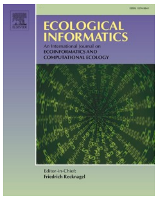
Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Leveraging transfer learning and active learning for data annotation in passive acoustic monitoring of wildlife

Hannes Kath[a,b,*], Patricia P. Serafini[c,d], Ivan B. Campos[c,e], Thiago S. Gouvêa[a,b], Daniel Sonntag[a,b]

[a] Interactive Machine Leraning, German Research Center for Artificial Intelligence (DFKI), Oldenburg, Germany
[b] Applied Artificial Intelligence, Carl von Ossietzky University of Oldenburg, Germany
[c] National Center for Wild Bird Conservation and Research (CEMAVE), Chico Mendes Institute for Biodiversity Conservation (ICMBio), Brazil
[d] Universidade Federal de Santa Catarina (UFSC), Brazil
[e] Departamento de Biologia Geral, Universidade Federal de Minas Gerais (UFMG), Brazil

ARTICLE INFO

Keywords:
Active learning
Transfer learning
Passive acoustic monitoring
BirdNet

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 with minimal disturbance of natural habitats (Sugai et al., 2019; Sugai and Llusia, 2019). PAM systems make it possible to continuously record environmental sounds (soundscapes), offering valuable insights into animal behaviour, species richness, and ecosystem health, with important applications in ecosystem management, rapid assessments of biodiversity, and basic research (Ross et al., 2023; Sueur et al., 2008). However, effectively utilising the vast amount of data generated by these systems still poses significant challenges, limiting adoption of PAM methods for biodiversity monitoring.

Acoustic indices are often used as a computationally efficient strategy for summarising and making sense of large soundscape datasets (Campos et al., 2021; Machado et al., 2017; Sueur et al., 2014). However, these methods are controversial and have been shown to misrepresent biodiversity in some cases (Bicudo et al., 2023; Sethi et al., 2023). Therefore, the detection and identification of discrete sound events (e.g., at the species level), while more costly, plays an essential role in extracting ecologically relevant information from soundscape datasets.

1.1. Sound event detection

In the field of machine learning, the task of detecting and identifying

[*] Corresponding author at: Interactive Machine Leraning, German Research Center for Artificial Intelligence (DFKI), Oldenburg, Germany.
E-mail addresses: hannes.kath@dfki.de (H. Kath), patricia.serafini@icmbio.gov.br (P.P. Serafini), ivan.campos@icmbio.gov.br (I.B. Campos), thiago.gouvea@dfki.de (T.S. Gouvêa), daniel.sonntag@dfki.de (D. Sonntag).

https://doi.org/10.1016/j.ecoinf.2024.102710
Received 22 December 2023; Received in revised form 26 June 2024; Accepted 28 June 2024
Available online 10 July 2024

discrete events in acoustic data is known as *sound event detection*, a challenge well suited to the capabilities of convolutional neural networks (CNNs) (Hershey et al., 2017; Nolasco et al., 2023). Due to the possible simultaneous occurrence of sounds from multiple species in a soundscape, species identification is better described as a *multi-label* sound event detection task. As with any other supervised learning task, sound event detection requires the training data to be annotated with class labels (e.g., species identity and times of occurrence), and obtaining or generating those annotations can be very time consuming.

While the idea of using CNNs for species identification in bioacoustics is not new (see LeBien et al. (2020) for an early example, and 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* focal recordings (see Kahl et al. (2021) for a prominent example), neglecting the multi-label character of soundscapes. Furthermore, focal recordings differ from PAM in that they are normally carried out with directional, professional-grade recorders actively pointed to the sound source by an expert in loco, thus yielding recordings of high quality and signal-to-noise ratio. For models intended for application to soundscape data, the use of focal recordings as training data constitutes a form of domain-shift, with recognised deleterious effects on performance (Kahl et al., 2021). The alternative, however, is costly: before training automated classifiers, experts would have to annotate PAM datasets for the label classes of interest, a process that can take over 10 min of analysis labour per minute of recorded data (Lüers et al., 2024).

1.2. Transfer learning

Therefore, practical few-shot learning methods for PAM are needed. *Transfer learning* is a key technique in few-shot learning (Wang et al., 2020), and consists of the transfer of knowledge learned from one task to another, often resulting in improved efficiency and performance in the target task. The basic idea is that a model trained on a large and diverse dataset for one source task can capture useful features and patterns that are applicable to a related target task.

