Feature embeddings from large-scale acoustic bird classifiers enable few-shot transfer learning

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

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

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

[4]Chemnitz University of Technology

July 13, 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 large-scale 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 (Sugai et al., 2019). PAM involves recording sound in nature and has been used to study a wide range of species, including whales and dolphins (Estabrook et al., 2022; Fouda et al., 2018), pinnipeds (Van Opzeeland et al., 2010; Crance et al., 2022), birds (Wood et al., 2019; Symes et al., 2022a), insects (Symes et al., 2022b; Mankin et al., 2011), fish (Rountree et al., 2006; Desiderà et al., 2019), frogs (Nelson and Garcia, 2017; Measey et al., 2017), and terrestrial mammals (Clink et al., 2023; Swider et al., 2022). 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 (Stowell, 2022). 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 (Brunk et al., 2023). However, the development of these tools typically depends on the availability of well-annotated training data. Obtaining sufficient training data can be a

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

major challenge. While there are sufficient amounts of training data available for some 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 (Stowell et al., 2019). 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 (Ghani, 2021). 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 machine learning model (Stowell, 2022).

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 (Catchpole and Slater, 2008). 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.

We can view feature embeddings as a lossy compression of the input data. For instance, in terms of raw data, the embedding produced by Google's bird classification model (called Perch) 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 perspective is typified by data augmentation techniques, in which we apply transformations of the inputs irrelevant to the desired function outputs, thus training the classifier to ignore the augmentations.

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 (Turian et al., 2022). 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 (Gemmeke et al., 2017a) (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-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

[†] https://xeno-canto.org

[‡] https://www.macaulaylibrary.org

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 (Oswald et al., 2022). 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. 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.

2 Related work

2.1 Transfer learning

In 2014 Oquab et al. (2014) observed that pre-trained CNN layers could be used as general feature extractors for novel tasks by training a new output layer for the target task. Meanwhile, Yosinski et al. (2014) demonstrated that using pre-trained features leads to more general models and experimented with different combinations of freezing and fine-tuning layers in the network. This strategy, where pre-trained models are utilized as foundational building blocks to extract robust features for new tasks and potentially fine-tuned for specific target tasks, is known as transfer learning (Chollet, 2017). Chu et al. (2016); Pittaras et al. (2017) have compared different transfer learning strategies in CNNs. This technique has proven to be extremely effective, especially when the available training data for the target task is limited. Lasseck (2018) employed DCNN models that were pre-trained on ImageNet dataset (Deng et al., 2009) to subsequently fine-tune them for bird sound detection. Similarly, Sevilla and Glotin (2017) fine-tune an Inception-v4 CNN that was previously trained on ImageNet to perform audio bird classification. In (Dufourq et al., 2022), the authors investigate the utility of transfer learning, specifically adapting existing convolutional neural networks (CNNs) pretrained on the ImageNet dataset, in the realm of bioacoustics research. Their study, which compares 12 modern CNN arctitectures across four passive acoustic datasets, reveals promising results that suggests transfer learning could make bioacoustic model design more accessible, efficient, and accurate, particularly in scenarios with limited data.

2.2 Few-shot learning

Few-shot learning (Wang et al., 2020) attempts to learn new classes from a small amount of training data; pre-trained embeddings are core to many few-shot learning strategies. 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 (Alain and Bengio, 2016). In 2021-2023, the DCASE competition included a few-shot bioacoustic classification task on multiple taxa, utilizing exactly five 'shots' (training examples) for each class of interest (Nolasco et al., 2023b). In the 2023 iteration of the competition (Nolasco et al., 2023a) one team used a pre-trained transformer model trained on AudioSet, while all other teams used only the 21 hours of provided training data. By contrast, in this study we focus on the relative utility of pre-trained embeddings, including embeddings from bird classifiers, and demonstrate the relative utility of additional training data for each dataset we work with, which is useful for practitioners.

