Leveraging Transfer Learning and Active Learning for Sound Event Detection in Passive Acoustic Monitoring of Wildlife

Hannes Kath[1,2], **Patricia P. Serafini**[3,4], **Ivan B. Campos**[3,5], **Thiago S. Gouvêa**[1,2], **Daniel Sonntag**[1,2]

[1]Interactive Machine Leraning, German Research Center for Artificial Intelligence (DFKI), Oldenburg, Germany
[2]Applied Artificial Intelligence, Carl von Ossietzky University of Oldenburg, Germany
[3]National Center for Wild Bird Conservation and Research (CEMAVE), Chico Mendes Institute for Biodiversity Conservation (ICMBio), Brazil
[4]Universidade Federal de Santa Catarina (UFSC), Brazil
[5]Departamento de Biologia Geral, Universidade Federal de Minas Gerais (UFMG), Brazil
{hannes.kath, thiago.gouvea, daniel.sonntag}@dfki.de
{patricia.serafini, ivan.campos}@icmbio.gov.br

Abstract

Passive Acoustic Monitoring (PAM) has emerged as a pivotal technology for wildlife monitoring, generating vast amounts of acoustic data. However, the successful application of machine learning methods for sound event detection in PAM datasets heavily relies on the availability of annotated data, which can be laborious to acquire. In this study, we investigate the effectiveness of transfer learning and active learning techniques to address the data annotation challenge in PAM. Transfer learning allows us to use pre-trained models from related tasks or datasets to bootstrap the learning process for sound event detection. Furthermore, active learning promises strategic selection of the most informative samples for annotation, effectively reducing the annotation cost and improving model performance. We evaluate an approach that combines transfer learning and active learning to efficiently exploit existing annotated data and optimize the annotation process for PAM datasets. Our transfer learning observations show that embeddings produced by BirdNet, a model trained on high signal-to-noise recordings of bird vocalisations, can be effectively used for predicting anurans in PAM data: a linear classifier constructed using these embeddings outperforms the benchmark by 21.7%. Our results indicate that active learning is superior to random sampling, although no clear winner emerges among the strategies employed. The proposed method holds promise for facilitating broader adoption of machine learning techniques in PAM and advancing our understanding of biodiversity dynamics through acoustic data analysis.

1 Introduction

Passive Acoustic Monitoring (PAM) has emerged as a powerful technology for wildlife monitoring, allowing researchers and biodiversity managers to gather extensive acoustic data without disturbing natural habitats (Sugai et al. 2019; Sugai and Llusia 2019). PAM systems continuously record sounds from various environments, offering valuable insights into animal behavior, species richness, and ecosystem health, with important applications in ecosystem management, rapid assessments of biodiversity (Sueur et al. 2008), and basic research (Ross et al. 2023). However, effectively utilizing this vast amount of data for sound event detection poses significant challenges due to the need for annotated data to train machine learning models.

The annotation of PAM datasets is a laborious and time-consuming process carried out by experts. This bottleneck hampers the rapid adoption of machine learning techniques and impedes the exploration of acoustic data's full potential. While previous projects, e.g. (Gouvêa et al. 2023), focus on annotating entire datasets (Kath, Gouvêa, and Sonntag 2023), our method uses transfer and active learning to optimise sound event detection in PAM datasets without necessarily examining the entire dataset.

Transfer learning shows remarkable success in various domains, where models pre-trained on a large dataset can be fine-tuned to perform specific tasks with limited labeled data. By adapting knowledge from related audio tasks or datasets, we can efficiently initialize and enhance sound event detection models for PAM, mitigating the requirement for extensive annotation efforts.

In addition to transfer learning, active learning offers a strategic way to prioritize the most informative samples for annotation (Kadir, Alam, and Sonntag 2023). Instead of randomly labeling all data points, an active learning algorithm seeks to intelligently selects samples that are most uncertain or challenging to the model, enabling faster convergence with fewer annotations.

This study explores the combination of transfer learning and active learning as a means to facilitate the annotation of PAM datasets, visualised in figure 1. Comparing 5 standard embedding models trained on data with different relationships to PAM, we find that BirdNet (Kahl et al. 2021), a neural model trained on bird songs most closely related to PAM, performs best. Using the embeddings of the penultimate layer of BirdNet for several active learning strategies, we find that most strategies outperform random sampling. While we haven't identified a single strategy that consistently outperforms all others, our results show that active learning significantly reduces the annotation cost. We believe that this work can serve as a significant step forward in the automation of sound event detection in PAM, leading to a deeper understanding of biodiversity dynamics and better-informed wildlife conservation strategies.

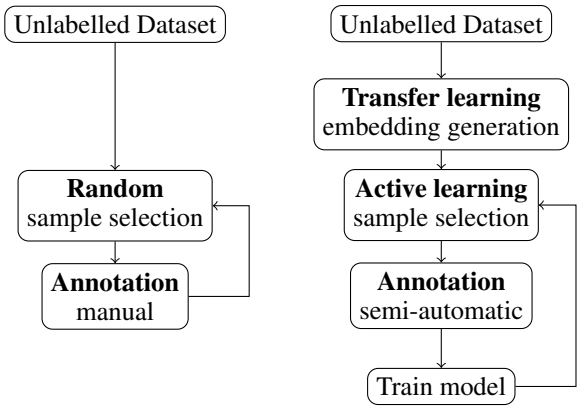


Figure 1: Workflow for annotating passive acoustic monitoring datasets, comparing the conventional approach (left) with the proposed approach (right).

2 Related Work

Sound Event Detection in PAM Data analysis is recognized as one of the bottlenecks in adoption of PAM methods for biodiversity monitoring (Sugai and Llusia 2019). While acoustic indices are widely used in acoustic monitoring of wildlife (Sueur et al. 2014; Campos et al. 2021), these are controversial and have been recently shown to misrepresent biodiversity in some cases (Bicudo et al. 2023; Sethi et al. 2023). Species identification, while more costly, plays an essential role in extracting ecologically relevant information from bioacoustic datasets. Due to the possible simultaneous occurrence of sounds from multiple species in PAM datasets, species identification is a multi-label sound event detection task. This machine learning challenge is well suited to the capabilities of convolutional neural networks (CNNs), as demonstrated in previous work (Hershey et al. 2017). While the idea of using CNNs for species identification in bioacoustics is not new (see Stowell 2022 for a survey), real-world applications are often limited by the lack of annotated multi-label data. In fact, deep learning models for species detection in PAM datasets are often trained using single-label annotated focal recordings (e.g., Kahl et al. 2021), neglecting the multi-label aspect of PAM data. Furthermore, focal recordings differ from PAM in that they are normally carried out with directional, professional-grade recorders actively pointed to the sound source (i.e., the specimens) by an expert *in loco*. These recordings tend to be of high quality and signal-to-noise ratio. For models trained with focal recordings for later use for inference in PAM datasets, this difference in data acquisition methods constitutes a form of domain-shift with recognized deleterious effects on performance (Kahl et al. 2021). The alternative is to have experts annotate PAM datasets, a laborious process. Therefore, practical few-shot learning approaches for PAM are needed.