Along these lines, LeBien et al. (2020) build a pipeline for frog and bird species identification from acoustic data using as feature extraction a ResNet50 pre-trained on ImageNet, a large image dataset (He et al., 2016). Florentin et al. (2020) and Dufourq et al. (2022) explore a wider range of CNN architectures pre-trained on ImageNet for single-species detection in PAM datasets. Tsalera et al. (2021) compare the performance of CNNs pre-trained on ImageNet or AudioSet, a large acoustic dataset, and find that models pre-trained on the audio domain are better at detecting sound events. Çoban et al. (2020) use VGGish, a CNN pre-trained on AudioSet, to detect coarse grained sound events (e.g., songbird, waterbird, insect) in a PAM dataset. Ghani et al. (2023) compare 5 models pre-trained on audio data on 6 bioacoustics datasets and find that Perch[1] and BirdNet (Kahl et al., 2021), which differ only slightly regarding their training data, perform best at species identification; evaluation was done on focal and citizen-science datasets. Swaminathan et al. (2024) extend the observation to attention-based architectures pre-trained on human speech. Lauha et al. (2022) show that transfer learning can be helpful also for small networks trained from scratch on small, targeted datasets gathered from online resources such as Macaulay Library[2] (as opposed to foundation models trained on large datasets). While the works cited above evaluate transfer learning models based on classification performance, McGinn et al. (2023) take a different approach and investigate the topology of fine grained, sub-species sound events in the embedding space afforded by BirdNet; 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 reflects species identity rather than sound morphology (i.e., the space immediately around a given cluster contains different calls of the same species, rather than similar calls from distinct species).

Therefore, to our knowledge, only one of the studies investigated large transfer learning models pre-trained on data from the bioacoustics domain (Ghani et al., 2023), and none assessed their performance particularly in soundscape type (omnidirectional, multi-label) datasets.

1.3. Active learning

While transfer learning can provide a solid starting point for sound event detection models, it does not eliminate the need for annotated data. *Active learning* is a machine learning strategy that consists of selecting and labelling first the most informative samples. The core idea is to make the learning process more efficient by selecting first the instances that are expected to provide the greatest reduction in uncertainty or error, rather than labelling a randomly selected subset of instances or all available data exhaustively. This is particularly useful in situations where labelling data is expensive, time-consuming, or otherwise resource-intensive (Kadir et al., 2023, e.g.).

Wang et al. (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. In two early bioacoustics applications, 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), while Kholghi et al. (2018) perform coarse-grained classification on omnidirectional recordings; in both cases, classifiers operate on low level descriptors and acoustic indices, both of which are hand-designed feature extractors that afford lower performance than representations learned by deep neural networks. Allen et al. (2021) use active learning and deep learning to detect humpback whale songs (single species) in a very large PAM dataset (187,000 h); they use a randomly initialised 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. Similarly, van Osta et al. (2023) use transfer learning (ResNet seemingly pre-trained on ImageNet) and an active learning strategy to train a classifier for a single cryptic bird species, but do not compare different active learning strategies.

In summary, while the applicability of active learning to the domain of bioacoustics has been demonstrated, none of the active learning studies we are aware of make use of state-of-the-art feature spaces (e.g., transfer learning models); in addition, the particularities of soundscape type (omnidirectional, multi-label) data have also not been addressed.

1.4. Contribution

This study explores the combination of transfer learning and active learning as a means to reduce the amount of time needed to annotate PAM datasets (Fig. 1). Specifically, we compare 5 standard embedding models pre-trained on large datasets from domains with varying degrees of proximity to PAM (namely, images, generic audio, or bioacoustics) and evaluate them with linear classifiers on three soundscape datasets covering different taxa. For the unparalleled multi-label PAM dataset AnuraSet, our simple linear model (applied to features extracted with a transfer learning model) surpasses the convolutional baseline (Cañas et al., 2023) by a remarkable margin of 21.7%. After identifying the feature space (i.e. transfer learning model) that gives the best classification accuracy for sound event detection with soundscape data, we investigate the potential of active learning to accelerate learning in this space. In contrast to previous literature, we perform a comparative study of different active learning strategies. We investigate a range of sampling strategies: uncertainty and diversity based methods, myopic (greedy) and adaptive (batch mode) methods, and combinations thereof. Finally, we evaluate the resulting learning curves through the lens of

