EDANSA-2019: THE ECOACOUSTIC DATASET FROM ARCTIC NORTH SLOPE ALASKA

Enis Berk Çoban[1], *Megan Perra*[2], *Dara Pir*[3], *Michael I Mandel*[1,4]

[1] The Graduate Center, CUNY, New York, NY, USA, ecoban@gradcenter.cuny.edu
[2] Institute of Arctic Biology, UAF, Fairbanks, AK, USA, meperra@alaska.edu
[3] Guttman Community College, CUNY, New York, NY, USA, dpir@gradcenter.cuny.edu
[4] Brooklyn College, CUNY, Brooklyn, NY, USA, mim@sci.brooklyn.cuny.edu

ABSTRACT

The arctic is warming at three times the rate of the global average, affecting the habitat and lifecycles of migratory species that reproduce there, like birds and caribou. Ecoacoustic monitoring can help efficiently track changes in animal phenology and behavior over large areas so that the impacts of climate change on these species can be better understood and potentially mitigated. We introduce here the Ecoacoustic Dataset from Arctic North Slope Alaska (EDANSA-2019), a dataset collected by a network of 100 autonomous recording units covering an area of 9000 square miles over the course of the 2019 summer season on the North Slope of Alaska and neighboring regions. We labeled over 27 hours of this dataset according to 28 tags with enough instances of 9 important environmental classes to train baseline convolutional recognizers. We are releasing this dataset and the corresponding baseline to the community to accelerate the recognition of these sounds and facilitate automated analyses of large-scale ecoacoustic databases.

Index Terms— Ecoacoustics, audio dataset, labeled data, baseline, biophony, anthrophony, geophony, convolutional network

1. INTRODUCTION

The Arctic Coastal Plain is an ecosystem in northern Alaska and Canada that hosts over 180 migratory bird species from nearly every continent on the planet. The health of this ecosystem is inextricably linked to other habitats across the globe [1] and is undergoing rapid change due to global warming [2, 3] and land-use change [4]. This region has rich oil and gas resources; extraction and transportation of these fossil fuels increase the usage of machinery and vehicles. As a result, anthrophony from industrial activity or aircraft overflights may change the acoustic environment in the area. Aircraft overflights associated with this activity have been a community concern in the region. One village on the Coastal Plain, Nuiqsut, experiences air traffic equivalent to a city 95 times its size [5]. Past acoustic monitoring studies in Alaska have been smaller in geographic scope and utilized coarse acoustic indices or manual labeling [6, 7]. While some of this research has addressed anthrophony [7], our dataset is the first to account for both developed and undeveloped regions across the Arctic Coastal Plain. Such recordings are valuable for understanding the natural state of the Arctic acoustic environment, how development changes that state, and how that change affects wildlife.

Passive acoustic monitoring is an effective tool to monitor this system—and many others—because acoustic data can tell us about changes in wildlife populations, including phenology [8], biodiversity [9], community structure [10], and distribution [11]. Because the volume of data produced by acoustic studies makes manual data processing prohibitively expensive, researchers have recently employed convolutional neural networks (CNN) to label the contents of large ecoacoustic datasets [12, 13, 14]. To train a CNN in a supervised fashion, researchers must label a small subset of data to train the model, and that labeled data is what we have provided in this paper. We used an earlier version of this dataset (batch-1, described in Subsection 2.3) in our research to understand the advantages of self-supervised learning and data valuation for audio classification [15]. We are providing the best performing model from that work as the baseline in this paper, which utilizes data augmentation [16, 17, 18] and global temporal pooling [19].

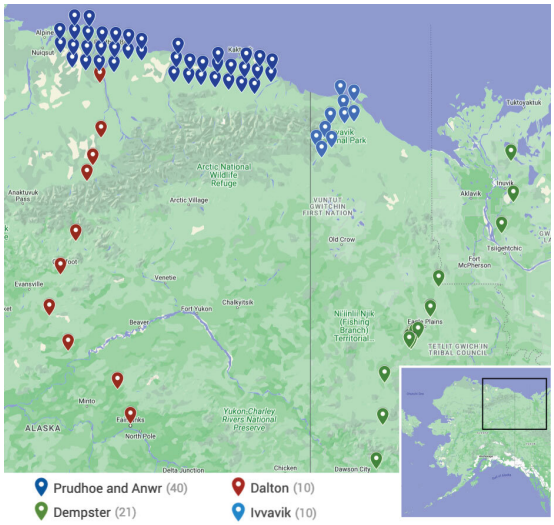

Figure 1: Audio recording device locations.

