Global birdsong embeddings enable superior transfer learning for bioacoustic classification

Burooj Ghani[1⋆], Tom Denton[2⋆], Stefan Kahl[3,4], Holger Klinck[3]

[1]Naturalis Biodiversity Center, [2]Google Research

[3]K. Lisa Yang Center for Conservation Bioacoustics, Cornell Lab of Ornithology, Cornell University

[4]Chemnitz University of Technology

November 20, 2023

Abstract

Automated bioacoustic analysis aids understanding and protection of both marine and terrestrial animals and their habitats across extensive spatiotemporal scales, and typically involves analyzing vast collections of acoustic data. With the advent of deep learning models, classification of important signals from these datasets has markedly improved. These models power critical data analyses for research and decision-making in biodiversity monitoring, animal behaviour studies, and natural resource management. However, deep learning models are often data-hungry and require a significant amount of labeled training data to perform well. While sufficient training data is available for certain taxonomic groups (e.g., common bird species), many classes (such as rare and endangered species, many non-bird taxa, and call-type) lack enough data to train a robust model from scratch. This study investigates the utility of feature embeddings extracted from audio classification models to identify bioacoustic classes other than the ones these models were originally trained on. We evaluate models on diverse datasets, including different bird calls and dialect types, bat calls, marine mammals calls, and amphibians calls. The embeddings extracted from the models trained on bird vocalization data consistently allowed higher quality classification than the embeddings trained on general audio datasets. The results of this study indicate that high-quality feature embeddings from large-scale acoustic bird classifiers can be harnessed for few-shot transfer learning, enabling the learning of new classes from a limited quantity of training data. Our findings reveal the potential for efficient analyses of novel bioacoustic tasks, even in scenarios where available training data is limited to a few samples.

Keywords: Deep learning, feature embeddings, bioacoustics, classification, few-shot learning, transfer learning, passive acoustic monitoring

1 Introduction

Bioacoustic analysis provides a rich window into biodiversity, animal behavior and ecosystem health. Passive acoustic monitoring (PAM) in particular has become a widely used tool for wildlife conservation. PAM uses battery-operated autonomous recording devices (ARUs) that collect vast amounts of acoustic data, containing a wealth of information about biological, geophysical, and anthropogenic activities in the deployment area. It allows researchers to study and protect animals and their habitats non-invasively at ecologically-relevant temporal and spatial scales [1]. PAM involves recording sound in nature and has been used to study a wide range of species, including whales and dolphins [2, 3], pinnipeds [4, 5], birds [6, 7], insects [8, 9], fish [10, 11], frogs [12, 13], and terrestrial mammals [14, 15]. In recent years, many automated deep learning-based analysis tools have been developed that are now commonly used to analyze long-term acoustic data efficiently [16]. By utilizing these tools, researchers can automatically detect and categorize animal vocalizations, saving them a significant amount of time and effort and facilitating the investigation of less researched species [17]. However, the development of these tools typically depends on the availability of well-annotated training data. Obtaining sufficient training data can be a major challenge. While there are sufficient amounts of training data available for some

[*] The two authors contributed equally to this paper and share first authorship.
(Email: burooj.ghani@naturalis.nl; tomdenton@google.com)

taxonomic groups, including common bird species (e.g., through community collections like Xeno-canto[†] or the Macaulay Library[‡]), training data is often lacking for rare and endangered species, which are often the prime target of conservation efforts [18]. In addition, traditional approaches to species-level classification may not be suitable for all applications. For example, a fixed set of classes may not be desirable in cases where researchers are interested in the fine-grained classification of vocalizations, such as identifying specific call types rather than simply identifying the presence or absence of a species [19]. Call types and the associated behaviors (e.g., foraging or breeding) can provide critically important cues on habitat use and inform, for example, land management decisions.

One way to address the challenge of data deficiencies is to utilize learned feature embeddings for *few-shot transfer learning*. In the context of machine learning, *feature embeddings* are vectors obtained from some intermediate layer of a trained machine learning model [16]. Since 2014, numerous studies have found that embeddings from pre-trained models can allow more efficient learning of novel tasks [20, 21, 22]. *Few-shot learning* refers to a wide array of methods attempting to produce strong models with little training data, inspired by the apparent human ability to learn new classes from a handful of examples [23]. Feature embeddings often figure prominently in few-shot learning designs, allowing *transfer* of learned features to new tasks and domains.

High-quality feature embeddings offer several benefits over traditional approaches to species-level classification. First, feature embeddings can help to differentiate between classes of acoustic events that are very similar and differ only in subtle details. For instance, songbirds can display local variations (also called dialects) in their song patterns, which may lead to slight differences in note sequences [24]. Feature embeddings can capture these nuances and enable more precise classification. Additionally, embeddings facilitate transfer learning between species, enabling researchers to train models on data from more commonly occurring or extensively studied species and then apply that knowledge to a target species, which may have insufficient training data. This approach also saves researchers time and effort that would otherwise be needed to train a dedicated classifier from scratch while enhancing the accuracy of classification results. Furthermore, cross-taxa classification based on feature embeddings is also possible when such embeddings can generalize across acoustic domains and events.

Because the relevant features for different problems may vary, we hypothesize that models trained on a problem closely related to the target problem will often outperform models trained on very different problems. In fact, the recent HEAR Benchmark competition found that no single model dominated across event detection, music transcription, and speech recognition tasks [25]. However, as mentioned earlier, many problems lack sufficient data for training a robust classifier from scratch. In these cases, re-using the feature embeddings from a pre-trained model allows learning the new task efficiently, so long as the embeddings are sufficiently relevant.

In this study, we investigate the use of various large-scale acoustic classifiers to produce feature embeddings that can be used to perform fine-grained classification of bird calls and dialect types, and out-of-scope but related identification of acoustic events (non-bird animal calls) that these models have not been trained on. Furthermore, we include in our analysis classifiers that are either trained on AudioSet dataset [26] (a broad spectrum of audio data extracted from YouTube clips) or on extensive datasets of bird vocalisations from around the world. In doing so, we are able to compare the effectiveness of these embeddings derived from different classifiers, evaluating their capacity to generalize and detect a variety of bioacoustic events.

The paper aims to provide a simple method for species-agnostic classification across taxonomic groups by leveraging transfer learning capabilities of selected classifiers. The effectiveness of the approach is demonstrated by evaluating on a diverse set of data sources covering birds, bats, marine mammals, and amphibians. Overall, our study suggests that the proposed approach can help to advance automated analysis in passive acoustic monitoring by solving the problem of species and call type recognition in low- and medium-data regimes. The use of transfer learning capabilities of selected classifiers provides a practical and effective way to classify a wide range of acoustic events across different taxa and can help to improve the accuracy and efficiency of PAM analysis efforts.

Our approach – utilizing fixed, pre-trained embeddings for novel problems – also suggests a more efficient workflow for large-scale bioacoustic data sets. Large PAM deployments may accumulate tens to hundreds of terabytes of data during a single field season [27]. This makes model inference tasks especially time-consuming and potentially expensive. Given a model which produces generally useful feature embeddings, the practitioner may embed their entire data set once and then use the pre-computed embeddings for a wide range of subsequent analysis tasks.

[†] https://xeno-canto.org

[‡] https://www.macaulaylibrary.org

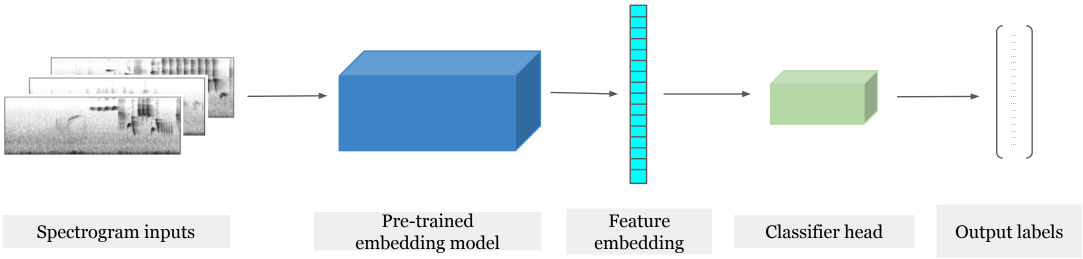

Figure 1: Workflow illustrating the process of downstream classification for various bioacoustic tasks. Spectrograms are processed by pre-trained models, producing embeddings. These embeddings are then passed to a classifier head—a fully-connected feed forward neural network—which is trained for final classification.

