COMPARING WITHOUT CLASSIFYING: EVALUATING EMBEDDINGS FROM BIOACOUSTIC DEEP LEARNING FEATURE EXTRACTORS

Vincent S. Kather[1,2*] **Burooj Ghani**[2] **Dan Stowell**[1,2]
[1] Department of Cognitive Science and Artificial Intelligence, Tilburg University, Netherlands
[2] Naturalis Biodiversity Center, Leiden, Netherlands

ABSTRACT

In computational bioacoustics, deep learning models are composed of feature extractors and classifiers. The feature extractors generate vector representations of the input sound segments, called embeddings. While benchmarking of classification scores provides insight into specific performance statistics, it is limited to species that are included in the models training data. Furthermore, it makes it impossible to compare models trained on very different taxonomic groups. This paper aims to address this gap by analyzing the embeddings generated by the feature extractor of more than 15 bioacoustic models spanning a wide range of setups (model architectures, training data, training paradigms). We evaluate and compare different ways how models structure embedding spaces through clustering and linear classification, which allows us to focus our comparison on feature extractors independent of classifiers. We believe that this approach lets us evaluate the adaptability and generalization potential of models going beyond the classes they were trained on.

Keywords: *deep learning, bioacoustics, embeddings*

1. INTRODUCTION

Human-driven climate change and deforestation have caused a rapid decline in global biodiversity [1]. Using sensor arrays for biodiversity monitoring, ecologists can gather information on environments and investigate how human pressures affect biodiversity and how we can halt its decline [2]. Passive acoustic monitoring (PAM) is one method that provides a low-cost and non-invasive way to monitor biodiversity [3]. The vast amount of data generated by PAM sensors has led to the rapid development of bioacoustic deep learning models to help researchers reduce the annotation effort [4]. The use of these state-of-the-art models has proven valuable in ecological studies, for example in general species assessments [5,6] or detection of endangered species [7].

While bioacoustic deep learning models are a useful asset in ecology, it is crucial to understand the model's limitations based on their training setup. The two main training strategies to develop bioacoustic deep learning models are supervised learning and self-supervised learning. Supervised learning models require large amounts of annotated data to be trained [8]. The models classify sounds based on a fixed number of predefined classes representing annotated sounds in the dataset. While supervised learning can have the benefit of instructing models to differentiate between species vocalizations, it requires annotated datasets. This limits supervised learning models to known and annotated classes and makes them sensitive to class imbalance and label quality. Self-supervised learning can be executed in different ways. In a paradigm referred to as masked prediction, the model is trained to predict a masked portion of the audio, thereby modelling bioacoustic characteristics without supervision [9, 10]. With growing annotated databases supervised learning models like Birdnet [11] improve in performance, yet recent developments in self-supervised learning on models like animal2vec [12] indicate a promising new direction for the field requiring less manual annotation. Not requiring annotations, self-supervised learning mod-

[*] ***Corresponding author***: *vincent.kather@naturalis.nl.*

els can be trained on far larger datasets, however there is no control of what is being learned. The model might learn to differentiate between sounds based on very different characteristics than the species that produces them.

To evaluate bioacoustic deep learning models, a comparison of only the classifier performance obscures the fine-grained differences between models and how they analyze input sounds. Bioacoustic deep learning models consist of artificial neural networks which can be subdivided into feature extractors and classifiers. The feature extractor creates an embedding (vector representation) of an input sound and the classifier (which corresponds to the final dense layer of the model) maps the embedding onto classes. Commonly, in bioacoustics, a suite of established benchmarks is used to compare the classifier performance of state-of-the-art models [13]. This requires the models to have been trained on the classes present in the benchmarking datasets. However, there is an alternative: using the embeddings created by the feature extractors, the generated embedding spaces can be analyzed in regard to their structural characteristics, irrespective of what the classifiers were trained on.