2. CORPUS

2.1. Monitoring sites

Samples were taken at latitudes between 64° and 70° N, and longitudes between 139° to 150° W, covering predominately the Arctic Coastal Plain but also spanning tundra, shrub, and boreal forest ecosystems on the north and south of the Brooks Mountain Range in northern Alaska. A map of the recording sites is shown in Figure 1. We used 40 recording devices to cover the Prudhoe Bay oilfields and the 1002 portions of the Arctic National Wildlife Refuge (ANWR) in a grid separating locations by 20 km with a random

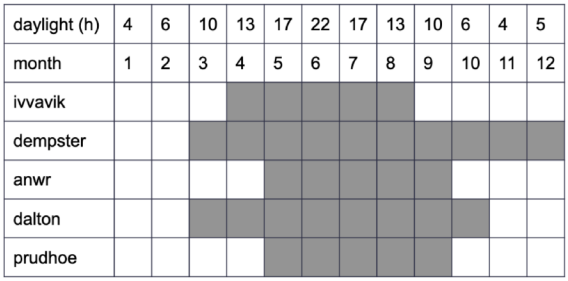

Figure 2: Audio recording times for monitors in each region. The top row displays the number of daylight hours for each corresponding month shown below it on the second row. The colored cells show the months where recordings were made for each region.

Label	b-1	b-2	b-3	b-4	Total
Biophony	4107	500	776	886	6269
Bird	3821	493	78	52	4444
Songbird	2210	238	5	6	2459
Waterfowl	573	126	2	3	704
Upland Bird	386	44	2	2	434
Insect	372	36	734	846	1988
Anthrophony	328	217	1367	1165	3077
Aircraft	100	93	731	769	1693
Silence	1146	325	32	19	1522
Total	5566	1045	2133	2038	10782

Table 1: Number of samples per label in each of the four batches, batch-1 (b-1) through batch-4 (b-4). The last row shows the total number of samples in each batch.

offset. We also had 20 recording devices at sites along the length of the Dempster and Dalton Highways (10 devices each) in the Yukon Territories and Alaska, respectively, and an additional 10 recording devices placed at existing wildlife monitoring sites throughout Ivvavik National Park in the Yukon. Devices are deployed over 3-5 days in each region, starting from the 16th and 23rd of March, 2nd, 3rd, and 4th of May, respectively, in Dalton, Dempster, Ivvavik, Prudhoe, and Anwr. Recordings are saved locally on the device and collected manually. The acoustic recording units (ARUs) were SM4 wildlife recorders from Wildlife Acoustics, sampling at a rate of 48 kHz with gains set to 16 dB. ARUs recorded 150 minutes of audio at a time, with rotating breaks of 120 to 150 minutes in order to cycle through every hour of the day within a 4-day period. In total, devices recorded 2,161 days of audio data throughout 2019.

Recording periods for each region are shown in Figure 2. Recordings from the Prudhoe Bay Oilfields, ANWR, and Ivvavik were able to capture wildlife activity on the Arctic Coastal Plain and in the foothills of the Brooks Range, which serve as the primary breeding grounds for a majority of migratory species in the area. Sites along the Dempster and Dalton Highway captured the arrival and departure of those migratory birds that use the major north-south flyways that converge on the Coastal Plain.

2.2. Taxonomy

Labels of our dataset are members of three taxonomic ranks: coarse, medium, and fine. The coarse rank is the highest rank of the taxonomic tree and contains the four most general labels: *"anthrophony"*, *"biophony"*, *"geophony"*, and *"silence"*. A higher rank contains more general labels compared to the lower ranks. The medium rank contains more specific labels for each of the coarse rank labels and includes *"bird"*, *"insect"*, and *"aircraft"*. The fine rank consists of the most specific labels, each belonging to one of the medium rank categories and includes *"songbird"*, *"waterfowl"* (which in our dataset includes ducks, geese, and swans), and *"upland bird"* (including grouse and ptarmigan). Labels used in our baseline system are shown on the leftmost column of Table 1.