Training and inference with small models over fixed embeddings are much faster than training entirely new models: Training a high-quality classifier from scratch can take many days of GPU time, but training small linear classifiers over fixed embeddings, which we discuss in this paper, can take less than a minute to train on a modern workstation. This allows fast experimentation with different analysis techniques and quickly iterating with human-in-the-loop active learning techniques.

1.1 Relationship to Previous Work

Previous works have investigated transfer learning for novel bioacoustic tasks, including the use of pre-trained global bird embeddings ([28], [29, 30]). Many of these prior works focus on a single use-case, leaving the question of breadth of generalization unanswered.

Other studies on bioacoustic transfer learning use only embeddings from VGGish [31], a general audio model. This is an older model, and even recent bioacoustic studies ([32, 33, 34, 35]) continue to use it, despite long being surpassed on general audio benchmarks. In this work, we provide comparisons of bioacoustic transfer learning from VGGish to more recent SOTA models such as AudioMAE (a self-supervised transformer), YAMNet, and PSLA (more recent convolutional models).

Our study provides a far more complete comparison of global bird embeddings against general audio models, across a diverse array of datasets, demonstrating robust generalization.

2 Methods

In this work, we focus on the extraction of feature embeddings from six CNN models and one transformer model, described below in Section 3.1. These *embedding models* are trained on either general YouTube data or global data sets of bird vocalizations. All models map spectrograms (visual representation of sound) to their class labels. Using logistic regression, we train classifiers on the feature embeddings extracted from each model as described in Section 2.1. Fig. 1 provides an overview of the classification pipeline we employed for our experiments.

Spectrograms serve as the input data for our framework. The pre-trained embedding model, which is essentially the large-scale classifier without the classifier head, processes the spectrograms and produces an embedding. The embedding can be seen as a compact representation capturing the salient features of the input. This embedding is then forwarded to the classifier head, which is implemented as a fully connected layer. The classifier head applies a linear transformation to the embedding, followed by a sigmoid function to obtain class probabilities, and is trained via standard logistic regression. In summary, this architecture, comprising the embedding model, fully connected layer, and sigmoid activation, enables the extraction of relevant features from spectrograms and the subsequent generation of probability estimates for downstream classification purposes.

By employing simple logistic regression, we are able to judge the direct utility of each model's pre-trained embedding to a range of problems. Additionally, we save an immense amount of training effort by pre-computing the embeddings for each dataset.

2.1 Linear Probes on Feature Embeddings

In this work, we consider the case of keeping the entire pre-trained embedding frozen and learning a single linear layer for the new tasks. This method is essentially a *linear probe* of the selected embeddings, which allows assessment of the availability of desired task-specific information in the embeddings [36]. If the downstream task can be performed well with just a linear layer on top of these embeddings, it suggests that the embeddings already have the necessary information encoded in them.

For each pairing of model and data set, we first calculate the model embeddings for the full data set. Each model has a native sample rate and window size, chosen independently of any of the datasets under consideration. Each audio sample is resampled to the model's native sample rate (though we experiment with alternatives in 4.4). When an example is shorter than the model's window size, we apply centered zero-padding to obtain the target length. When a model's window size is shorter than a target example, we frame the audio according to the model's window size, create an embedding for each frame, and then average the results. In the end, each example is associated with a single embedding vector. We then randomly choose a fixed number k of examples from each class, using a seeded random shuffle to ensure that the same training examples are used for every model. The k examples are used to train a linear classifier over the pre-computed embeddings, and all remaining examples are used for evaluating the trained classifier. We use a binary cross entropy (BCE) loss, with sigmoid activation, and train the classifier to convergence. This process is repeated five times with different random seeds for each combination of model, dataset, and k, using the same set of five random seeds for each combination. We do this to report a reliable estimate of the classification performance [37]. By default, we choose $k = 32$.

For evaluation we compute (1) macro-averaged ROC-AUC (computing ROC-AUC for each class, and then averaging over all classes) and (2) Top-1 Accuracy. Reported metrics are averaged over the five training runs.

2.2 Experiment: Few-Shot Learning

By varying the amount of training data, we can further understand the resilience of each embedding to reduced data availability. As it turns out, linear probes of pretrained embeddings is a surprisingly successful strategy for few-shot learning, as described in [38], which found this baseline method to be competitive with far more complicated approaches such as meta-learning. Other subsequent works have confirmed that 'frustratingly simple' approaches with strong embeddings are sufficient for many few-shot learning tasks [39] [40] [41]. We leave comparison of more involved few-shot techniques for the future work.

We vary k as powers-of-two between 4 and either 32 or 256, depending on the size of the target dataset. In the results, we provide results for all tested values of k in Figure 2 and specifically for $k = 32$ in Table 3.

2.3 Experiment: Visualizing Embedding Spaces

We can also observe the geometry of the embedding space using a t-SNE transformation of the model embeddings [42]. The t-SNE transformation attempts to preserve distances in the embedding space while projecting to two dimensions. In Fig. 3 we plot t-SNE transforms for YAMNet, AudioMAE, BirdNET and Perch. Note that t-SNE plots can be tricky to interpret appropriately [43], though points which are close in the original space tend to be close after applying the t-SNE transform. We choose t-SNE over other alternatives specifically because it uses a shallow model, and we wish to preserve as closely as possible the structure of the embedding spaces.

2.4 Experiment: Deep-Dive into AudioMAE

In recent years, transformer and self-supervised models have taken a dominant position in machine learning research. Therefore, it may be surprising that AudioMAE - a self-supervised transformer - under-performed the humble EfficientNet-B1 architecture. We therefore performed a number of additional experiments to discover whether additional tweaking of the experimental setup would uncover hidden performance gains for the AudioMAE embeddings. We applied three different treatments on all six datasets, attempting to find a higher-performing operating point for the AudioMAE.

First, we compared quality of the pre-trained unsupervised embedding and the embedding obtained from supervised fine-tuning on AudioSet. Because the unsupervised objective is spectrogram reconstruction, one would expect that all relevant information should be present in the pre-trained embedding, but possibly suppressed by fine-tuning on the irrelevant AudioSet label-space. In fact, using the pre-trained or fine-tuned embedding does change the metrics, but not in a predictable way.

Second, we experimented with ignoring the audio sample rate when loading the target audio. Because the AudioMAE consumes 16kHz audio, any significant features above the Nyquist frequency of 8kHz will be lost when audio is resampled to the model's input rate. Instead of resampling, we may instead load the audio at its native sample rate and feed it directly to the model as though it were 16kHz.

Finally, we tried using a two-layer network with the pre-trained model, under the hypothesis that the raw self-supervised embedding may not be well aligned for classification tasks. The two-layer network consists of batch-normalization, a hidden layer with 2048 units (double the embedding dimensionality), a ReLU activation, and an output layer.

2.5 Experiment: Varying Embedding Size

We can view feature embeddings as a lossy compression of the input data. For instance, in terms of raw data, the embedding produced by the Perch model contains only 1.6% of the data of the raw audio (a 1280-dimensional 32-bit float vector, derived from 5 seconds of 32 kHz audio encoded as 16-bit integers). Yet these embeddings enable efficient recognition of a wide range of global bird species. For this to work well, the classifier must learn features relevant to the classification problem while allowing irrelevant data to be discarded. This compression viewpoint suggests that embedding size may play a role in overall embedding quality.

We ran an additional ablation on embedding size while investigating the difference between BirdNET 2.2 and Perch models, which had embedding sizes 320 and 1280, respectively. We investigate the role of embedding size by comparing two versions of BirdNET with different embedding sizes (320 and 1024) and six versions of Perch (160, 320, 640, 960, 1280, 2560), using the same methodology and datasets described above.

Since we do not have ready access to the training infrastucture for the general audio models, we exclude them from this ablation.

2.6 Experiment: No-Pretraining Baseline

It is reasonable to ask how a custom-trained model would perform on the novel tasks we consider in this study. However, it is difficult to train a model from scratch with reasonable performance on the amount of data provided for the tasks. To provide this baseline, we extract MFCC features from the target audio and train a small two-layer neural network with 2048 hidden features on each dataset, mirroring a baseline considered in the BEANS benchmark [33].

2.7 Experiment: Comparison to a Regional Model

One might also ask how transfer of the global bird embeddings compares to transfer of embeddings from a model trained on a smaller collection of bioacoustic classes. For this, we compare the transfer performance of a regional model trained for identification of 89 species in the Sierra Nevadas.

