Highlights

Acoustic Recognition of Individual Animals in the Presence of Unknown Individuals

Lifi Huang, Rohan Clarke, Daniella Teixeira, André Chiaradia, Bernd Meyer

- We show that features learned by models pretrained on bird species data can be transferred to individual classification tasks with minimal effort.
- We define and explore the out-of-distribution classification problem on individual animal vocalisations and address various subtleties and ecological use cases.
- We compile and contribute two additional datasets to facilitate further research on individual acoustic recognition.

Acoustic Recognition of Individual Animals in the Presence of Unknown Individuals

Lifi Huang[a,*], Rohan Clarke[b], Daniella Teixeira[c,d], André Chiaradia[e], Bernd Meyer[a]

[a] *Department of Data Science and AI, Monash University, Melbourne, Victoria, Australia*
[b] *School of Biological Sciences, Monash University, Melbourne, Victoria, Australia*
[c] *School of Biology and Environmental Science, Queensland University of Technology, Brisbane, Queensland, Australia*
[d] *Bush Heritage Australia, Melbourne, Victoria, Australia*
[e] *Conservation Department, Phillip Island Nature Parks, Cowes, Victoria, Australia*

Abstract

Passive acoustic monitoring is firmly established as an effective non-invasive technique for wildlife monitoring. The analysis of animal vocalizations recorded in their natural habitats is commonly used to monitor species occupancy, distribution mapping and community composition. The ability to distinguish between individual animals, however, remains underexplored and presents an exciting opportunity to study individual animal behavior and population demographics in more detail. In this work, we investigate bioacoustic individual-level recognition. In contrast to existing work, we focus on settings where only a subset of the existing population is known and labeled. This is crucial because wildlife populations are constantly changing so that solutions operating only within a known set of individuals are not realistically applicable in the wild. Using two novel datasets, we show that models initially trained to classify only known individuals can also be extended to detect new, previously unseen, individuals that are not part of the training set. We demonstrate that feature extractors pretrained on species classification can be successfully adapted for this task. Extending individual-level recognition to unknown individuals, so-called out-of-distribution classification, is a crucial step towards making individual recognition a realistic possibility in the wild.

*Corresponding author: lifi.huang@monash.edu

Keywords: bioacoustics, open set, feature embeddings, transfer learning, passive acoustic monitoring

1. Introduction

Monitoring individual animals within wildlife populations can facilitate conservation efforts, ecological research and wildlife management. Traditionally, the tracking of individuals has generally been undertaken by physical marking techniques like bird banding [1] or appendage clipping [2]. These individual-level data can then feed into analytical frameworks like mark-recapture studies to estimate population parameters such as population size, abundance, survival rates and movement patterns. However, traditional physical marking approaches have a number of limitations - They are often labor-intensive and costly to implement and potentially harmful to the animals being studied [3, 4, 5, 6]. Furthermore, many physical marking techniques do not endure when the individuals are especially long-lived [7]. In recent years, non-invasive identification techniques have emerged as promising alternatives that address many of the concerns associated with physical marking while still allowing individual-level tracking. This includes photo identification relying on stable external marks or patterns [8] as well as molecular screening of eDNA (e.g. scats or other sources [9]). Bioacoustic monitoring has emerged as an another promising technique and instead uses audio recording devices to capture and analyse animal sounds. The rise of machine learning and deep learning has greatly benefitted the processing and analyzing of the substantial amount of data generated and is widely applied using sophisticated neural network architectures to automate the acoustic detection and recognition tasks [10, 11, 12].

To date, the majority of work using such audio data has focused on species-level monitoring. This has proven valuable for biodiversity assessments, habitat surveys and tracking broad population trends [8, 10]. Yet, few studies exist that analyse and apply bioacoustic recordings for individual-level tracking [10].

Individuals from numerous species are known to produce individually unique vocalisations that serve a variety of biological purposes including mate recognition, territorial neighbor recognition and parent-offspring recognition [13]. Furthermore, vocalisation characteristics may also differ on the individual-level without any adaptive function, for example, due to anatomy.

In bioacoustic monitoring, these vocalisations have the potential to serve as acoustic fingerprints and may offer a new dimension of insights and benefits for wildlife monitoring [10]. The ability to non-invasively obtain large amounts of individual-level data has the potential to facilitate more accurate population estimates, deeper insights into social interactions and movement patterns of specific individuals without the need for physical capture or tagging.

While the theoretical framework suggests promising applications, individual acoustic recognition has not gained as much attention as species-level bioacoustic monitoring and remains in its early stages, primarily due to its inherent complexities and technical challenges [10, 14]. It requires sophisticated techniques to analyse subtle variations in acoustic features that are often not even noticeable for humans, for example deep learning. Importantly, these methods often require large datasets of labelled calls from known individuals for training, which can be impractical, time-consuming and labor-intensive to obtain. This is especially challenging for wild animal populations, where individual identities are typically not already established. As a result, testing individual acoustic recognition has largely been confined to controlled settings where each individual is known, significantly simplifying the recognition task [10, 14]. These proof-of-concept studies typically involve recording a small, fixed set of animals (e.g. a captive population or tagged subset of the total population) under optimal conditions to confirm that the individual-specific acoustic features can be reliably detected and classified.

The transfer of these methods to natural environments presents significant challenges. Wildlife populations are typically both open and dynamic, with individuals routinely entering and leaving the monitored area. This scenario presents a much more difficult problem as, instead of identifying known individuals, a functioning system must now detect and distinguish vocalisations from both known and unknown individuals, which is termed out-of-distribution (OOD) classification. This study explores the challenges of out-of-distribution classification on the individual level. First, we collect and collate two novel datasets for individual acoustic recognition and train a deep learning model to empirically confirm that the vocalisations of these two species are indeed individually distinct and distinguishable by an automatic classifier. Next, we simulate a fluctuating population by withholding a subset of individuals during training and only introducing them during testing. We show that our method can be adjusted to perform OOD classification and to detect these as novel individuals without having been trained

on them. Finally, we explore performance improvements and discuss the ecological applicability of our results.

2. Related works

Prior research has confirmed individually unique vocalisations in a range of species, primarily from birds and mammals [13, 15]. A limited number of studies on automatically classifying individuals has emerged in recent years using a range of different methods including random forests [16], various forms of unsupervised clustering [17], support vector machines [18], neural networks pretrained on images [19] and task-specific neural networks [20, 21]. Notably, the vast majority of studies have focused on a closed-set/in-distribution (ID) scenario (93.3% as surveyed in [10]). The remaining 6.7% works that focused on an OOD scenario have primarily been performed either through manual classification or using traditional machine learning methods [10], with only a single example leveraging the advances in deep learning [22].

OOD classification has garnered significant attention in the deep learning community due to its ability to handle scenarios where models encounter previously unseen classes during inference [23, 24, 25]. The significant amount of existing OOD work in other domains presents a compelling opportunity to extend and adapt these methods to individual wildlife recognition applications. These can be grouped into four overarching groups [25]:

Classification-based methods are methods that directly use a trained model's output, e.g. the class probability to classify OOD samples. The training paradigms for such models can be categorised into two distinct approaches: those that incorporate a dedicated OOD-specific training objective [26, 27] and those that do not [28, 29]. The former explicitly optimises the model's performance on OOD detection tasks, while the latter relies on the model's inherent generalisation capabilities acquired through standard training procedures. Both approaches may incorporate explicit examples of OOD samples during training, disjoint to those used for testing, to facilitate the learning of the ID/OOD separation [30, 31]. Previous work has shown that the efficacy is significantly impacted by the correlations between the OOD samples provided during training and those encountered during testing [32].

Density-based methods use probabilistic models to first characterise the ID data and subsequently identify test samples in low-density regions of the modeled distribution as OOD instances. While theoretically sound, these methods often present significant challenges in terms of training and optimi-

sation complexity. As a result, their empirical performance frequently fails to surpass that of classification-based approaches [25].

Distance-based methods classify samples as OOD explicitly based on a distance measure to stored ID samples. Such methods include those memory-efficient ones that store the centroid or prototypes of ID samples [33] as well as nearest neighbor classifiers, which store all ID samples in memory [34]. The effectiveness of these methods heavily depends on the quality of feature space representation. Many distance-based approaches assume that known classes form tight clusters in the feature space, which may not always be the case. Distance-based methods are especially impacted by high-dimensional feature spaces, where the difference between minimum and maximum distances becomes indiscernible [35] and irrelevant attributes in the feature space can impede the separability of different distributions.