2.3 Feature embeddings for bioacoustics tasks

BirdNET* embeddings have been previously used to distinguish adult and juvenile owls and woodpecker call types (McGinn et al., 2023), and to improve pre-trained model performance on downstream bioacoustic tasks given additional unlabeled data (Tolkova et al., 2021; Boudiaf et al., 2023). The BEANS Benchmark Hagiwara et al. (2022) applies pre-trained image classification models (ResNets He et al. (2016)) and audio event classifiers (VGGish (Hershey et al., 2017)) to a range of bioacoustic tasks. However, the strongest model considered in the benchmark (VGGish) is an older general audio event classification model. In (Lauha et al., 2022) the authors illustrate the process of developing a global identification model that can be refined for optimal performance on

*https://birdnet.cornell.edu

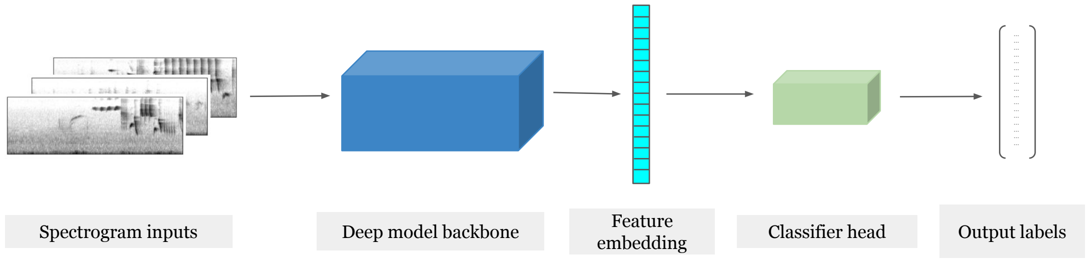

Figure 1: Pipeline for Downstream Classification

localized data from specific regions, employing a strategy that aligns closely with transfer learning. The methodology consists of constructing a model trained on a global data set, then fine-tuning it using local data. This allows the model to be highly adaptable, providing effective classification of bird vocalizations across different regions. In (Sethi et al., 2020) feature embeddings from the VGGish model are employed to transform soundscapes from various ecosystems into a common acoustic space. By using the feature embeddings, the researchers were able to extract meaningful features from complex eco-acoustic data, providing them with a useful tool for monitoring ecosystems in an efficient and scalable manner. In (Sethi et al., 2022) a classification model is trained using embeddings derived from the VGGish model to study the utility of soundscapes to predict species occurrence in tropical forests. Çoban et al. (2020) employed VGGish feature embeddings to train models to identify eight acoustic event categories, including vocalizations of songbirds and insects as well as non-biological signals from recordings made in Northern Alaska. The study found that the performance of classification models was significantly influenced by the choice of acoustic index. Models using AudioSet embeddings demonstrated substantially superior accuracy, precision, and recall, improving by 12%–16%. The authors therefore recommend the use of AudioSet embeddings in soundscape analysis due to its consistent performance, even with limited data sets, and its ability to be compressed efficiently, facilitating the use of restricted data storage without impacting the comparability of results between different studies. Heath et al. (2021) analysed soundscape recordings from North-Eastern Borneo using analytical indices and VGGish-based feature embeddings and reported consistent and superior performance on small pools of data using the latter.

3 Method

In this work, we have focused on the extraction of feature embeddings from four CNN models and one transformer model, described below in Section 3.3. These models are trained on either general YouTube data or global data sets of bird vocalizations. All of these are large-scale audio classification models that map spectrograms (visual representation of sound) to their class labels. We train linear models (fully-connected feed forward NNs) on the feature embeddings, extracted from these models, using different amounts of training data, as described in Section 3.2. Fig. 1 provides an overview of the classification pipeline we employed for our experiments. Spectrograms serve as the input data for our framework. The deep backbone, 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 deep backbone, 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.

3.1 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 1 presents an overview of all the datasets used in this work.

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

Table 1: Summary of target dataset characteristics. For the Bats dataset, frequency shifting was applied to move signal into the audible range.

Yellowhammer Dialects (YD): The YD dataset comprises two dialects of Yellowhammer songs, denoted as X and B (Petrusková et al., 2015), 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. Initially, Connor Wood performed preliminary annotations, which were later refined by Pavel Pipek specializing 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 audio files have been frequency-shifted to place the bats in the audible range. The dataset is sourced from two origins: 1) Training dataset for NABat Machine Learning V1.0 (Gotthold et al., 2022), 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. After this pre-processing step, all of the datasets featured a sampling rate of 44.1 kHz.

Rainforest Connection Kaggle dataset (RFCX): This is the training data from the 2021 Species Audio Detection challenge, consisting of recordings of Puerto Rican birds and frogs. This data set has weak negative labels. 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 are present 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, and the results for this data set indicate the ability of the BirdNET and Perch embeddings to separate species ID for under-trained species.

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 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 (Murphy et al., 2022). 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 (Sayigh et al., 2016; Watkins, 2021). 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).