3 Models and Datasets

3.1 Model descriptions

We compare three models trained on bird data (BirdNET, Perch, and Sierras) to four models trained on variants of AudioSet [44] for general audio event detection (AudioMAE, PSLA, YAMNet, VGGish).

BirdNET and Perch are similar models, differing mostly in their training data. While Perch is trained exclusively on bird sounds data, BirdNET's training dataset also comprises of a relatively small fraction of non-birds sound data. AudioSet comprises an extensive compilation of over 2 million audio clips, each 10 seconds in duration. These clips are derived from YouTube videos and are categorically labeled according to the type of sound they contain, with a total of 527 unique classes. The classes include 'wild animals,' but the associated labels are very coarse (bird, frog, roaring cat) and constitute only about 2% of the total dataset. To elaborate further, the specifications of the models are detailed as follows:

Perch[*] is an EfficientNet B1 [45] trained on the full corpus of bird song recordings from Xeno-

[*] https://tfhub.dev/google/bird-vocalization-classifier/4

	Architecture	Training Data	Window (s)	Embedding Size	CPU(ms/s)
Google Perch	EfficientNet B1	XenoCanto	5.0	1280	24.3
BirdNET 2.2 / 2.3	EfficientNet B1	XC+ML+Custom	3.0	320 / 1024	10.0 / 11.1
Sierras Birds	EfficientNet B0	XenoCanto+Custom	5.0	1280	12.2
AudioMAE	MAE (Large)	AudioSet	10.0	1024	78.2
PSLA	EfficientNet B2	AudioSet	10.0	1408	246.0*
YAMNet	MobileNet v1	AudioSet	0.96	1024	7.7
VGGish	Modified VGG	YouTube 8M	0.96	128	2.8
BEANS Baseline	MFCCs + MLP	N/A	N/A	160	N/A

Table 1: Summary of Embedding Model Characteristics. CPU(ms/s) is the benchmarked run-time for evaluating one audio window with the model, divided by model's window size. Models were benchmarked on a 4.3GHz AMD CPU with 12 cores. We expect that PSLA inference can be optimized to under 50ms.

Canto (XC) downloaded in July, 2022. Because XC is weakly labeled (a single label for an entire file), we use an activity detector to select training windows from each file, as described in [46]. During training we augment with MixUp [47], random gain adjustment, and random time-shifting of up to one second. The model is trained to classify all levels of the taxonomy for each recording simultaneously (species, genus, family, order). The base Google Perch model and further evaluation statistics are available at TFHub and supporting code is available on GitHub[†].

BirdNET [48] also uses an EfficientNet architecture, but does not use taxonomic outputs. BirdNET has a broader training set, including XC, the Macaulay Library, and labeled soundscape data from around the world, ultimately targeting many thousands of bird species. Additionally, BirdNET is trained to identify human speech, dogs, and many species of frogs. To enable a range of downstream use-cases, BirdNET trades off some accuracy for efficient computation. We report on BirdNET 2.2 and 2.3. Version 2.3 features a higher embedding dimension (see Section 4.5) and is trained on a larger bird species list. The BirdNET code is available on GitHub[‡], and includes support for training small classifiers on embeddings.

Sierra Birds [46] is a 'regional' bird classification model, trained on 89 bird species found in the California Sierra Nevada mountains. The model uses an EfficientNet B0 architecture, marginally smaller than the B1 architecture used by the BirdNET and Perch models, and was primarily trained on XC recordings. The model training data was augmented with additional non-bird noise data, as described in [46].

AudioMAE [49] is a more recent general audio model built with a transformer architecture. The model is trained on AudioSet with a self-supervision task, reconstructing masked spectrograms. The model consists of an encoder (which produces embeddings of patches of the spectrogram) and a decoder (which reconstructs the spectrogram from the patch embeddings). For this study, we use the embeddings produced by the encoder and discard the decoder. A 1024-dimensional embedding is obtained by averaging the per-patch embeddings, as is typical when using AudioMAE for classification tasks. We evaluated a re-implementation of AudioMAE, using the 'Large' model with 300M parameters, provided by Eduardo Fonseca [50]. This model obtains a mAP of 46.4 on AudioSet-2M after fine-tuning, comparable to the original AudioMAE's reported mAP of 47.3. We experimented with many configurations of AudioMAE, as described in Section 4.4. AudioMAE training consists of a pre-training stage, where it is trained only for reconstruction of masked spectrograms, and a fine-tuning stage, where it is trained for supervised classification. None of these methods was consistently better than all others, so for brevity, we report results for the fine-tuned model with averaged embeddings unless otherwise noted. The original AudioMAE code can be accessed on GitHub[§].

PSLA [51] is an EfficientNet model trained on AudioSet, which we include for comparison with the EfficientNet-based global bird models. The model is competitive with SOTA models, obtaining a mAP score of 47.4 on Audioset-2M. Unlike the bird models, it uses an attention layer over the final embeddings to produce its predictions. We replace this attention layer with an averaging of the final embeddings to produce a summarized embedding for an audio segment, suitable for transfer learning.

YAMNet and **VGGish** are both convolutional models trained to predict AudioSet classes. YAMNet uses a MobileNetV1 architecture [52]. VGGish is an older audio event-detection model,

[†] https://github.com/google-research/perch

[‡] https://github.com/kahst/BirdNET-Analyzer

[§] https://github.com/facebookresearch/AudioMAE using a variant of the VGG architecture and trained on an earlier version of AudioSet [31]. Both of these models process audio frames of 0.96 seconds. While the YAMNet model generates a feature embedding vector of 1024 dimensions, the VGGish embedding size is limited to 128 dimensions. The YAMNet[¶] and VGGish[‖] codes can be accessed on GitHub.

BEANS Baseline is not a pretrained model. Following the method in the BEANS Benchmark [33], we extract a 160-channel MFCC representation of the input audio. This frequency information is then mean-pooled for the audio example, and a two-layer fully connected neural network is trained to predict the target class. This provides a simple baseline without pre-training for each task.

3.2 Evaluation datasets

We use a range of datasets for our analysis. These datasets were constructed by different groups with different goals and methodologies, and therefore vary in their characteristics. For instance, the RFCX and Watkins datasets contain cross-class contamination — examples of a specific class where another unlabeled class is present. The bat species and Watkins datasets have variable clip length, whereas the other datasets have a fixed clip length. Table 2 presents an overview of all the datasets used in this work.

	# Classes	Mean Class Size	Smallest Class	Sample Rate	Clip Length (s)
Godwit Calls	5	1343	628	44.1kHz	3.0
Yellowhammer Dialects	2	772	444	48kHz	3.5
Bats	4	887	360	44.1kHz*	1.0-13.0
Watkins Marine Mammals	32	60	35	22.05kHz	0.1-10.0
RFCX Frog Species	12	50	37	48kHz	5.0
RFCX Bird Species	13	53	34	48kHz	5.0

Table 2: Summary of target dataset characteristics. For the Bats dataset, frequency shifting was applied to move signal into the audible range.

Godwit Calls (GC): The GC dataset contains five different calls of Black-tailed Godwit. The recordings were made by Ondrej Belfin as part of his masters thesis at the University of Groningen in the Netherlands [53]. All recordings are 3 seconds long and are annotated by Ondrej Belfin himself.

Yellowhammer Dialects (YD): The YD dataset comprises two dialects of Yellowhammer songs, denoted as X and B [54], derived from audio recordings of Yellowhammer vocalizations. The two dialects are characterized based on variations of elements in the terminal phrase of the song. These recordings were sourced from submissions made through the BirdNET App, captured with various mobile phone microphones. Recordings were annotated in a two-step process. Connor Wood performed preliminary annotations, which were later refined by Pavel Pipek, a specialist in yellowhammer dialects at the Department of Ecology, Charles University in Prague. All recordings were acquired in 2020. Each audio recording within the data set has a duration of three seconds, facilitating a comprehensive analysis of the yellowhammer vocalizations. These dialects have a duration in the range 2.2-2.7 seconds and a fundamental frequency in the range of 5-6 kHz.

Bats (BT): The BT dataset contains four species of North American bats. The eastern red bat (*Lasiurus borealis*, LABO) with 1,124 recordings, the little brown bat (*Myotis lucifugus*, MYLU) with 1,119 recordings, the northern long-eared bat (*Myotis septentrionalis*, MYSE) with 360 recordings, and the tricolored bat (*Perimyotis subflavus*, PESU) with 948 recordings. The dataset is sourced from two origins: 1) Training dataset for NABat Machine Learning V1.0 [55], and 2) Dr. Patrick Wolff, US Army ERDC-CERL. The datasets were collected at ultrasonic sampling rates. We applied pitch shifting via sample rate conversion to these datasets, to bring the bat vocalizations into the audible range. After this pre-processing step, all audio has a sampling rate of 44.1 kHz.

