relevant as a baseline for this task. The second system proposed is based on prototypical networks which remain the state of the art for few-shot learning [4]. The changes from last year's system are the use of a ResNet, and adapting segment size depending on the target class in the query set. These changes mainly address the problem of high variation of event lengths and create a more adaptive system.

The evaluation of this task is based on an event-level F-measure with macro-averaged metric across all classes [5]. A positive match between predicted events and reference is found by applying the Intersection over Union (IoU) with 30% minimum overlap, followed by Hopcroft-Karp-Karzanov algorithm [18] for bipartite graph matching. True Positives (TP), False Positives (FP), and False Negatives (FN) can be computed after the matching step. These are defined as: TP - predicted events that match ground truth events; FP - predicted events that do not match any ground truth events;

[5] https://www.preparedmindslab.org/home

Team code	Code	Eval set: F-score % (95% CI)	Val set F-score %	Main characteristics
Du_NERCSLIP	(A)	60.22 (59.66-60.70)	74.4	CNN+ProtoNet; Frame-level embeddings; PCEN;
Liu_Surrey	(B)	48.52 (48.18-48.85)	50.03	CNN+ProtoNet; extra data; PCEN+$\triangle MFCC$; various post-process.
Martinsson_RISE	(C)	47.97 (47.48-48.40)	60	ResNet+ProtoNet; Ensemble(15) based input size; logMel+PCEN
Hertkorn_ZF	(D)	44.98 (44.44-45.42)	61.76	CNN; Frequency resolution preserving pooling; various post-process
Liu_BIT-SRCB	(E)	44.26 (43.85-44.62)	64.77	CNN+ProtoNet; Transductive inference
Wu_SHNU	(F)	40.93 (40.48-41.30)	53.88	ResNet+ProtoNet; Continual-learning; spectrogram
Zgorzynski_SRPOL	(G)	33.24 (32.69-33.69)	57.2	CNN+Siamese Networks; Emsemble (3) average event-length;
Mariajohn_DSPC	(H)	25.66 (25.40-25.91)	43.89	CNN+ProtoNet; logMel; augmentation with time-shifting and mirroring
Wilbo_RISE	(I)	21.67 (21.32-21.97)	47.94	ResNet+ProtoNEt; Semi-supervised; Melspect+PCEN; various post-process
Zou_PKU	(J)	19.20 (18.88-19.51)	51.99	CNN+protoNet; mutual information loss; time frequency masking + mixup
Huang_SCUT	(K)	18.29 (18.01-18.56)	54.63	Transductive inference + Adapted central difference convolution
Tan_WHU	(L)	17.22 (16.82-17.55)	54.53	CNN+ProtoNet pretrained; transductive inference; task adaptive features
Li_QMUL	(M)	15.49 (15.16-15.77)	47.88	CNN+protoNet; PCEN; time, frequency masking + time warping
baseline-TempMatch	[5]	12.35 (11.52-12.75)	3.37	Spectrogram Cross correlation
baseline-ProtoNet	[5]	5.3 (5.1-5.2)	28.45	ResNet+ProtoNet
Zhang_CQU	(N)	4.34 (3.74-4.56)	44.17	CNN+protoNet; Fine tunning with MIMI; PCEN
Kang_ET	(O)	2.82 (2.76-2.87)	-	CNN+ProtoNEt; pretrained ECAPA-TDNN; Fine-tuning; Specaugment

Table 2: F-score results per team (best scoring system) on evaluation and validation sets, and summary of system characteristics. 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 [17].

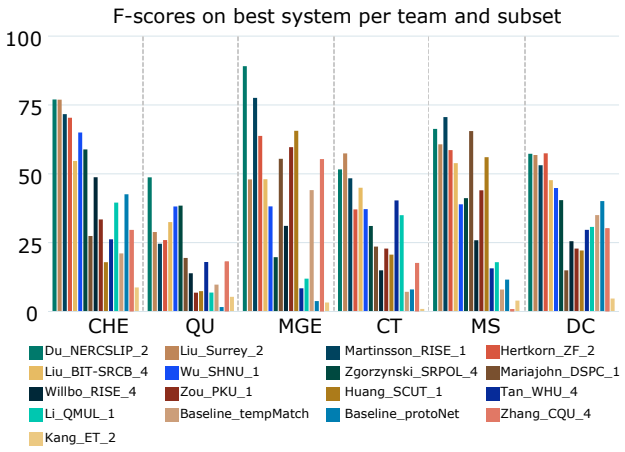

Figure 1: F-Score results by dataset. Systems are ordered by highest scoring rank on the evaluation set.

FN - ground truth events that are not predicted. Matches to UNK events are ignored from these counts as to not negatively impact the systems that predict these events. Finally, the F-score metric is computed per dataset in the evaluation set and the harmonic mean over all is reported.

