



HAL
open science

Mixture of Mixups for Multi-label Classification of Rare Anuran Sounds

Ilyass Moummad, Nicolas Farrugia, Romain Serizel, Jeremy Froidevaux,
Vincent Lostanlen

► **To cite this version:**

Ilyass Moummad, Nicolas Farrugia, Romain Serizel, Jeremy Froidevaux, Vincent Lostanlen. Mixture of Mixups for Multi-label Classification of Rare Anuran Sounds. EUSIPCO 2024, Aug 2024, Lyon, France. hal-04620733

HAL Id: hal-04620733

<https://imt.hal.science/hal-04620733>

Submitted on 21 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0
International License

Mixture of Mixups for Multi-label Classification of Rare Anuran Sounds

Ilyass Moummad
IMT Atlantique, CNRS, Lab-STICC
ilyass.moummad@imt-atlantique.fr

Nicolas Farrugia
IMT Atlantique, CNRS, Lab-STICC
nicolas.farrugia@imt-atlantique.fr

Romain Serizel
University of Lorraine, INRIA, LORIA
romain.serizel@loria.fr

Jeremy Froidevaux
CREA Mont-Blanc, MNHN
jeremy.froidevaux@mnhn.fr

Vincent Lostanlen
Ecole Centrale Nantes, CNRS, LS2N
vincent.lostanlen@ls2n.fr

Abstract—Multi-label imbalanced classification poses a significant challenge in machine learning, particularly evident in bioacoustics where animal sounds often co-occur, and certain sounds are much less frequent than others. This paper focuses on the specific case of classifying anuran species sounds using the dataset AnuraSet, that contains both class imbalance and multi-label examples. To address these challenges, we introduce Mixture of Mixups (Mix2), a framework that leverages mixing regularization methods Mixup, Manifold Mixup, and MultiMix. Experimental results show that these methods, individually, may lead to suboptimal results; however, when applied randomly, with one selected at each training iteration, they prove effective in addressing the mentioned challenges, particularly for rare classes with few occurrences. Further analysis reveals that the model trained using Mix2 is also proficient in classifying sounds across various levels of class co-occurrences.¹

Index Terms—Bioacoustics, deep learning, mixup, multi-label classification, class imbalance.

I. INTRODUCTION

Machine learning faces a significant challenge in the domain of multi-label classification, where instances can belong to multiple classes simultaneously [1]. The issue of class imbalance, wherein certain labels are disproportionately represented is potentially adding to the complexity. This latter problem can result in overfitting of frequent classes and underfitting of rare ones [2]. Such challenges are notably prominent within the field of bioacoustics, where multi-label scenarios and class imbalance are frequently encountered [3], [4].

Bioacoustics is a cross-disciplinary science that combines biology and acoustics to study sounds produced by animals. It provides insights into animal behaviour, physiological status, abundance and distribution and, at the community level, information on species diversity and interactions [5], [6]. Birds have been extensively studied in the field of bioacoustics due to their diverse and easily detectable vocalizations in the audible range, making them a prominent and convenient model taxon for research in this field. Many bioacoustic benchmarks, tailored for machine learning applications and species detection/recognition, focus on bird sounds to address various

challenges within the field, such as BirdCLEF for species classification [7], MetaAudio for few-shot classification [8], DCASE 2022 and 2023 for bioacoustic event detection [9], [10], or BIRB [4], a large scale generalization benchmark.

Avian species are pivotal subjects in bioacoustics research and passive acoustic monitoring. Yet, other vocalizing taxa hold significant importance particularly in the broader context of wildlife monitoring and conservation efforts amidst the ongoing biodiversity crisis [11]. This is particularly true for Amphibians, which are one of the most endangered taxa worldwide, with almost half at the risk of extinction [12], [13]. Amphibians are highly sensitive to both land-use and climate changes, further compounded by the devastating effects of the pathogenic fungal disease chytridiomycosis [14], [15]. Standardized, wide-scale monitoring is crucial to facilitate the formulation and implementation of targeted conservation strategies. In this regard, the use of acoustic tools offer promising prospects for anuran monitoring and conservation programmes. However, automatic identification of anuran species presents a significant challenge due to their nocturnal, elusive behavior, diverse habitats, seasonal influences, species diversity, limited vocalization windows, and often remote locations. While a few studies leverage machine learning for the classification of anuran sounds [16]–[19], most of them rely on privately collected recordings, posing challenges for comparisons between different works.