Reconstruction-based methods use an encoder-decoder framework trained on ID data and exploit the fact that reconstruction models trained only on ID data cannot recover OOD data well and leverage this discrepancy in model performance to detect OOD samples. In practice, such models are difficult to train due to the problematic trade-off between reconstruction quality and OOD detection capability, where improving one aspect often comes at the cost of degrading the other. This type of architecture faces inherent tensions between compression and feature preservation, while also struggling with scalability issues in high-dimensional spaces and datasets with many categories [36].

One application area of deep learning that already considers OOD scenarios and techniques in an ecological context is the visual re-identification of individual animals [37, 38]. It involves using computer vision and machine learning algorithms to recognise specific animals across different images or video frames, even when the visual conditions or contexts vary significantly from the original training data. The general approach here is to train a feature embedder on ID samples and to compare the test sample embeddings to those of ID samples. Distance-based methods in the form of similarity scores and learned thresholds are then used to either classify the individuals or label them as new, unseen individuals [39, 40, 41].

3. Material and methods

3.1. Data collection

We used five datasets gathered from open and unconstrained wildlife populations for the performance evaluations of our computational methods in this work (Table 1). The south-eastern red-tailed black cockatoo (*Calyptorhynchus banksii graptogyne*) and little penguin (*Eudyptula minor*) data are novel and were collected by the authors for this work. The little owl (*Athene noctua*), tree pipit (*Anthus trivialis*) and chiffchaff (*Phylloscopus collybita*) data were obtained from [42].

	Number of individuals (known/unknown)	Number of vocalisations (known/unknown)
Red-tailed black cockatoo (female begging call)	16 (12/4)	1785 (1433/352)
Little penguin (exhalant display call)	15 (10/5)	2429 (1958/471)
Little owl (territorial call)	16 (12/4)	952 (749/203)
Tree pipit (song)	10 (7/3)	1364 (1062/302)
Chiffchaff (song)	13 (9/4)	6238 (5075/1163)

Table 1: Summary of audio data used. While individuals are all known during the closed-set experiments, individuals were split into known and unknown subsets for the out-of-distribution experiments. The unknown individuals were chosen such that their vocalisations total roughly 20% of the total data.

The datasets used vary in vocal complexity. Generally, we focus for each species on the call type that is most suitable for individual-level recognition. The red-tailed black cockatoo (RTBC) adult female begging call and the little owl (LO) territorial call consist of a single syllable, while the little penguin (LP) display call consists of an exhalant and inhalant phase. The tree pipit (TP) and chiffchaff (CC) songs are more complex and contain (on average) 11 and 9-24 syllable types repeated in phrases and variably combined respectively [16]. Example spectrograms of the different vocalisations are shown in Figure 1.

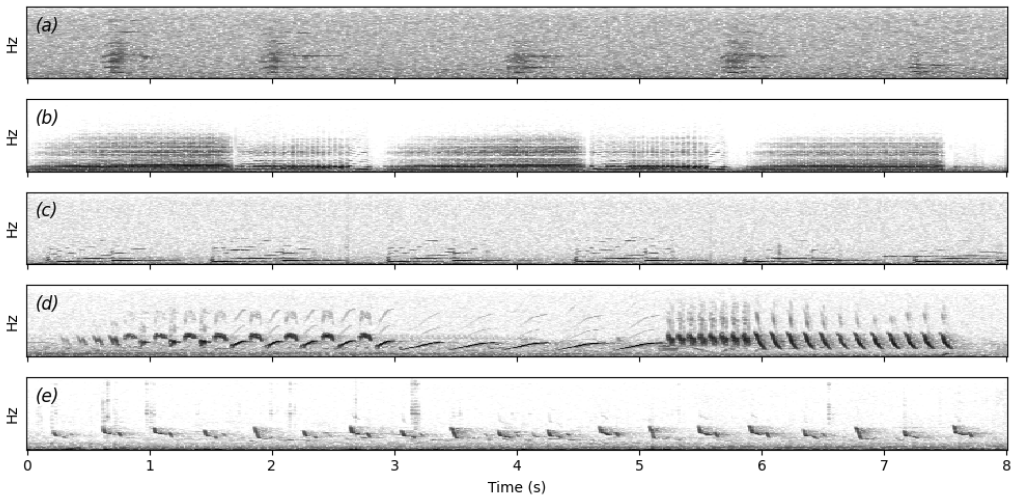

Figure 1: Example spectrograms representing a a) south-eastern red-tailed black cockatoo adult female begging call, b) little penguin display call (3 exhale and 2 inhale components), c) little owl territorial call, d) tree pipit song and e) chiffchaff song.

3.1.1. Little penguins

Little penguins were recorded at St Kilda Breakwater, VIC, Australia (37.86° S, 144.96° E). A combination of Wildlife Acoustics Songmeter Minis and Micros (44.1 kHz, 16-bit, PCM with 6 dB gain) were placed in occupied nests and used to continuously record overnight from 19:00 to 07:00 (all times reported as local time) in the nesting period between April and May 2024. We restrict ourselves to audio from a single night for each nest to minimise the chance of recording unwanted individuals (e.g. if individuals moved during the nest site selection phase). Individual LPs were identified by correlating the vocalisation amplitudes in recordings with the locations of known nesting sites. Calls less than 80% of the maximum amplitude were regarded as originating from neighboring nests or penguins passing by and deemed not clearly identifiable. We note that because the LP nest in pairs and are visually indistinguishable, the vocalisation label we infer is actually a nest-level identity (i.e. a vocalisation from partner 1 or 2 inside nest A is labelled as identity A). However, the distinguishing features of a nest are inherently a combination of characteristics from both individuals and successful nest-level classification would indicate that the combination of traits from individuals in each nest is unique. Consequently, we argue that

classification on a nest-level is comparable to classification on an individual-level in the context of determining the ability to classify individuals.

The vocalisations were extracted individually from the continuous recordings and validated by author LH using Raven Pro 1.6.5. We target the exhale component of the display call (see Figure 1b), as it has exhibited signs for individual distinctiveness [43].

3.1.2. Red-tailed black cockatoos

Audio recordings for the south-eastern red-tailed black-cockatoo were obtained for 16 wild nests in South Australia and Victoria. Most of these recordings were acquired from a large, ongoing nest monitoring program run by the South-eastern Red-tailed Black-Cockatoo Recovery Program [44]. Data from two nests were acquired from an existing database of recordings [45]. Recordings used in the current study were collected between 2017 - 2022.

Of the 16 nests, three were natural nests in tree hollows and 13 were nests in artificial nest boxes. Audio recordings were collected using either Frontier Labs BAR recorders (n = 2 nests), AudioMoth recorders (n = 7 nests) or SWIFT recorders (n = 7 nests). Sound recorders were attached to nest trees or, if that was not possible, a nearby structure. Recorders collected continuous audio for 3 – 4 hours each day, commencing roughly 2.5 hours before sunset, which is the time of peak calling activity at nests. Sampling rate varied by recorder type. BAR recordings were made at 44.1 kHz while AudioMoth and SWIFT recordings were made at 32 kHz. All recordings were made in .wav format.

To build the acoustic dataset, sample calls were labelled and extracted from the audio recordings using Raven Pro 1.6 software. Some audio recordings were first processed using a custom call recogniser to detect the species, while others were processed without the aid of any automation. In both cases, audio recordings were manually inspected in Raven Pro by one observer (DT) and vocalisations were labelled to the level of the call type. Only female begging calls were extracted for the current study. Although red-tailed black-cockatoos give many call types at nests [45], the female begging call was chosen as the best proof-of-concept call type because it is reliably given at active nests, which allows sufficient sample sizes of calls, and is not given by males, which reduces the chance of the observer assigning a call to the wrong individual. Begging calls are highly variable within and between individuals, which may or may not aid individual classification. Red-tailed

black-cockatoos nest in loose colonies, meaning that begging calls from non-target females may be captured in recordings. To minimise the chance of mislabelling, calls were only extracted if they were loud and clear, indicative of a bird close to the microphone. Because nest hollows may be used by different individuals during a nesting period (e.g. after one nest fledges or fails, another bird may nest there), we only extracted calls where we were confident that the nest was being used by the same individual.

3.1.3. Little owls

This dataset comprises Little owls recorded from sunset until midnight between March and April of 2013—2014 in two Central European farmland areas. The territorial calls were recorded for up to 3 minutes after a short playback provocation (1 min) from a maximum of 50 meters. For full details, see [42, 46].

3.1.4. Tree pipits

This dataset comprises songs of tree pipit males that were recorded throughout the entire day from mid-April to mid-July between 2011 and 2013 in the Czech Republic. Only spontaneously singing males were recorded. For full details, see [42, 47].

3.1.5. Chiffchaffs

This dataset comprises songs of chiffchaff males recorded between 2008 and 2011 in the Czech Republic. Only spontaneously singing males were recorded from within about 5–15 m distance. Because the syllable repertoires of males are reported to differ almost completely between two years [48, 49], we only consider the "within-year" CC data from [42] for the context of this work. For full details, see [42].