Watkins Marine Mammal Sounds Database (WMMSD): The WMMSD dataset covers 60 species of marine mammals but we employ the 'best of' category enlisted in the database as the species with higher quality and lower noise recordings. The taxonomical representation encompasses species from the Odontocete and Mysticete suborders within the order Cetacea, in

[¶] https://github.com/tensorflow/models/tree/master/research/audioset/yamnet

[‖] https://github.com/tensorflow/models/tree/master/research/audioset/vggish addition to the Phocid and Otariid families, which are part of the clade Pinnipedia. The auditory documentation, spanning a substantial time period of seven decades, encapsulates a diverse range of recording methodologies, ambient acoustical conditions, and sampling frequencies [56]. The compilation of this auditory data was accomplished and annotated by several researchers including William Watkins, William Schevill, G. C. Ray, D. Wartzok, D. and M. Caldwell, K. Norris, and T. Poulte, and is openly accessible for academic use [57, 58]. The audio examples are cropped to the length of the actual vocalization, which means that the lengths of the audio files vary greatly by species. We exclude five classes for which there are a fewer than 32 examples provided, and two additional species which are characterized by very low frequency vocalizations (fin whale and northern right whale).

Rainforest Connection Kaggle dataset (RFCX Frogs, RFCX Birds): This is the training data from the 2021 Species Audio Detection challenge, consisting of recordings of Puerto Rican birds and frogs. Both birds and frogs are present in the class list; to understand model performance on these taxa, we present results on each taxa separately, and all together.

The bird species in the RFCX data appear in the training data for both the Perch and BirdNET models, but most of these species have very limited training data. As of this writing, the median number of Xeno-Canto recordings for these thirteen species is just 17, and only two species have more than 50 recordings (the Bananaquit with 579 recordings, and the Black-Whiskered Vireo with 68 recordings). Thus, these are largely low-data species for these models.

3.2.1 Dataset Limitations

Each of the datasets we work with presents distinct difficulties.

First, our methodology does not create an ideal train/test split when multiple examples originate from the same original recording. Ideally, different source recordings or entire recording sites would appear consistently as train or test data to reflect model generalization to new conditions. We do not have sufficient metadata available for all datasets to perform such a split, and so results may overestimate model generalization on the target tasks. Instead, we treat each example independently, and create a train/test split over the examples we have. We believe this issue affects only the Bats, RFCX, and a subset of the Watkins species.

Secondly, some recordings contain additional unlabeled vocalizations, which may lead to underestimation of model quality. This is especially the case for the Watkins and RFCX frog datasets. (See Table 6 for some analysis of the Watkins dataset.)

4 Results

4.1 Linear Probes on Feature Embeddings

Our study delves into the classification performance by employing a variety of embeddings with linear probes for novel bioacoustic tasks. For an in-depth look, refer to Table 3, which elucidates the results when training linear probes with $k = 32$ examples per class. Additionally, Figure 2 provides a visual representation of the results across diverse training data sizes, ranging from 4 to either 32 or 256 examples per class, contingent on the dataset size.

Model	GC		YD		BT		WMMSD		RFCX Frogs		RFCX Birds	
	Top-1	AUC	Top-1	AUC	Top-1	AUC	Top-1	AUC	Top-1	AUC	Top-1	AUC
Perch	**0.92**	**0.99**	**0.87**	*0.91*	**0.86**	**0.97**	**0.83**	**0.98**	**0.74**	*0.96*	**0.83**	**0.97**
BirdNET 2.3	0.91	0.99	0.84	0.91	0.85	0.96	0.81	0.98	0.73	0.95	0.78	0.96
Sierras	0.76	0.93	0.56	0.57	0.77	0.93	0.72	0.96	0.65	0.92	0.69	0.94
AudioMAE	0.85	0.96	0.61	0.66	0.63	0.85	0.74	0.96	0.56	0.89	0.43	0.85
PSLA	0.20	0.80	0.65	0.51	0.27	0.57	0.06	0.76	0.10	0.59	0.06	0.60
YamNet	0.71	0.91	0.54	0.55	0.61	0.83	0.69	0.96	0.48	0.86	0.43	0.84
VGGish	0.63	0.86	0.51	0.51	0.57	0.80	0.04	0.56	0.48	0.85	0.39	0.81
BEANS Baseline	0.14	0.53	0.72	0.51	0.24	0.53	0.04	0.56	0.11	0.52	0.05	0.58

Table 3: Table of Results. We report the top-1 accuracy and ROC-AUC score of the linear classifiers, averaged over five runs, for each data set. All results are for 32 training examples per species. Entries are bold-faced if the model scored highest on all five runs, and italic if highest on four of five runs.

The Perch and BirdNET 2.3 models obtain similar performance. However, Perch achieved the highest Top-1 accuracy and AUC across all the datasets, making it the most consistent performer. It performed particularly well with “Godwit Calls” and “Bat Species”, with AUCs of 0.99 and 0.97, respectively. Similarly, BirdNET 2.3 exhibited a good performance, especially with Godwit Calls (GC) and Bat species (BT) (0.99, 0.96).

Both bird models significantly outperform the AudioSet models on all tasks (VGGish, YAMNet, PSLA, and AudioMAE). The macro-averaged ROC-AUC scores are typically high, suggesting good binary classification on each class individually. In case of AudioMAE, the performance dropped significantly, especially noticeable with the Yellowhammer dialects (YD) and RFCX birds datasets, which had lower AUCs of 0.66 and 0.78, respectively. The performance declined further using the YamNet model. The Top-1 accuracy was relatively low across datasets, and AUCs were significantly lower, particularly for the YD dataset. The VGGish model had the lowest performance across all datasets, notably underperforming on the WMMSD dataset with a very low Top-1 accuracy of 0.04 and AUC of 0.52. In summary, the Perch and BirdNET 2.3 models outperformed the others in terms of both Top-1 accuracy and AUC, demonstrating superior generalizability across various bioacoustic datasets. On the other hand, VGGish showed the weakest performance. Among the three models trained on the AudioSet, the transformer-based AudioMAE model outperformed the CNN-based VGGish, YAMNet, and PSLA models across all datasets except for the RFCX Birds dataset in which YAMNet performed slightly better. The performance gain for YD and GC datasets was significant.

4.2 Experiment: Few-shot learning

In Fig. 2 we show results with varying amounts of training data per class. We again find that using transfer learning with global bird models (BirdNET and Perch) consistently outperforms general event-detection models trained on YouTube data (AudioMAE, Yamnet, and VGGish).

In all cases, the bird models have an ROC-AUC significantly greater than 0.5 even with only 4 training examples.