The recent introduction of the AnuraSet dataset [13] addresses this issue by providing an extensive compilation of Neotropical anuran vocalizations. This dataset currently stands as the largest open-source dataset in anuran sound research. Encompassing two Brazilian biomes, it features a diverse collection of 42 anuran species, demonstrating strong co-occurrence, including certain underrepresented species. In addition, certain rare classes exhibit no overlap between the training and test sets. AnuraSet presents inherent challenges characterized by imbalanced distributions, multi-label instances, and noisy conditions in the natural environment. It provides a significant benchmark for addressing these challenges in the application of anuran sound classification.

In order to address such difficult generalization problems in

This work is supported by AI@IMT, OSO-AI, GdR IASIS, WeAMEC project PETREL, and CNRS MITI CAPTEO.

¹Our code is released at: <https://github.com/ilyassmoummad/Mix2>

deep learning, promising solutions are mixing regularization methods, such as Mixup [20], that dynamically generate synthetic examples during training. Mixup penalizes the loss accordingly to prevent the model from overfitting the training data and to smooth decision boundaries from class to class. This leads to better prediction outside the training examples. These methods regularize neural networks by promoting linear behavior between training samples in either input space or latent space [20]–[22]. In order to address class imbalance, Remix [23] relaxes the mixing factor by pushing the decision boundaries towards majority classes. Regarding multi-label classification, FlowMixup [24] decouples the extracted features by adding constraints to the hidden states of the models.

In this work, we investigate three mixing regularization methods: Mixup [20], Manifold Mixup [21], and Multi-Mix [22], to address multi-label imbalanced classification in AnuraSet dataset. Our findings show that alternating between these methods during training leads to significant improvement in classification performance, especially for rare classes.

II. METHOD

In Figure 1, we present a comprehensive overview of our proposed system, Mix2. Subsequently, we elaborate on the distinct components of Mix2.

A. Mixup

Mixup [20] is a technique that involves training on convex combinations of pairs of examples and their corresponding labels. This method induces decision boundaries that linearly transition from one class to another, resulting in a smoother estimation of uncertainty. Although its effectiveness is not fully understood, Mixup is widely employed as a regularization technique in deep learning for supervised classification problems. Consider two randomly sampled instances from the dataset x_i and x_j with their corresponding labels y_i and y_j . A synthetic example \tilde{x} and its label \tilde{y} are generated via the formula:

$$\begin{aligned}\tilde{x} &= \lambda x_i + (1 - \lambda)x_j \\ \tilde{y} &= \lambda y_i + (1 - \lambda)y_j\end{aligned}$$

where λ represents a random interpolation coefficient sampled from a Beta distribution.

B. Manifold Mixup

Manifold Mixup [21] extends the concept of Mixup to the embeddings of a neural network, introducing interpolation at this level to smooth decision boundaries across multiple levels of representation. The synthetic embedding, \tilde{h} , is computed using the formulas:

$$\begin{aligned}\tilde{h} &= \lambda h_i + (1 - \lambda)h_j \\ \tilde{y} &= \lambda y_i + (1 - \lambda)y_j\end{aligned}$$

Here, \tilde{h}_i and \tilde{h}_j represent the embeddings of two instances, y_i and y_j are their corresponding labels.

C. MultiMix

Both Mixup and Manifold Mixup involve interpolating two training samples and their corresponding labels. In contrast, MultiMix [22] extends this concept to the entire mini-batch in the embedding space. The synthetic batch of embeddings, \tilde{H} , is obtained through the formulas:

$$\begin{aligned}\tilde{H} &= \Lambda H \\ \tilde{Y} &= \Lambda Y\end{aligned}$$

Here, H is the matrix of embeddings for the original batch, Λ is a matrix of interpolation coefficients sampled from a Dirichlet distribution, Y is the matrix of corresponding one-hot encoded labels for the original batch, and \tilde{Y} is the synthetic batch of labels. The sampling occurs on the entire convex hull of the mini-batch, rather than linear segments between pairs of examples. The matrix product ΛH performs the interpolation between all samples in the batch. Likewise, ΛY interpolates between the corresponding labels.

