Acoustic identification of individual animals with hierarchical contrastive learning

Ines Nolasco, Ilyass Moummad, Dan Stowell, Emmanouil Benetos

Acoustic identification of individual animals with hierarchical contrastive learning

Ines Nolasco[†] Ilyass Moummad[◇] Dan Stowell[‡] Emmanouil Benetos[†]

[†]*Centre for Digital Music, Queen Mary University of London, London, UK*
[◇]*IMT Atlantique, CNRS, Lab-STICC, Brest, France*
[‡]*Department of Cognitive Science and Artificial Intelligence, Tilburg University, Tilburg, Netherlands*

***Abstract*—Acoustic identification of individual animals (AIID) is closely related to audio-based species classification but requires a finer level of detail to distinguish between individual animals within the same species. In this work, we frame AIID as a hierarchical multi-label classification task and propose the use of hierarchy-aware loss functions to learn robust representations of individual identities that maintain the hierarchical relationships among species and taxa. Our results demonstrate that hierarchical embeddings not only enhance identification accuracy at the individual level but also at higher taxonomic levels, effectively preserving the hierarchical structure in the learned representations. By comparing our approach with non-hierarchical models, we highlight the advantage of enforcing this structure in the embedding space. Additionally, we extend the evaluation to the classification of novel individual classes, demonstrating the potential of our method in open-set classification scenarios.**

***Index Terms*—Bioacoustics, contrastive loss, hierarchical classification, representation learning, open set.**

I. Introduction

Acoustic identification of individual animals (AIID) refers to the automatic differentiation of vocalisations among different individuals within a group. Many animal species exhibit distinct Acoustic Signatures —unique vocal characteristics that can be leveraged for identification. Traditionally, AIID has been approached in scenarios limited to a single species or a small, predefined group of individuals, [1]–[3]. However, this approach restricts the applicability of AIID in real-world settings where multiple species coexist and the set of individuals is not fixed, [4]–[6]. The AIID problem across multiple species and taxa can be framed as a constrained multi-label classification task, where each instance requires the prediction of three labels: individual identity, species, and taxon. Importantly, there is a constraint such that knowledge of a label at a lower level (e.g., a specific individual) inherently provides information about the higher levels.

We propose that hierarchical classification is particularly well-suited for animal-related acoustic tasks due to the strong phylogenetic relationships that influence their physical and behavioural traits, [7]. In this context, several related bioacoustic tasks can be organised within a hierarchical structure, akin to animal taxonomy. For example, while species classification and AIID both seek to distinguish vocal characteristics, they differ primarily in the granularity of detail required. Species classification focuses on identifying differences between species, whereas AIID targets the distinction of individuals within a species. This concept extends to broader taxonomic classifications, such as differentiating between mammals and birds, which generally requires more generalised characteristics. Nevertheless, all these tasks share a common foundation in signal representation, with the critical difference being the level of detail at which the algorithm analyses the signals. Hierarchical classification is often implemented by constraining the possible classes at lower levels of the hierarchy based on the prediction of the previous level [8], [9]. Here instead, the hierarchical structure in the label space is leveraged to guide the learning of the embeddings and highlight the similarities and the hierarchical relationships between features. By utilising hierarchical information at the feature level, we are able to obtain a representation of each ID that preserves the information regarding species, and taxonomic group. Preserving the "full picture" of our data examples is an important step towards classification of previously unseen classes.

Distance-based methods are particularly suitable in open-set scenarios, as they enable the system to learn data representations that can accommodate the inclusion of novel classes. To this end, we adapt the hierarchical contrastive learning approach proposed in [10] to AIID, aiming to create a more generalisable AIID system. In this implementation, the multi-label constraint is enforced by producing one prediction for each level of the hierarchy considered. Further refinements address the confidence level at each hierarchical level, given the confidence in the lower levels.

The primary contributions of this work are: 1) To the best of our knowledge, this is the first work to apply contrastive learning to AIID, leveraging hierarchical contrastive loss to create robust representations for this task. 2) Evaluation of hierarchical and contrastively-trained embeddings for the open-set scenario in which new "leaf" classes are encountered.