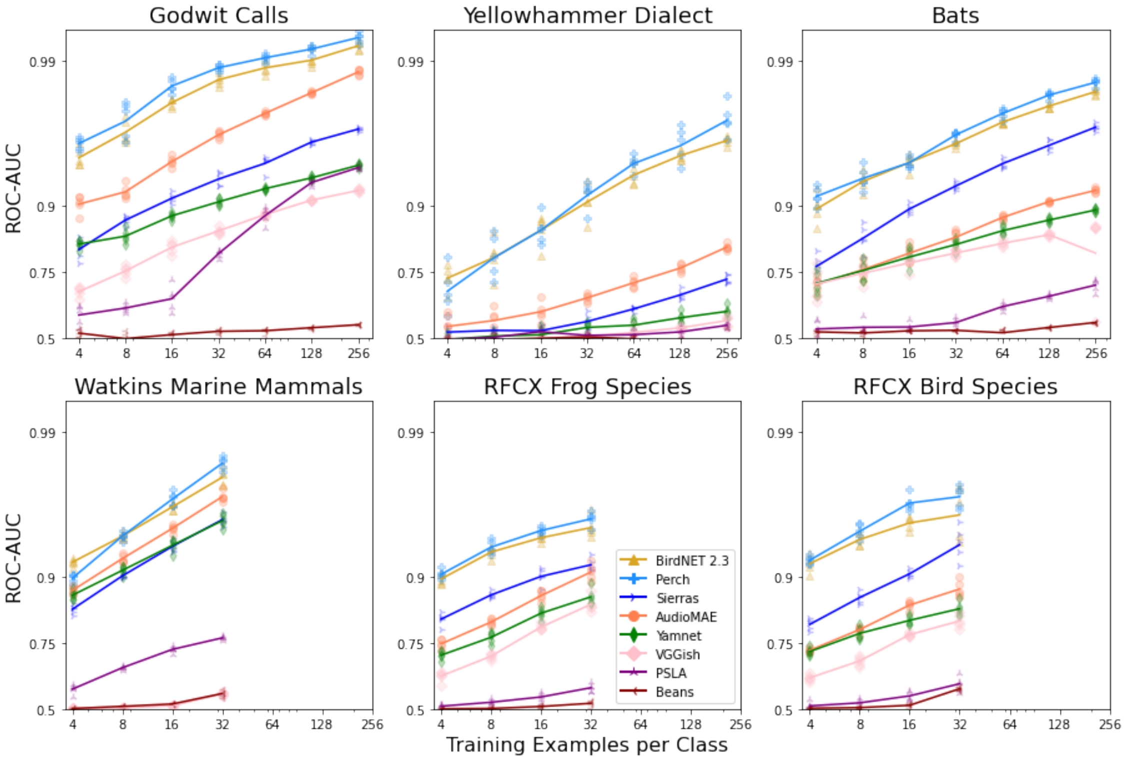

Figure 2: Results of Few-Shot Transfer Learning Tasks. ROC-AUC scores are plotted with log-odds scaling. A point is plotted for each experiment, and the curve connects the average quality for experiments at each number of training examples per class.

4.3 Visualizing Embedding Spaces

In the easier Godwit problem (Fig. 3), we observe cleaner clustering of labeled data in the Perch embeddings, with large margins suggesting easy linear separability of the classes. By contrast, there are no clean margins between classes in the YAMNet embeddings, and smaller, noisier margins for the AudioMAE embeddings.

For the more difficult Yellowhammer problem, we observe a complete intermixing of the two classes for YAMNet, explaining the model's inability to linearly separate the classes. For AudioMAE, which performs marginally better, we can observe a couple pockets of concentrated blue points, but no clear clustering. For Perch and BirdNET we see some clustering, but still a great deal of inter-mixed data, explaining the slow improvement on this task.

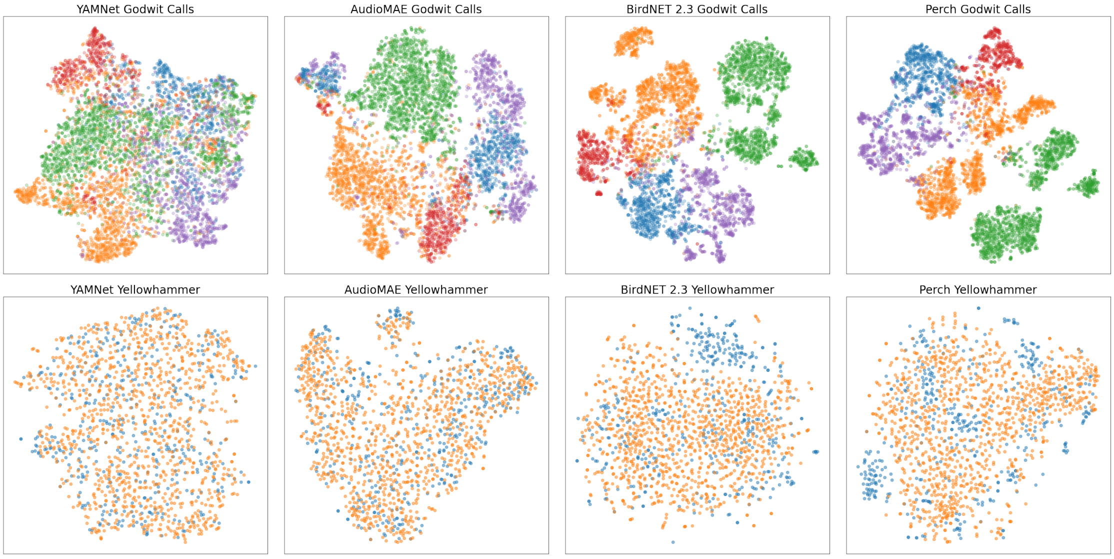

Figure 3: t-SNE plots of Godwit and Yellowhammer embeddings. Points are colored by class.

4.4 Experiment: Deep-Dive into AudioMAE

The best overall AudioMAE performance was obtained by using a 2-layer perceptron and no audio resampling with the pre-trained embeddings. Despite substantial effort, we found the global bird embeddings - with no additional tweaking - uniformly outperformed the AudioMAE model. PSLA benefitted significantly from a two-layer probe, but still underperformed both AudioMAE and the global bird classifiers.

4.5 Experiment: Varying Embedding Size

An ablation over the embedding dimension is summarized in Fig. 4. The Top-1 Accuracy and ROC-AUC scores on different datasets using various embedding sizes are shown in Table 5. Perch with a 320-dimensional embedding (matching BirdNET 2.2) has significantly degraded quality in all tasks. Doubling the base Perch embedding dimension to 2560 yields a further increase in model performance for some downstream tasks.

Increasing the size of the BirdNET embedding to 1024 led to similar performance as the Perch model in most downstream tasks.

4.6 Experiment: No-Pretraining Baseline and Regional Model

The results for the BEANS MFCC baseline and Sierra Birds model are included in Table 3 and Figure 2. We find that the MFCC model barely exceeds chance, significantly underperforming transfer from VGGish, the weakest of the audio domain models.

We find that the global bird models significantly outperform the 'regional' Sierra Birds model. Furthermore, the regional model underperforms general audio embeddings on some datasets.

Model	Probe	RS?	GC		YD		BT		WMMSD		RFCX-F		RFCX-B	
			Top-1	AUC	Top-1	AUC	Top-1	AUC	Top-1	AUC	Top-1	AUC	Top-1	AUC
Perch	LR	Y	0.92	0.99	**0.87**	0.91	0.86	0.97	**0.83**	**0.98**	0.74	0.96	0.83	0.97
Perch	LR	N	0.91	0.99	0.86	**0.93**	0.80	0.94	0.80	0.98	**0.76**	0.96	0.83	0.98
Perch	2LP	Y	**0.92**	0.99	0.87	0.92	**0.86**	**0.97**	0.81	0.98	0.73	0.96	**0.83**	0.97
Perch	2LP	N	0.92	**0.99**	0.84	0.92	0.81	0.95	0.79	0.96	0.76	**0.96**	0.83	**0.98**
BN2.3	LR	Y	0.91	0.99	0.84	0.91	0.84	0.96	0.81	0.98	0.73	0.95	0.78	0.96
MAE/p	LR	Y	0.72	0.91	0.57	0.60	0.68	0.88	0.60	0.93	0.59	0.91	0.53	0.90
MAE/p	LR	N	0.80	0.95	0.59	0.62	0.65	0.87	0.68	0.95	0.64	0.93	0.65	0.94
MAE/p	2LP	Y	0.82	0.96	*0.66*	0.64	*0.74*	*0.91*	0.78	0.97	0.69	0.94	0.64	0.94
MAE/p	2LP	N	0.84	0.97	0.63	0.65	0.73	0.91	*0.81*	*0.97*	*0.70*	*0.94*	*0.72*	*0.96*
MAE/f	LR	Y	0.85	0.96	0.61	0.66	0.63	0.85	0.74	0.96	0.56	0.89	0.43	0.85
MAE/f	LR	N	0.86	0.97	0.62	0.66	0.61	0.84	0.76	0.97	0.62	0.91	0.55	0.88
MAE/f	2LP	Y	0.85	0.97	0.63	0.67	0.65	0.86	0.79	0.97	0.60	0.89	0.49	0.86
MAE/f	2LP	N	*0.87*	*0.97*	0.64	*0.67*	0.64	0.86	0.80	0.97	0.60	0.89	0.62	0.90
PSLA	LR	N	0.20	0.80	0.65	0.51	0.27	0.57	0.06	0.76	0.10	0.59	0.06	0.60
PSLA	2LP	N	0.82	0.96	0.46	0.58	0.46	0.76	0.63	0.94	0.35	0.77	0.40	0.81

Table 4: Results of AudioMAE Steel-man Experiments. Results are all for 32 examples per class. Probe is LR for linear regression or 2LP for two-layer perceptron. 'RS' indicates whether the audio was resampled to the embedding model's preferred sample rate. 'MAE/p' is the pretrained unsupervised embedding model, and 'MAE/f' is fine-tuned with supervision on AudioSet. The highest score in each column is bold-faced, and the highest AudioMAE score is in italic.