Transfer Learning One key technique in few-shot learning is to transfer the knowledge and representations learned from one task to another, often resulting in improved efficiency and performance in the target task. The basic idea behind transfer learning is that a model trained on a large and diverse dataset for a source task can capture useful features and patterns that are applicable to a related target task. Instead of training a new model from scratch on the target task, which might require a significant amount of labeled data and computational resources, transfer learning allows building upon the existing knowledge of the source model.

Along these lines, Tsalera, Papadakis, and Samarakou (2021) use foundation CNNs pre-trained on large image (ImageNet) and audio (AudioSet) datasets to solve sound event detection tasks, and find that models pre-trained on the audio domain perform better. Dufourq et al. (2022) apply transfer learning to PAM datasets. They compare 12 different CNN architectures pre-trained on ImageNet as feature extractors for single-species detection in PAM datasets, and find that ResNets (101V2, 152V2) (He et al. 2016) performs best, followed by VGG16 (Simonyan and Zisserman 2015); Dufourq et al. did not explore models pre-trained on audio datasets. Çoban et al. (2020) use VGGish, a VGG variant pre-trained on AudioSet, to detect sound events in a PAM dataset; the event classes are coarse grained (e.g., 'songbird', 'waterbird', and 'insect') as opposed to fine grained (e.g., species identity). McGinn et al. (2023) investigate the topology of fine grained, sub-species sound events in the embedding space afforded by BirdNet, a CNN trained on focal recordings of bird vocalizations labelled at the species level (Kahl et al. 2021); they find that different call types of a same species (e.g., drumming versus vocalization) form distinct clusters, and that the vicinity of each such cluster contains different calls of the same species, rather than similar calls from distinct species. Ghani et al. (2023) compare 5 models pre-trained on audio data on 6 PAM datasets and find that Perch[1] and BirdNet, which differ mainly in their training data, perform best.

Active Learning While transfer learning can provide a solid starting point for sound event detection models, it does not do away with the need for human-annotated data. Active learning is a machine learning strategy that involves selecting and labeling first the most informative or uncertain examples in a dataset in order to improve the performance of a model while minimizing the amount of labeled data required. The core idea is to make the learning process more efficient by selecting the instances that are expected to provide the greatest reduction in uncertainty or error, rather than labeling a randomly selected subset of instances or all available data exhaustively. This is particularly useful in situations where labeling data is expensive, time-consuming, or otherwise resource-intensive.

Wang, Cartwright, and Bello (2022) use a synthetic dataset built by recombining environmental sounds with urban soundscape background to study how active learning can improve upon random selection in the context of prototype based classification with models trained with few-shot episodes. Qian et al. (2017) use active learning to improve on the data efficiency of bird species classifiers applied to a museum sound collection (likely focal recordings); their classifiers operate on low level descriptors,

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

which are interpretable feature extractors that might afford lower performance than deep learning methods. Allen et al. (2021) use active learning to detect humpback whale songs (single species) in a very large PAM dataset (187 000 h); they use a randomly initialized ResNet-50 variant (no transfer learning), and the small size of their validation set (6.25 h, or 0.003 % of the data) precludes comparing different active learning methods. Active learning is a central element of human-in-the-loop machine learning workflows (Monarch 2021). In a related application, Ryazanov et al. (2021) implement a human-in-the-loop system for marine acoustic event detection in which a human expert oversees and validates novel training samples synthesized by sampling the latent space of a variational autoencoder (a form of data augmentation).

3 Methods

Datasets While other studies mostly use single-label datasets, e.g. (Ghani et al. 2023), we take advantage of AnuraSet, a recently released real-world benchmark multi-label PAM dataset consisting of 27 h of audio plus manually created expert annotations for 42 species of anurans (frogs and toads) from two different biomes (Cañas et al. 2023). Following the original authors, in our transfer learning experiments we examine the overall performance of AnuraSet and its partitions based on the number of positive samples: frequent (>10 000), common (5 000-10 000) and rare (<5 000). Our active learning experiments focus on the common partition. In addition, we use a novel, small, manually multi-label annotated portion of a multi-year PAM dataset collected in Fernando de Noronha, Brazil. The selected part, referred to here as the Noronha set, consists of 1.25 h annotated by an expert for 5 species of oceanic birds. For all experiments, we divide each audio file into 3-second segments referred to as 'samples'. A sample is considered positive for a given event class whenever event

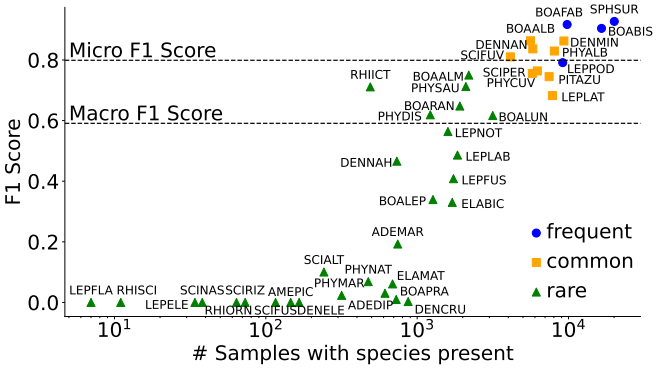


Figure 2: Transfer learning applied to AnuraSet using features extracted from the last layer before the classification layer of BirdNet. A linear classifier (logistic regression) is used. The resulting F1 score for each species is plotted against the number of samples containing that species. Frequent, common and rare species are defined according to (Cañas et al. 2023).