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 (Belfin, 2022). All recordings are 3 seconds long and are annotated by Ondrej Belfin himself. The author of the dataset is in the process of publishing the dataset, a link to the published dataset will be added to the final version of the paper.

3.2 Experimental methodology

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 4.5). 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 choose a fixed number k of examples from each class at random, by randomly shuffling the list of examples and picking the first k examples of each class in the shuffled list. We use 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 (Ghani and Hallerberg, 2021).

Note that our use of BCE loss could be replaced with categorical cross entropy (CCE) with a softmax output. We found that this produces somewhat higher model quality scores, but is not reflective of real-world requirements, where models may encounter simultaneous vocalizations. Training with BCE loss is equivalent to training independent binary classifiers for each class. For each experiment, we compute (1) macro-averaged ROC-AUC (computing ROC-AUC for each class, and then averaging over all classes) and (2) Top-1 Accuracy.

Dataset Limitations: Each of the datasets we work with presents different 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 entirely 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. 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 4 for some analysis of the Watkins dataset.)

3.3 Model descriptions

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. We compare these bird models to three models trained for general audio event detection, using variants of AudioSet (Gemmeke et al., 2017b). 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 (Tan and Le, 2019) trained on the full corpus of bird song recordings from Xeno-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 (Denton et al., 2022). During training we augment with MixUp (Zhang et al., 2017), random gain adjustment, and random time-shifting of up to one second. The model is

[†] https://tfhub.dev/google/bird-vocalization-classifier/2

	Architecture	Training Data	Window (s)	Embedding Size	CPU(ms/s)
BirdNET 2.2 / 2.3	EfficientNet B1	XC+ML+Custom	3.0	320 / 1024	10.0 / 11.1
Google Perch	EfficientNet B1	XenoCanto	5.0	1280	24.3
AudioMAE	MAE (Large)	AudioSet	10.0	1024	78.2
YAMNet	MobileNet v1	AudioSet	0.96	1024	7.7
VGGish	Modified VGG	YouTube 8M	0.96	128	2.8

Table 2: 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.

trained to classify all levels of the taxonomy for each recording simultaneously (species, genus, family, order). Note that the model in (Denton et al., 2022) was a regional model trained on 89 species, while the Perch model is trained on all Xeno-Canto species. This new single model obtains a cMAP score of 0.49 on the Caples data set, where the regional model obtained a score of 0.34 using a combination of ensembling and source separation. The base Google Perch model and further evaluation statistics are available at TFHub and supporting code is available on GitHub[‡].

BirdNET (Kahl et al., 2021) 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 over 3,000 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, which differ only in the dimensionality of the embedding (see Section 4.3). The BirdNET code is available on GitHub[§], and includes support for training small classifiers on embeddings.

YAMNet and **VGGish** are both convolutional models trained to predict AudioSet classes. YAMNet uses a MobileNetV1 architecture (Howard et al., 2017). VGGish is an older audio event-detection model, using a variant of the VGG architecture and trained on an earlier version of AudioSet (Hershey et al., 2017). 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.

AudioMAE (Huang et al., 2022) 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 (Georgescu et al., 2022). 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 4.5. 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[**].

4 Results

4.1 Classification performance on novel bioacoustic tasks

Table 3 presents the classification performance using a range of embeddings with linear probes on novel bioacoustic tasks. The results presented in the table correspond to an experiment in which

[‡] https://github.com/google-research/chirp/tree/main/chirp/inference

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

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

[‖] https://github.com/tensorflow/models/tree/master/research/audioset/vggish

[**] https://github.com/facebookresearch/AudioMAE

the models are trained on 32 audio samples. Figure 2 shows results at various amounts of training data, from 4 to 32 or 256 examples per class (depending on dataset size).

Model	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	**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
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
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

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, 0.97). Both bird models significantly outperform the AudioSet models on all tasks (VGGish, YAMNet, 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 and YAMNet 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. It's important to note that these results are average values over five runs.

4.2 Varying amount of training data: 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. This suggests these models can be used for active learning on novel tasks, starting even from a handful of examples.

Lower Top-1 accuracy scores suggest that inter-class calibration may still be a difficulty for simple linear probes, though *unlabelled vocalizations* in the eval 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 4). Both species are known to overlap in range and are frequently recorded together, especially during the late spring and early summer months (Chou et al., 2020). 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 (Rankin et al., 2017) 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

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 4: 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.

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.