Model	Size	Godwit Calls		Yellowhammer		Bat Species		Watkins		RFCX Frogs		RFCX Birds	
		Top-1	AUC	Top-1	AUC	Top-1	AUC	Top-1	AUC	Top-1	AUC	Top-1	AUC
Perch	2560	0.91	0.99	0.92	0.96	0.85	0.96	0.83	0.98	0.75	0.96	0.82	0.97
Perch	1280	0.91	0.99	0.88	0.93	0.85	0.96	0.80	0.98	0.73	0.95	0.83	0.97
Perch	960	0.91	0.98	0.88	0.93	0.85	0.96	0.80	0.98	0.74	0.95	0.82	0.97
Perch	640	0.90	0.98	0.87	0.92	0.84	0.96	0.74	0.97	0.73	0.95	0.81	0.97
Perch	320	0.89	0.98	0.80	0.87	0.80	0.94	0.71	0.96	0.71	0.94	0.81	0.97
Perch	160	0.88	0.97	0.80	0.84	0.79	0.93	0.66	0.95	0.68	0.93	0.78	0.96
BirdNET 2.3	1024	0.91	0.99	0.84	0.91	0.85	0.96	0.81	0.98	0.73	0.95	0.78	0.96
BirdNET 2.2	320	0.90	0.98	0.83	0.88	0.83	0.95	0.79	0.98	0.75	0.96	0.79	0.96

Table 5: Results of Embedding Size Ablation. We report the top-1 accuracy and ROC-AUC score of the linear classifiers, averaged over five runs, for each data set. All results are for 32 training examples per species. All Perch models were trained for this ablation from scratch.

5 Discussion

The performance results displayed in Fig. 2 underscore the value of transfer learning with global bird models such as BirdNET and Perch. These models consistently outperformed general event-detection models trained on broader auditory data, such as YouTube-sourced data utilized by AudioMAE, PSLA, YAMNet, and VGGish. This observation is pivotal as it suggests that models specifically trained on bird data possess a heightened capacity for generalization, successfully identifying and analyzing previously not encountered bioacoustic patterns.

In our study, we employed a linear probe method to evaluate the pre-trained embeddings of various models, including BirdNET and Perch. This method, essentially a linear layer placed on top of the selected embeddings, serves as a tool for assessing the availability of desired task-specific information within these embeddings. Consequently, our findings, as demonstrated in Fig. 2 and Table 3, suggest that the embeddings derived from global bird models contain rich, task-specific information conducive to bioacoustic analysis. This is in contrast to general auditory event-detection models, which although trained on a broader spectrum of auditory data, do not exhibit the same level of performance. This superior generalization capability of deep embeddings from global bird models is an important finding, as it highlights the potential of these specialized models in providing a more robust and adaptable framework for varied bioacoustic tasks by learning good quality embeddings from data. In the realm of bioacoustic sound event detection, the ability to generalize across distinct taxonomic categories and acoustic characteristics is invaluable, as it facilitates the fine-grained classification of call types, song dialects, and out-of-scope identification of acoustic events.

This finding might be attributed to the inherent diversity and complexity found in bird vocalizations. Bird songs and calls occupy a broad range both temporally and in the spectral domain, exhibiting diverse frequency modulations, harmonic structures, and rhythmic patterns. This wide array of acoustic characteristics provides a rich and versatile training data set for models such as BirdNET and Perch. The comprehensive nature of these vocalizations may have facilitated the models' ability to learn more generalized representations of bioacoustic patterns.

This versatility in bird vocalizations has a dual implication. Firstly, it enriches the training dataset, providing varied instances for the model to learn from, and subsequently, it enables the model to capture a broader range of acoustic patterns, improving its ability to generalize to novel categories. Secondly, the acoustic diversity among bird species might mimic the bioacoustic variability encountered in other taxa, thus further enhancing the model's generalization capabilities when applied to sounds from different taxa. Indeed, previous work has found a collection of mechanisms, termed MEAD, by which all birds and mammals vocalize [59], and which may help explain the successful transfer of bird features to other taxa. This hypothesis provides an intriguing direction for future research – exploring the specific characteristics of bird vocalizations that contribute to these superior generalization capabilities. Understanding these characteristics could guide the collection and selection of training data for future bioacoustic models, with the aim of maximizing their generalization potential.

The extensive diversity inherent in bird vocalizations, both in terms of acoustic characteristics and species diversity, is not just a theoretical advantage but also a practical one. The availability of a vast array of bird species audio data provides an advantageous basis for model training. Recent work has shown that training on a larger number of diverse classes helps model generalization more than adding data from a smaller number of classes [60]. To make this concrete, we include a comparison to a regional bird classification model trained on 89 bird species from the Sierra Nevadas. The global bird models, which were trained on a larger variety of classes, exhibited superior performance compared to the Sierra Birds model. This underperformance of the regional model is further highlighted when compared to general audio embeddings, where it falls short in some datasets. This aligns with findings from the above paper which emphasized that a greater number of training classes enhances the generalization ability of models. The regional model's limited scope in terms of training classes could be a key factor in its lower effectiveness compared to the more broadly trained global bird models.

We designed an experiment aimed to uncover any potential performance gains hidden within the architecture of the AudioMAE, which is SOTA and widely believed to have a superior architecture. Despite the general dominance of transformer and self-supervised models in machine learning in the

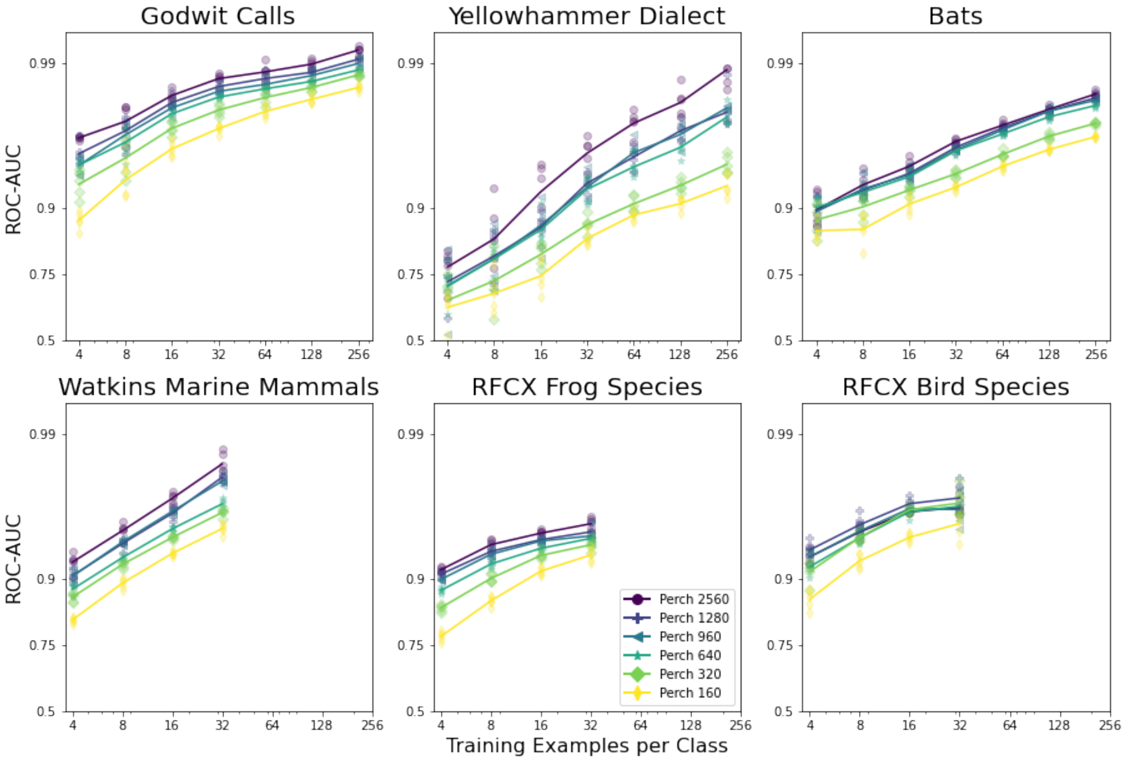

Figure 4: Results of Embedding Size Ablation Test. ROC-AUC scores are plotted with log-odds scaling.