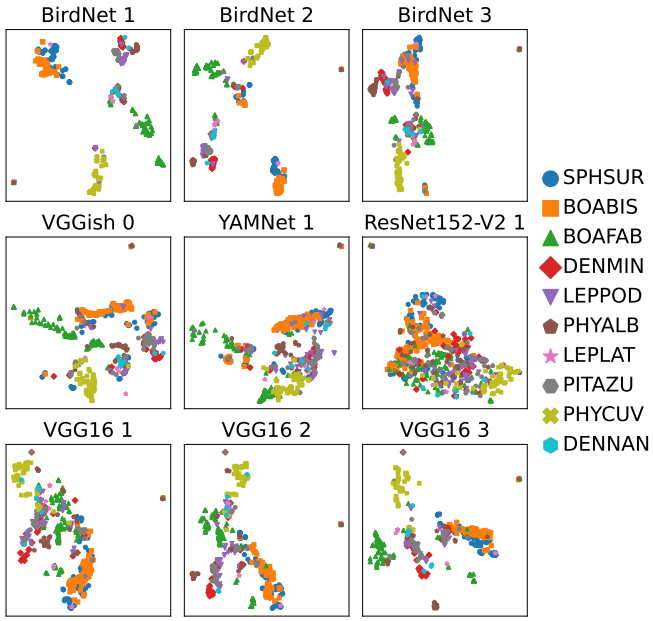


Figure 3: UMAP plots for diffenerent embedding layers of different embedding models for AnuraSet. Colors indicate top 10 classes. For UMAP generation, we select randomly 5 000 samples. In the plot we show only samples that are aligned to exactly one class.

occurrence overlaps with the sample, even if only partially and briefly. All performance metrics reported are computed on a held-out evaluation set except when otherwise stated. For AnuraSet, the evaluation set is that of the original study (Cañas et al. 2023); for the Noronha set, we randomly select a third of the dataset.

Transfer Learning We explore the potential of several standard pre-trained CNNs as feature extractors for sound event detection at the species level in PAM (table 1). The CNNs used here were trained on datasets from different domains and modalities, with varying degrees of similarity to the target modality (audio) and domain (multiple anuran species in the PAM data). Following Dufourq et al. (2022), we test ResNet152-V2 (He et al. 2016) and VGG16 (Simonyan and Zisserman 2015); these are CNNs pre-trained on ImageNet (Deng et al. 2009), a dataset on the visual modality. VGGish[2], a variant of VGG11A (Simonyan and Zisserman 2015), and YAMNet[3], a MobileNet-V1 network (Howard et al. 2017), were pre-trained on AudioSet (Gemmeke et al. 2017), a dataset from the same target modality (audio) but a different domain (YouTube sound clips). BirdNet (Kahl et al. 2021) was trained on data from the target modality (audio) and a related domain (bird vocalizations in focal recordings, also at species level).

Deep neural networks learn multiple representations of different levels of abstraction: the first layers reflect low level input features, while that last layers capture structure

[2]https://tfhub.dev/google/vggish/1
[3]https://tfhub.dev/google/yamnet/1

				AnuraSet								Noronha set	
		Layer		Frequent		Common		Rare		All			
Model	**Pre-Training**	# from last	Size	Mic F1	Mac F1	Mic F1	Mac F1	Mic F1	Mac F1	Mic F1	Mac F1	Mic F1	Mac F1
BirdNet	Bird vocalisations	1	1 024	**0.901**	**0.888**	**0.788**	**0.786**	**0.488**	0.402	**0.799**	**0.591**	0.455	0.527
		2	6 144	0.866	0.847	0.746	0.743	0.468	**0.440**	0.755	0.555	**0.727**	**0.554**
		3	4 608	0.876	0.856	0.749	0.759	0.484	0.431	0.772	0.572	0.704	0.490
VGGish	AudioSet	0	128	0.609	0.564	0.266	0.224	0.005	0.025	0.414	0.326	0.077	0.094
YAMNet	AudioSet	1	1 024	0.750	0.706	0.482	0.444	0.083	0.135	0.560	0.412	0.347	0.412
VGG16	ImageNet	1	4 096	0.680	0.577	0.410	0.400	0.031	0.087	0.466	0.332	0.112	0.128
		2	4 096	0.716	0.642	0.466	0.458	0.049	0.137	0.476	0.359	0.130	0.158
		3	25 088	0.851	0.817	0.701	0.684	0.373	0.341	0.726	0.528	0.313	0.371
ResNet152-V2	ImageNet	1	2 048	0.699	0.620	0.049	0.066	0.001	0.007	0.159	0.128	0.020	0.024

Table 1: Size and performance of embedding layers from different transfer learning models. The layers are labelled in reverse order excluding the classification layer, with layer 1 being the last layer before the classification layer. We analysed the frequent, common and rare part as well as the whole data set of AnuraSet, and the Noronha set. We provide micro (Mic) and macro (Mac) F1 scores calculated for the evaluation set. Each score represents the average result of multiple independent runs (5 for all AnuraSet experiments, 30 for Noronha set). The standard deviation is constantly less than 0.06.

more directly related to the predictions it makes (Bengio 2009). We evaluate embeddings at different layers within the CNNs (table 1). For VGG16 we investigate the last three layers before the final classification layer ('fc2', 'fc1', and 'flatten'). For ResNet152-V2 we only investigate the last embedding layer ('avg_pool'). Considering our future goal of implementing a real-time pipeline with transfer learning and active learning, we decide not to explore further layers of both visual domain models due to their large dimensionality (100 352 for both models). As the models pre-trained on AudioSet were designed to be used as feature extractors, we only use their last embedding layer. For BirdNet we investigate the last three embedding layers, batch normalization and dropout layers excluded ('GLOBAL_AVG_POOL', 'POST_CONV_1', and 'BLOCK_4-4_ADD'); the latter layer is a convergence point of a branched architecture, so we do not investigate further layers. We refer to each layer by natural numbers reflecting distance from the classification layer (e.g.,'BirdNet-1' denotes the last layer before the classification layer of the BirdNet model).

Figure 3 shows low-dimensional representations of the AnuraSet embeddings in each model/layer combination that we generate using UMAP, a neighbour-embedding method that attempts to preserve in the low-dimensional representation the same distance between points as observed in the high-dimensional embedding space (McInnes, Healy, and Melville 2020). We compute UMAP embeddings for a subset of 5 000 random samples from the top 10 classes of AnuraSet.