For example, a sample that contains a birdsong event is annotated with the *"biophony"*, *"bird"*, and *"songbird"* tags representing labels from the coarse, medium, and fine ranks, respectively. Annotating a sample with a child label will automatically annotate it with the parent label, however, it is possible for a sample to be annotated only with a parent label. Our annotators could assign nearly all samples a designation from the coarse labels. Fine labels under the bird category were not uncommon, but fine labels under other categories were relatively rare or absent.

2.3. Sampling and labeling

Our dataset is composed of four batches differentiated by the sampling and labeling methods used, which are described below. We initially labeled data with broad coverage of time and space in our dataset so as to capture as many sound classes as possible without bias. We pulled our initial batch, batch-1, of random samples from each site within the Arctic National Wildlife Refuge and the adjacent Oilfields (sites 11-50). We selected a random 150-minute contiguous recording within each site, visually examined each recording's spectrogram in Audacity [20], and labeled all visually identified sounds present in that recording via listening. We excluded sound recordings that were inaudible due to wind-related clipping. We describe the details of four examples in Subsection 2.4.

The same expert labeler labeled sound clips in all batches based on the taxonomy described in Subsection 2.2 In almost all cases, sounds could be identified at the coarsest scale (e.g., *"anthrophony"* or *"biophony"*), and more specific labels were added as they could be identified. This generated 3083 separately labeled sounds that ranged in length from a few seconds to a few minutes. To generate equal-length samples, they were then split into 5566 non-overlapping, 10-second clips. However, not all clips' lengths were divisible by 10, generating clips less than 10 seconds. Clips less than 2 seconds were discarded, and clips between 2 and 10 seconds were zero-padded to the full 10 seconds.

We divided the whole set of samples into training, validation, and test sets so that all samples from a given site were confined to one of these three sets rather than split between them. To determine which site went to which set, we used a multiple knapsack problem detailed in [15]. This split ensured that we were measuring generalization across sites, and thus to future recordings.

In order to increase the number of examples in the rare class *"anthrophony"*, we labeled more samples that we expected to be relevant from sites where this class was more common using predictions of a model trained on the training set of batch-1. At those sites, we pulled 500 10-second clips not tied to model confidence and 500 clips where model confidence for anthrophony was 0.75 or higher. We split the season into 12 weekly periods starting on May 7 and pulled $\sim$80 samples from each weekly period, 40 of which were not tied to model confidence and were just the first 40 samples from that week, and the other 40 of which were tied to model confidence and tended to be distributed more throughout the week.

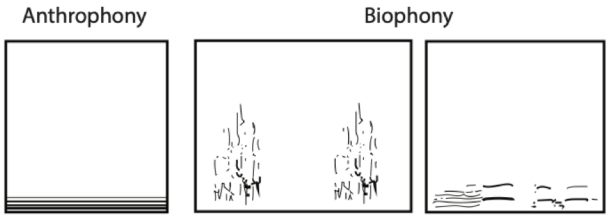

Figure 3: The depictions in this figure highlight the visual characteristics of spectrograms from four examples produced in Audacity. From left to right, these represent a truck driving on a road, a bird singing, and a loon singing near an acoustic recording unit (in sites 26, 50, and 11, respectively).

For each week, we randomly chose one of the training sites (30, 29, 25, 24, 22, 21, 19, 16, 13, 49, 48, 44, 41, 40, 38, 37, 33) to pull samples from. The proportion of all 1000 samples taken from each training site depended on the proportion of anthrophony that was identified at that site in batch-1. So if 30% of all anthrophony in batch-1 had come from one site, 30% of our samples for batch-2 were also pulled from that site. This meant that, in some cases, we had to use multiple sites to fill our quota of samples for a given week. This process created a second batch of clips, batch-2. All of the clips from this 1000-sample set were labeled by trained undergraduate students, and then their labels were reviewed and corrected by an expert labeler before being finalized.