Output dimensions of feature extractors vary greatly, but it is uncertain if their dimensionality correlates to the downstream classifier performance. Dimensionality reduction algorithms are useful to standardize the dimensionality of different feature extractors, as well as help visualize the high dimensional embedding spaces. However, there is little investigation of how reducing the embedding space affects performance. We therefore compare performance in both original and reduced embedding spaces.

Due to their size, datasets that are used to train bioacoustic models for bird detection are often based on citizen science data (e.g. xeno-canto [14]). The majority of these recordings are focal recordings of individuals which are weakly labeled with little polyphony. When models get applied to large PAM datasets, these feature very different recording conditions. To accurately evaluate the models in this study, we use a PAM bird vocalization dataset from Colombia and Costa Rica [15], as well as a dataset of frog vocalizations in tropical rainforests of Brazil [16]. Both datasets are comprised of PAM recordings in noisy environments and can therefore be considered as challenging datasets. Through the selection of challenging datasets, we hope to on the one hand emphasize the performance differences between the models and on the other hand produce results that will reflect in real world applications.

This study aims to showcase the potential for evaluating and especially comparing bioacoustic deep learning models in regard to their training paradigm and training data. This evaluation is based on the structuring capabilities of their respective feature extractors analyzed through clustering and classification performance. Classification in this case refers to recognition of novel classes, as none of the models have been pretrained on the classes selected here. This way single classification layers are attached to each feature extractor, all of which are trained on the same evaluation sets (bird and frog vocalizations), i.e. same classes and same data, allowing us to compare their performance. Classification is done both in a linear and a k-nearest neighbor (kNN) approach. We perform our analysis in both the original embedding space and a reduced dimensional embedding space. That way we ensure the dimension is standardized for the second evaluation, whilst we can investigate performance differences between the original embedding space and the reduced space. This method of analysis opens up the possibilities for a fair comparison of deep learning feature extractors guiding the field to a better understanding how training configurations affect downstream performance.

2. METHODS

To incorporate a variety of training setups, covering popular models as well as models targeted to various species groups, we compare a total of 15 pre-trained bioacoustic different feature extractors. Table 1 shows the different feature extractors along with their model specific training setup. As can be seen both self-supervised and supervised learning feature extractors are represented. Furthermore, large variations in input length, embedding dimension and training data provide a landscape of feature extractors, allowing us to analyze performance of differently structured embedding spaces.

2.1 Dataset

The evaluation dataset that was used for this study is a collection of soundscape recordings from neotropical coffee farms in Colombia and Costa Rica [15]. The recordings in this dataset feature challenging soundscape recordings with overlapping vocalizers and noisy environments. The annotations are made for bird species. The dataset has been reduced from its original size to only include sound events corresponding to classes with more than 150 annotations. This results in 11 vocalizing bird species ranging in annotation count from 153 to over 4000 (see legend in Fig. 1). We intentionally selected soundscape recordings with gradual changes of background noise and overlapping species vocalizations to amplify the differences between the feature extractors' capabilities to structure the data.

2.2 Data pipeline

For each of the feature extractors, the respective model code base was cloned, and the model was stripped of its classifier. For both animal2vec feature extractors, outputs from the attention heads and input lengths are averaged, resulting in one embedding per input segment (as is the case with all other feature extractors). Data is imported from the sound files, resampled to the model specific sample rate and padded to fit the model specific input length. All the necessary code to reproduce the computations can be found in the repository **bacpipe**[1] (**bioa**coustic **c**ollection **pipe**line).

2.3 Evaluating dimensionality reduction

Our primary focus is the comparison of the two paradigms: supervised learning and self-supervised learning. Furthermore, we are looking into how the data chosen for training affects the clustering capabilities of different feature extractors.

The clustering is computed using KMeans with the same number of clusters as classes in the ground truth. Clustering performance will be evaluated using Adjusted Mutual Information (AMI) [21] to compare the KMeans clustering with the ground truth. Adjusted Rand Index is not included in this study, as it focuses on how well data points are grouped in a clustering, whereas we are primarily interested how well the KMeans clustering agrees with the ground truth labels. Silhouette Score is also not included in this comparison as the challenging dataset yielded very low performance and variance, making a meaningful comparison impossible.