II. Related Work

Central to AIID in natural contexts are the challenges of generalisation and open set classification. Recently, this task has gained attention due to advancements in deep learning and foundation models for bioacoustics [6], [11]. Leveraging

the hierarchical structure of labels to improve AIID, was first explored in [12], who proposed a hierarchy-aware loss function to guide the learning of embeddings that preserve the hierarchical relationships and enhance individual classification across species. Hierarchy can also be useful in the open set classification, as shown in [13].

In this work, we adopt the hierarchical contrastive learning loss from [10] to further these goals. Contrastive learning has emerged as a powerful approach for learning robust feature representations. In its self-supervised form [14], contrastive learning aims to minimise the distance between positive pairs of samples, typically augmentations of the same instance, while maximising the separation from negative pairs, enabling models to learn meaningful feature spaces without explicit labels. This paradigm has proven effective in a variety of domains [15]. The extension to supervised contrastive learning [16] incorporates label information to further enhance discriminative power by grouping samples from the same class closer together, thus improving the learned feature space for tasks such as classification. Supervised contrastive learning has shown promising results in bioacoustics for few-shot capabilities in classification [17] and detection [18].

Recently, hierarchical contrastive learning [10] has emerged as an advancement in visual representation learning, addressing the need for models to recognise not only coarse-level categories but also fine-grained hierarchical relationships within the data. By incorporating hierarchical structures, such methods can create more structured and context-aware feature spaces, which are particularly useful for tasks where data exhibits multi-level or nested class relationships, as often seen in bioacoustics.

III. Dataset

The dataset is a collection of short recordings from vocalisations of animals of different species (see summary in table I). The data is sourced from different research initiatives focusing on animal communication. Recordings are made in the animal's natural environment, by experts that follow the individuals and annotate their ID. Due to the use of different acquisition methods, the recordings present high variation in acoustic characteristics. The dataset represents a natural setting in which systems for AIID need to operate.

Data is split into training, validation and test sets, which contain examples of 66 individuals from the various species. Additionally, an unseen ID set is defined which contains 3 novel IDs for each of the species in the training set (21 novel ID classes). In total there are 6055 examples for training, 1529 for validation, 1912 for test and 4799 examples of unseen ID classes.

Species	**Taxon**	**# Ids**
Chiffchaffs (CHF) [19]	Birds	23
Tree pipits (TP) [19]	Birds	10
Little Owls (LO) [19]	Birds	16
Eurasia eagle owls (EEO)	Birds	7
Spotted hyenas (SH) [20]	Mammals	5
Hyrax (HY) [21]	Mammals	19
Grey wolves (GW) [22]	Mammals	7
Total number of recordings	14295	

TABLE I
SUMMARY OF DATASET.

IV. Methods

A. Representation learning

In contrastive learning approaches, the model architecture comprises two important components: A feature extractor designed to map the input data into an abstract latent representation; and a shallow neural network called projector, which projects the features to a low dimensional space where the contrastive loss is computed. The projector is primarily used to train the feature extractor and is discarded after training (Fig. 1). Here, we describe the supervised contrastive losses explored in our experiments.

1) Supervised contrastive loss: The supervised contrastive loss (SupCon) [16] is formulated as:

$$L_{\text{SupCon}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau)}{\sum_{n \in N(i)} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_n / \tau)} \tag{1}$$

where $i \in I$ indexes an augmented sample within a batch, $P(i) = \{p \in I : y_p = y_i\}$ is the set of positive samples sharing the same label as i, $N(i) = I \setminus \{i\}$ represents the set of negative samples, and τ is a temperature scaling parameter.