To ensure the accuracy of our original samples, plus the additional anthrophony samples, we built a user interface (UI) in Python that allowed users to quickly listen to a sample, view its spectrogram, and label it. The expert labeler checked or unchecked boxes next to each possible label to validate the original labels associated with a sample. All 6616 10-second clips that we had previously labeled were reviewed using this process. Note that all numbers provided in Table 1 are after this relabeling.

The UI made it considerably more efficient to review clips, allowing us to label additional samples from the *"aircraft"* sound class by selecting high-confidence predictions from the baseline model trained on batches 1 and 2. This created our third batch, batch-3. Note that batch-3 is only used for training, so we are less concerned that this selection process might bias the labels. Using this process, we were able to label an additional 2133 clips.

Performance was still below what we had hoped for the sound classes *"insect"*, *"anthrophony"* and *"aircraft"*, so we decided to collect a final batch of data, batch-4. In this batch, we labeled samples for the validation and test sets. Since selecting samples to label using a single model's confidence scores could lead to choosing only the type of samples that are successfully recognized by this model, we used an ensemble of 7 different iterations of our sound labeling model, including architectural and training variants, developed with earlier versions of the dataset. We normalized the confidence scores of each model across time to be between 0 and 1. Then assigned the maximum confidence across models to each label on each clip. Again, we pulled clips where this combined confidence was high, though this differed for each sound class as follows: 0.7 for anthrophony, 0.25 for aircraft, and 0.99 for the insect. The expert labeler used the UI to review these clips and was able to identify an additional 1,874 instances of labels so that the final set of label-clip associations was over 10,000. Table 1 displays the number of samples per label in each of the four batches of our dataset.

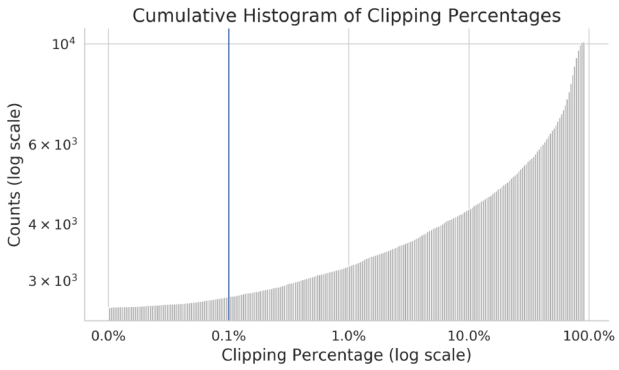

Figure 4: Number of clips out of 10,000 randomly sampled that have at most a particular clipping percentage (percentage of samples at maximum or minimum value).

2.4. Labeling examples

When labeling batch-1, we viewed the 150-minute contiguous recordings as spectrograms in Audacity so we could locate sounds visually. The spectrogram was shown up to 24 kHz so that all possible sounds were visible. While the harmonics of birdsong could extend to 10-12 kHz, a majority of the visible signal on the spectrogram was at 9 kHz or below. The presence of anthrophony was generally indicated by lower frequency, uniform partials that tend to have a consistent pitch and long duration. The presence of biophony presents as harmonic signals with clear partials that have a wide frequency range and short duration, though they are often part of a larger temporal pattern, or song. Rain was the most common form of geophony and was clearly indicated by extremely brief, nonharmonic signals on the spectrogram that look similar to 'clicks', i.e., they have a vertical, broadband form that does not follow an intentional temporal pattern.

Figure 3 displays depictions that highlight the visual characteristics of spectrograms from three examples produced in Audacity. From left to right, the first depiction, labeled *"anthrophony"*, represents a truck driving on a road, near an acoustic recording unit (site 26) in the oilfields. The lines highlight the visible, flat partials at the lower frequencies that are characteristic of anthrophony sounds. The second depiction represents a bird singing near an acoustic recording unit (site 50) in the Arctic National Wildlife Refuge and is labeled *"songbird"*. The patterns highlight the signals created by the birdsong, which show a complex pattern of partials that extends from low to high frequencies; this complex song is repeated multiple times to produce a 'singing bout'. The third depiction represents a loon singing near an acoustic recording unit (site 11) in the oilfields and is labeled *"waterfowl"*. The patterns highlight partials of the loon call, which show a definitive temporal pattern.