Clustering performance is evaluated in both the original embedding spaces and an embedding space reduced to 300 dimensions. This way the embedding dimension is standardized and performance can be compared while controlling for this factor. It also allows us to compare the performance of each model in their high dimensional original embedding space, as well as in a reduced dimension. To preserve relative distances between data points, Principal Component Analysis (PCA), a linear dimensionality reduction is selected. To visualize the embeddings in two dimensions, a non-linear dimensionality reduction algorithm, uniform manifold approximation projection (UMAP) [22] is selected.

Performance is also evaluated by training a linear classifier on each of the embedding spaces. Data is split into train, validation and test set in the ratio 0.65:0.15:0.2. The classifier is trained on the 11 classes for 10 epochs with a batch size of 64 and a learning rate of 0.001. Performance is evaluated using a balanced macro accuracy score [23] to handle the imbalance in class size.

3. RESULTS

Two dimensional UMAP embeddings are shown in Fig. 1. The worst performing feature extractors, produce large unstructured clouds of mixed color, indicating that no significant clustering is achieved. In the first and second row, feature extractors can be seen to separate the embeddings into meaningful clusters. It is noticable that some feature extractors such as AvesEcho_PaSST and ProtoCLR seem to generate more subclusters than most other feature extractors. The seven best performing feature extractors are all trained using supervised learning and the top three additionally trained on bird vocalizations. All three of the AVES models (BirdAVES, AVES and NonBioAVES) reach similar performances in spite of big differences in their fine-tuning datasets [8].

To investigate how training setup and training data affect performance, Fig. 2 shows a scatterplot of the different feature extractors. Performance is evaluated by macro accuracy of linear classification on the x-axis and AMI of clustering on the y-axis.

When focussing on the y-axis, all self-supervised learning feature extractors (in red) reach clustering performances under 0.31. Performance by linear classification is more equally distributed, however, again supervised learning feature extractors reach the three highest values. Furthermore, Animal2vec_XC, the only self-supervised learning feature extractor that was not fine-tuned, performs poorly by both clustering and linear classification. Google_Whale represents the only supervised learning feature extractor performing very poorly by clustering.

Comparing by training data, feature extractors trained on only or including bird datasets outperform the other feature extractors by linear classification and even more so by clustering. When looking at the combined performance by clustering and linear classification, Animal2vec_XC and ProtoCLR are the only two feature extractors trained

[1] `github.com/bioacoustic-ai/bacpipe`

Table 1. List of feature extractors compared in this study. Columns ”abbrev.” shows the an abbreviated name used in Figures 2 and 3. ”training” shows the training setup chosen during training, i.e. ssl for self-supervised learning, sup l for supervised learning and ft for fine-tuning. The ”architecture” column more specifically describes the model architecture used. ”dimension” shows the output dimension of the feature extractor. ”trained on” summarizes the species group that was used to train the model. ”ref.” provides the respective publication for a given model.

name	abbrev.	training	architecture	dimension	trained on	ref.
Animal2vec_XC	brdnet	ssl	d2v2.0	768	birds	[12]
Animal2vec_MK	a2v_xc	ssl + ft	d2v2.0	1024	meerkats	[12]
AudioMAE	a2v_mk	ssl + ft	ViT	768	general	[10]
AVES_ESpecies	aud_mae	ssl + ft	HuBERT	768	general + animals	[8]
AvesEcho_PaSST	aves	sup l	PaSST	768	birds	[17]
BioLingual	bioling	mm sup l	CLAP	512	animals + birds	[18]
BirdAVES_ESpecies	birdaves	ssl + ft	HuBERT	1024	general + birds	[8]
BirdNET	aecho	sup l	EffNetB0	1024	birds	[11]
Google_Whale	i66	sup l	EffNetB0	1280	whales	-
Insect459NET	i459	sup l	EffNetv2s	1280	insects	-
Insect66NET	perch	sup l	EffNetv2s	1280	insects	-
NonBioAVES_ESpecies	p_clr	ssl + ft	HuBERT	1024	general + non-bio	[8]
Perch_Bird	s_perch	sup l	EffNetB0	1280	birds	-
ProtoCLR	g_whale	sup l	CvT-13	384	birds	[19]
SurfPerch	nonbioaves	sup l	EffNetB0	1280	coral reefs + birds	[20]

on birds that perform poorly. Biolingual, which was trained on large bird databases using a multi-modal approach performs well by clustering, but poorly by linear classification.