For all of these methods, in practice, the mixing of targets is performed at the loss level.

D. Novel contribution: Mixture of Mixups (Mix2)

We now present Mix2, a probabilistic mixture of the three aforementioned methods. At each SGD training iteration, Mix2 employs random selection from three mixing regularization options: Mixup, Manifold Mixup, or Multimix. This dynamic approach ensures that the model learns from a diverse range of augmented samples, promoting robustness and generalization.

III. EVALUATION

A. Dataset

AnuraSet [13] constitutes a multi-label classification dataset comprising 93,378 3-second segments capturing anuran calls from 42 distinct species. Derived from 1,612 annotated 1-minute audio recordings, equivalent to 26.87 hours of acoustic documentation, the dataset is generated using fixed-length 3-second windows with a 1-second sliding window (75% overlap between samples). To prevent data leakage between training and testing, the dataset is partitioned into a 2/3 train set and a 1/3 test set at the audio recording level. The upper part of Figure 3 illustrates the polyphony levels within the dataset, while Figure 2 categorizes the species into three groups: frequent (more than 10k), common (between 5k and 10k), and rare (fewer than 5k).

Given the absence of a dedicated validation set, designing one from the training set is challenging due to the overlap between examples. The split should be conducted at the audio recording level, from which examples were extracted using a sliding window. Additionally, due to the long-tail distribution, splitting classes with few occurrences into distinct subsets is difficult; this problem is already evident between the train and test sets. Consequently, creating a proper validation set poses a challenge. In line with standard machine learning benchmarks, we conduct model validation on the test set.

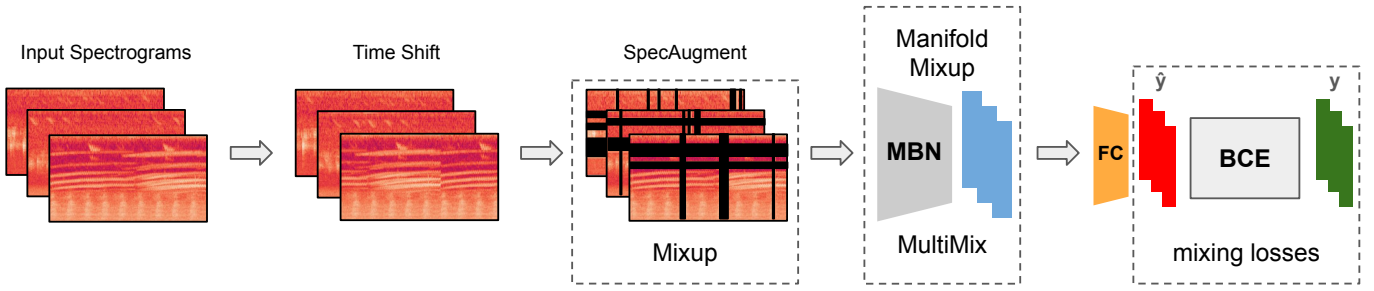


Fig. 1. Overview of our system: Mix2. MBN, FC and BCE stand for MobileNetV3-Large, Fully Connected, and Binary Cross-Entropy, respectively.

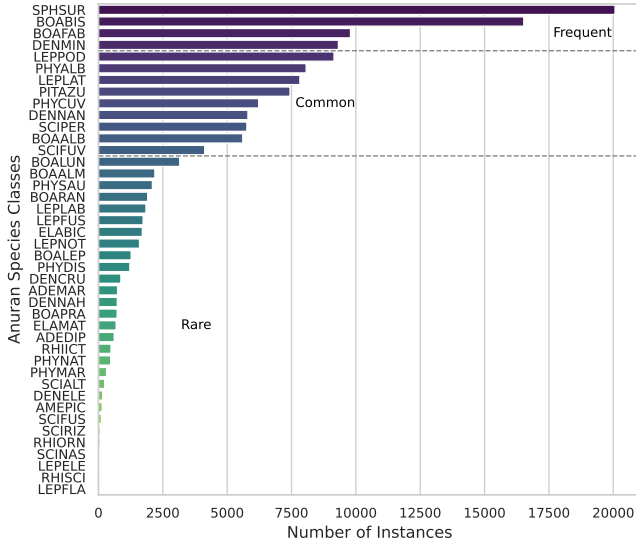


Fig. 2. Histogram of class imbalance in AnuraSet. Dashed lines denote “frequent” species (more than 10k instances), “common” species (5k–10k instances), and “rare” species (fewer than 5k instances).