3.2. Problem formulation

We first show that the individuals from all five species have individually distinct vocalisations in a closed-set (in-distribution) scenario. Consider a labelled training dataset with N samples:

$$\mathcal{D}_{train,closed} = \{(\mathbf{x}_i, \tau_i)\}_{i=1}^{N} \subset \mathcal{X} \times \mathcal{C}$$

where $\mathcal{X}$ is the input space and $\mathcal{C} = \{1, ...C\}$ the set of known classes with each class representing one individual. Given a previously unheard vocalisation, the model should return the probability of it originating from a known

individual i. Formally, given a sample $\mathbf{x}$ from a test dataset in which all M samples are drawn from the same Cartesian product $\mathcal{X} \times \mathcal{C}$ as the training dataset the model should return a probability distribution over the known classes $p(y|\mathbf{x})$.

Next, we extend our problem to an out-of-distribution scenario by withholding a subset of the individuals during training and treating them as novel individuals during the testing phase. Formally, we split the set of previously known classes $\mathcal{C}$ into a new set of known $\mathcal{K}$ and unknown $\mathcal{U}$ classes:

$$\mathcal{K} = \{c \in \mathcal{C} | c <= K\}$$

$$\mathcal{U} = \{c \in \mathcal{C} | c > K\}$$

where K and $U = C - K$ are the number of known and unknown classes, respectively. The new training dataset is then given by

$$\mathcal{D}_{train,open} = \{(\mathbf{x}_i, \tau_i)\}_{i=1}^{N'} \subset \mathcal{X} \times \mathcal{K}$$

Contrary to the closed-set scenario, the test samples may now also originate from unknown individuals, resulting in

$$\mathcal{D}_{test,open} = \{(\mathbf{x}_i, \tau_i)\}_{i=1}^{M'} \subset \mathcal{X} \times (\mathcal{K} \cup \mathcal{U})$$

$$\mathcal{D}_{test,open} = \{(\mathbf{x}_i, \tau_i)\}_{i=1}^{M'} \subset \mathcal{X} \times \mathcal{C}$$

Following the framework defined in [25], we focus on two OOD subtasks with ecological relevance: novelty detection and open-set recognition. Novelty detection requires a model to produce a score $\mathcal{S}(y \in \mathcal{K}|\mathbf{x})$ for a sample $\mathbf{x}$ to belong to the known classes $\mathcal{K}$. A score threshold is then chosen to classify samples as ODD, i.e. as belonging to either $\mathcal{K}$ or $\mathcal{U}$. Open-set recognition additionally requires the model to return a probability distribution over the known classes $p(y|\mathbf{x})$ for all samples assumed to belong to $\mathcal{K}$.

3.3. Closed-set classification

Convolutional neural networks (CNN) have demonstrated robust performance in diverse audio classification tasks, exhibiting their efficacy as a versatile framework for acoustic signal analysis [50]. We train a recent general end-to-end model based on CNNs (AemNet [51]) as the baseline for our closed-set classification experiments. This model skips the conventional spectrogram extraction step and instead converts the raw audio input into a 2D

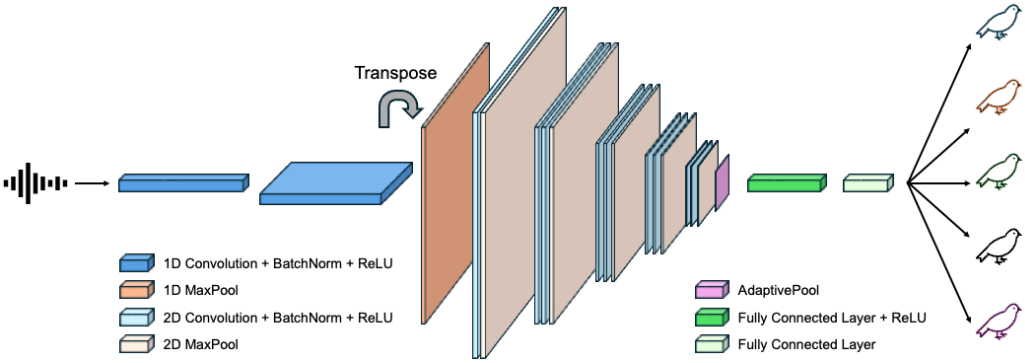

Figure 2: AemNet architecture. The raw audio input is converted through a series of 1D convolutions, batch normalisations, ReLUs and max pooling operations into a 2D tensor, which is in turn fed into a classification CNN.

tensor through a series of trainable 1D convolutions. The result is then fed to a classification CNN (Figure 2).

An adaptive pool in front of the fully connected layers produces a vector of fixed size, allowing the model to incorporate arbitrary length inputs with no architectural changes (the exact architecture can be found in Appendix A). We apply time-masking with a minimum and maximum silent length of 0 and 30% of the total sound length and random gain with a minimum and maximum gain of $\pm$ 6 dB. We start the mixup [52] augmentation process after 5 epochs to reduce confusion in the early stages of learning and randomly sample our mixup lambdas from a Beta-distribution with $\alpha = 5$ and $\beta = 2$. In every epoch, each of these three augmentations is independently applied to a sample with 50% probability.

We then compare the AemNet performance against BirdNET (V2.4) [11] and Google-Perch (V1.4) [12], two CNN-based models widely used for species classification [12, 53, 54]. Although these two models are pretrained on bird species data, they have been shown to perform well on previously unseen bioacoustic classes such as marine mammals and frogs [55]. We use these pretrained models as feature embedders by removing their classification heads and train our own two-layer linear classifier separated by a ReLU on top of the extracted features.

Because each focal species has varying vocal complexities and call durations, we set the respective input audio lengths for AemNet to be longer than the majority of the individual calls, up to a maximum of 5 seconds: RTBC

(2 s), LP (3 s), LO (1 s), TP (5 s), CC (5 s). BirdNET and Google-Perch use constant input lengths of 3 and 5 seconds respectively. Vocalisations that are shorter are randomly zero-padded and longer bouts are cut by selecting a random starting point from which the appropriate length audio is extracted for all models.

All experiments use a 80/20 train/test data split. Note that for the LO, TP and CC data, we deviate from the original work's train/test split to provide a more consistent comparison between all species results. We use a base learning rate of 0.001, cosine annealing and the AdamW optimiser [56] for all experiments. We train AemNet and the linear classifiers for 200 and 100 epochs respectively. We use five-fold stratified cross-validation for all model training and testing and report the respective average performances. Wilcoxon signed-rank tests are conducted for all relevant experimental comparisons. Given the small sample size of five cross-validation runs, we consider paired samples with a p-value of 0.0625 (i.e. the smallest possible p-value) as significant.

3.4. Out-of-distribution classification

We investigate the OOD classification task on individual animals, focussing on novelty detection and open-set recognition. As with the closed-set classification, we use an 80/20 train/test split, but, as mentioned in Section 3.2, additionally exclude a selection of individuals and their vocalisations totalling approximately 20% of the total data to use as the unknown class (see Table 1). This ensures that at test time, we have approximately a 50/50 split between known and unknown test individuals. We retrain each of the three models on the remaining data using the procedure from Section 3.3 and evaluate their OOD classification performances.

For the novelty detection task, we restrict our experiments to post-hoc methods, i.e. output-based methods that do not need modifying of the training procedure and objective (here: categorical cross-entropy (CE) loss) [25]. This property is particularly important in environments where retraining is either not possible [11] or prohibitively expensive.

We detect novel individuals using two methods: The MSP is a commonly used baseline that uses the maximum softmax probability of a model trained on a closed-set classification as the novelty detection score $\mathcal{S}(y \in \mathcal{K}|\mathbf{x}) = max_{y \in \mathcal{K}} p(y|\mathbf{x})$. It has produced state-of-the-art results on various novelty detection benchmarks by leveraging the strong correlation between a model's closed- and open-set performance [57]. By contrast, the Simplified

Hopfield Energy (SHE) method instead uses a simplified version of the Modern Hopfield Network's energy function to measure the discrepancy between ID and OOD samples [29]. Explicitly, it calculates and stores the average of the trained model's ID samples' patterns as $\bar{S}_i \in \mathbb{R}^d$:

$$\bar{S}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} s_{i_j}$$

where $s_{i_j} \in \mathbb{R}^d$ and N_i denote the stored ID patterns and cardinality of class i respectively. The similarity of a test sample $\mathbf{x}$'s pattern $\xi \in \mathbb{R}^d$ to the stored averages is then computed using the dot product and used to determine its novelty.

$$\mathcal{S}(y \in \mathcal{K}|\mathbf{x}) = \text{SHE}(\xi) = \sum_{i}^{K} \xi^T \bar{S}_i$$

Effectively, SHE is simply the average cosine distance between the sample's feature vector and all stored feature vectors of the respective class. We use the model's logits, i.e. the output of the last layer before classification, as the pattern. With its lack of hyperparameters and computational efficiency, this method has resulted in state-of-the-art results on multiple OOD vision recognition benchmarks [29].