4.3 Embedding size

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. Increasing the size of the BirdNET embedding to 1024 led to similar performance as the Perch model in most downstream tasks; the new BirdNET 2.3 has a larger embedding as a result.

An ablation over the embedding dimension is summarized in Fig. 3 that shows the ROC-AUC scores. The Top-1 Accuracy and ROC-AUC scores on different datasets using various embedding sizes are shown in Table 5. For this, we varied the final embedding size in the Perch model, keeping the EfficientNet B1 architecture otherwise unchanged. The 320-dimensional embedding (matching BirdNET 2.2) has significantly degraded quality in all tasks. Doubling the base Perch embedding

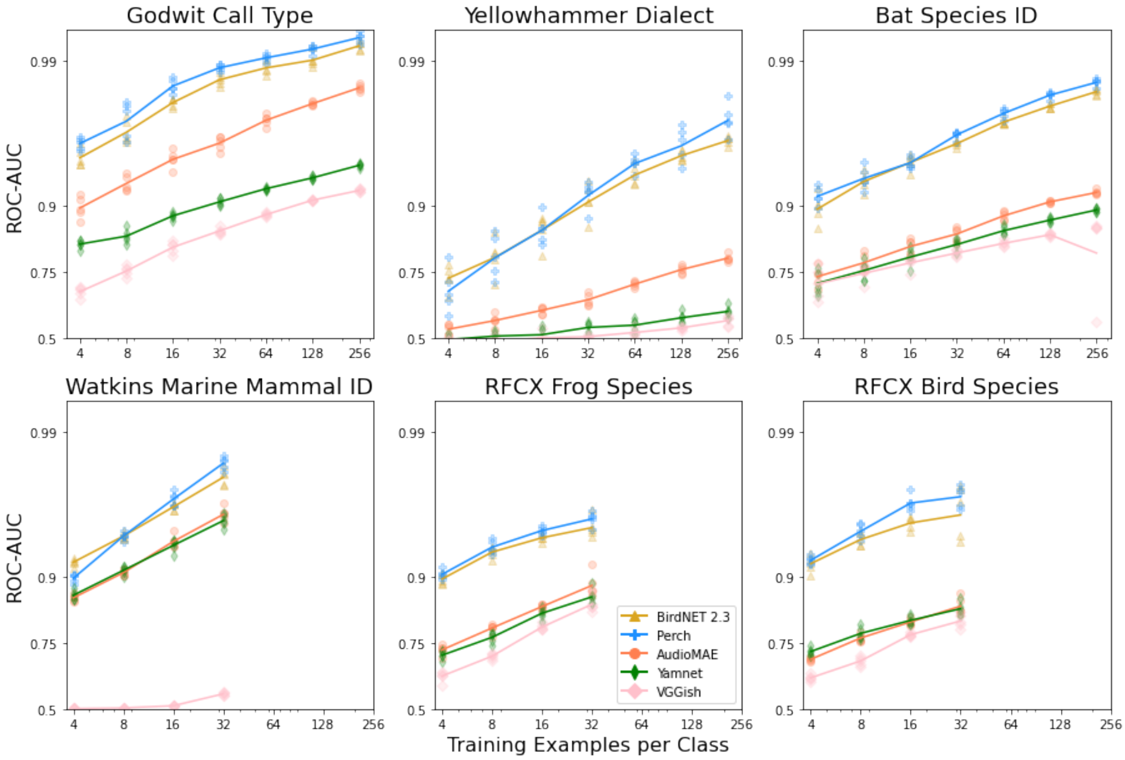

Figure 2: Results of Few-Shot Transfer Learning Tasks. ROC-AUC scores are plotted with log-odds scaling. A point is plotted for each experiment, and the curve connects the average quality for experiments at each number of training examples per class.

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.

dimension to 2560 yields a further increase in model performance for some downstream tasks. The larger embedding size substantially increases model size (because the large classification output layer doubles in size) and increases the storage footprint for the embeddings themselves. However, the impact on overall model runtime (as reported in Table 2) is modest because most computation time is spent in the early layers.

4.4 Visualizing pre-trained embeddings

We can also observe the geometry of the embedding space using a t-SNE transformation of the model embeddings (Hinton and Roweis, 2002). The t-SNE transformation attempts to preserve distances in the embedding space while projecting to two dimensions. In Fig. 4 we plot t-SNE transforms for YAMNet, AudioMAE, and Perch. Note that t-SNE plots can be tricky to interpret appropriately (Wattenberg et al., 2016), though points which are close in the original space tend to be close after applying the t-SNE transform.