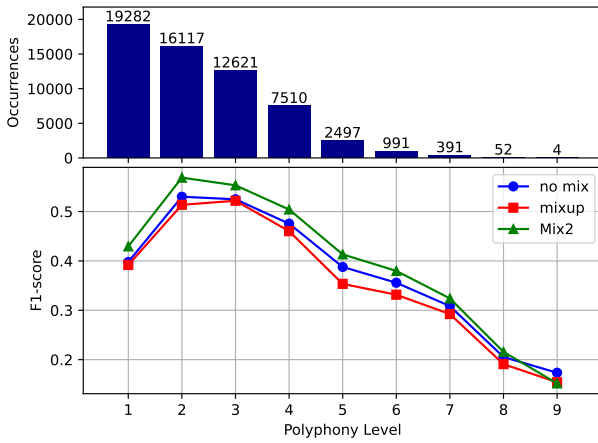


Fig. 3. In AnuraSet, Mixup leads to a performance decline across various polyphony levels compared to the absence of mixing. Conversely, Mix2 demonstrates an improvement in performance across different polyphony levels.

B. Metric

The benchmark employs the macro F -score, assigning equal importance to all species. It is computed as the mean of the F -score for each class.

C. Model

MobileNetV3, or MBN for short, is a well-established series of convolutional neural network architecture for embedded computing applications [25]. The architectural refinements of MobileNetV3 include: squeeze-and-excitation blocks, residual connections, depthwise convolutions, and atrous spatial pyramid pooling. In this paper, we propose to train a MobileNetV3-Large, which has three million parameters.

D. Time–frequency representation

We resample AnuraSet to 16 kHz. Following the AnuraSet baseline, we compute a mel-frequency spectrogram with 128 filters, a window size of 512, and a hop size of 128.

E. Data augmentation

Following the AnuraSet baseline, we apply SpecAugment [26], a data augmentation technique which sets time–frequency magnitudes to zero over multiple random rectangular regions of the mel-frequency spectrogram. In addition, we apply random circular time shift following a previous study on machine learning for birdsong classification [27].

F. Training

For all our experiments, we train MobileNetV3-Large [25] from scratch using AdamW optimizer with a batch size of 128, a learning rate of 10^{-2} , and a weight decay of 10^{-6} for 100 epochs. We train all our systems for 100 epochs, with the exception of the first row (10 epochs) in Table I for a fair comparison with the baseline systems of AnuraSet, which are trained for 10 epochs. When we train the model using Mix2, we select Mixup, Manifold Mixup, or Multimix, with probabilities of 25%, 50%, and 25%, respectively. Manifold Mixup is selected more frequently as it was the best-performing single regularization method.

TABLE I
MACRO F -SCORES (IN %) ON ANURASET.

WE REPORT MEANS AND STANDARD DEVIATIONS OVER THREE RUNS. FREQUENT, COMMON, AND RARE CLASSES ARE DEFINED IN FIGURE 2.

Method	Frequent	Common	Rare	All
State of the art: 11M–116M parameters, ImageNet pre-training [13]				
ResNet18	61.6	52.0	14.8	34.9
ResNet50	62.3	53.9	9.9	33.2
ResNet152	68.4	56.8	15.7	37.8
Ours: 3M parameters, no pre-training				
MobileNetV3-Large (10 epochs)	86.3	72.4	35.4	48.2
	± 0.2	± 0.3	± 1.0	± 0.6
+90 epochs	88.9	75.5	45.6	56.2
	± 0.2	± 0.3	± 0.6	± 0.5
+SpecAug	86.7	74.41	47.0	56.6
	± 0.4	± 0.4	± 1.0	± 0.7
+TimeShift	87.0	74.2	46.9	56.6
	± 0.2	± 0.0	± 1.0	± 0.7
+Regularization				
(1 method)				
Mixup	87.4	74.3	44.9	55.2
	± 0.2	± 0.5	± 0.4	± 0.4
Manifold Mixup	86.8	74.3	47.0	56.6
	± 0.4	± 0.4	± 0.8	± 0.5
MultiMix	88.0	75.1	46.1	56.3
	± 0.4	± 0.1	± 0.2	± 0.2
(2 methods)				
Mixup + Manifold Mixup	88.6	76.6	51.7	60.5
	± 0.1	± 0.2	± 1.1	± 0.8
Mixup + MultiMix	88.5	75.7	47.8	57.7
	± 1.1	± 0.9	± 2.1	± 1.7
Manifold Mixup + MultiMix	88.1	75.7	48.4	58.0
	± 0.4	± 0.1	± 2.3	± 1.7
(3 methods)				
Mixture of Mixups (Mix2)	89.2	77.1	51.8	60.8
	± 0.1	± 0.1	± 0.4	± 0.3