Conceptually, it is important to provide the model training with *some* information on what samples outside the known distribution look like if OOD recognition is required. To do so, we adopt Entropic Open-Set loss (EOS) [30, 58]. This loss incorporates additional samples from unknown classes during training to improve the generalisation beyond the closed-set classification task. Intuitively, when presented with an input from an unknown class, a sensible approach is to maintain maximum uncertainty regarding its class membership. This is achieved by enforcing a uniform probability distribution across all known classes, thereby maximising the entropy of the output. We follow the approach in [58] and use a one-hot encoded target vectors $t_i \in \mathbb{R}^K$ for each known class $c \leq K$

$$t_{i,j} = \begin{cases} 1 & j = c \\ 0 & \text{otherwise} \end{cases}$$

and target vectors with identical values for the unknown classes $i > K$:

$$t_{i,j} = \frac{1}{K}$$

To ensure a diverse coverage of distinctly different sounds, we use the ESC-10 [59] dataset as the negative sample class.

We calculate the area under the ROC curve (AU-ROC) as well as the Open-Set Classification Rate (OSCR) [30, 58] to quantify the novelty detection and open-set recognition performances, respectively.[1] We require two different metrics, as the AU-ROC considers only the binary classification task and does not provide information on the known class classification performance. The OSCR is defined as the area under the "Correct Classification Rate (CCR) vs False Positive Rate (FPR)" curve, which, similar to the ROC, can be plotted by varying the θ threshhold from 0 to 1. The CCR and FPR are given by:

$$\mathrm{CCR}(\theta) = \frac{|\{x_n \mid \tau_n \leq K \wedge \tau_n = \underset{1 \leq c \leq K}{\arg\max}(y_{n,c} | y_{n,c} > \theta)\}|}{N_K}$$

$$\mathrm{FPR}(\theta) = \frac{|\{x_n \mid \tau_n > K \wedge \max_{1 \leq c \leq K} y_{n,c} > \theta\}|}{N_U}$$

where x_n denotes the sample n, $y_{n,c}$ the softmax output probability of class c for sample n, and N_K and N_U the total numbers of known and unknown test samples. Informally, the CCR is the fraction of known samples that have been correctly classified with a softmax probability above a threshold θ and the FPR is the fraction of unknown samples that were wrongly classified as a known sample with a softmax probability above the same threshold θ. We use five-fold stratified cross-validation for model training and testing and report the respective average performances as well as their standard deviations. Wilcoxon signed-rank tests are conducted for all relevant experimental comparisons. Given the small sample size of five cross-validation runs, we consider paired samples with a p-value of 0.0625 (i.e. the smallest possible p-value) as significant. Effectively, this p-value means that the performance of the better method was superior for *all* cases of the cross-validation.

[1] We use the roc_auc_score function from scikit-learn 1.3.0 to compute the AU-ROC. To compute the OSCR, we first calculate the CCR and FPR for n equidistant θ-values ranging from the lowest to the highest experimental softmax probabilities, where n is the total number of test samples. We then use the trapezoidal rule to compute the area under the CCR/FPR curve, i.e. the OSCR.

4. Results

4.1. Closed-set classification

All three models were able to perform closed-set classification with very high accuracy for all five species (Table 2) on the individual level (RTBC, LO, TP, CC) and on the nest level (LP), respectively. No significant differences were found between BirdNET and Google-Perch on the simple, one-syllable songs (RTBC, LP, LO). The models pretrained on bird species data performed significantly better than the custom model on LP and LO and insignificantly better on the RTBC vocalisations. For the longer TP and CC songs, the custom model and BirdNET performed similarly and significantly better than Google-Perch in both cases.

	Model	ACC	AU-ROC
Red-tailed black cockatoo	AemNet	95.41±1.10	99.91±0.04
	BirdNET	97.48±0.18	99.95±0.02
	Google-Perch	97.25±0.48	99.95±0.01
Little penguin	AemNet	93.63±0.78	99.76±0.04
	BirdNET	**97.03±0.61**	**99.88±0.05**
	Google-Perch	**95.61±0.86**	**99.82±0.05**
Little owl	AemNet	92.04±1.57	99.61±0.27
	BirdNET	**97.59±0.71**	**99.96±0.02**
	Google-Perch	**98.64±0.97**	**99.96±0.05**
Tree pipit	AemNet	**94.80±1.45**	**99.80±0.07**
	BirdNET	**92.53±1.82**	**99.55±0.19**
	Google-Perch	89.30±1.30	99.05±0.23
Chiffchaff	AemNet	**97.68±0.21**	**99.96±0.01**
	BirdNET	**97.28±0.48**	**99.96±0.01**
	Google-Perch	94.92±0.84	99.82±0.07

Table 2: Closed-set classification results. For each species, AemNet was trained from scratch, while BirdNET and Google-Perch were finetuned using a two-layer linear classifier. Significant results are shown in bold. Significance is only marked for testing between the respective tested method and the lowest performing method. For example, for Little owls, Perch and BirdNet both perform significantly better than AemNet but no assertion is made about their performance relative to each other.

4.2. Out-of-distribution classification

Our experiments showed that the AU-ROC obtained by using the MSP was consistently higher compared to using the SHE. Furthermore, we found that using EOS loss improved both the AU-ROC and the OSCR across all datasets for both pretrained models. AemNet benefited only in some cases from the inclusion of negative samples and its performance degraded on some datasets (Figure 3). We thus focus on the best combinations of methods for

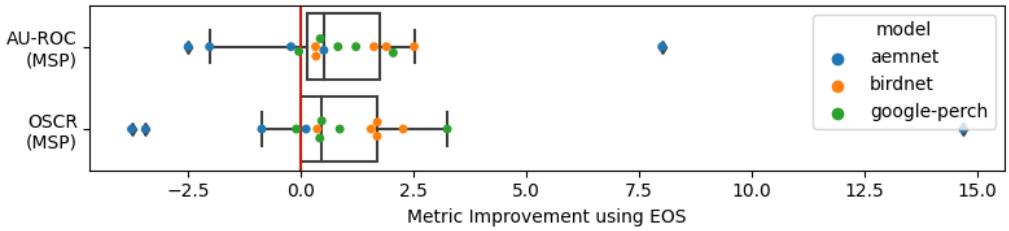

Figure 3: Absolute AU-ROC (MSP) and OSCR (MSP) improvements through training with EOS compared to CE.

the OOD classification results in Table 3. Complete results are available in Appendix B, Table B.6.

BirdNET performed the best across almost all datasets and tasks, while AemNet and Google-Perch were either similar or worse. The more complex TP songs appeared to be the most challenging to classify with lower scores across all models and tasks, with no clear favorite. Considering many of the TP vocalisations were longer than the fixed input length of BirdNET, we repeated our prior experiments with a minor change (Table 4): As we could not increase the input length of BirdNET, we instead decreased the input lengths for both AemNet and Google-Perch to 3 seconds to obtain more comparable OOD classification results (the inputs to Google-Perch were zero-padded to 5 seconds snippets after trimming to ensure model compatibility). Both the AU-ROC as well as the OSCR decreased for Google-Perch and increased for AemNet, albeit insignificantly. Additionally, we retrained our custom model with a 7 second input length (more than 90% of the samples are shorter) and observed a comparable change in performance.

A positive correlation (Pearson correlation coefficient: 0.78) between the closed-set and the open-set performance was found (Figure 4).

	Model + Loss	AU-ROC (MSP)	OSCR (MSP)
Red-tailed black cockatoo	AemNet + CE	85.88±2.10	85.12±2.68
	BirdNET + EOS	**93.59±1.50**	**92.66±1.31**
	Google-Perch + EOS	**92.73±0.62**	**91.55±0.64**
Little penguin	AemNet + EOS	**89.75±1.26**	85.26±2.32
	BirdNET + EOS	**89.64±0.69**	84.82±1.44
	Google-Perch + EOS	87.97±0.97	85.41±1.48
Little owl	AemNet + EOS	84.36±1.83	82.15±1.41
	BirdNET + EOS	**93.03±1.01**	**90.58±0.67**
	Google-Perch + EOS	**92.53±1.29**	**90.06±1.48**
Tree pipit	AemNet + CE	77.53±4.24	77.04±4.27
	BirdNET + EOS	77.78±3.94	72.99±4.69
	Google-Perch + EOS	76.89±1.92	73.67±1.57
Chiffchaff	AemNet + CE	**93.11±0.83**	91.04±0.75
	BirdNET + EOS	**93.63±0.31**	**93.78±0.41**
	Google-Perch + EOS	90.96±0.54	90.13±0.52

Table 3: Out-of-distribution classification results. For each species, AemNet was trained from scratch, while BirdNET and Google-Perch were finetuned using a two-layer linear classifier. The best performances for each model and species are reported. Models trained with an EOS loss used samples of the ESC-10 dataset as the negative class. The AU-ROC is reported for the novelty detection task, while the OSCR is reported for the open-set task. Both metrics are reported using the MSP. Significant results are shown in bold.