Sound event detection performance of each embedding is evaluated using a linear classifier (single fully connected layer). As the samples may contain overlapping calls from different species, we implement a multi-label classifier with logistic activation and a binary cross-entropy loss function. Linear classifiers are trained on frozen embeddings (no fine tuning) for up to 1 000 epochs with early stopping based on validation loss (minimum delta of 0.1, patience of 10 epochs, with reinstatement of best weights). When the embedding model outputs an array of time points for each input sample, we treat it as a multiple instance learning problem (Wang, Li, and Metze 2019) by applying the classifier to each time point, and then pooling predictions with the exponential softmax function $y = \frac{\sum_i y_i \exp(y_i)}{\sum_i \exp(y_i)}$, facilitating the training process. We report the metrics micro and macro F1 scores (see table 1). The results represent the mean computed over multiple runs with different random seeds.

Active Learning We explore a range of sampling strategies: uncertainty and diversity based, myopic (greedy) and adaptive (batch mode), and combinations thereof. In all cases, 5 % of the samples are selected at random.

$$\Phi_{LC\_bi}(y) = 1 - |2y - 1| \tag{1}$$

$$\Phi_{RC\_bi}(y) = \frac{0.5 - |y - 0.5|}{0.5 + |y - 0.5|} \tag{2}$$

$$\Phi_{EN\_bi}(y) = -y \log_2(y) - (1 - y) \log_2(1 - y) \tag{3}$$

Uncertainty sampling strategies compute uncertainty scores for each unlabelled sample and select those with the highest scores. Multi-label tasks with n classes use n binary classifiers, resulting in n uncertainty scores per sample. Following Monarch (2021), we implement 'least confidence' (equation (1)), 'ratio' (equation (2)) and 'entropy' (equation (3)), where $\Phi_{**\_bi}(y)$ is the uncertainty score from method ** for binary classifiers and y is the classifier output. To derive a single uncertainty score from the n scores assigned to each sample, we explore averaging and selecting the maximum value. $\Phi_{LC\_bi}(y)$, $\Phi_{RC\_bi}(y)$ and $\Phi_{EN\_bi}(y)$ have a strictly monotonic increase in the range $[0; 0.5]$ and a strictly monotonic decrease in the range

$[0.5; 1]$. Consequently, using the maximum score yields the same selected sample. Therefore, we use a singular method with maximum score aggregation and choose $\Phi_{RC\_bi}(y)$.

Diversity sampling strategies aim for comprehensive coverage of the data space, ensuring even distribution and avoiding class imbalance. Diversity sampling selects samples directly, ignoring class-specific scores and existing labels, thus eliminating the need to combine scores, unlike uncertainty sampling. We implement k-means clustering using the Euclidean distance measure. Within each cluster, we select the centroid (the sample with the smallest distance to the cluster centre), an outlier (the sample farthest from the nearest cluster centre) and three random samples. The number of clusters is inversely determined; e.g., to annotate 20 samples at a rate of 5 samples per cluster, we use 4 clusters (Monarch 2021, chapter 3).

Adaptive sampling strategies reduce the redundancy within the selected batch of samples for an iteration. Adaptive uncertainty sampling uses the predictions of the trained model to relabel the validation set as 'correct' or 'incorrect'. The model's last layer is replaced by a single node and retrained using the generated labels. Iteratively, the unlabelled set is fed into the model, samples that are likely to be 'incorrect' are selected, added to the 'correct' labelled validation set and the model is retrained (Konyushkova, Sznitman, and Fua 2017). Adaptive diversity sampling minimises the distribution gap between training and unlabelled data. After labelling the validation set 'validation' and the unlabelled set 'unlabelled', the model's last layer is replaced with a single node and retrained using the generated labels. Iteratively, the unlabelled subset is fed into the model, samples likely to be 'unlabelled' are selected. They are iteratively added to the validation set (Monarch 2021, chapter 5). Both adaptive strategies use 5 iterations in our implementation.

Combining active learning strategies addresses the limitations of pure strategies. Uncertainty sampling selects samples close to the decision boundaries, but may introduce redundancy. Diversity sampling covers the entire input space, but may miss critical regions. We therefore investigate methods that combine uncertainty and diversity strategies. Filtering pre-selects 50 % of the samples by diversity sampling and uses uncertainty sampling to sample from this pre-selection. We use this method for 'combi: ratio max + clustering'. Hybrid sampling selects 50 % of the samples from each of the two methods. All other combination methods use hybrid sampling.

Class labels are available for all samples used in this study, and an active learning scenario is emulated by hiding all labels from the classifier at first and incrementally revealing the ones for each batch of samples queried by the sampling methods. We use a batch size of 20 samples. The classifier heads are identical to those from the transfer learning experiment, always applied to data embeddings with the same selected model (BirdNet-1, see section 4).

4 Results

An annotated PAM dataset typically serves one of two primary purposes: as a resource for training new machine learning models for later deployment for inference in a related domain (e.g., geographical region, taxa), or as an end product in itself for subsequent analysis of ecological phenomena within the same domain. In this study, we explore the potential of combining transfer learning and active learning to accelerate the annotation of species-level sound events in PAM datasets for both purposes.

Transfer Learning

We start by testing different pre-trained models as feature extractors for species-level sound event detection in AnuraSet. In order to gain intuition on the potential of each embedding model, we generate low dimensional neighbor embedding visualizations for high dimensional embeddings of samples from the top 10 classes of AnuraSet (figure 3). The BirdNet embeddings show a clear separation between class clusters, with more pronounced differentiation in layers closer to the final layer. VGGish and YAMNet show effective cluster separation for only a subset of clusters, while ResNet152-V2 embeddings appear as a continuum, salt-and-pepper pattern in the low dimensional representation. Cluster separation is visible for VGG16, with more apparent separation for layers further away from the top.