Species	Confused Species	Confusion Rate
Bearded Seal	Bowhead Whale	0.186
Pantropical Spotted Dolphin	Spinner Dolphin	0.097
Common Dolphin	Striped Dolphin	0.091
Frasers Dolphin	Pantropical Spotted Dolphin	0.082
Killer Whale	Narwhal	0.067

Table 6: Top five marine mammal species confusions, averaged over five runs with the Perch model, using 32 examples per class. Bearded Seal and Bowhead Whale often appear in the same recording, though only one is labeled.

recent years, AudioMAE was outperformed by the more basic EfficientNet-B1. The investigation involved three key experiments: examining the quality of pre-trained vs. fine-tuned embeddings, bypassing audio sample rate adjustments, and employing a two-layer network to align the self-supervised embedding for classification tasks. Surprisingly, the best performance of AudioMAE was achieved using the pre-trained embeddings with a two-layer perceptron and without audio resampling. However, even with these adjustments, the global bird embeddings, without any tweaking, consistently outperformed the AudioMAE model. We also found that PSLA, a near-SOTA AudioSet model with an EfficientNet architecture, performed poorly on our target tasks. Together, these findings demonstrate that the data effects dominate the architecture effects.

To demonstrate the robustness of each embedding in the context of reduced data availability, we ran experiments to train the linear probes by varying the amount of training data. Our results, as seen in Fig. 2, have shown promising prospects for bioacoustic recognition tasks even when faced with as little as 4 training samples. The global bird models once again outperform other models in this experiment. This shows that these models can be used for active learning on novel tasks, starting even from a handful of examples. Lower Top-1 accuracy scores (see Table 3) suggest that inter-class calibration may still be a difficulty for simple linear probes, though *unlabelled vocalizations* in the test set may account for some difficulty. For the Watkins dataset, a significant amount of confusions (18.6%) occurred between bearded seals and bowhead whales, two highly vocal Arctic marine mammal species (see Table 6). Both species are known to overlap in range and are frequently recorded together, especially during the late spring and early summer months [61]. This is also the case for the weakly-labeled training data we used, which explains the comparatively high degree of confusion. More sophisticated pre-processing of the training data and adding some strongly labeled data would help to increase the classification performance for these two species. The confusion between co-occurring dolphin species is also not surprising. First, these data were downsampled to the audible frequency range, which will cutoff higher frequency components of the vocalizations. In addition, dolphin species are generally difficult to classify acoustically [62] because they produce highly variable vocalizations including whistle, echolocation clicks, and burst pulses. Lastly, dolphins also occur in mixed species groups which can make it challenging to obtain clean training data.

We also see a particularly high variance in model quality for the YD dataset in the low-data regime. Since this is only a two-class problem, there are fewer total examples used for training in the low-data regime. However, this is also a subtle problem: The Yellowhammer dialect is distinguished by the order of the last two notes of the song: mid-then-high versus high-then-mid. Other variations in timbre of the initial portion of the song and up- or down-sweep in the high note do not distinguish between the two dialects. The subtlety of the problem apparently makes it easy to over-generalize from few examples.

To demonstrate the efficacy of embeddings as condensed ’fingerprints’ of raw audio, we conducted experiments that revealed an enhancement in separability for bioacoustic tasks with the increase in embedding size. However, this improvement comes at the cost of increased model size and reduced inference speed. Newer versions of BirdNET have a larger embedding size as a result of these observations.

6 Conclusions

Our study explored generalizablility of feature embeddings within the bioacoustics domain, focusing on the application of large-scale audio classification models to previously unencountered taxonomic groups such as marine mammals, bats, and frogs, in addition to intraspecific calls and dialects of a bird species. Our study supports the hypothesis that feature embeddings, especially those derived from bird data, can effectively represent high-dimensional categorical or discrete features as a low-dimensional continuous vector space. We report that embeddings derived from models specifically trained using bird sounds data consistently facilitated superior classification quality compared to those trained on broader audio datasets. This study's findings also suggest that feature embeddings from global bird acoustic classifiers can be effectively utilized for few-shot transfer learning. This enables the acquisition of new classes with a minimal amount of training data. Our empirical findings have significant implications for Passive Acoustic Monitoring, potentially enhancing the methods by which we detect and classify animal species based on their sounds. This could revolutionize the application of PAM, particularly in low-data regimes, by enabling more effective transfer learning between coarse-level classification and more fine-grained vocalization classification.

7 Acknowledgements

The German Federal Ministry of Education and Research is funding the development of BirdNET through the project "BirdNET+" (FKZ 01—S22072). Additionally, Federal Ministry of Environment, Nature Conservation and Nuclear Safety is funding the development of BirdNET through the project "DeepBirdDetect" (FKZ 67KI31040E). BirdNET is also supported by Jake Holshuh (Cornell class of '69) and The Arthur Vining Davis Foundations. The work in the K. Lisa Yang Center for Conservation Bioacoustics is made possible by the generosity of K. Lisa Yang to advance innovative conservation technologies to inspire and inform the conservation of wildlife and habitats. BG acknowledges funding from the Landesforschungsförderung Hamburg within the Au-Tag BeoFisch (LFF-FV91) project while working as a researcher at the Hamburg University of Applied Sciences. We also acknowledge the rest of the Google Perch team (Vincent Dumoulin, Eleni Triantafillou, Bart van Merrienboer, and Jenny Hamer) who contributed to the development of the Perch model and offered numerous helpful suggestions on this work.

References

[1] Larissa Sayuri Moreira Sugai, Thiago Sanna Freire Silva, José Wagner Ribeiro Jr, and Diego Llusia. Terrestrial passive acoustic monitoring: review and perspectives. *BioScience*, 69(1): 15–25, 2019.

[2] Bobbi J Estabrook, Jamey T Tielens, Ashakur Rahaman, Dimitri W Ponirakis, Christopher W Clark, and Aaron N Rice. Dynamic spatiotemporal acoustic occurrence of north atlantic right whales in the offshore rhode island and massachusetts wind energy areas. *Endangered Species Research*, 49:115–133, 2022.

[3] Leila Fouda, Jessica E Wingfield, Amber D Fandel, Aran Garrod, Kristin B Hodge, Aaron N Rice, and Helen Bailey. Dolphins simplify their vocal calls in response to increased ambient noise. *Biology letters*, 14(10):20180484, 2018.

[4] Ilse Van Opzeeland, Sofie Van Parijs, Horst Bornemann, Stephan Frickenhaus, Lars Kindermann, Holger Klinck, Joachim Plötz, and Olaf Boebel. Acoustic ecology of antarctic pinnipeds. *Marine Ecology Progress Series*, 414:267–291, 2010.

[5] Jessica L Crance, Catherine L Berchok, Brynn M Kimber, Jenna M Harlacher, Eric K Braen, and Megan C Ferguson. Year-round distribution of bearded seals, erignathus barbatus, throughout the alaskan chukchi and northern bering sea. *Deep Sea Research Part II: Topical Studies in Oceanography*, 206:105215, 2022.

[6] Connor M Wood, Ralph J Gutiérrez, and M Zachariah Peery. Acoustic monitoring reveals a diverse forest owl community, illustrating its potential for basic and applied ecology. *Ecology*, 100(9):1–3, 2019.

[7] Laurel B Symes, Kyle D Kittelberger, Sophia M Stone, Richard T Holmes, Jessica S Jones, Itzel P Castaneda Ruvalcaba, Michael S Webster, and Matthew P Ayres. Analytical approaches for evaluating passive acoustic monitoring data: A case study of avian vocalizations. *Ecology and Evolution*, 12(4):e8797, 2022.

[8] Laurel B Symes, Shyam Madhusudhana, Sharon J Martinson, Ciara E Kernan, Kristin B Hodge, Daniel P Salisbury, Holger Klinck, and Hannah ter Hofstede. estimation of katydid calling activity from soundscape recordings. *Journal of Orthoptera Research*, 31(2):173–180, 2022.

[9] Richard W Mankin, David W Hagstrum, Mark T Smith, AL Roda, and Moses TK Kairo. Perspective and promise: a century of insect acoustic detection and monitoring. *American Entomologist*, 57(1):30–44, 2011.

[10] Rodney A Rountree, R Grant Gilmore, Clifford A Goudey, Anthony D Hawkins, Joseph J Luczkovich, and David A Mann. Listening to fish: applications of passive acoustics to fisheries science. *Fisheries*, 31(9):433–446, 2006.