5. Discussion and future work

Our work shows that models pretrained on bird species classification can reliably classify vocalisations of known individual birds and detect unknown individuals across five different bird species. Unknown individuals of species with longer and more complex vocalisations were more difficult to detect, hinting at a limitation of conventional input length models. Nevertheless, our results support the use of pretrained models for individual acoustic recognition tasks.

5.1. Closed-set classification

We demonstrated the presence of individually-specific vocal signatures in the red-tailed black cockatoo begging call and exhalant phase of the little

	Model + Loss	AU-ROC (MSP)	OSCR (MSP)
Tree pipit (3 s)	AemNet + CE	80.33±2.35	79.66±2.33
	BirdNET + EOS	77.78±3.94	72.99±4.69
	Google-Perch + EOS	73.30±2.55	67.97±1.97
Tree pipit (5 s)	AemNet + CE	77.53±4.24	77.04±4.27
	Google-Perch + EOS	76.89±1.92	73.67±1.57
Tree pipit (7 s)	AemNet + CE	79.23±2.84	78.80±1.86

Table 4: Pipit out-of-distribution classification results using various input lengths (the 3 second input for Google-Perch was zero-padded to 5 seconds to ensure model compatibility).

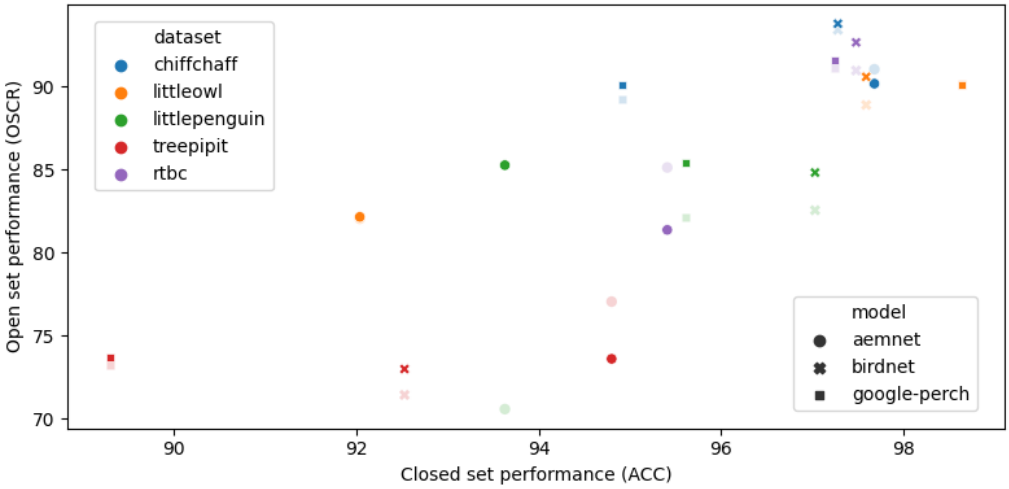

Figure 4: Correlation between closed-set performance (ACC) and open-set performance (OSCR). Results obtained using EOS are shown in bold, while CE results are shown faint.

penguin display call. This adds to the growing list of taxa with confirmed individually distinct vocalisations and establishes the foundation for subsequent analyses of intra-species vocalisation patterns.

Our experiments reveal an interesting finding regarding the efficacy of pretrained deep learning models compared to custom-built models for individual acoustic classification. The strong performance of the pretrained models demonstrates that transfer learning from species classification to individual

classification works surprisingly well, even without extensive finetuning. This may appear counter-intuitive: During pretraining of these feature embedders, recordings from different individuals of the same species were designated as the same class and differences between individuals thus explicitly neglected[2]. It is therefore particularly interesting that intra-individual differences were retained in the embeddings generated by these models.

One possible explanation may be that the acoustic features needed to distinguish between species are generally similar to those that vary between individuals within a species. If this is correct, feature embedders trained on a broad variety of species, such as these, can be expected to develop feature sets powerful enough for individual-level classification. An alternative explanation may be that sufficient residual variation between individuals remains in the embeddings per chance, not because this is functionally equivalent to what is needed for species classification. That such variation could be used for classification tasks beyond the species level would be consistent with the fact that even randomly defined features can deliver surprisingly accurate classification results [61].

Notwithstanding these theoretical considerations, the ability to use pretrained models offers significant advantages from a practical perspective in terms of resource efficiency and accessibility. Not only do pretrained models reduce the computational resources, they are also considerably easier to use and to adapt for non-domain experts without extensive machine learning expertise. This enables a broader range of scientists to leverage deep learning in their acoustic monitoring work, including those working in individual identification.

While the authors in [55] showed that models trained on bird vocalisations learn a wide range of acoustic features generalisable to novel categories, our work demonstrates that these features also generalise to intra-species categories, at least in the case of our test species. As our test species are not distinguished in any obvious way, it stands to reasons that these features may

[2]Both pre-trained models were primarily trained using Xeno-Canto data [60]. While Xeno-Canto contains audio samples for all our five target species, only 16 LP and 74 RTBC recordings were available at the time of pretraining, while over 1000 recordings were available for LO, TP and CC. The Xeno-Canto LP recordings do not distinguish between the inhalant and exhalant phase and the RTBC data labels do not specify the call types, as the majority of recordings were uploaded before the call types were defined in [45].

be useful for individual recognition in a much wider range of bird species and possibly other wildlife.

5.2. Out-of-distribution classification

Our work differs from the only other work using deep learning in a bioacoustic out-of-distribution setting [22] in several ways. Instead of building and training a task-specific model, we showed that existing models pretrained on closed-set species classification can be extended to individual out-of-distribution classification tasks with strong performances and minimal effort. We applied this approach approach across more individuals and several species, demonstrating its viability and robustness. We formalized two out-of-distribution classification tasks and used task-relevant metrics to quantify our model performances, allowing for comparison between models as well as improvement suggestions. Our results on these tasks demonstrate two key findings:

First, exposing the models to negative samples during training improved both the AU-ROC as well as the OSCR in the majority of our cases. Interestingly, AemNet's TP and CC performance declined with the inclusion of negative samples in the training process. This may be due to the fact that AemNet first needed to learn the new feature set of the more complex songs and the negative samples introduced too many confounding factors to an already more difficult task, whereas the pretrained methods were only required to learn a classification task from a known feature set. We note that the classes used were entirely unrelated to the task and that the effectiveness of this approach may be dependent on the nature and diversity of the negative class. As a result, we propose that in addition to targeting labelled individuals during the audio recording process, additional unlabelled individuals should be recorded and included as negative samples during training as well. These are individuals whose exact identity cannot be inferred but are known to be disjoint to the set of known training individuals. This can be achieved e.g. by recording visually different individuals (e.g. unbanded individuals) or recording individuals in a different location (e.g. in a different section of the recording site). Because the EOS loss does not discriminate between the negative classes, the only requirement is that the negative class individuals are not part of the supervised training data. We note that this data is far easier to obtain as no individual labelling is required. This usage of more relevant out-of-distribution inputs may help the models to learn more subtle but relevant differences.

Second, the pretrained models performing comparable to or better than our model trained from scratch on the closed-set task suggests they may have developed more robust and generalisable features for individual identification, which is particularly important for the out-of-distribution classification task. Our results on the simple vocalisations align with this conjecture, however, as in the closed-set case, their performances on the TP vocalisations were similar to or worse than that of AemNet. One reason this may have occurred is that the smaller context window of BirdNET (3 s) may not have been sufficient to cover all of the relevant individuality information of a sample, where the median and average sample lengths are 3.5 and 4.0 seconds respectively. Our results from the extended evaluation of the TP vocalisations (Table 4) showed, however, an upper-bound in performance, suggesting that conventional fixed length architectures may be suboptimal for capturing the full spectro-temporal complexity of long multi-syllable bird songs in the context of individual out-of-distribution classification.