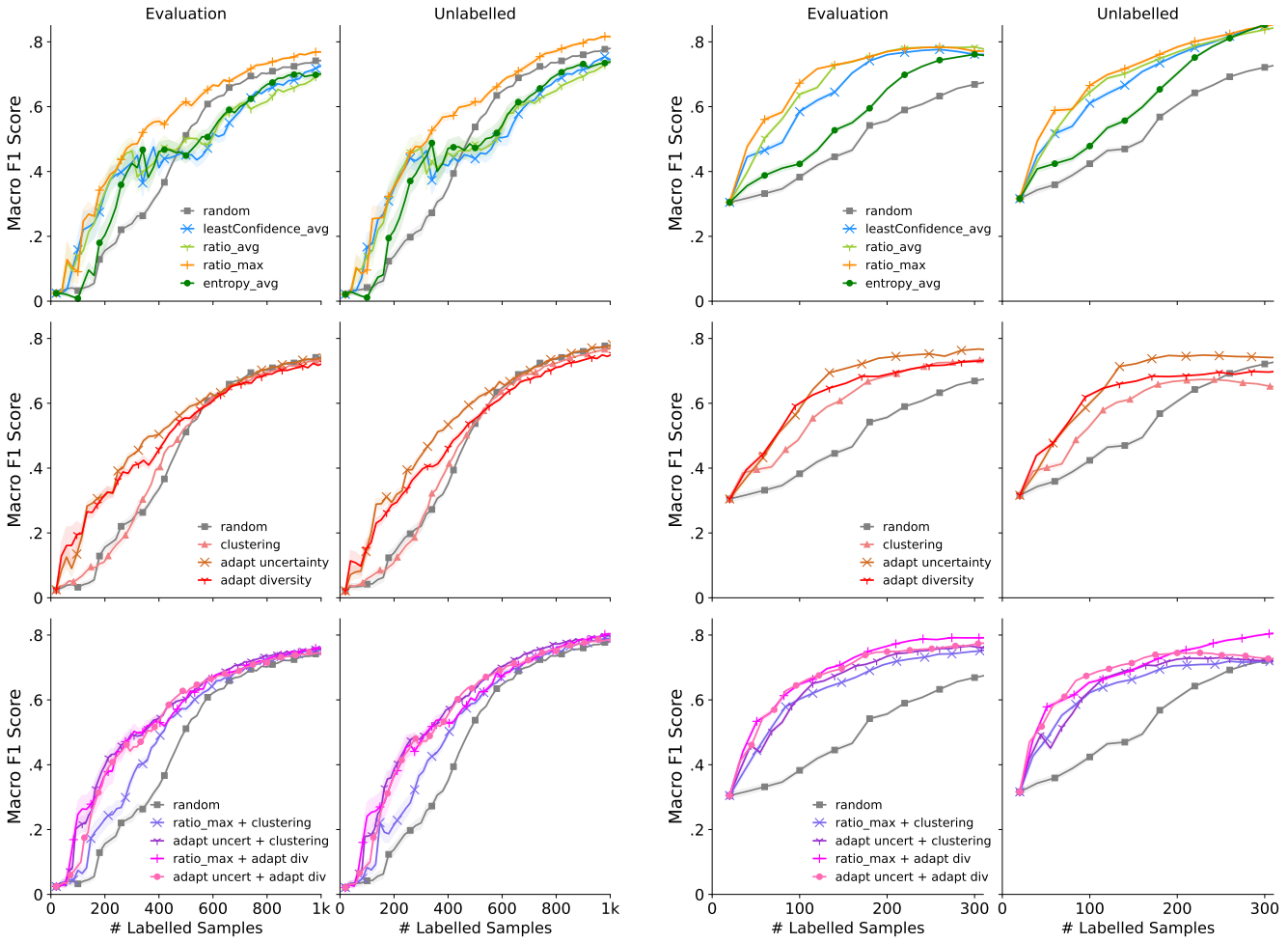
We then train linear classifiers on embeddings of the AnuraSet (frequent, common, rare and all) and Noronha datasets using all pre-trained models. The quantitative results largely match the intuitions afforded by neighbor embedding visualizations, with BirdNet performing best, followed by intermediate layers of VGG16 (albeit with a much smaller dimensionality, see table 1).

Overall, we find that BirdNet-1 performes best as a feature extractor for multi-label classification for the PAM datasets AnuraSet and Noronha set, resulting in the best macro F1 scores. The analysis of the frequent, common and rare parts of AnuraSet shows that this result is independent of the number of positive samples. Figure 2 shows the single class F1 score for BirdNet-1 for each of the 42 classes from AnuraSet. As reported in the original paper (Cañas et al. 2023), one can observe a strong correlation between F1 score and class size, and consequently a wide gap between macro and micro F1 scores. BirdNet-1 is used as a feature extractor for all subsequent active learning experiments.

Active Learning

We investigate the effect of active learning by emulating the annotation of the common partition of the AnuraSet and the Noronha set.

From a machine learning perspective, the two objectives outlined in the beginning of section 4 diverge in the data distribution. A machine learning model aims to classify new data that comes from the same distribution as the original dataset. Therefore, we construct evaluation sets that reflected the distribution of the original training data. For the AnuraSet, we use the identical evaluation set used by the original authors (Cañas et al. 2023). For the Noronha dataset, we randomly select a third of the data to form the evaluation set. As illustrated in figure 1, the process of annotating an entire dataset using active



(a) AnuraSet common, 5 independent runs

(b) Noronha set, 30 independent runs

Figure 4: Active learning on the common partition of AnuraSet and Noronha set, using the embeddings of BirdNet-1. Macro F1 score computed on evaluation data (left), and on the portion of the training data that remains unlabelled (right). Mean ± SEM across multiple independent runs. *Top:* uncertainty-based sampling strategies ('least confidence', 'ratio' and 'entropy') and score aggregation methods ('max' and 'average'). *Center:* diversity-based sampling strategy ('clustering') and two adaptive strategies ('uncertainty' and 'diversity'). *Bottom:* mixed diversity- and uncertainty-based sampling strategies.

learning is an iterative process that relies on careful sample selection strategies. This process leads to a distinction in the distribution between the entire dataset and the remaining unlabelled subset. Consequently, we will present results specific to this remaining unlabelled subset. In all active learning experiments, we use the embeddings generated by BirdNet-1, which show the best transfer learning performance. Due to the significant imbalance of classes in the datasets, we used the macro F1 score as the evaluation metric, and provide the corresponding macro precision and macro recall values in appendix A. As a baseline for active learning, all figures show the performance of random sampling.

We investigate the uncertainty sampling strategies 'least confidence', 'ratio' and 'entropy' with the score aggregation methods 'max' and 'average' ('avg'). The top row of figure 4 shows the results of the F1 score. For both datasets, for both the evaluation set and the unlabelled set, the score aggregation method 'max' consistently outperforms 'average,' surpassing random sampling and converging just below an F1 score of 0.8.