In the easier Godwit problem (Fig. 4), 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

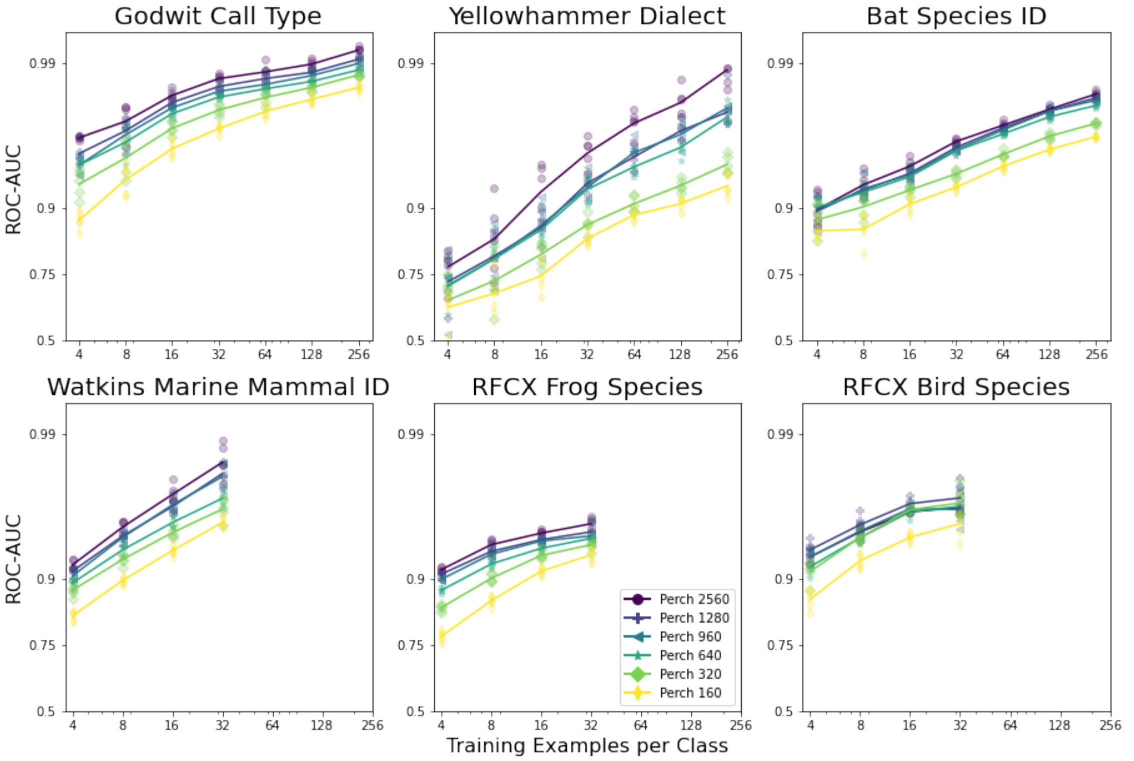

Figure 3: Results of Embedding Size Ablation Test. ROC-AUC scores are plotted with log-odds scaling.