[11] Elena Desiderà, Paolo Guidetti, Pieraugusto Panzalis, Augusto Navone, Cathy-Anna Valentini-Poirrier, Pierre Boissery, Cédric Gervaise, and Lucia Di Iorio. Acoustic fish communities: sound diversity of rocky habitats reflects fish species diversity. *Marine Ecology Progress Series*, 608:183–197, 2019.

[12] Danielle V Nelson and Tiffany S Garcia. Seasonal and diel vocal behavior of the northern red-legged frog, rana aurora. *Northwestern Naturalist*, 98(1):33–38, 2017.

[13] G John Measey, Ben C Stevenson, Tanya Scott, Res Altwegg, and David L Borchers. Counting chirps: acoustic monitoring of cryptic frogs. *Journal of Applied Ecology*, 54(3):894–902, 2017.

[14] Dena Jane Clink, Isabel Comella, Abdul Hamid Ahmad, and Holger Klinck. A workflow for the automated detection and classification of female gibbon calls from long-term acoustic recordings. *Frontiers in Ecology and Evolution*, 11:28, 2023.

[15] Colin R Swider, Christopher F Gemelli, Peter H Wrege, and Susan E Parks. Passive acoustic monitoring reveals behavioural response of african forest elephants to gunfire events. *African Journal of Ecology*, 2022.

[16] Dan Stowell. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10:e13152, 2022.

[17] Kristin M Brunk, RJ Gutiérrez, M Zachariah Peery, C Alina Cansler, Stefan Kahl, and Connor M Wood. Quail on fire: changing fire regimes may benefit mountain quail in fire-adapted forests. *Fire Ecology*, 19(1):19, 2023.

[18] Dan Stowell, Michael D Wood, Hanna Pamuła, Yannis Stylianou, and Hervé Glotin. Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. *Methods in Ecology and Evolution*, 10(3):368–380, 2019.

[19] Burooj Ghani. Machine learning-based analysis of bird vocalizations. https://ediss.uni-goettingen.de/handle/11858/13959?locale-attribute=en, 2021.

[20] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014.

[21] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[22] Francois Chollet. The limitations of deep learning. *Deep learning with Python*, 2017.

[23] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

[24] Clive K. Catchpole and Peter J.B. Slater. *Bird song: Biological Themes and Variations (2nd edition)*. Cambridge University Press, 2008.

[25] 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*, pages 125–145. PMLR, 2022.

[26] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[27] Julie N Oswald, Amy M Van Cise, Angela Dassow, Taffeta Elliott, Michael T Johnson, Andrea Ravignani, and Jeffrey Podos. A collection of best practices for the collection and analysis of bioacoustic data. *Applied Sciences*, 12(23):12046, 2022.

[28] Kate McGinn, Stefan Kahl, M Zachariah Peery, Holger Klinck, and Connor M Wood. Feature embeddings from the birdnet algorithm provide insights into avian ecology. *Ecological Informatics*, page 101995, 2023.

[29] Irina Tolkova, Brian Chu, Marcel Hedman, Stefan Kahl, and Holger Klinck. Parsing birdsong with deep audio embeddings. *arXiv preprint arXiv:2108.09203*, 2021.

[30] Malik Boudiaf, Tom Denton, Bart van Merriënboer, Vincent Dumoulin, and Eleni Triantafillou. In search for a generalizable method for source free domain adaptation. *arXiv preprint arXiv:2302.06658*, 2023.

[31] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification. https://arxiv.org/abs/1609.09430, 2017.

[32] Sarab S Sethi, Robert M Ewers, Nick S Jones, Jani Sleutel, Adi Shabrani, Nursyamin Zulkifli, and Lorenzo Picinali. Soundscapes predict species occurrence in tropical forests. *Oikos*, 2022 (3):e08525, 2022.

[33] Masato Hagiwara, Benjamin Hoffman, Jen-Yu Liu, Maddie Cusimano, Felix Effenberger, and Katie Zacarian. Beans: The benchmark of animal sounds. *arXiv preprint arXiv:2210.12300*, 2022.

[34] Becky E Heath, Sarab S Sethi, C David L Orme, Robert M Ewers, and Lorenzo Picinali. How index selection, compression, and recording schedule impact the description of ecological soundscapes. *Ecology and Evolution*, 11(19):13206–13217, 2021.

[35] Enis Berk Çoban, Dara Pir, Richard So, and Michael I Mandel. Transfer learning from youtube soundtracks to tag arctic ecoacoustic recordings. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 726–730. IEEE, 2020.

[36] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

[37] Burooj Ghani and Sarah Hallerberg. A randomized bag-of-birds approach to study robustness of automated audio based bird species classification. *Applied Sciences*, 11(19):9226, 2021.

[38] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2019.

[39] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E. Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[40] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: A good embedding is all you need? In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV*, page 266–282, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58567-9. doi: 10.1007/978-3-030-58568-6_16. URL https://doi.org/10.1007/978-3-030-58568-6_16.

[41] Kazufumi Hosoda, Keigo Nishida, Shigeto Seno, Tomohiro Mashita, Hideki Kashioka, and Izumi Ohzawa. It's DONE: direct one-shot learning without training optimization. *CoRR*, abs/2204.13361, 2022. doi: 10.48550/arXiv.2204.13361. URL https://doi.org/10.48550/arXiv.2204.13361.

[42] Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15, 2002.

[43] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 1(10):e2, 2016.

[44] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[45] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[46] Tom Denton, Scott Wisdom, and John R Hershey. Improving bird classification with unsupervised sound separation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 636–640. IEEE, 2022.

[47] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[48] Stefan Kahl, Connor M Wood, Maximilian Eibl, and Holger Klinck. Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021.

[49] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *NeurIPS*, 2022.

[50] Mariana-Iuliana Georgescu, Eduardo Fonseca, Radu Tudor Ionescu, Mario Lucic, Cordelia Schmid, and Anurag Arnab. Audiovisual masked autoencoders. *arXiv preprint arXiv:2212.05922*, 2022.

[51] Yuan Gong, Yu-An Chung, and James Glass. Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3292–3306, 2021.

[52] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[53] Ondrej Belfin. Vocalizations of black-tailed godwit. https://fse.studenttheses.ub.rug.nl/id/eprint/26433, 2022.

[54] Tereza Petrusková, Lucie Diblíková, Pavel Pipek, Eckehard Frauendorf, Petr Procházka, and Adam Petrusek. A review of the distribution of yellowhammer (emberiza citrinella) dialects in europe reveals the lack of a clear macrogeographic pattern. *Journal of Ornithology*, 156:263–273, 2015.

[55] B. Gotthold, A. Khalighifar, B.R. Straw, and B.E. Reichert. Training dataset for nabat machine learning v1.0. https://doi.org/10.5066/P969TX8F, 2022.

[56] Daniel T Murphy, Elias Ioup, Md Tamjidul Hoque, and Mahdi Abdelguerfi. Residual learning for marine mammal classification. *IEEE Access*, 10:118409–118418, 2022.

[57] Laela Sayigh, Mary Ann Daher, Julie Allen, Helen Gordon, Katherine Joyce, Claire Stuhlmann, and Peter Tyack. The watkins marine mammal sound database: an online, freely accessible resource. In *Proceedings of Meetings on Acoustics 4ENAL*, volume 27, page 040013. Acoustical Society of America, 2016.

[58] William Watkins. Watkins marine mammal sound database, 2021.

[59] Coen PH Elemans, Jeppe Have Rasmussen, Christian T Herbst, Daniel Normen Düring, Sue Anne Zollinger, Henrik Brumm, K Srivastava, Niels Svane, Ming Ding, Ole Naesbye Larsen, et al. Universal mechanisms of sound production and control in birds and mammals. *Nature communications*, 6(1):8978, 2015.

[60] Xu Luo, Hao Wu, Ji Zhang, Lianli Gao, Jing Xu, and Jingkuan Song. A closer look at few-shot classification again. *arXiv preprint arXiv:2301.12246*, 2023.

[61] Emily Chou, Ricardo Antunes, Stephanie Sardelis, Kathleen M Stafford, Leigh West, Christopher Spagnoli, Brandon L Southall, Martin Robards, and Howard C Rosenbaum. Seasonal variation in arctic marine mammal acoustic detection in the northern bering sea. *Marine Mammal Science*, 36(2):522–547, 2020.

[62] Shannon Rankin, Frederick Archer, Jennifer L Keating, Julie N Oswald, Michael Oswald, Alex Curtis, and Jay Barlow. Acoustic classification of dolphins in the california current using whistles, echolocation clicks, and burst pulses. *Marine Mammal Science*, 33(2):520–540, 2017.