2.5. Clipping

Clipping is distortion caused by loud sounds. Recorders have a high and low limit for the amplitude of the sounds they can process; if these thresholds are passed, the data gets corrupted. We count the proportion of sample values in a 10-second interval that are at the maximum or minimum observed levels in a given recording.

An acceptable clipping percentage depends on the specific application. To pick a threshold of acceptable clipping level, we lis-

tened to random 10-second clips and observed that a clipping proportion less than 0.1% is almost unnoticeable to a listener. We therefore removed samples with more than 0.1% clipping before labeling in batch-2, batch-3, and batch-4 of our dataset. Figure 4 shows for 10,000 randomly selected clips, how many had clipping less than a given percentage. In a similar fashion, for batch-1, the expert labeler checked the recordings' spectrogram and only labeled those without any visible clipping artifacts. 80% of these samples from batch-1 have a clipping percentage lower than the 0.1% threshold.

3. BASELINE

Our initial labeling included examples of 28 unique categories in all levels of the hierarchy, out of 41 we thought we might encounter. Of these, only 9 had more than 100 examples in batch-1 and were used to train our baseline model: *"biophony"*, *"bird"*, *"songbird"*, *"waterfowl"*, *"upland bird"*, *"insect"*, *"athrophony"*, and *"silence"*.

As a baseline system[1], we provide the best system described in [15]. Our baseline system employs the Ecoacoustic Dataset from Arctic North Slope Alaska (EDANSA-2019)[2] and uses convolutional neural networks (CNNs) together with global temporal pooling and data augmentation. We share our code with MIT and our dataset under Creative Commons 4.0 licenses, which are highly permissive. We decided to use CNNs with hyperparameters inherited from AlexNet [21] due to their common success in sound event detection experiments [22]. Each sample is a 10-second clip, preprocessed and turned into a mel-spectrogram with a hop size of 23 ms, a window size of 42 ms, and 128 mel-frequency bins. We use a stack of 4 convolutional layers where all kernels are 5×5, followed by two fully connected layers. We train our model for 1600 epochs and keep the one with the highest mean AUC score over all labels on the validation set. Table 2 shows the AUC per label of the baseline model on the validation and test sets.

Label	Validation	Test
Biophony	0.95	0.96
Bird	0.96	0.98
Songbird	0.90	0.96
Waterfowl	0.87	0.90
Upland bird	0.87	0.93
Insect	0.90	0.83
Anthrophony	0.88	0.88
Aircraft	0.96	0.88
Silence	0.96	0.93
Average	0.92	0.92

Table 2: AUC per label of the baseline on validation and test sets.

4. PREVIOUS WORK

There are a number of open-source datasets, similar to ours, shared along with their research findings. The CityNet dataset, which is collected from London, has diverse anthropogenic classes but the biophony classes are limited to only general labels like *"bird"*, *"insect"*, *"vegetation"*, and *"wing beats"* [23]. Another soundscape dataset consists of 5 hours of recordings collected from Sonoma County, California, USA and samples are labeled with *"anthropophony"*, *"biophony"*, *"geophony"*, *"quiet"*, and *"interference"* [14]. The main difference between these datasets and ours is that ours is recorded in remote locations and over a much larger area. Our dataset consists of 29 hours of labeled data, compared to 19 hours in CityNet and 5 hours from Sonoma County.

There are large datasets focusing on bird calls, which are challenging to model and of high scientific interest. BIRDCLEF is a family of such datasets focusing on short targeted recordings as opposed to long-term continuous recordings. It consists of sound recordings collected by the Xeno-canto community and new versions with different purposes have been released every year since 2014. The latest, 2022 version, consists of 15k recordings, totaling over 190 hours covering 152 species from Hawaii, specially designed for modeling calls of rare and endangered bird species with small amounts of training data [24]. Another dataset with 385 minutes of dawn chorus recordings was collected from Eastern North America, including 48 species and 16,052 annotations [25]. Some of the other datasets with bird calls are BirdVox [26], Nips4Bplus [27], Freefield1010 [28], Warblrb10k and PolandNFC [29].