2) Hierarchical Multi-label Contrastive Learning: SupCon [16] only leverages one hierarchy level. To leverage hierarchical class structures, we adopt a hierarchical multi-label contrastive learning framework that incorporates hierarchy-aware losses [10]:

HiMulCon (Hierarchical multi-label contrastive loss):

$$L_{\text{HiMulCon}} = \sum_{l \in L} \frac{1}{|L|} \sum_{i \in I} \frac{-\lambda_l}{|P_l(i)|} \sum_{p \in P_l(i)} L_{\text{pair}}(i, p_i^l) \tag{2}$$

where L represents different levels in the hierarchy, λ_l is a level-dependent penalty factor, $P_l(i)$ is the set of positive pairs at level l, and L_{pair} calculates the contrastive loss for a specific pair.

3) Hierarchical Multi-label Contrastive constraint enforcing: An additional constraint is included into the previous loss in order to ensure that losses at each level do not decrease with increasing hierarchy depth:

$$L_{\text{HiMulConE}} = \sum_{l \in L} \frac{1}{|L|} \sum_{i \in I} \frac{-\lambda_l}{|P(i)|} \sum_{p \in P_l(i)} \max(L_{\text{pair}}(i, p_i^l), L_{\text{max}}^{\text{pair}}(l-1)) \tag{3}$$

the term $L_{\text{max}}^{\text{pair}}(l-1)$ is the maximum loss computed from the previous level and defines the absolute minimum value of loss each level can achieve. HiMulCon consists on an independent penalty defined on each level, whereas the added constraint is a dependent penalty that is defined in relation to the losses computed at the lower levels. Both combined form the **HiMulConE** loss.

B. Classification

We assess the effectiveness of the feature extractors learned using various loss functions with a k-Nearest Neighbor (kNN) classifier, a common evaluation protocol for learned representations [15]. The evaluation is conducted in two scenarios: 1) classification of new examples from the same classes used during training, and 2) classification of data from previously unseen classes.

In the first scenario, for each test sample, the predicted label is determined by finding the k nearest neighbors from the training set and assigning the majority label from among those neighbors. This method allows us to evaluate how well the learned features generalize to new samples within the same classes.

In the second scenario, we assess the ability of the learned representations to generalize to previously unseen classes. Here, we consider both the training examples and all examples from the unseen classes when determining the nearest neighbors for a test sample. This approach allows us to evaluate how new classes are represented in the learned feature space, even though the model has not encountered them during pretraining. To further assess the generalisation capabilities of the models in an open-set scenario, we implement a one-shot classification setting. In this setup, a support set containing only one example from the unseen classes is used to classify query samples. The nearest neighbors for each query are identified from both the support set and the training set (with the latter acting as distractors). This evaluation highlights the ability of the models to learn from very few labeled examples of new classes in a more realistic scenario.

V. Experiments

A. Experimental setup

The losses are applied into the training of the network described in Fig. 1. This is defined as a hierarchical network, which contains one head for each level of the hierarchy. Embeddings from the OpenL3 pretrained model are computed using available code[1] from the HEAR challenge [23].

The experiments are defined as:

1) **SC:** for the non-hierarchical baseline we train a modified version of the network in figure1, where only the ID projector is kept. This network is trained with the SupCon loss, as defined in eq.(1).
2) **HC:** Hierarchy is included by training the complete network with the HiMulCon loss function (eq.(2)). In this experiment we define equal contribution of each level of the hierarchy into the final loss. (λ_l =1/3)
3) **HCλ:** In order to test the effect of the combining factor λ_l of eq.(2) on the performance, we experiment with various values. Fine-tuning on validation set shown the best values as $\lambda_{ID} = 10$, $\lambda_{species} = 5$, and $\lambda_{taxa} = 1$.
4) **HCE:** the hierarchy is further enforced on the learning process by applying the HiMulConE loss function (eq.(4)).