G. Tasks

Certain classes exhibit no overlap between the training and test sets, with 4 found only in the training set and 2 only in the test set. Moreover, a substantial portion of the dataset lacks any active class (33,913 out of 93,378). In evaluating the benchmark, the assessment is exclusively conducted on the active classes. To gain insights, we create two subsets of the AnuraSet. The first subset involves removing the non-overlapping classes between training and testing, referred to as AnuraSet-36N (N stands for negative as it also contain examples with no class activated), with the same number of examples as AnuraSet but containing 36 classes instead of 42. The second subset, extends beyond the non-overlapping classes by also eliminating examples with silence, referred to as AnuraSet-36 (with 59,465 examples and 36 classes). In Table II, we compare the performance of Mix2 with Mixup and no mixing on these two subsets.

IV. RESULTS AND DISCUSSIONS

We conduct a comprehensive ablation study to demonstrate the effect of each individual component of our proposed method that contributes to the improvement in F -score.

The upper part of Table I shows the three proposed baselines from the AnuraSet paper [13]. These baselines involve fine-tuning three residual convolutional architectures ResNet18, ResNet50, ResNet152 pre-trained on ImageNet [28], using the training set of AnuraSet sampled at a rate of 22,050 Hz. The

TABLE II
MACRO F -SCORES (IN %) ON ANURASET-36N AND ANURASET-36. WE REPORT MEANS AND STANDARD DEVIATIONS OVER THREE RUNS. FREQUENT, COMMON, AND RARE CLASSES ARE DEFINED IN FIGURE 2.

Method	Frequent	Common	Rare	All
AnuraSet-36N				
no mixing	87.4	83.9	56.3	65.9
	± 0.3	± 0.2	± 0.5	± 0.3
Mixup	87.2	83.5	55.0	64.9
	± 0.4	± 0.3	± 0.7	± 0.5
Mix2	88.9	86.6	62.0	70.4
	± 0.2	± 0.3	± 0.5	± 0.3
AnuraSet-36				
no mixing	88.8	84.4	60.6	69.0
	± 0.3	± 0.2	± 0.7	± 0.4
Mixup	88.5	83.9	58.8	67.7
	± 0.3	± 0.2	± 0.7	± 0.4
Mix2	89.5	86.8	67.4	74.2
	± 0.1	± 0.2	± 0.6	± 0.5

fine-tuning is conducted for 10 epochs, using SpecAugment with a masking strategy covering up to 60 time bands and 8 frequency bands. The second part of the table displays our results, where each row illustrates the impact of adding a specific technique on top of all the previous ones. Training longer (100 vs 10 epochs) contributes to higher performance. The addition of SpecAugment provides a slight boost in scores, particularly for rare classes. Introducing a single mixing regularization negatively affects performance. Conversely, combining two methods, especially Mixup and Manifold Mixup, results in a substantial performance boost. Similarly, Mix2, incorporating all three methods, achieves the highest score with more than 4% gain in macro F -score. The gain in F -score is also evident in the rare group, highlighting the efficiency of our proposed method to handle class imbalance in multi-label classification.

This phenomenon is intriguing, given that the individual regularization methods, when applied alone, tend to decrease performance (with the exception of Manifold Mixup). We speculate that combining different mixing regularizations with a certain probability leads to a more diverse set of augmented training examples, capturing a broader range of data variations and enhancing the model’s robustness.