Larger general-purpose datasets have been extracted from YouTube such as Audio Set [30] and VGGSOUND [31] and include bioacoustic classes as a small part of their corpus. There are also continuously recorded open-source sound datasets without bioacoustic labels, such as SONYC-UST-V2, which is the output of an urban noise monitoring project and it is a multi-labeled [32]. This dataset is $\sim$51 hours long in total and labeled with 8 main tags that are common in city environments, such as engine, music, and the human voice.

5. CONCLUDING REMARKS

This paper presented the Ecoacoustic Dataset from Arctic North Slope Alaska (EDANSA-2019), collected by autonomous recording units during the summer of 2019, and its corresponding baseline. We provided detail on the recordings and the sampling and labeling methods used to generate the four batches of our dataset. This work should help facilitate the analysis of large-scale ecoacoustic recordings made in arctic conditions, and it would be interesting to examine the extent to which models trained on this data can generalize to data collected in other environments and ecosystems.

6. ACKNOWLEDGMENT

We are grateful to Scott Leorna, Dr. Todd Brinkman, and Dr. Natalie T. Boelman for help with data collection and Eleanor Davol for help with labeling. This work is supported by the National Science Foundation (NSF) grant OPP-1839185. Any opinions, findings, and conclusions or recommendations are those of the author(s) and do not necessarily reflect the views of the NSF.

7. REFERENCES

[1] B. K. Sullender. (2019, Apr) Migratory birds in the heart of the arctic: A journey through the arctic national wildlife refuge.

[2] J. E. Walsh and B. Brettschneider, "Attribution of recent warming in alaska," *Polar Sci.*, vol. 21, pp. 101–109, 2019.

[3] P. Fauchald, T. Park, H. Tømmervik, R. Myneni, and V. H. Hausner, "Arctic greening from warming promotes declines

[1] https://github.com/speechLabBcCuny/EDANSA-2019

[2] https://zenodo.org/record/6824272

in caribou populations," *Science Advances*, vol. 3, no. 4, p. e1601365, 2017.

[4] M. K. Raynolds, D. A. Walker, K. J. Ambrosius, J. Brown, K. R. Everett, M. Kanevskiy, G. P. Kofinas, V. E. Romanovsky, Y. Shur, and P. J. Webber, "Cumulative geoecological effects of 62 years of infrastructure and climate change in ice-rich permafrost landscapes, prudhoe bay oilfield, alaska," *Glob. Ch. Bio.*, vol. 20, no. 4, pp. 1211–1224, 2014.

[5] T. R. Stinchcomb, T. J. Brinkman, and D. Betchkal, "Extensive aircraft activity impacts subsistence areas: acoustic evidence from arctic alaska," *Environmental Research Letters*, vol. 15, no. 11, p. 115005, oct 2020.

[6] R. Buxton, E. Brown, L. Sharmin, C. M. Gabriele, and M. F. McKenna, "Using bioacoustics to examine shifts in songbird phenology," *Eco. & Evol.*, vol. 6, pp. 4697–4710, 05 2016.

[7] H. Vincelette, R. Buxton, N. Kleist, M. F. McKenna, D. Betchkal, and G. Wittemyer, "Insights on the effect of aircraft traffic on avian vocal activity," *Ibis*, vol. 163, no. 2, pp. 353–365, 2021.

[8] R. Y. Oliver, D. P. W. Ellis, H. E. Chmura, J. S. Krause, J. H. Pérez, S. K. Sweet, L. Gough, J. C. Wingfield, and N. T. Boelman, "Eavesdropping on the Arctic: Automated bioacoustics reveal dynamics in songbird breeding phenology," *Sci Adv*, vol. 4, no. 6, p. eaaq1084, 06 2018.

[9] D. Proppe, C. Sturdy, and C. St. Clair, "Anthropogenic noise decreases urban songbird diversity and may contribute to homogenization," *Glob. Ch. Bio.*, vol. 19, pp. 1075–84, 04 2013.