[1]https://github.com/hearbenchmark/hear-baseline

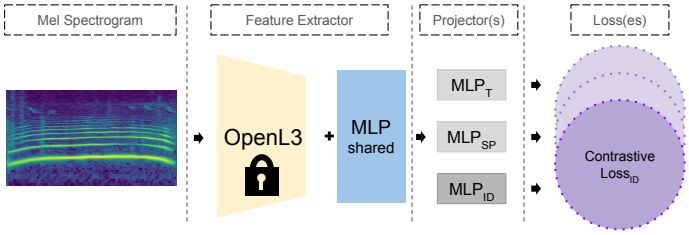

Fig. 1. Overview of our pretraining pipeline: audio recordings of animal calls are first processed through a frozen, pretrained OpenL3 model to extract high-level representations on each 25ms segment of the audio file. the final embeddings is the average of these across the whole call. The final openl3 embedding is then passed through a MLP to adapt the features specifically for bioacoustic sounds. The adapted features are subsequently fed into a projector to perform supervised contrastive learning for individual identification (ID). For hierarchical contrastive learning, two additional projectors are included for the species (SP) and taxa (T) classification.

5) **HCEλ:** Similar to experiment 3) various values for λ_l are experimented with and through tuning on the validation set we select the combination $\lambda_{ID} = 10$, $\lambda_{species} = 1$, and $\lambda_{taxa} = 1$.

Several hyperparameters, such as temperature, learning rate, weight decay, batch size, and λ values, are tuned using a sweep process across a defined range of values. The best parameters are selected which produce the highest accuracy on the validation set. For validation, accuracy is measured by applying *KNN* with $k = 1$ nearest neighbours to the embeddings extracted from the shared layer of the network.

B. Evaluation

The evaluation process is designed to assess three key aspects of the models: 1) How effectively can the models classify individual IDs; 2) Do the learned embedding spaces correctly capture the hierarchical relationships between labels? And 3), Can the models generalise to unseen classes at the ID level? To these effects, first the trained models are evaluated for their classification accuracy across all three levels of the hierarchy. Classification is performed by applying the *KNN* classification (see section **??**) to the embeddings extracted from the shared layer of the network (see Fig. 1). Due to the imbalanced nature of the dataset, the accuracy values are computed using the balanced_accuracy_score implementation from *sklearn* [24]. This ensures that classes with fewer examples do not disproportionately affect the overall performance evaluation.

Secondly, We define two types of hierarchical inconsistency errors: **species/ID** - predicted ID does not belong to the predicted species; **taxon/species** - predicted species does not belong to the predicted taxonomic group. Here the evaluation focus not on how correct the model is, but instead if it is consistent in its predictions accordingly to the hierarchy.

And finally we assess the generalisation capability to classify unseen ID classes in two ways, first we evaluate if the trained embedding space is suitable to represent novel classes by employing NN classification using the combination of train and unseen ID set as the reference set, (see section IV-B).

Secondly, because this is not a real use case of classification on novel classes, we also test the models' ability to identify a new class based on one single example. The 1-shot classification results allow us to understand the potential use of these models in the open set scenario.

C. Results

All the trained models are evaluated on both test and unseen ID set. Accuracy values are reported in Tables II and III respectively. Regarding the analysis on hierarchical inconsistency errors, results obtained indicate that all models produce consistent predictions regarding the hierarchy. Since no errors of these types were found we are led to observe that all misclassifications occur within the correct parent class for the ID and Species levels.

Test Set	SC	HC	HCλ	HCE	HCEλ
CHF					
Species	99.8	**100**	**100**	**100**	**100**
ID	90.9	94.4	**94.7**	94.4	93.8
TP					
Species	92.8	**100**	98.7	97.8	98.9
ID	56.8	**73.7**	62.8	67.8	69.9
LO					
Species	98.6	**100**	**100**	99.3	**100**
ID	44.7	**70.3**	69.7	53.3	69.7
EEO					
Species	98.1	98.0	98.1	98.1	**100**
ID	53.8	**55.8**	61.5	44.2	53.8
SH					
Species	98.8	**100**	**100**	**100**	96.6
ID	96.5	**100**	97.7	98.9	94.3
HY					
Species	99.4	**100**	**100**	**100**	99.4
ID	49.7	57.1	**60.4**	44.6	59.9
GW					
Species	**100**	**100**	**100**	**100**	**100**
ID	92.3	88.4	**96.1**	92.7	92.3
Balanced acc					
Taxon	99.7	**100**	**100**	**100**	99.8
Species	98.2	**99.7**	94.5	99.3	99.3
ID	61.0	**73.2**	72.2	64.2	72.3

TABLE II
BALANCED ACCURACY, OVERALL AND FOR EACH SPECIES, ON THE TEST SET ACROSS ALL EVALUATED MODELS.