When referring back to Table 1 embedding dimension does not correlate with clustering or linear classification performance. Furthermore, the only two feature extractors trained on marine sounds, Google_Whale and SurfPerch (trained on birds and marine sounds) reach very different performances.

To account for the differences in embedding dimension, we visualized the change in performance between the original embedding space and a standardized embedding dimension of 300, to which all embedding spaces were reduced to using PCA. The results are shown in Fig. 3. While the poorly performing models ProtoCLR and Animal2vec_MK are able to increase their linear classification performance, Insect459, Insect66 and Google_Whale slightly improve linear classification and clustering performance. While linear classification performance remains similar, SurfPerch, BirdNET and AvesEcho_PaSST all decrease in clustering performance. For the remaining feature extractors, standardizing the dimension, does not affect performance significantly.

4. DISCUSSION

Although the self-supervised feature extractors represented in this study are trained on very large datasets, their combined performance is inferior to most of the supervised learning feature extractors including those trained on non-bird categories. Perch and BirdNET [11], both of which are trained on thousands of classes of bird vocalizations vastly outperform the other feature extractors. Similarly, SurfPerch [20] and AvesEcho_PaSST [17], both of which have largely benefited from Perch or BirdNET pretraining, perform well by both clustering and linear classification. More generally, the embedding spaces of feature extractors trained on very large annotated bird song datasets seem to yield good separation between clusters. Interestingly, Insetct66 and Insect459 both of which were trained on solely insect sounds and far fewer classes than the bird-trained biolingual, vastly outperform it in