[10] C. D. Francis, C. P. Ortega, and A. Cruz, "Noise pollution filters bird communities based on vocal frequency," *PLOS ONE*, vol. 6, no. 11, pp. 1–8, 11 2011.

[11] N. J. Kleist, R. P. Guralnick, A. Cruz, and C. D. Francis, "Sound settlement: noise surpasses land cover in explaining breeding habitat selection of secondary cavity-nesting birds," *Ecol Appl*, vol. 27, no. 1, pp. 260–273, 01 2017.

[12] S. S. Sethi, N. S. Jones, B. D. Fulcher, L. Picinali, D. J. Clink, H. Klinck, C. D. L. Orme, P. H. Wrege, and R. M. Ewers, "Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set," *PNAS*, vol. 117, no. 29, pp. 17 049–17 055, 2020.

[13] J. LeBien, M. Zhong, M. Campos-Cerqueira, J. P. Velev, R. Dodhia, J. L. Ferres, and T. M. Aide, "A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network," *Ecological Informatics*, vol. 59, p. 101113, 2020.

[14] C. A. Quinn, P. Burns, G. Gill, S. Baligar, R. L. Snyder, L. Salas, S. J. Goetz, and M. L. Clark, "Soundscape classification with convolutional neural networks reveals temporal and geographic patterns in ecoacoustic data," *Ecological Indicators*, vol. 138, p. 108831, 2022.

[15] E. B. Çoban, A. R. Syed, D. Pir, and M. I. Mandel, "Towards large scale ecoacoustic monitoring with small amounts of labeled data," in *Proc. WASPAA*, 2021, pp. 181–185.

[16] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. INTERSPEECH*, 2015, pp. 3586–3589.

[17] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[18] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[19] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *Tr. ASLP*, vol. 28, pp. 2880–2894, 2020.

[20] A. Team, 2019. [Online]. Available: https://www.audacityteam.org

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *NeurIPS*, vol. 25, pp. 1097–1105, 2012.

[22] K. Choi, G. Fazekas, and M. B. Sandler, "Automatic tagging using deep convolutional neural networks," in *ISMIR*, 2016.

[23] A. J. Fairbrass, M. Firman, C. Williams, G. J. Brostow, H. Titheridge, and K. E. Jones, "Citynet—deep learning tools for urban ecoacoustic assessment," *Methods in Ecology and Evolution*, vol. 10, no. 2, pp. 186–197, 2019.

[24] S. Kahl, A. Navine, T. Denton, H. Klinck, P. Hart, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, and A. Joly, "Overview of birdclef 2022: Endangered bird species recognition in soundscape recordings," *Working Notes of CLEF*, 2022.

[25] L. M. Chronister, T. A. Rhinehart, A. Place, and J. Kitzes, "An annotated set of audio recordings of eastern north american birds containing frequency, time, and species information," *Ecology*, vol. 102, 2021.

[26] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, and J. P. Bello, "Birdvox-full-night: A dataset and benchmark for avian flight call detection," in *ICASSP*, 2018, pp. 266–270.

[27] V. Morfi, Y. Bas, H. Pamuła, H. Glotin, and D. Stowell, "Nips4bplus: a richly annotated birdsong audio dataset," *PeerJ Computer Science*, vol. 5, p. e223, 2019.

[28] D. Stowell and M. D. Plumbley, "An open dataset for research on audio field recording archives: freefield1010," *arXiv preprint arXiv:1309.5275*, 2013.

[29] F. Berger, W. Freillinger, P. Primus, and W. Reisinger, "Bird audio detection-dcase 2018," *DCASE*, 2018.

[30] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017, pp. 776–780.

[31] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP*, 2020, pp. 721–725.

[32] M. Cartwright, J. Cramer, A. E. M. Mendez, Y. Wang, H.-H. Wu, V. Lostanlen, M. Fuentes, G. Dove, C. Mydlarz, J. Salamon, *et al.*, "Sonyc-ust-v2: An urban sound tagging dataset with spatiotemporal context," *arXiv preprint arXiv:2009.05188*, 2020.