In addition, we experimented with weighted sampling by oversampling rare combinations of classes. However, this approach yielded worse results than the baseline and was therefore excluded from the study.

In Figure 3, we study the macro F -score score of examples with different polyphony levels ranging from 1 to 9 for Mix2, Mixup, and when no mixing is applied. We observe a similar trend where Mixup performs worse than when it is not applied, and Mix2 noticeably improves the performance, a finding worthy of future investigation. Error bars are shown at each point, with the standard deviation being very small across runs.

Table II compares the performance of Mix2 with Mixup and no mixing the impact of removing the non-overlapping classes from the dataset (AnuraSet-36N) in the first part, as well as when also the segments with no class activated are removed (AnuraSet-36) in the second part. Similar to Table I, this

result confirms that Mix2 effectively addresses the challenges associated with multi-label classification and class imbalance, rather than uncovering any hidden issues within the inherently ambiguous classification problem of AnuraSet (examples with no class and non-overlapping train/test classes).

V. CONCLUSIONS AND PERSPECTIVES

This study addresses the challenges of multi-label imbalanced classification, particularly within the domain of bioacoustics focusing on anuran species sound classification using the AnuraSet dataset. The proposed Mix2 system, combining mixing regularization methods Mixup, Manifold Mixup, and MultiMix, to capture a wide range of data variation, has demonstrated significant efficacy in addressing imbalance and multi-label.

For future work, we want to investigate whether combining different mixing strategies learn better representations in self-supervised learning, and whether the learned representations generalize well in out-of-distribution settings. Furthermore, we aim to integrate ecological information on diel temporal activity patterns of different species and account for potential temporal acoustic niche segregation [29] to only mix sounds of species co-occurring spatially and temporally.

The presence of non-overlapping rare classes in the train/test sets poses a challenge. One promising avenue for addressing this challenge is the exploration of few-shot learning, where the model is trained with limited annotations for these rare classes. Additionally, zero-shot learning, leveraging multi-modal approaches such as language, provides another interesting direction for research in mitigating the impact of non-overlapping rare classes. Another direction is open-set classification, where the objective is to classify events from unknown classes [30].