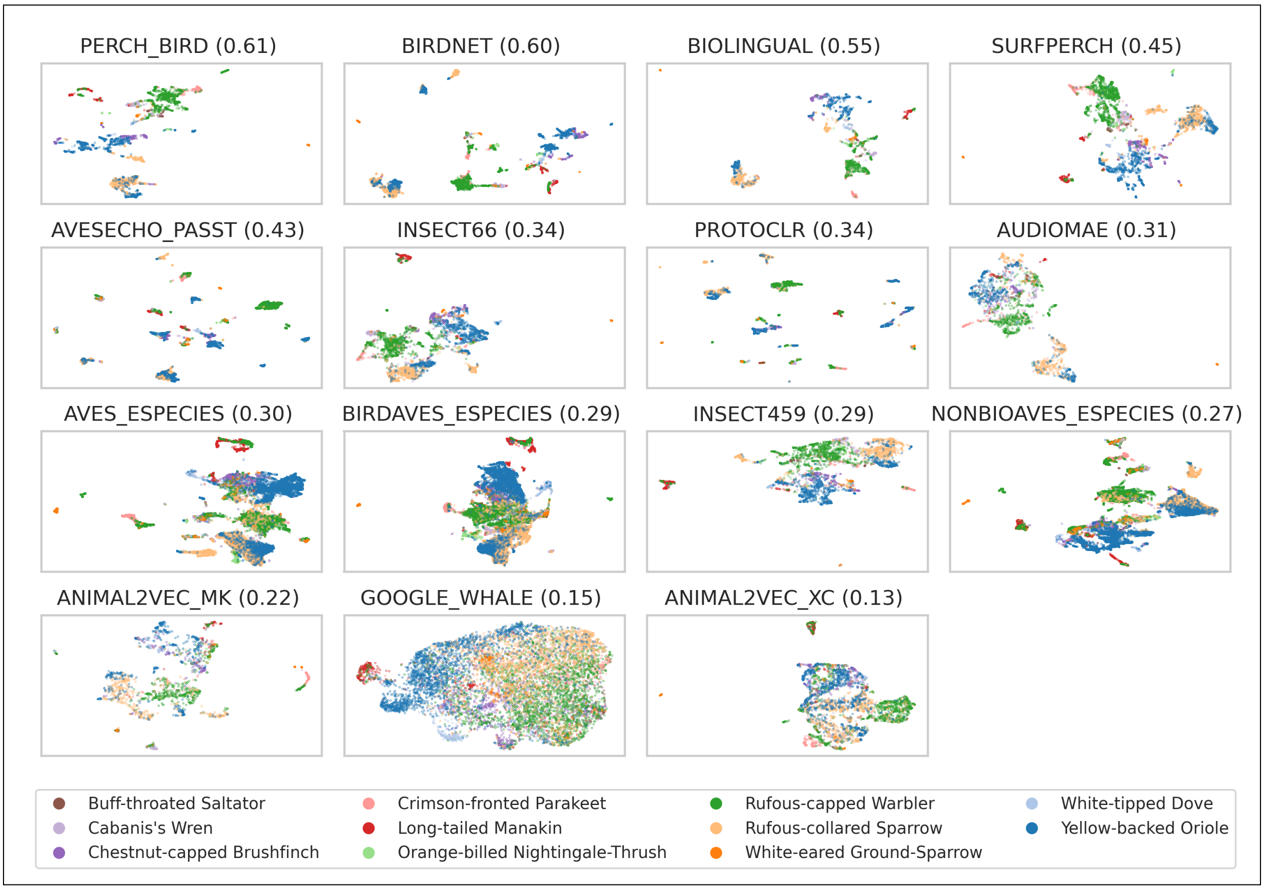

Figure 1. Two-dimensional embedding spaces of all feature extractors, sorted descending by their clustering performance of AMI values (indicated next to their name) from top left to bottom right. Given the different input lengths of the feature extractors, the number of embeddings vary significantly. Colors correspond to the class labels, which are 11 different tropical bird species.

linear classification. As stated in the introduction, self-supervised learning models lack supervision and might therefore learn classes not meaningful to differentiate between species vocalizations. When comparing Figures 1 and 2 we observe poor clustering both in terms of AMI performance and by qualitative visual analysis of the embedding separation, which could be resulting from non-meaningful classes. However, for the three AVES feature extractors, the linear classifier is nontheless able to learn a meaningful differentiation between the classes. This could be attributed to the fact, that while they are self-supervised, the training data consisted of curated and non-sparse sound events, thereby increasing the likelihood that meaningful classes are learned.

For the self-supervised learning feature extractors, fine-tuning seems to only marginally improve performance. The three feature extractors based on the AVES models all share the same general audio pretraining and architecture but differ largely in fine-tuning. The similarity in performance indicates that the dominant influence on the structuring of embeddings is defined by either the pretraining or the architecture. Animal2vec_XC and Animal2vec_MK, the latter of which is fine-tuned, largely share the same architecture but were trained on very different datasets. Yet, both models reach similarly bad performance, this is especially surprising for the fine-tuned Animal2vec_MK. For the supervised learning feature extractor SurfPerch, which was developed for marine data in