VI. DISCUSSION

In this work, we framed the problem of acoustic identification of individuals as a multi-label hierarchical task. Through the use of contrastive learning-based methods, we investigated how learning an embedding space that captures the hierarchical structure of the labels can enhance individual identification. Additionally, we explored the limits of these models as encoders for performing classification of novel classes at the ID level.

The accuracy results on the test set (see table II) demonstrate that guiding the models to jointly optimise distances between embeddings across all hierarchical levels improves performance, not only at the fine-grained ID level but also at higher levels of the hierarchy, such as taxon and species. This confirms the value of preserving hierarchical relationships in the learned representations. Additionally, per-species accuracy results indicate the heterogeneous nature of the problem which contributes to the challenge of performing AIID for multiple species.

Unseen IDs	SC	HC	HCλ	HCE	HCEλ
NN					
Taxon	99.1	**99.7**	**99.7**	99.6	99.1
Species	96.9	**99.1**	**99.1**	98.3	97.8
ID	80.9	85.7	**88.1**	86.0	84.7
1-shot					
Taxon	92.2	**97.7**	96.5	97.3	95.6
Species	84.2	92.3	**93.8**	89.3	88.2
ID	6.0	9.6	**15.3**	9.8	13.22

TABLE III
BALANCED ACCURACY RESULTS ON THE UNSEEN ID SET FOLLOWING TWO CLASSIFICATION PROCESSES: NN CLASSIFICATION AND 1-SHOT CLASSIFICATION.

When the models are applied to novel ID classes, the results indicate that the new classes are well represented in the learnt embedding space. Also the advantage of applying hierarchy aware training losses is clear in this novel class scenario. In a real scenario, a novel ID class scenario presents more closely as an open set classification problem, where we need to identify a new class from the first moment an example appears without having a complete representation from other examples of the new class. This situation is approximated here by employing a 1-shot classification approach. All the models exhibited a notable drop in accuracy, however showing a clear advantage of the models that included hierarchy. Furthermore, despite this decline, the models consistently preserved the hierarchical structure, as evidenced by the strong performance at the taxon and species levels. Moreover, the absence of consistency errors suggests that most misclassifications occur within the correct parent class—errors at the ID level involve confusion between IDs of the same species, rather than across species.

Overall, this work shows the potential of hierarchical embeddings for improving identification accuracy across multiple levels of an hierarchical problem. And while few-shot learning remains a challenging task, especially for novel classes at the individual ID level, the preservation of hierarchical integrity indicates that our approach provides a robust framework for AIID. Although we employ frozen features in this work, future research could investigate feature adaptation techniques to enhance classification performance and explore methodologies for open-set classification.

ACKNOWLEDGEMENTS

I. Nolasco is supported by the Engineering and Physical Sciences Research Council [grant number EP/R513106/1]. E. Benetos is supported by a RAEng/Leverhulme Trust Research Fellowship [grant number LTRF2223-19-106].

REFERENCES

[1] Sougata Sadhukhan, Holly Root-Gutteridge, and Bilal Habib, "Identifying unknown Indian wolves by their distinctive howls: its potential as a non-invasive survey method," *Scientific Reports*, vol. 11, no. 1, pp. 1–13, 2021.

[2] Sophia Yin and Brenda McCowan, "Barking in domestic dogs: Context specificity and individual identification," *Animal Behaviour*, vol. 68, no. 2, pp. 343–355, 2004.

[3] Matthew Wijers, Paul Trethowan, Byron Du Preez, Simon Chamaillé-Jammes, Andrew J Loveridge, David W Macdonald, and Andrew Markham, "Vocal discrimination of african lions and its potential for collar-free tracking," *Bioacoustics*, vol. 30, no. 5, pp. 575–593, 2021.