Considering the strong closed-set performances of these models, long-range dependencies important for individuality may have been discarded during training. More flexible computational frameworks that can handle long-range dependencies, such as transformers [62] and structured state space sequence models [63], may be needed to better handle such dependencies within more complex songs.

5.3. Ecological applicability

Our results demonstrate robust performance in cases where each individual is known, i.e. closed-set recognition tasks. However, the real-world applicability of these findings is limited, as in most cases, conservation monitoring more closely resembles out-of-distribution problems, i.e. where not all individuals are known.

While the techniques presented in this work provide a solid foundation for identifying known individuals and differentiating between known and unknown individuals, we posit that integrating capabilities to count individuals could substantially enhance the impact of these models. The ability to produce count estimates of unknown individuals in a recording, effectively providing local abundance estimates around a recording device, would be a substantial step forward in estimating population trends from bioacoustic data. This is crucial not only for biodiversity and environmental assessments, but also conservation planning, resource management and decision-

making. Source count estimation of unlabelled wild populations presents inherent challenges, primarily due to the number of individuals. The transition from a controlled setting, where the system is exposed to a limited number of individuals, to inference in the wild, where the number of unknown individuals may be much greater, is not straightforward. A system performing well on five unknown individuals in the lab may not be guaranteed to perform well on a large number of unknown individuals in the field. However, addressing these obstacles could yield transformative results.

There are two primary use cases for source count estimation in an out-of-distribution scenario. The first is to confidently identify, or separate, individuals in order to measure abundance at a point in time. Actual identity is not important *per se* and any slow change to vocal signatures over longer time scales is not problematic. The second use case is to identify and label individuals, so that capture-recapture models could be used to measure, for instance, population size. Shift in individual vocal signatures, for example due to ageing and growth, may be problematic in this case. Further study is required, specifically whether domain adaptation techniques can address this.

When individuals are vocally unique, one approach to obtain source count estimates could leverage the strong performance of pretrained models. As the pretrained feature spaces contain sufficient information to distinguish individuals, it appears likely that clustering in the feature space can be used to (approximately) count individuals. Specifically, deep clustering techniques could be employed, wherein the pretrained embeddings are used to feed into a clustering objective rather than a classification loss. One would expect that similar vocalisations (i.e. those originating from the same individual) cluster together and separate from dissimilar ones (i.e. those originating from different individuals). In an ideal model, the number of clusters generated would represent the number of sources. An alternative to clustering could be density estimation in the feature space. Different to many existing methods of estimating abundance from sound (e.g. cue counting [64]), these methods do not rely on the number of detected vocalisations, but rather the differences in their content. The strength of this approach is that it does not require additional, difficult to obtain (and often unreliable) information, such as cue rates. We intend to pursue this idea in future work by training such a network on our existing labeled dataset and testing its efficacy in the field by comparing this bioacoustic estimate to one obtained by another method, such as cue counting [64], counting using unmanned aerial vehicles [65, 66]

or manually counting the individuals within the recording area.

By focusing research efforts on refining and implementing such techniques, we can potentially develop more comprehensive and accurate models that not only classify and separate individuals, but also provide crucial population-level insights. These advancements could significantly enhance our understanding of ecosystem dynamics, support conservation efforts and contribute to more informed ecological decision-making processes.

Availability

Our code will be available at `https://github.com/lifihuang/individual_acoustic_recognition_ood`. Our data will be available at Zenodo.

Acknowledgements

Research on little penguins was conducted under Monash University animal ethics approval SBS_39695 and Parks Victoria access approval AA-0000400. Research on red-tailed black cockatoos was conducted under University of Queensland ethics approval SBS/076/17/VIC.

We would like to thank Flossy Sperring for assisting in the data collection of the little penguin vocalisations. For contributing data for the red-tailed black-cockatoo, we thank BirdLife Australia, the Victorian Government, the University of Queensland and the South-eastern Red-tailed Black-Cockatoo Recovery Team. BM would like to thank Leo op den Brouw for the conversations that inspired this work.

Author contributions

Lifi Huang: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Resources; Software; Validation; Visualization; Writing – original draft **Rohan Clarke**: Conceptualization; Methodology; Project administration; Resources; Supervision; Writing – review and editing **Daniella Teixeira**: Data curation; Writing – original draft **André Chiaradia**: Conceptualization; Methodology; Project administration; Supervision; Writing – review and editing **Bernd Meyer**: Conceptualization; Funding acquisition; Methodology; Project administration; Resources; Supervision; Writing – review and editing

Declaration of competing interest

None

Appendix A. AemNet architecture

AemNet
Conv1D(1, 64, 9, 2), BatchNorm1D, ReLU Conv1D(64, 128, 5, 2), BatchNorm1D, ReLU
MaxPool1D(80, padding=4)
Conv2D(1, 16, (3,3), 1), BatchNorm2D, ReLU MaxPool2D((2,2))
Conv2D(16, 32, (3,3), 1), BatchNorm2D, ReLU Conv2D(32, 32, (3,3), 1), BatchNorm2D, ReLU MaxPool2D((2,2))
Conv2D(32, 64, (3,3), 1), BatchNorm2D, ReLU Conv2D(64, 64, (3,3), 1), BatchNorm2D, ReLU MaxPool2D((2,2))
Conv2D(64, 128, (3,3), 1), BatchNorm2D, ReLU Conv2D(128, 128, (3,3), 1), BatchNorm2D, ReLU MaxPool2D((2,2))
Conv2D(128, 256, (3,3), 1), BatchNorm2D, ReLU Conv2D(256, 256, (3,3), 1), BatchNorm2D, ReLU MaxPool2D((2,2))
AdaptivePool2D((1,1))
Dropout(0.2)
Conv2D(256, 128, (1,1), 1), ReLU
Conv2D(128, num_classes, (1,1), 1)

Table A.5: AemNet architecture. Unless stated otherwise, default PyTorch 2.0.1 argument order and values are used for all functions.

Appendix B. Further results

	Model + Loss	AU-ROC (MSP)	AU-ROC (SHE)	OSCR (MSP)
Red-tailed black cockatoo	AemNet + CE	85.88±2.10	63.73±8.78	85.12±2.68
	AemNet + EOS	83.38±2.85	82.62±0.70	81.26±2.39
	BirdNET + CE	93.27±1.68	88.08±6.72	90.96±1.48
	BirdNET + EOS	93.59±1.50	90.89±1.79	92.66±1.31
	Google-Perch + CE	92.77±0.30	91.10±0.95	91.13±0.59
	Google-Perch + EOS	92.73±0.62	90.18±0.55	91.55±0.64
Little penguin	AemNet + CE	81.74±2.47	54.31±10.99	70.58±3.63
	AemNet + EOS	89.75±1.26	83.28±4.70	85.26±2.32
	BirdNET + CE	87.14±0.79	55.96±3.27	82.55±1.57
	BirdNET + EOS	89.64±0.69	75.53±2.75	84.82±1.44
	Google-Perch + CE	85.92±1.07	65.03±6.04	82.16±0.56
	Google-Perch + EOS	87.97±0.97	71.12±3.71	85.41±1.48
Little owl	AemNet + CE	83.85±2.83	79.59±5.60	82.03±3.29
	AemNet + EOS	84.36±1.83	76.44±4.80	82.15±1.41
	BirdNET + CE	91.41±1.47	87.63±2.43	88.88±1.50
	BirdNET + EOS	93.03±1.01	87.93±3.10	90.58±0.67
	Google-Perch + CE	91.30±1.81	84.11±4.86	90.16±1.81
	Google-Perch + EOS	92.53±1.29	83.35±2.16	90.06±1.48
Tree pipit	AemNet + CE	77.53±4.24	77.69±3.79	77.04±4.27
	AemNet + EOS	75.50±7.17	73.37±8.33	73.60±6.66
	BirdNET + CE	75.89±2.17	68.22±1.54	71.43±2.96
	BirdNET + EOS	77.78±3.94	69.41±2.44	72.99±4.69
	Google-Perch + CE	76.07±2.19	75.99±1.26	73.21±1.29
	Google-Perch + EOS	76.89±1.92	76.62±3.33	73.67±1.57
Chiffchaff	AemNet + CE	93.11±0.83	91.40±2.74	91.04±0.75
	AemNet + EOS	92.89±1.01	90.64±1.99	90.17±1.47
	BirdNET + CE	93.29±0.31	91.68±0.72	93.41±0.22
	BirdNET + EOS	93.63±0.31	91.32±1.00	93.78±0.41
	Google-Perch + CE	90.53±0.50	91.51±0.51	89.26±0.48
	Google-Perch + EOS	90.96±0.54	91.21±0.88	90.13±0.52

Table B.6: Out-of-distribution classification results. For each species, AemNet was trained from scratch, while BirdNET and Google-Perch were finetuned using a two-layer linear classifier. All models were trained with a CE and an EOS loss using samples of the ESC-10 dataset as the negative class. The AU-ROCs based on the MSP as well as the SHE are reported for the novelty detection task, while the OSCR is reported for the open-set task.