We further investigate the diversity sampling strategy 'clustering' and explore two adaptive approaches – one for uncertainty and the other for diversity. The results of the F1 score are shown in the center row of figure 4. For both datasets, the adaptive uncertainty method shows a slight performance advantage over other methods, both for the evaluation set and the unlabelled set, with all methods outperforming random sampling.

The bottom row of figure 4 shows the F1 score for the combined sampling strategies. We choose the 'ratio max' uncertainty sampling strategy for the combination due to

the superior performance of the 'max' versions and the simplicity of calculating the ratio. For both datasets, for both the evaluation set and the unlabelled set, all combined methods clearly outperform random sampling, with no single method emerging as the clear best.

Looking at precision and recall in figure S1 and figure S2, we observe a rapid convergence of precision around 0.9 for all methods. On the other hand, for most of the methods, recall does not show any convergence and remains significantly lower than precision, around 0.6 for the Common part of the Anuraset and around 0.2 for the Noronha set. While the choice of sampling method seems to have a limited effect on precision, there is a clearly visible effect on recall, where most methods clearly outperform random sampling, leading to the ranking of F1 score performance.

5 Discussion

In this investigation, we explored the application of knowledge transfer from large models pre-trained on diverse domains to the challenge of sound event detection in multi-species PAM datasets. We found that the final embedding layer of BirdNet, a CNN trained on focal recordings of bird vocalisations (Kahl et al. 2021), provides the most effective feature space. Notably, the linear classifier using BirdNet embeddings outperforms the models examined by (Dufourq et al. 2022) and (Cañas et al. 2023), beating the latter by 21.7 %.

Our findings unveil the effectiveness of active learning in the realm of multi-label sound event detection for PAM, combined with transfer learning. While previous active learning efforts in sound event detection for PAM (Qian et al. 2017; Allen et al. 2021) have not utilized features extracted with transfer learning, our study pioneered this intersection. In our exploration of uncertainty-based sampling strategies, originally designed for multi-class classification, we noted their superior performance over random sampling in our multi-label (multiple binary) classification scenario. It's pertinent to emphasize that the absence of a decisive winner was expected given our focus on multi-label tasks, in contrast to the multi-class setup these strategies were designed for.

Although the creation of a fully functional data annotation application falls beyond our current scope, we made a deliberate inclusion of a dataset in this study that directly aligns with the objectives of our methods. The Noronha set, afflicted by the same challenge our methods aim to mitigate—tedious data annotation—features a relatively limited number of labels. Despite its scale, we deemed it relevant to incorporate. We underscore the equivalence in real-world context between the datasets used here; while AnuraSet serves as a pivotal benchmark, it is crucial to recognize its authenticity as well. The ongoing nature of the AnuraSet project and its planned expansions further attest to its practicality, despite the common bottleneck of data annotation. Our aspiration is to contribute to the enhancement of annotation efficiency.

The discussion around precision, characterized by notable highs, juxtaposed against low recall demands attention. The latter raised concerns since, within the active learning framework, unattended events (false negatives) are irrevocably lost unless manually verified. A potential remedy could involve a workflow that mirrors medical tests, starting with heightened sensitivity to false negatives followed by a phase emphasizing specificity to false positives. In our methodology, a similar approach could be realized by adjusting learning to penalize false negatives, possibly via weighted binary cross entropy loss or custom loss functions as in (Tian et al. 2022).

While the observed low recall necessitates careful consideration, it's important to clarify that the scope of this study didn't encompass the optimization for accuracy metrics, exemplified by F1 Score. Instead, our primary goal was to identify efficient strategies that synergize transfer learning and active learning. To potentially elevate accuracy, strategies such as applying Per-Channel Energy Normalization (PCEN) (Lostanlen, Salamon, and Cartwright 2019), refining spectrogram feature engineering (Dufourq et al. 2022), or employing transfer learning with fine-tuning could be explored.

In our future endeavors, we intend to harness the methodologies examined herein to drive the development of a PAM data annotation tool. This endeavor will necessitate evaluations of computational costs, such as matmul operations, in conjunction with the metrics discussed in this paper. Furthermore, empowering users with control over the specificity/sensitivity trade-off could provide a customizable solution to match their needs.

Acknowledgements

This research is part of the Computational Sustainability & Technology project field[4], and has been supported by the Ministry for Science and Culture of Lower Saxony (MWK), the Endowed Chair of Applied Artificial Intelligence, Oldenburg University, and DFKI.

References

Allen, A.; Harvey, M.; Harrell, L.; et al. 2021. A Convolutional Neural Network for Automated Detection of Humpback Whale Song in a Diverse, Long-Term Passive Acoustic Dataset. *Frontiers in Marine Science*, 8.

Bengio, Y. 2009. Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning*, 2(1): 1–127.

Bicudo, T.; Llusia, D.; Anciães, M.; and Gil, D. 2023. Poor performance of acoustic indices as proxies for bird diversity in a fragmented Amazonian landscape. *Ecological Informatics*, 77: 102241.

Campos, I.; Fewster, R.; Truskinger, A.; et al. 2021. Assessing the potential of acoustic indices for protected area monitoring in the Serra do Cipó National Park, Brazil. *Ecological Indicators*, 120: 106953.

Cañas, J.; Toro-Gómez, M.; Sugai, L.; et al. 2023. A dataset for benchmarking Neotropical anuran calls identification in passive acoustic monitoring. *Scientific Data*, 10(1): 771.

[4]https://cst.dfki.de/

Deng, J.; Dong, W.; Socher, R.; et al. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255.

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

Gemmeke, J.; Ellis, D.; Freedman, D.; et al. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780.

Ghani, B.; Denton, T.; Kahl, S.; and Klinck, H. 2023. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *CoRR*, abs/2307.06292.

Gouvêa, T.; Kath, H.; Troshani, I.; et al. 2023. Interactive Machine Learning Solutions for Acoustic Monitoring of Animal Wildlife in Biosphere Reserves. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 6405–6413. Macau, SAR China.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Identity Mappings in Deep Residual Networks. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, 630–645.

Hershey, S.; Chaudhuri, S.; Ellis, D.; et al. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, 131–135.

Howard, A.; Zhu, M.; Chen, B.; et al. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *CoRR*, abs/1704.04861.

Kadir, A.; Alam, H.; and Sonntag, D. 2023. EdgeAL: An Edge Estimation Based Active Learning Approach for OCT Segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, 79–89.