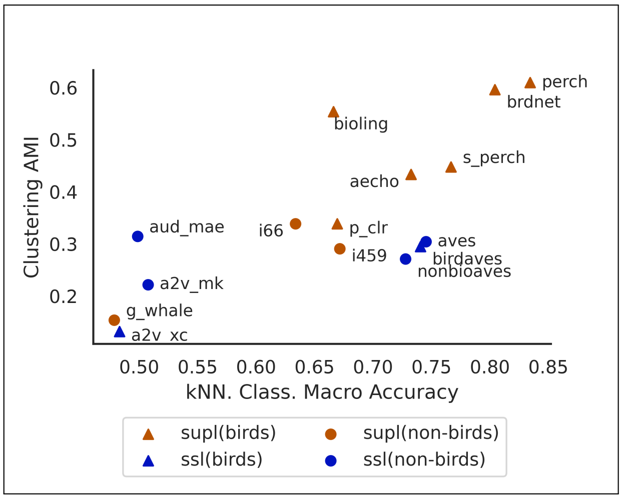

Figure 2. Comparison of feature extractors by learning paradigm and training data. Abbreviated names correspond to abbrev. column in Tab. 1. Differences in color correspond to training paradigm and differences in symbols correspond to training data. The x-axis shows clustering results of AMI while the y-axis shows macro accuracy results of linear classification. Colors correspond to supervised learning and self-supervised learning feature extractors, while symbols separate bird and non-bird training data.

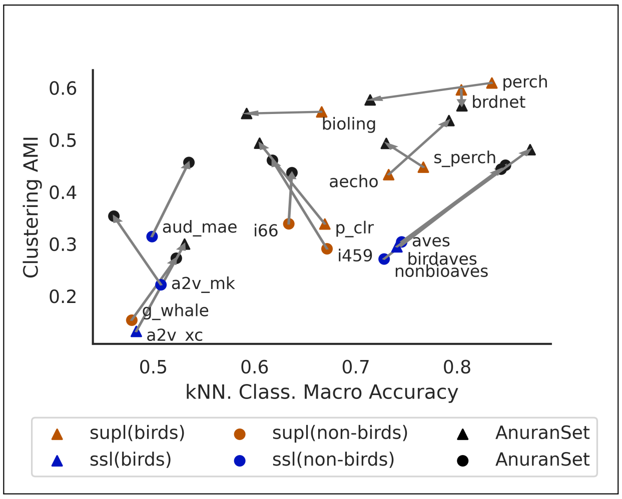

Figure 3. Comparison of feature extractor performance in original high dimensional embedding space and reduced 300 dimensional space using PCA. The x-axis shows clustering results of AMI while the y-axis shows macro accuracy results of linear classification. Symbols denote supervised or self-supervised learning, while red and green correspond to bird and non-bird training data. The grey line and black markers indicate performance in the reduced dimensional embedding space.

coral reefs, bird training data from Perch was mixed with coral reef sounds during pretraining. While the target domain is very different from the bird sound dataset used in this evaluation study, SurfPerch still reaches very high performance. Again, this indicates that including data of the target domain is more effective during pretraining than fine-tuning.

While the dimensions of the feature extractors vary greatly, performance does not correlate with dimension. Nonetheless, in Fig. 3 we demonstrated that using a standardized embedding space only marginally impacts performance and does not change the hierarchy of performance among most of the models. The performance differences that can be observed are predominantly in linear classification, indicating that the graph structure, which KMeans builds for the clustering, does not change much through a linear dimensionality reduction.

This analysis underlines the high quality of embeddings created by large supervised learning feature extractors like BirdNET and Perch. While their training domain aligned with the domain of the evaluation dataset, so did that of biolingual, Animal2vec_XC and BirdAVES_ESpecies, none of which reached performance metrics similar to BirdNET and Perch. Nonetheless, reproducing this comparison with an evaluation set different from the training domains of these models would be very interesting. This study is meant to present a workflow for a more in-depth analysis of embedding spaces, which can be reproduced with the provided repository bacpipe [2] . By establishing a default analyses of feature extractors alongside the common classification benchmarks, bioacoustic research can accelerate towards a better understanding of what training characteristics are beneficial in this domain.

5. CONCLUSION

In this study we have compared a variety of different state-of-the-art bioacoustic deep learning models, representing

[2] `github.com/bioacoustic-ai/bacpipe`

different training paradigms and training domains. To compare the models, we have isolated their feature extractors and used them to generate embeddings of a curated evaluation dataset consisting of annotated bird sounds from recordings in Costa Rica and Colombia. The aim of this study was to firstly present a large comparison of very different bioacoustic deep learning feature extractors and to evaluate how training paradigms and training domains affect performance. Performance was evaluated through clustering using an AMI score and through linear classification using a macro accuracy score.