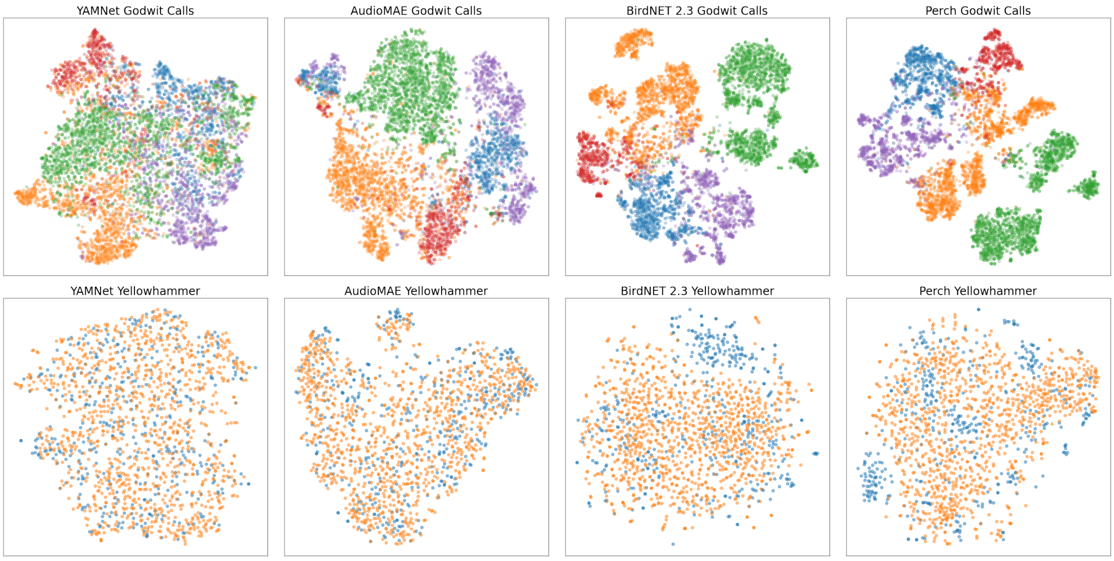

Figure 4: t-SNE plots of Godwit and Yellowhammer embeddings. Points are colored by class.

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, we see some clustering, but still a great deal of inter-mixed data.

4.5 Additional AudioMAE Investigation

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 performed a number of additional experiments to discover whether additional tweaking of the experimental setup would uncover hidden performance gains for the AudioMAE embeddings. In Table 6 we give results for three different treatments on all six datasets.

First, we compared embedding quality for between the pre-trained unspervised 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.

One significant improvement was obtained by 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. This change almost always improved the AudioMAE metrics.

We also 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.

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 bird models - with no additional tweaking - uniformly outperformed the AudioMAE model.

5 Discussion

Our study explored generalizable feature representations (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

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

Table 6: 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.

calls and dialects of a bird species. Our empirical findings have significant implications for Passive Acoustic Monitoring (PAM), potentially enhancing the methods by which we detect and classify animal species based on their sounds.

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, 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. 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. 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.

This superior generalization capability of deep embeddings from bird models is an interesting 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 bioacoutic 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. Our results have shown promising prospects for bioacoustic recognition tasks even when faced with limited training data. Such good quality feature embeddings can be utilized toward few-shot transfer learning to learn new classes from a small amount of training data.

Furthermore, 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. This could revolutionize the application of PAM, particularly in low-data regimes, by enabling more effective transfer learning between species or coarse-level classification and more fine-grained vocalization classification.

6 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. Our 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 AuTag BeoFisch (LFF-FV91) project while working as a researcher at the Hamburg University of Applied Sciences.

References

Alain, G. & Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Belfin, O. (2022). Vocalizations of black-tailed godwit. https://fse.studenttheses.ub.rug.nl/id/eprint/26433.

Boudiaf, M., Denton, T., van Merriënboer, B., Dumoulin, V., & Triantafillou, E. (2023). In search for a generalizable method for source free domain adaptation. *arXiv preprint arXiv:2302.06658*.

Brunk, K. M., Gutiérrez, R., Peery, M. Z., Cansler, C. A., Kahl, S., & Wood, C. M. (2023). Quail on fire: changing fire regimes may benefit mountain quail in fire-adapted forests. *Fire Ecology*, *19*(1), 19.

Catchpole, C. K. & Slater, P. J. (2008). *Bird song: Biological Themes and Variations (2nd edition)*. Cambridge University Press.

Chollet, F. (2017). The limitations of deep learning. *Deep learning with Python*.

Chou, E., et al. (2020). Seasonal variation in arctic marine mammal acoustic detection in the northern bering sea. *Marine Mammal Science*, *36*(2), 522–547.

Chu, B., Madhavan, V., Beijbom, O., Hoffman, J., & Darrell, T. (2016). Best practices for fine-tuning visual classifiers to new domains. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14* (pp. 435–442).: Springer.

Clink, D. J., Comella, I., Ahmad, A. H., & Klinck, H. (2023). A workflow for the automated detection and classification of female gibbon calls from long-term acoustic recordings. *Frontiers in Ecology and Evolution*, *11*, 28.

Çoban, E. B., Pir, D., So, R., & Mandel, M. I. (2020). Transfer learning from youtube soundtracks to tag arctic ecoacoustic recordings. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 726–730).: IEEE.

Crance, J. L., Berchok, C. L., Kimber, B. M., Harlacher, J. M., Braen, E. K., & Ferguson, M. C. (2022). 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.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255).: Ieee.

Denton, T., Wisdom, S., & Hershey, J. R. (2022). Improving bird classification with unsupervised sound separation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 636–640).: IEEE.