Kahl, S.; Wood, C.; Eibl, M.; and Klinck, H. 2021. BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61: 101236.

Kath, H.; Gouvêa, T.; and Sonntag, D. 2023. A Human-in-the-Loop Tool for Annotating Passive Acoustic Monitoring Datasets. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 7140–7144. Macau, SAR China.

Konyushkova, K.; Sznitman, R.; and Fua, P. 2017. Learning Active Learning from Data. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, 4225–4235.

Lostanlen, V.; Salamon, J.; and Cartwright, o. 2019. Per-Channel Energy Normalization: Why and How. *IEEE Signal Processing Letters*, 26(1): 39–43.

McGinn, K.; Kahl, S.; Peery, M.; et al. 2023. Feature embeddings from the BirdNET algorithm provide insights into avian ecology. *Ecological Informatics*, 74: 101995.

McInnes, L.; Healy, J.; and Melville, J. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *CoRR*, abs/1802.03426.

Monarch, R. 2021. *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-centered AI*. Simon and Schuster. ISBN 978-1-61729-674-1.

Qian, K.; Zhang, Z.; Baird, A.; and Schuller, B. 2017. Active learning for bird sound classification via a kernel-based extreme learning machine. *The Journal of the Acoustical Society of America*, 142(4): 1796–1804.

Ross, S.; O'Connell, D.; Deichmann, J.; et al. 2023. Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions. *Functional Ecology*, 37(4): 959–975.

Ryazanov, I.; Nylund, A.; Basu, D.; et al. 2021. Deep Learning for Deep Waters: An Expert-in-the-Loop Machine Learning Framework for Marine Sciences. *Journal of Marine Science and Engineering*, 9(2): 169.

Sethi, S.; Bick, A.; Ewers, R.; et al. 2023. Limits to the accurate and generalizable use of soundscapes to monitor biodiversity. *Nature Ecology & Evolution*, 1–6.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*, abs/1409.1556.

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

Sueur, J.; Farina, A.; Gasc, A.; et al. 2014. Acoustic Indices for Biodiversity Assessment and Landscape Investigation. *Acta Acustica united with Acustica*, 100(4): 772–781.

Sueur, J.; Pavoine, S.; Hamerlynck, O.; and Duvail, S. 2008. Rapid Acoustic Survey for Biodiversity Appraisal. *PLOS ONE*, 3(12): e4065.

Sugai, L.; and Llusia, D. 2019. Bioacoustic time capsules: Using acoustic monitoring to document biodiversity. *Ecological Indicators*, 99: 149–152.

Sugai, L.; Silva, T.; Ribeiro, J.; and Llusia, D. 2019. Terrestrial Passive Acoustic Monitoring: Review and Perspectives. *BioScience*, 69(1): 15–25.

Tian, J.; Mithun, N.; Seymour, Z.; et al. 2022. Striking the Right Balance: Recall Loss for Semantic Segmentation. In *2022 International Conference on Robotics and Automation (ICRA)*, 5063–5069.

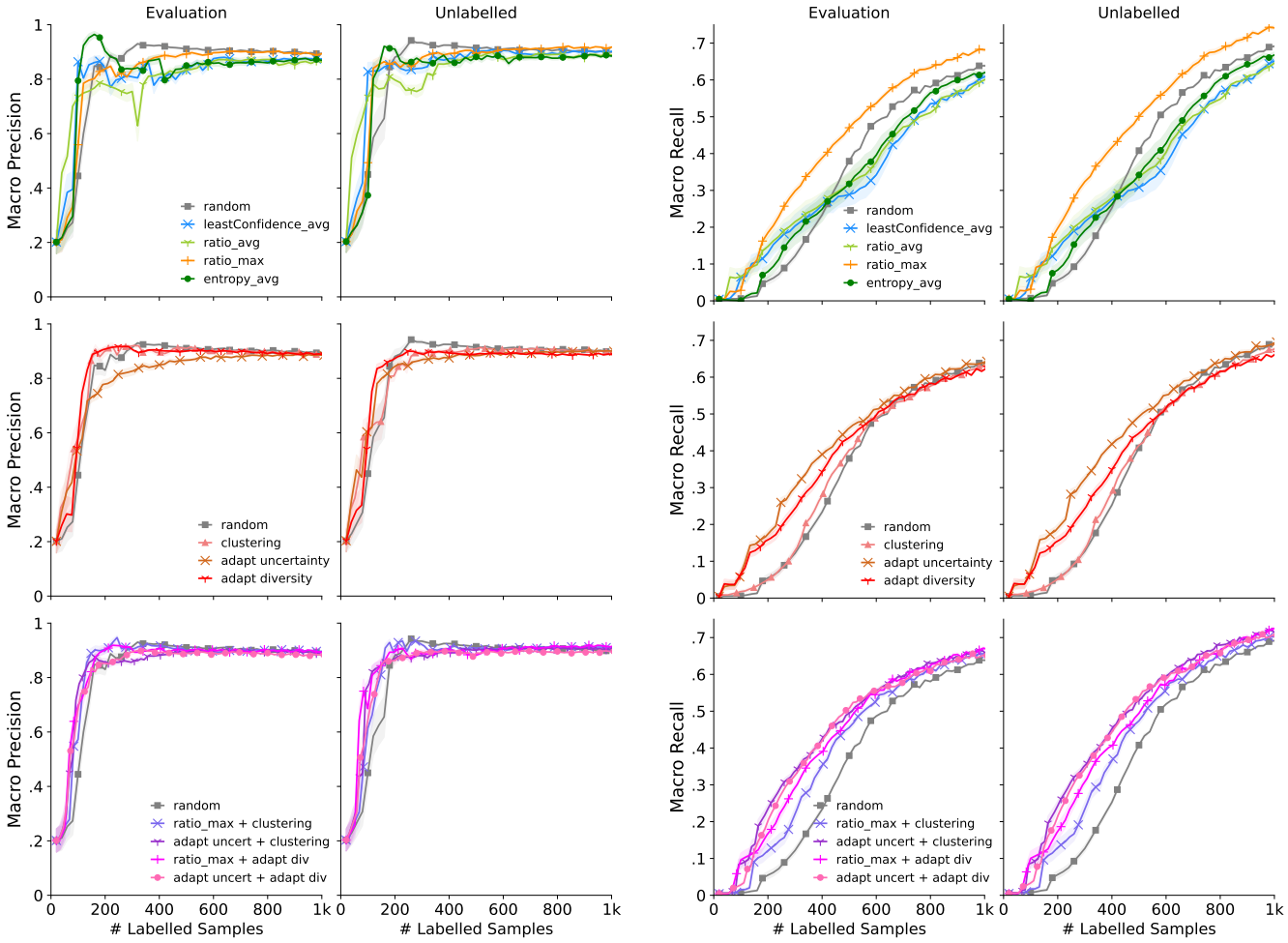
Tsalera, E.; Papadakis, A.; and Samarakou, M. 2021. Comparison of Pre-Trained CNNs for Audio Classification Using Transfer Learning. *J. Sens. Actuator Networks*, 10(4): 72.