4. RESULTS

For the 2022 edition, 15 teams participated submitting a total of 46 systems. The results for the highest scoring submission for each team are presented in Table 2, together with the reported F-scores on the validation set and summary of the system characteristics. Fig. 1 presents the F-scores obtained by each team on each subset of the evaluation data. The majority of systems adopted a prototypical network approach. Similar to last year's results, 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 report improved results (B, C, F, I). Transductive inference, the method used by the past edition's winning team, was also applied here by several participants (B, M, L, J). The highest scoring system implements a frame-level embedding learning approach which confers to the system a high time resolution capability (A). The system was particularly effective on the QU and MGE dataset (Fig.1). This confirms that good time precision is fundamental, particularly for classes with events of very short duration as the ones in these datasets. The system ranked in second place implements a novel approach designed to optimise the contrast between positive events and negative prototypes (B). 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 (C) and (G) implemented an ensemble approach where each individual model focuses on a different input size range. In (E) this is explored through a multi-scale ResNet, and in (I) with a wide ResNet containing many channels. Finally, it is worth mentioning the system in (D). Their few-shot adaptation was based on fine-tuning alone. The innovation here is related to simple modifications to a CNN-based architecture in order to optimise the use of information, particularly in the frequency axis. Furthermore, by allowing the network to overfit (up to a degree) to the 5 shots, the system achieves surprisingly good performance across all the datasets of the evaluation set.

Overall, this edition saw some novel ideas being implemented that tried to address previously identified 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. We believe these remain relevant questions for our goal and for SED in general, and that the collective work developed here helped pushing few-shot bioacoustic sound event detection into DCASE central stage.

5. REFERENCES

[1] D. Stowell, "Computational bioacoustics with deep learning: a review and roadmap," *PeerJ*, vol. 10, 2022.

[2] W.-P. Vellinga and R. Planqué, "The xeno-canto collection and its relation to sound recognition and classification," in *CLEF (Working Notes)*, 2015.

[3] A. E. Méndez Méndez, M. Cartwright, and J. P. Bello, "Machine–crowd–expert model for increasing user engagement and annotation quality," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–6.

[4] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[5] V. Morfi, I. Nolasco, V. Lostanlen, S. Singh, A. Strandburg-Peshkin, L. F. Gill, H. Pamula, D. Benvent, and D. Stowell, "Few-shot bioacoustic event detection: A new task at the dcase 2021 challenge." in *DCASE*, 2021, pp. 145–149.

[6] Y.-X. Wang, R. B. Girshick, M. Hebert, and B. Hariharan, "Low-shot learning from imaginary data," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7278–7286, 2018.

[7] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *ArXiv*, vol. abs/1803.02999, 2018.

[8] Y. Wang, J. Salamon, M. Cartwright, N. J. Bryan, and J. P. Bello, "Few-shot drum transcription in polyphonic music," *CoRR*, vol. abs/2008.02791, 2020.

[9] Y. Shiu, K. Palmer, M. A. Roch, E. Fleishman, X. Liu, E.-M. Nosal, T. Helble, D. Cholewiak, D. Gillespie, and H. Klinck, "Deep neural networks for automated detection of marine mammal species," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[10] P. Wolters, C. Daw, B. Hutchinson, and L. Phillips, "Proposal-based few-shot sound event detection for speech and environmental sounds with perceivers," *arXiv preprint arXiv:2107.13616*, 2021.

[11] D. Yang, H. Wang, Z. Ye, and Y. Zou, "Few-shot bioacoustic event detection = a good transductive inference is all you need," DCASE2021 Challenge, Tech. Rep., June 2021.

[12] K. D. S. Lehmann, "Communication and cooperation in silico and nature," Ph.D. dissertation, Michigan State University, 2020.

[13] J. Gómez-Gómez, E. Vidaña-Vila, and X. Sevillano, "Western mediterranean wetlands bird species classification: evaluating small-footprint deep learning approaches on a new annotated dataset," 2022. [Online]. Available: https://arxiv.org/abs/2207.05393

[14] I. Kiskin, D. Zilli, Y. Li, M. Sinka, K. Willis, and S. Roberts, "Bioacoustic detection with wavelet-conditioned convolutional neural networks," *Neural Computing and Applications*, vol. 32, no. 4, pp. 915–927, 2020.

[15] I. Kiskin, M. Sinka, A. D. Cobb, W. Rafique, L. Wang, D. Zilli, B. Gutteridge, R. Dam, T. Marinos, Y. Li, *et al.*, "Humbugdb: a large-scale acoustic mosquito dataset," *arXiv preprint arXiv:2110.07607*, 2021.

[16] Y. Li, I. Kiskin, M. Sinka, D. Zilli, H. Chan, E. Herreros-Moya, T. Chareonviriyaphap, R. Tisgratog, K. Willis, and S. Roberts, "Fast mosquito acoustic detection with field cup recordings: an initial investigation." in *DCASE*, 2018, pp. 153–157.

[17] "DCASE challenge 2022 Few-shot bioacoustic event detection task - results page," accessed: 2022-09-27. [Online]. Available: https://dcase.community/challenge2022/task-few-shot-bioacoustic-event-detection-results

[18] J. E. Hopcroft and R. M. Karp, "An n5/2 algorithm for maximum matchings in bipartite graphs," *SIAM J. Comput.*, vol. 2, pp. 225–231, 1973.