Desiderà, E., et al. (2019). Acoustic fish communities: sound diversity of rocky habitats reflects fish species diversity. *Marine Ecology Progress Series*, *608*, 183–197.

Dufourq, E., Batist, C., Foquet, R., & Durbach, I. (2022). Passive acoustic monitoring of animal populations with transfer learning. *Ecological Informatics*, *70*, 101688.

Estabrook, B. J., Tielens, J. T., Rahaman, A., Ponirakis, D. W., Clark, C. W., & Rice, A. N. (2022). 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.

Fouda, L., Wingfield, J. E., Fandel, A. D., Garrod, A., Hodge, K. B., Rice, A. N., & Bailey, H. (2018). Dolphins simplify their vocal calls in response to increased ambient noise. *Biology letters*, *14*(10), 20180484.

Gemmeke, J. F., et al. (2017a). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 776–780).: IEEE.

Gemmeke, J. F., et al. (2017b). Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017* New Orleans, LA.

Georgescu, M.-I., Fonseca, E., Ionescu, R. T., Lucic, M., Schmid, C., & Arnab, A. (2022). Audiovisual masked autoencoders. *arXiv preprint arXiv:2212.05922*.

Ghani, B. (2021). Machine learning-based analysis of bird vocalizations. https://ediss.uni-goettingen.de/handle/11858/13959?locale-attribute=en.

Ghani, B. & Hallerberg, S. (2021). A randomized bag-of-birds approach to study robustness of automated audio based bird species classification. *Applied Sciences*, *11*(19), 9226.

Gotthold, B., Khalighifar, A., Straw, B., & Reichert, B. (2022). Training dataset for nabat machine learning v1.0. https://doi.org/10.5066/P969TX8F.

Hagiwara, M., Hoffman, B., Liu, J.-Y., Cusimano, M., Effenberger, F., & Zacarian, K. (2022). Beans: The benchmark of animal sounds. *arXiv preprint arXiv:2210.12300*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Heath, B. E., Sethi, S. S., Orme, C. D. L., Ewers, R. M., & Picinali, L. (2021). How index selection, compression, and recording schedule impact the description of ecological soundscapes. *Ecology and Evolution*, *11*(19), 13206–13217.

Hershey, S., et al. (2017). Cnn architectures for large-scale audio classification. https://arxiv.org/abs/1609.09430.

Hinton, G. E. & Roweis, S. (2002). Stochastic neighbor embedding. *Advances in neural information processing systems*, *15*.

Howard, A. G., et al. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Huang, P.-Y., et al. (2022). Masked autoencoders that listen. In *NeurIPS*.

Kahl, S., Wood, C. M., Eibl, M., & Klinck, H. (2021). Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, *61*, 101236.

Lasseck, M. (2018). Acoustic bird detection with deep convolutional neural networks. In *DCASE* (pp. 143–147).

Lauha, P., Somervuo, P., Lehikoinen, P., Geres, L., Richter, T., Seibold, S., & Ovaskainen, O. (2022). Domain-specific neural networks improve automated bird sound recognition already with small amount of local data. *Methods in Ecology and Evolution*, *13*(12), 2799–2810.

Mankin, R. W., Hagstrum, D. W., Smith, M. T., Roda, A., & Kairo, M. T. (2011). Perspective and promise: a century of insect acoustic detection and monitoring. *American Entomologist*, *57*(1), 30–44.

McGinn, K., Kahl, S., Peery, M. Z., Klinck, H., & Wood, C. M. (2023). Feature embeddings from the birdnet algorithm provide insights into avian ecology. *Ecological Informatics*, (pp. 101995).

Measey, G. J., Stevenson, B. C., Scott, T., Altwegg, R., & Borchers, D. L. (2017). Counting chirps: acoustic monitoring of cryptic frogs. *Journal of Applied Ecology*, *54*(3), 894–902.

Murphy, D. T., Ioup, E., Hoque, M. T., & Abdelguerfi, M. (2022). Residual learning for marine mammal classification. *IEEE Access*, *10*, 118409–118418.

Nelson, D. V. & Garcia, T. S. (2017). Seasonal and diel vocal behavior of the northern red-legged frog, rana aurora. *Northwestern Naturalist*, *98*(1), 33–38.

Nolasco, I., et al. (2023a). Few-shot bioacoustic event detection at the dcase 2023 challenge.

Nolasco, I., et al. (2023b). Learning to detect an animal sound from five examples. *arXiv preprint arXiv:2305.13210*.