Wang, Y.; Cartwright, M.; and Bello, J. 2022. Active Few-Shot Learning for Sound Event Detection. In *Interspeech 2022*, 1551–1555.

Wang, Y.; Li, J.; and Metze, F. 2019. A Comparison of Five Multiple Instance Learning Pooling Functions for Sound Event Detection with Weak Labeling. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 31–35.

Çoban, E.; Pir, D.; So, R.; and Mandel, M. 2020. Transfer Learning from Youtube Soundtracks to Tag Arctic Ecoacoustic Recordings. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 726–730.

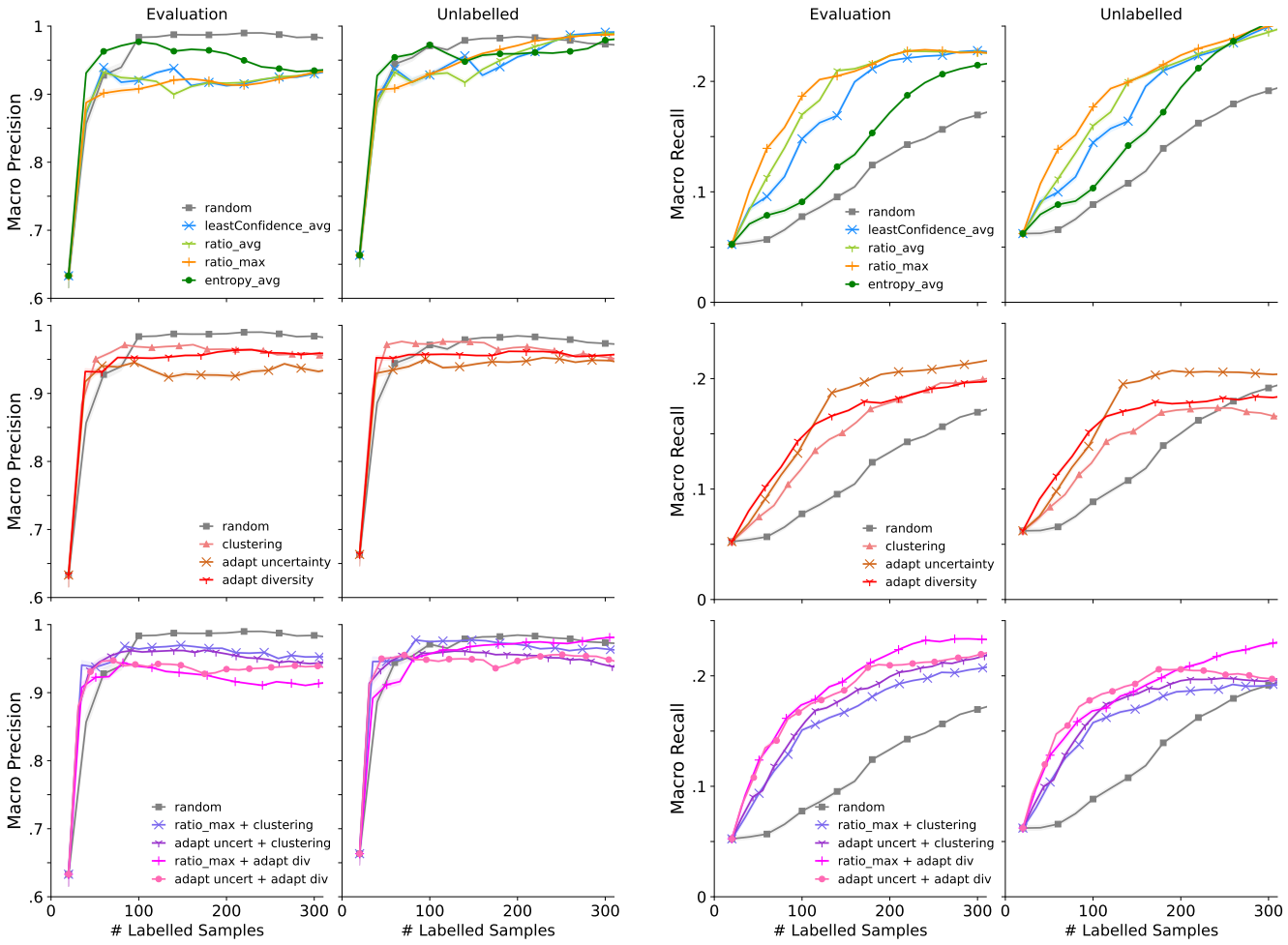
A Supplementary Figures



(a) Precision

(b) Recall

Figure S1: Active learning on the common partition of AnuraSet, using the embeddings of BirdNet-1.. Macro precision/recall score computed on evaluation data (left), and on the portion of the training data that remains unlabelled (right). Mean ± SEM across multiple independent runs. *Top:* uncertainty-based sampling strategies ('least confidence', 'ratio' and 'entropy') and score aggregation methods ('max' and 'average'). *Center:* diversity-based sampling strategy ('clustering') and two adaptive strategies ('uncertainty' and 'diversity'). *Bottom:* mixed diversity- and uncertainty-based sampling strategies.



(a) Precision

(b) Recall

Figure S2: Active learning on Noronha set, using the embeddings of BirdNet-1.. Macro precision/recall score computed on evaluation data (left), and on the portion of the training data that remains unlabelled (right). Mean ± SEM across multiple independent runs. *Top:* uncertainty-based sampling strategies ('least confidence', 'ratio' and 'entropy') and score aggregation methods ('max' and 'average'). *Center:* diversity-based sampling strategy ('clustering') and two adaptive strategies ('uncertainty' and 'diversity'). *Bottom:* mixed diversity- and uncertainty-based sampling strategies.