[4] Ladislav Ptacek, Lukas MacHlica, Pavel Linhart, Pavel Jaška, and Ludek Muller, "Automatic recognition of bird individuals on an open set using as-is recordings," *Bioacoustics*, vol. 25, no. 1, pp. 55–73, 2016.

[5] Stavros Ntalampiras and Ilyas Potamitis, "Acoustic detection of unknown bird species and individuals," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 3, pp. 291–300, 2021.

[6] Elly Knight, Tessa Rhinehart, Devin R de Zwaan, Matthew J Weldy, Mark Cartwright, Scott H Hawley, Jeffery L Larkin, Damon Lesmeister, Erin Bayne, and Justin Kitzes, "Individual identification in acoustic recordings," *Trends in Ecology & Evolution*, 2024.

[7] Jozsef Arato and W Tecumseh Fitch, "Phylogenetic signal in the vocalizations of vocal learning and vocal non-learning birds," *Philosophical transactions of the royal society B*, vol. 376, no. 1836, pp. 20200241, 2021.

[8] Jason Cramer, Vincent Lostanlen, Andrew Farnsworth, Justin Salamon, and Juan Pablo Bello, "Chirping up the right tree: Incorporating biological taxonomies into deep bioacoustic classifiers," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 901–905.

[9] Dongliang Chang, Kaiyue Pang, Yixiao Zheng, Zhanyu Ma, Yi-Zhe Song, and Jun Guo, "Your" flamingo" is my" bird": Fine-grained, or not," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11476–11485.

[10] Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah, "Use all the labels: A hierarchical multi-label contrastive learning framework," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16660–16669.

[11] Masato Hagiwara, "Aves: Animal vocalization encoder based on self-supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[12] Inês Nolasco and Dan Stowell, "Rank-based loss for learning hierarchical representations," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3623–3627.

[13] Nico Lang, Vésteinn Snæbjarnarson, Elijah Cole, Oisin Mac Aodha, Christian Igel, and Serge Belongie, "From coarse to fine-grained open-set recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17804–17814.

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1597–1607.

[15] Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum, "A Cookbook of Self-Supervised Learning," *arXiv preprint arXiv:2304.12210*, 2023.

[16] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised Contrastive Learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18661–18673, 2020.

[17] Ilyass Moummad, Nicolas Farrugia, and Romain Serizel, "Self-Supervised Learning for Few-Shot Bird Sound Classification," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2024, pp. 600–604.

[18] Ilyass Moummad, Nicolas Farrugia, and Romain Serizel, "Regularized Contrastive Pre-training for Few-shot Bioacoustic Sound Detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1436–1440.

[19] Dan Stowell, Tereza Petrusková, Martin Šálek, and Pavel Linhart, "Datasets for automatic acoustic identification of individual birds," Oct. 2018.

[20] Kenna DS Lehmann, Frants H Jensen, Andrew S Gersick, Ariana Strandburg-Peshkin, and Kay E Holekamp, "Long-distance vocalizations of spotted hyenas contain individual, but not group, signatures," *Proceedings of the Royal Society B*, vol. 289, no. 1979, pp. 20220548, 2022.

[21] Vlad Demartsev, Arik Kershenbaum, Amiyaal Ilany, Adi Barocas, Yishai Weissman, Lee Koren, and Eli Geffen, "Lifetime changes in vocal syntactic complexity of rock hyrax males are determined by social class," *Animal Behaviour*, vol. 153, pp. 151–158, 2019.

[22] Holly Root-Gutteridge, *Improving individual identification of wolves (Canis lupus) using the fundamental frequency and amplitude of their howls: a new survey method*, Nottingham Trent University (United Kingdom), 2013.

[23] Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W Schuller, Christian J Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, et al., "Hear: Holistic evaluation of audio representations," in *NeurIPS 2021 Competitions and Demonstrations Track*. PMLR, 2022, pp. 125–145.

[24] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann, "The balanced accuracy and its posterior distribution," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 3121–3124.