References

[1] B. Calvo, R. W. Furness, A review of the use and the effects of marks and devices on birds, Ringing & Migration 13 (3) (1992) 129–151. doi:10.1080/03078698.1992.9674036.
URL https://doi.org/10.1080/03078698.1992.9674036

[2] A. Hussek, C. Graf, K. Hackländer, Are ear notches an effective tool for monitoring individual rhino?, Pachyderm 60 (2019) 55–66. doi:10.69649/pachyderm.v60i.34.

[3] M. Griesser, N. A. Schneider, M.-A. Collis, A. Overs, M. Guppy, S. Guppy, N. Takeuchi, P. Collins, A. Peters, M. L. Hall, Causes of Ring-Related Leg Injuries in Birds – Evidence and Recommendations from Four Field Studies, PLOS ONE 7 (12) (2012) e51891.
URL https://doi.org/10.1371/journal.pone.0051891

[4] M. Mccarthy, K. Parris, Clarifying the effect of toe clipping on frogs with Bayesian statistics, Journal of Applied Ecology 41 (2004) 780–786. doi:10.1111/j.0021-8901.2004.00919.x.

[5] P. Fallow, A. Chiaradia, Y. Ropert-Coudert, A. Kato, R. Reina, Flipper Bands Modify the Short-Term Diving Behavior of Little Penguins, Journal of Wildlife Management 73 (2009) 1348–1354. doi:10.2193/2008-146.

[6] X. Meyer, A. J. J. MacIntosh, A. Kato, A. Chiaradia, Y. Ropert-Coudert, Hydrodynamic handicaps and organizational complexity in the foraging behavior of two free-ranging penguin species, Animal Biotelemetry 3 (1) (2015) 25. doi:10.1186/s40317-015-0061-8.
URL https://doi.org/10.1186/s40317-015-0061-8

[7] S. M. Baylis, D. Drynan, N. Clarke, M. Van, P. Sunnucks, R. H. Clarke, Estimates of wear rates in metal bird bands, with applications for survival and movement models of marked individuals, Journal of Field Ornithology 89 (4) (2018) 393–406. doi:https://doi.org/10.1111/jofo.12268.
URL https://doi.org/10.1111/jofo.12268

[8] D. Tuia, B. Kellenberger, S. Beery, B. R. Costelloe, S. Zuffi, B. Risse, A. Mathis, M. W. Mathis, F. van Langevelde, T. Burghardt, R. Kays,

H. Klinck, M. Wikelski, I. D. Couzin, G. van Horn, M. C. Crofoot, C. V. Stewart, T. Berger-Wolf, Perspectives in machine learning for wildlife conservation, Nature Communications 13 (1) (2022) 1–15. doi:10.1038/s41467-022-27980-y.

[9] J. Fediajevaite, V. Priestley, R. Arnold, V. Savolainen, Meta-analysis shows that environmental DNA outperforms traditional surveys, but warrants better reporting standards, Ecology and Evolution 11 (9) (2021) 4803–4815. doi:https://doi.org/10.1002/ece3.7382. URL https://doi.org/10.1002/ece3.7382

[10] E. Knight, T. Rhinehart, D. R. de Zwaan, M. J. Weldy, M. Cartwright, S. H. Hawley, J. L. Larkin, D. Lesmeister, E. Bayne, J. Kitzes, Individual identification in acoustic recordings (2024). doi:10.1016/j.tree.2024.05.007.

[11] S. Kahl, C. M. Wood, M. Eibl, H. Klinck, BirdNET: A deep learning solution for avian diversity monitoring, Ecological Informatics 61 (November 2020) (2021) 101236. doi:10.1016/j.ecoinf.2021.101236. URL https://doi.org/10.1016/j.ecoinf.2021.101236

[12] T. Denton, S. Wisdom, J. R. Hershey, Improving Bird Classification With Unsupervised Sound Separation, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings 2022-May (2022) 636–640. arXiv:2110.03209, doi:10.1109/ICASSP43922.2022.9747202.

[13] N. V. Carlson, E. M. K. Kelly, I. Couzin, Individual vocal recognition across taxa: A review of the literature and a look into the future: Individual vocal recognition, Philosophical Transactions of the Royal Society B: Biological Sciences 375 (1802) (2020). doi:10.1098/rstb.2019.0479.

[14] D. Stowell, Computational bioacoustics with deep learning: a review and roadmap, PeerJ (2022) 1–46arXiv:2112.06725, doi:10.7717/peerj.13152.

[15] P. Linhart, M. Mahamoud, I. Dan, D. T. Blumstein, The potential for acoustic individual identification in mammals, Mammalian Biology (March) (2022). doi:10.1007/s42991-021-00222-2. URL https://doi.org/10.1007/s42991-021-00222-2

[16] D. Stowell, T. Petrusková, M. Šálek, P. Linhart, Automatic acoustic identification of individuals in multiple species: Improving identification across recording conditions, Journal of the Royal Society Interface 16 (153) (2019). doi:10.1098/rsif.2018.0940.

[17] D. J. Clink, H. Klinck, Unsupervised acoustic classification of individual gibbon females and the implications for passive acoustic monitoring, Methods in Ecology and Evolution 12 (2) (2021) 328–341. doi:10.1111/2041-210X.13520.

[18] D. J. Clink, M. C. Crofoot, A. J. Marshall, Application of a semi-automated vocal fingerprinting approach to monitor Bornean gibbon females in an experimentally fragmented landscape in Sabah, Malaysia, Bioacoustics 28 (3) (2019) 193–209. doi:10.1080/09524622.2018.1426042.
URL http://doi.org/10.1080/09524622.2018.1426042

[19] M. Trapanotto, L. Nanni, S. Brahnam, X. Guo, Convolutional Neural Networks for the Identification of African Lions from Individual Vocalizations, Journal of Imaging 8 (4) (2022). doi:10.3390/jimaging8040096.

[20] K. Martin, O. Adam, N. Obin, V. Dufour, Rookognise: Acoustic detection and identification of individual rooks in field recordings using multi-task neural networks, Ecological Informatics 72 (2022) 1–39. doi:10.1016/j.ecoinf.2022.101818.

[21] H. Bocaccio, M. Domínguez, B. Mahler, J. C. Reboreda, G. Mindlin, Identification of dialects and individuals of globally threatened yellow cardinals using neural networks, Ecological Informatics 78 (May) (2023) 102372. doi:10.1016/j.ecoinf.2023.102372.
URL https://doi.org/10.1016/j.ecoinf.2023.102372

[22] C. L. Bedoya, L. E. Molles, Acoustic censusing and individual identification of birds in the wild, bioRxiv (19) (2021) 2021.10.29.466450.

[23] A. Bendale, T. E. Boult, Towards Open Set Deep Networks, in: Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on, 2015.

[24] M. Salehi, D. Hendrycks, A Unified Survey on Anomaly , Novelty , Open-Set , and Out- of-Distribution Detection : Solutions and Future Challenges (2022). arXiv:arXiv:2110.14051v5.

[25] J. Yang, K. Zhou, Y. Li, Z. Liu, Generalized Out-of-Distribution Detection : A Survey, arXiv preprint (Id) (2024). arXiv:arXiv:2110.11334v3.

[26] Y.-C. Hsu, Y. Shen, H. Jin, Z. Kira, Generalized {ODIN:} Detecting Out-of-distribution Image without Learning from Out-of-distribution Data, CVPR (2020).
URL https://arxiv.org/abs/2002.11297

[27] J. Bitterwolf, A. Meinke, M. Hein, Certifiably Adversarially Robust Detection of Out-of-Distribution Data, NeurIPS (2020).
URL https://arxiv.org/abs/2007.08473

[28] S. Liang, Y. Li, R. Srikant, Enhancing the reliability of out-of-distribution image detection in neural networks, in: 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings, 2018, pp. 1–15. arXiv:1706.02690.

[29] J. Zhang, Q. Fu, X. Chen, L. Du, Z. Li, G. Wang, X. Liu, S. Han, D. Zhang, Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy, ICLR (2023) 1–19.

[30] A. R. Dhamija, G. Manuel, T. E. Boult, Reducing Network Agnostophobia (NeurIPS) (2018). arXiv:arXiv:1811.04110v2.

[31] D. Hendrycks, M. Mazeika, T. G. Dietterich, Deep Anomaly Detection with Outlier Exposure, ICLR (2019).
URL http://arxiv.org/abs/1812.04606

[32] A. Shafaei, M. Schmidt, J. J. Little, A less biased evaluation of out-of-distribution sample detectors, 30th British Machine Vision Conference 2019, BMVC 2019 (2020). arXiv:1809.04729.