We have shown that in spite of recent improvements in bioacoustic self-supervised learning, performance is still inferior to that of supervised learning models. This performance difference is visible in both linear classification and even more in clustering. Furthermore, we have shown that alignment of training domain and target domain during pretraining impacts performance more, than during fine-tuning. This study presents a roadmap for a more in-depth performance evaluation of bioacoustic deep learning models, allowing for a better understanding of how training setup impacts downstream performance.

6. ACKNOWLEDGMENTS

A lot of work went into developing each of the feature extractors presented here, and we highly appreciate that they are made publicly available. We would also like to acknowledge Álvaro Vega-Hidalgo and his coauthors, who collected, assembled and annotated the dataset used to evaluate the feature extractors. This project is funded by the Marie Skłodowska-Curie doctoral network BioAcousticAI.

7. REFERENCES

[1] S. H. M. Butchart, M. Walpole, B. Collen, A. van Strien, J. P. W. Scharlemann, R. E. A. Almond, J. E. M. Baillie, B. Bomhard, C. Brown, J. Bruno, K. E. Carpenter, G. M. Carr, J. Chanson, A. M. Chenery, J. Csirke, N. C. Davidson, F. Dentener, M. Foster, A. Galli, J. N. Galloway, P. Genovesi, R. D. Gregory, M. Hockings, V. Kapos, J.-F. Lamarque, F. Leverington, J. Loh, M. A. McGeoch, L. McRae, A. Minasyan, M. H. Morcillo, T. E. E. Oldfield, D. Pauly, S. Quader, C. Revenga, J. R. Sauer, B. Skolnik, D. Spear, D. Stanwell-Smith, S. N. Stuart, A. Symes, M. Tierney, T. D. Tyrrell, J.-C. Vié, and R. Watson, "Global Biodiversity: Indicators of Recent Declines," *Science*, vol. 328, pp. 1164–1168, May 2010.

[2] D. S. Schmeller, M. Böhm, C. Arvanitidis, S. Barber-Meyer, N. Brummitt, M. Chandler, E. Chatzinikolaou, M. J. Costello, H. Ding, J. García-Moreno, M. Gill, P. Haase, M. Jones, R. Juillard, W. E. Magnusson, C. S. Martin, M. McGeoch, J.-B. Mihoub, N. Pettorelli, V. Proença, C. Peng, E. Regan, U. Schmiedel, J. P. Simaika, L. Weatherdon, C. Waterman, H. Xu, and J. Belnap, "Building capacity in biodiversity monitoring at the global scale," *Biodiversity and Conservation*, vol. 26, pp. 2765–2790, Nov. 2017.

[3] L. S. M. Sugai, T. S. F. Silva, J. W. Ribeiro, Jr, and D. Llusia, "Terrestrial Passive Acoustic Monitoring: Review and Perspectives," *BioScience*, vol. 69, pp. 15–25, Jan. 2019.

[4] D. Stowell, "Computational bioacoustics with deep learning: A review and roadmap," *PeerJ*, vol. 10, p. e13152, Mar. 2022.

[5] D. Tuia, B. Kellenberger, S. Beery, B. R. Costelloe, S. Zuffi, B. Risse, A. Mathis, M. W. Mathis, F. van Langevelde, T. Burghardt, R. Kays, H. Klinck, M. Wikelski, I. D. Couzin, G. van Horn, M. C. Crofoot, C. V. Stewart, and T. Berger-Wolf, "Perspectives in machine learning for wildlife conservation," *Nature Communications*, vol. 13, p. 792, Feb. 2022.

[6] A. Cowans, X. Lambin, D. Hare, and C. Sutherland, "Improving the integration of artificial intelligence into existing ecological inference workflows," *Methods in Ecology and Evolution*, vol. n/a, Dec. 2024.

[7] S. Allen-Ankins, S. Hoefer, J. Bartholomew, S. Brodie, and L. Schwarzkopf, "The use of BirdNET embeddings as a fast solution to find novel sound classes in audio recordings," *Frontiers in Ecology and Evolution*, vol. 12, Jan. 2025.