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1717–1724).

Oswald, J. N., Van Cise, A. M., Dassow, A., Elliott, T., Johnson, M. T., Ravignani, A., & Podos, J. (2022). A collection of best practices for the collection and analysis of bioacoustic data. *Applied Sciences*, *12*(23), 12046.

Petrusková, T., Diblíková, L., Pipek, P., Frauendorf, E., Procházka, P., & Petrusek, A. (2015). 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.

Pittaras, N., Markatopoulou, F., Mezaris, V., & Patras, I. (2017). Comparison of fine-tuning and extension strategies for deep convolutional neural networks. In *MultiMedia Modeling: 23rd International Conference, MMM 2017, Reykjavik, Iceland, January 4-6, 2017, Proceedings, Part I 23* (pp. 102–114).: Springer.

Rankin, S., Archer, F., Keating, J. L., Oswald, J. N., Oswald, M., Curtis, A., & Barlow, J. (2017). Acoustic classification of dolphins in the california current using whistles, echolocation clicks, and burst pulses. *Marine Mammal Science*, *33*(2), 520–540.

Rountree, R. A., Gilmore, R. G., Goudey, C. A., Hawkins, A. D., Luczkovich, J. J., & Mann, D. A. (2006). Listening to fish: applications of passive acoustics to fisheries science. *Fisheries*, *31*(9), 433–446.

Sayigh, L., Daher, M. A., Allen, J., Gordon, H., Joyce, K., Stuhlmann, C., & Tyack, P. (2016). The watkins marine mammal sound database: an online, freely accessible resource. In *Proceedings of Meetings on Acoustics 4ENAL*, volume 27 (pp. 040013).: Acoustical Society of America.

Sethi, S. S., Ewers, R. M., Jones, N. S., Sleutel, J., Shabrani, A., Zulkifli, N., & Picinali, L. (2022). Soundscapes predict species occurrence in tropical forests. *Oikos*, *2022*(3), e08525.

Sethi, S. S., et al. (2020). Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set. *Proceedings of the National Academy of Sciences*, *117*(29), 17049–17055.

Sevilla, A. & Glotin, H. (2017). Audio bird classification with inception-v4 extended with time and time-frequency attention mechanisms. *CLEF (Working Notes)*, *1866*, 1–8.

Stowell, D. (2022). Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, *10*, e13152.

Stowell, D., Wood, M. D., Pamuła, H., Stylianou, Y., & Glotin, H. (2019). Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. *Methods in Ecology and Evolution*, *10*(3), 368–380.

Sugai, L. S. M., Silva, T. S. F., Ribeiro Jr, J. W., & Llusia, D. (2019). Terrestrial passive acoustic monitoring: review and perspectives. *BioScience*, *69*(1), 15–25.

Swider, C. R., Gemelli, C. F., Wrege, P. H., & Parks, S. E. (2022). Passive acoustic monitoring reveals behavioural response of african forest elephants to gunfire events. *African Journal of Ecology*.

Symes, L. B., et al. (2022a). Analytical approaches for evaluating passive acoustic monitoring data: A case study of avian vocalizations. *Ecology and Evolution*, *12*(4), e8797.

Symes, L. B., et al. (2022b). estimation of katydid calling activity from soundscape recordings. *Journal of Orthoptera Research*, *31*(2), 173–180.

Tan, M. & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105–6114).: PMLR.

Tolkova, I., Chu, B., Hedman, M., Kahl, S., & Klinck, H. (2021). Parsing birdsong with deep audio embeddings. *arXiv preprint arXiv:2108.09203*.

Turian, J., et al. (2022). Hear: Holistic evaluation of audio representations. In *NeurIPS 2021 Competitions and Demonstrations Track* (pp. 125–145).: PMLR.

Van Opzeeland, I., et al. (2010). Acoustic ecology of antarctic pinnipeds. *Marine Ecology Progress Series*, *414*, 267–291.

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, *53*(3), 1–34.

Watkins, W. (2021). Watkins marine mammal sound database.

Wattenberg, M., Viégas, F., & Johnson, I. (2016). How to use t-sne effectively. *Distill*, *1*(10), e2.

Wood, C. M., Gutiérrez, R. J., & Peery, M. Z. (2019). Acoustic monitoring reveals a diverse forest owl community, illustrating its potential for basic and applied ecology. *Ecology*, *100*(9), 1–3.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in neural information processing systems*, *27*.

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.