[33] K. Lee, K. Lee, H. Lee, J. Shin, A simple unified framework for detecting out-of-distribution samples and adversarial attacks, Advances in Neural Information Processing Systems 2018-Decem (Lid) (2018) 7167–7177. arXiv:1807.03888.

[34] Y. Sun, Y. Ming, X. Zhu, Y. Li, Out-of-Distribution Detection with Deep Nearest Neighbors, Proceedings of Machine Learning Research 162 (2022) 20827–20840. arXiv:2204.06507.

[35] K. S. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, When Is "Nearest Neighbor" Meaningful?, in: International Conference on Database Theory, 1999.
URL https://api.semanticscholar.org/CorpusID:206634099

[36] R. Roady, T. L. Hayes, R. Kemker, A. Gonzales, C. Kanan, Are open set classification methods effective on large-scale datasets?, PloS one 15 (9) (2020) e0238302. doi:10.1371/journal.pone.0238302.

[37] L. Adam, V. Cermak, K. Papafitsoros, L. Picek, WildlifeReID-10k: Wildlife re-identification dataset with 10k individual animals, arXiv (2024) 1–12arXiv:arXiv:2406.09211v2.

[38] R. Shinoda, PetFace: A Large-Scale Dataset and Benchmark for Animal Identification, in: ECCV, 2024. arXiv:arXiv:2407.13555v2.

[39] V. Cermak, L. Picek, L. Adam, K. Papafitsoros, WildlifeDatasets: An open-source toolkit for animal re-identification, arXiv (2023) 1–12arXiv:arXiv:2311.09118v2.

[40] O. Wahltinez, S. J. Wahltinez, An open- source general purpose machine learning framework for individual animal re- identification using few-shot learning, Methods in Ecology and Evolution (15) (2024) 373–387. doi:10.1111/2041-210X.14278.

[41] B. Wang, X. Li, X. An, W. Duan, Y. Wang, D. Wang, J. Qi, Open-Set Recognition of Individual Cows Based on Spatial Feature Transformation and Metric Learning., Animals : an open access journal from MDPI 14 (8) (apr 2024). doi:10.3390/ani14081175.

[42] P. Stowell, D., Petrusková, T., Šálek, M., & Linhart, Datasets for automatic acoustic identification of individual birds (0.1) (2018). doi: https://doi.org/10.5281/zenodo.1413495.

[43] J. Burger Fadely, Vocal Recognition by Little Penguins (Eudyptula minor) on Phillip Island, Victoria, Australia, Masters thesis, San Jose State University (1991).

[44] South-eastern Red-tailed Black-Cockatoo National Recovery Team. URL https://www.redtail.com.au/recovery-team.html

[45] D. Teixeira, R. Hill, M. Barth, M. Maron, B. J. van Rensburg, Nest-associated vocal behaviours of the south-eastern red-tailed black cockatoo, Calyptorhynchus banksii graptogyne, and the Kangaroo Island glossy black cockatoo, C. lathami halmaturinus, Austral Ecology 45 (7) (2020) 990–1006. doi:10.1111/aec.12921.

[46] P. Linhart, M. Šálek, The assessment of biases in the acoustic discrimination of individuals, PLoS ONE 12 (5) (2017) 1–16. doi:10.1371/journal.pone.0177206.

[47] T. Petrusková, I. Pišvejcová, A. Kinštová, T. Brinke, A. Petrusek, Repertoire-based individual acoustic monitoring of a migratory passerine bird with complex song as an efficient tool for tracking territorial dynamics and annual return rates, Methods in Ecology and Evolution 7 (3) (2016) 274–284. doi:10.1111/2041-210X.12496.

[48] A. Průchová, P. Jaška, P. Linhart, Cues to individual identity in songs of songbirds: testing general song characteristics in Chiffchaffs Phylloscopus collybita, Journal of Ornithology 158 (4) (2017) 911–924. doi:10.1007/s10336-017-1455-6.

[49] L. Ptacek, L. MacHlica, P. Linhart, P. Jaska, L. Muller, Automatic recognition of bird individuals on an open set using as-is recordings, Bioacoustics 25 (1) (2016) 55–73. doi:10.1080/09524622.2015.1089524.

[50] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, K. Wilson, CNN Architectures for Large-Scale Audio Classification, in: International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.
URL https://arxiv.org/abs/1609.09430

[51] P. Lopez-Meyer, J. A. Del Hoyo Ontiveros, H. Lu, G. Stemmer, Efficient end-to-end audio embeddings generation for audio classification on target applications, ICASSP, IEEE International Conference on Acoustics,

Speech and Signal Processing - Proceedings 2021-June (2021) 601–605. doi:10.1109/ICASSP39728.2021.9414229.

[52] H. Zhang, M. Cisse, Y. Dauphin, D. Lopez-Paz, mixup: Beyond Empirical Risk Management, arXiv (2018) 1–13arXiv:arXiv:1710.09412v2.

[53] D. Funosas, L. Barbaro, L. Schillé, A. Elger, B. Castagneyrol, M. Cauchoix, Assessing the potential of BirdNET to infer European bird communities from large-scale ecoacoustic data, Ecological Indicators 164 (May) (2024). doi:10.1016/j.ecolind.2024.112146.

[54] C. M. Wood, H. Klinck, M. Gustafson, J. J. Keane, S. C. Sawyer, R. J. Gutiérrez, M. Z. Peery, Using the ecological significance of animal vocalizations to improve inference in acoustic monitoring programs, Conservation Biology 35 (1) (2021) 336–345. doi:https://doi.org/10.1111/cobi.13516.
URL https://conbio.onlinelibrary.wiley.com/doi/abs/10.1111/cobi.13516

[55] B. Ghani, T. Denton, S. Kahl, H. Klinck, Global birdsong embeddings enable superior transfer learning for bioacoustic classification (2023) 1–18arXiv:arXiv:2307.06292v2.

[56] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization (2019). arXiv:1711.05101.
URL https://arxiv.org/abs/1711.05101

[57] S. Vaze, K. Han, A. Vedaldi, A. Zisserman, Open-Set Recognition: a Good Closed-Set Classifier is All You Need? (2021) 1–24arXiv:2110.06207.
URL http://arxiv.org/abs/2110.06207

[58] A. Palechor, G. Manuel, Large-Scale Open-Set Classification Protocols for ImageNet (2023). arXiv:arXiv:2210.06789v2.

[59] K. J. Piczak, ESC: Dataset for environmental sound classification, MM 2015 - Proceedings of the 2015 ACM Multimedia Conference (2015) 1015–1018doi:10.1145/2733373.2806390.

[60] Xeno-canto, Xeno-Canto Foundation and Naturalis Biodiversity Center.
URL https:xeno-canto.org

[61] A. Dempster, F. Petitjean, G. I. Webb, ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels, Data Mining and Knowledge Discovery 34 (5) (2020) 1454–1495. doi:10.1007/s10618-020-00701-z.
URL https://doi.org/10.1007/s10618-020-00701-z

[62] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need (2017). arXiv:1706.03762.
URL https://arxiv.org/abs/1706.03762

[63] A. Gu, K. Goel, C. Ré, Efficiently Modeling Long Sequences with Structured State Spaces (2022). arXiv:2111.00396.
URL https://arxiv.org/abs/2111.00396

[64] T. A. Marques, L. Thomas, S. W. Martin, D. K. Mellinger, J. A. Ward, D. J. Moretti, D. Harris, P. L. Tyack, Estimating animal population density using passive acoustics, Biological Reviews 88 (2) (2013) 287–309. doi:https://doi.org/10.1111/brv.12001.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/brv.12001

[65] J. C. Hodgson, S. M. Baylis, R. Mott, A. Herrod, R. H. Clarke, Precision wildlife monitoring using unmanned aerial vehicles, Scientific Reports 6 (March) (2016) 1–7. doi:10.1038/srep22574.
URL http://dx.doi.org/10.1038/srep22574

[66] B. G. Weinstein, L. Garner, V. R. Saccomanno, A. Steinkraus, A. Ortega, K. Brush, G. Yenni, A. E. McKellar, R. Converse, C. D. Lippitt, A. Wegmann, N. D. Holmes, A. J. Edney, T. Hart, M. J. Jessopp, R. H. Clarke, D. Marchowski, H. Senyondo, R. Dotson, E. P. White, P. Frederick, S. K. M. Ernest, A general deep learning model for bird detection in high-resolution airborne imagery., Ecological applications : a publication of the Ecological Society of America 32 (8) (2022) e2694. doi:10.1002/eap.2694.