[8] M. Hagiwara, "AVES: Animal Vocalization Encoder based on Self-Supervision," Oct. 2022.

[9] A. Baevski, A. Babu, W.-N. Hsu, and M. Auli, "Efficient Self-supervised Learning with Contextualized Target Representations for Vision, Speech and Language," in *Proceedings of the 40th International Conference on Machine Learning*, pp. 1416–1429, PMLR, July 2023.

[10] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked

Autoencoders that Listen," *Advances in Neural Information Processing Systems*, vol. 35, pp. 28708–28720, Dec. 2022.

[11] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "BirdNET: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101236, Mar. 2021.

[12] J. C. Schäfer-Zimmermann, V. Demartsev, B. Averly, K. Dhanjal-Adams, M. Duteil, G. Gall, M. Faiß, L. Johnson-Ulrich, D. Stowell, M. B. Manser, M. A. Roch, and A. Strandburg-Peshkin, "Animal2vec and MeerKAT: A self-supervised transformer for rare-event raw audio input and a large-scale reference dataset for bioacoustics," July 2024.

[13] J. Hamer, E. Triantafillou, B. van Merriënboer, S. Kahl, H. Klinck, T. Denton, and V. Dumoulin, "BIRB: A Generalization Benchmark for Information Retrieval in Bioacoustics," Dec. 2023.

[14] xeno-canto, "Xeno-canto :: Sharing wildlife sounds from around the world." https://xeno-canto.org/, 2025.

[15] Á. Vega-Hidalgo, S. Kahl, L. B. Symes, V. Ruiz-Gutiérrez, I. Molina-Mora, F. Cediel, L. Sandoval, and H. Klinck, "A collection of fully-annotated soundscape recordings from neotropical coffee farms in Colombia and Costa Rica," Jan. 2023.

[16] J. S. Cañas, M. P. Toro-Gómez, L. S. M. Sugai, H. D. Benítez Restrepo, J. Rudas, B. Posso Bautista, L. F. Toledo, S. Dena, A. H. R. Domingos, F. L. de Souza, S. Neckel-Oliveira, A. da Rosa, V. Carvalho-Rocha, J. V. Bernardy, J. L. M. M. Sugai, C. E. dos Santos, R. P. Bastos, D. Llusia, and J. S. Ulloa, "A dataset for benchmarking Neotropical anuran calls identification in passive acoustic monitoring," *Scientific Data*, vol. 10, p. 771, Nov. 2023.

[17] B. Ghani, V. J. Kalkman, B. Planqué, W.-P. Vellinga, L. Gill, and D. Stowell, "Generalization in birdsong classification: Impact of transfer learning methods and dataset characteristics," Sept. 2024.

[18] D. Robinson, A. Robinson, and L. Akrapongpisak, "Transferable Models for Bioacoustics with Human Language Supervision," Aug. 2023.

[19] I. Moummad, R. Serizel, E. Benetos, and N. Farrugia, "Domain-Invariant Representation Learning of Bird Sounds," Sept. 2024.

[20] B. Williams, B. van Merriënboer, V. Dumoulin, J. Hamer, E. Triantafillou, A. B. Fleishman, M. McKown, J. E. Munger, A. N. Rice, A. Lillis, C. E. White, C. A. D. Hobbs, T. B. Razak, K. E. Jones, and T. Denton, "Leveraging tropical reef, bird and unrelated sounds for superior transfer learning in marine bioacoustics," May 2024.

[21] S. Romano, J. Bailey, V. Nguyen, and K. Verspoor, "Standardized Mutual Information for Clustering Comparisons: One Step Further in Adjustment for Chance," in *Proceedings of the 31st International Conference on Machine Learning*, pp. 1143–1151, PMLR, June 2014.

[22] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," Sept. 2020.

[23] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The Balanced Accuracy and Its Posterior Distribution," in *2010 20th International Conference on Pattern Recognition*, pp. 3121–3124, Aug. 2010.