REFERENCES

- [1] Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor W Tsang, "The emerging trends of multi-label learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- [2] Adane Nega Tarekegn, Mario Giacobini, and Krzysztof Michalak, "A review of methods for imbalanced multi-label classification," *Pattern Recognition*, 2021.
- [3] Forrest Briggs, Balaji Lakshminarayanan, Lawrence Neal, Xiaoli Z Fern, Raviv Raich, Sarah JK Hadley, Adam S Hadley, and Matthew G Betts, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," *JASA*, 2012.
- [4] Jenny Hamer, Eleni Triantafyllou, Bart van Merriënboer, Stefan Kahl, Holger Klinck, Tom Denton, and Vincent Dumoulin, "BIRB: A Generalization Benchmark for Information Retrieval in Bioacoustics," *arXiv preprint arXiv:2312.07439*, 2023.
- [5] Paola Laiolo, "The emerging significance of bioacoustics in animal species conservation," *Biological conservation*, 2010.
- [6] Weronika Penar, Angelika Magiera, and Czesław Klocek, "Applications of bioacoustics in animal ecology," *Ecological complexity*, 2020.
- [7] Stefan Kahl, Amanda Navine, Tom Denton, Holger Klinck, Patrick Hart, Hervé Glotin, Hervé Goëau, Willem-Pier Vellinga, Robert Planqué, and Alexis Joly, "Overview of BirdCLEF 2022: Endangered bird species recognition in soundscape recordings," in *CLEF (Working Notes)*, 2022.
- [8] Calum Heggan, Sam Budgett, Timothy Hospedales, and Mehrdad Yaghoobi, "MetaAudio: A few-shot audio classification benchmark," in *ICANN*, 2022.
- [9] Ines Nolasco, Shubhr Singh, E Vidana-Villa, Emily Grout, J Morford, M Emmerson, F Jensens, Helen Whitehead, Ivan Kiskin, Ariana Strandburg-Peshkin, et al., "Few-shot bioacoustic event detection at the DCASE 2023 challenge," *DCASE Workshop 2023*, 2023.
- [10] Ines Nolasco, Shubhr Singh, Veronica Morfi, Vincent Lostanlen, Ariana Strandburg-Peshkin, Ester Vidana-Vila, Lisa Gill, Hanna Pamula, Helen Whitehead, Ivan Kiskin, et al., "Learning to detect an animal sound from five examples," *Ecological informatics*, 2023.
- [11] José J Lahoz-Monfort and Michael JL Magrath, "A comprehensive overview of technologies for species and habitat monitoring and conservation," *BioScience*, 2021.
- [12] Jennifer A Luedtke, Janice Chanson, Kelsey Neam, Louise Hobin, Adriano O Maciel, Alessandro Catenazzi, Amaël Borzée, Amir Hamidy, Anchalee Aowphol, Anderson Jean, et al., "Ongoing declines for the world's amphibians in the face of emerging threats," *Nature*, 2023.
- [13] Juan Sebastián Cañas, María Paula Toro-Gómez, Larissa Sayuri Moreira Sugai, Hernán Darío Benítez Restrepo, Jorge Rudas, Breyner Posso Bautista, Luís Felipe Toledo, Simone Dena, Adão Henrique Rosa Domingos, Franco Leandro de Souza, et al., "A dataset for benchmarking neotropical anuran calls identification in passive acoustic monitoring," *Scientific Data*, 2023.
- [14] Christian Hof, Miguel B Araújo, Walter Jetz, and Carsten Rahbek, "Additive threats from pathogens, climate and land-use change for global amphibian diversity," *Nature*, 2011.
- [15] Josh West, "Importance of amphibians: A synthesis of their environmental functions, benefits to humans, and need for conservation," 2018.
- [16] Julia Strout, Bryce Rogan, SM Mahdi Seyednezhad, Katrina Smart, Mark Bush, and Eraldo Ribeiro, "Anuran call classification with deep learning," in *IEEE ICASSP*, 2017.
- [17] Miriam Alabi, "Classification of anuran frog species using machine learning," *arXiv preprint arXiv:2112.05148*, 2021.
- [18] Erhan Akbal, Prabal Datta Barua, Sengul Dogan, Turker Tuncer, and U Rajendra Acharya, "Explainable automated anuran sound classification using improved one-dimensional local binary pattern and tunable q wavelet transform techniques," *Expert Systems with Applications*, 2023.
- [19] Juan Colonna, Tanel Peet, Carlos Abreu Ferreira, Alípio M Jorge, Elsa Ferreira Gomes, and João Gama, "Automatic classification of anuran sounds using convolutional neural networks," in *Proceedings of the ninth international c* conference on computer science & software engineering*, 2016.
- [20] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.
- [21] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio, "Manifold mixup: Better representations by interpolating hidden states," in *ICML*. PMLR, 2022.
- [22] Shashanka Venkataramanan, Ewa Kijak, Yannis Avrithis, et al., "Embedding space interpolation beyond mini-batch, beyond pairs and beyond examples," *NeurIPS*, 2024.
- [23] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan, "Remix: rebalanced mixup," in *ECCV Workshops*. Springer, 2020.
- [24] Jintai Chen, Hongyun Yu, Ruiwei Feng, Danny Z Chen, et al., "Flow-mixup: Classifying multi-labeled medical images with corrupted labels," in *IEEE BIBM*, 2020.
- [25] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al., "Searching for MobileNetV3," in *ICCV*, 2019.
- [26] Daniel S Park, Yu Zhang, Chung-Cheng Chiu, Youzheng Chen, Bo Li, William Chan, Quoc V Le, and Yonghui Wu, "SpecAugment on large scale datasets," in *IEEE ICASSP*. IEEE, 2020.
- [27] Ilyass Moummad, Romain Serizel, and Nicolas Farrugia, "Self-Supervised Learning for Few-Shot Bird Sound Classification," in *ICASSPW SASB*. IEEE, 2024.
- [28] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE CVPR*, 2009.
- [29] Bernie Krause, "The niche hypothesis: How animals taught us to dance and sing," *Whole Earth Review*, 1987.
- [30] Sridhar Sripathi and Mark Cartwright, "Multi-label Open-set Audio Classification," *DCASE Workshop*, 2023.