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

[2] https://www.macaulaylibrary.org/

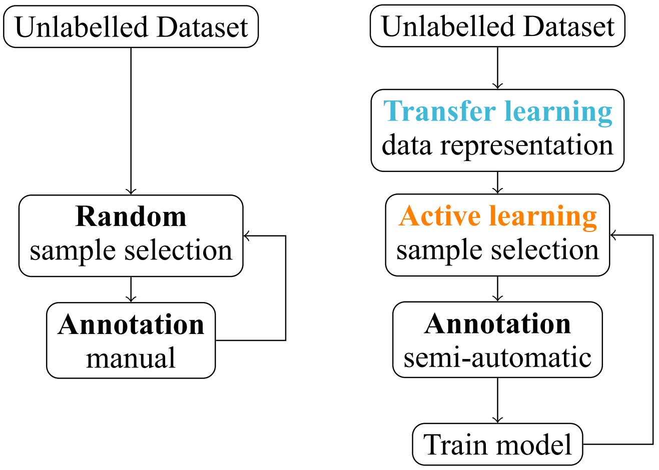

Unlabelled Dataset
Random sample selection
Annotation manual
Unlabelled Dataset
Transfer learning data representation
Active learning sample selection
Annotation semi-automatic
Train model

Fig. 1. Workflow for annotating passive acoustic monitoring datasets, comparing the conventional approach (left) with the proposed approach (right). We compare the transfer learning models BirdNet, VGGish, YAMNet, VGG16 and ResNet152-V2 for generating data representations. As active learning method we select uncertainty methods, diversity methods, adaptive methods and combinations of those.

different accuracy metrics (namely precision and recall) and discuss their practical implications.

2. Methods

2.1. Datasets

The affordability of recording equipment has led to the publication of many PAM datasets in recent years. However, it is noteworthy that the vast majority of these datasets provide single-label annotations, indicating that labels are mutually exclusive and cannot occur simultaneously. Other studies predominantly use single-label datasets, e.g. Ghani et al. (2023). As PAM recordings often capture several species calling simultaneously, we take advantage of more realistic PAM datasets with multi-label annotations, which provide information on the presence or absence of each species individually.

AnuraSet is a recently released real-world benchmark multi-label PAM dataset containing 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). The original authors divide the one-minute audio files recorded in four different areas into segments of three seconds each, with an overlap of two seconds. This segmentation approach resulted in 58 three-second audio files per minute, increasing the dataset to 77 h of audio. The sample rate is 22.05 kHz. To mitigate class imbalance, the authors implement a stratified training/evaluation split, allocating 30% of the data to the evaluation set. To prevent data leakage, the split was performed on the original one-minute files, ensuring that all corresponding three-second files were assigned exclusively to either the training or evaluation set. Due to class imbalance, AnuraSet comes in partitions based on the number of positive samples: frequent (>10,000), common (5000–10,000) and rare (<5000). We used the evaluation set and the partitions as defined by the original authors.

Noronha set is a novel, small, expert annotated multi-label dataset derived from a multi-year PAM program carried 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 seabirds. The sample rate is 48 kHz. As a first pre-processing step, we segmented the one-minute files into three second snippets with no overlap. A stratified split was used to generate an evaluation set containing one-third of the available data.

Watkins Marine Mammal Sound Database (Watkins) is a single-label dataset containing calls from 56 marine mammal species (Sayigh et al., 2016). To emulate a multi-label PAM dataset, we inserted sound samples from the Watkins database into a noisy background[3] at random positions, including the possibility of overlapping events. The sample rate is 48 kHz. For training, we included all classes with 500–1000 occurrences, for a total of 9 classes. The audio data was segmented into three second long files. From the total duration of 7.8 h of audio, the evaluation set (1.6 h) contains 20% of the events from each class.

2.2. 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. 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 species in PAM datasets). To ensure the robustness of the results, we averaged them over 30 independent runs.

2.2.1. Model architectures

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, a variant of VGG11A (Simonyan and Zisserman, 2015), and YAMNet, 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 vocalisations from focal recordings, also annotated at species level).

2.2.2. Model layers

Deep neural networks learn multiple representations of different levels of abstraction: the first layers reflect low level input features, while the last layers capture structure more directly related to the predictions output by the model (Bengio, 2009). We evaluate embeddings at different layers within the CNNs. 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). Since the models pre-trained on AudioSet were designed to be used as feature extractors, we only use the last layer for VGGish and the penultimate layer for YAMNet. 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.

2.2.3. Pre-processing

For all experiments, we use three-second long audio segments referred to as 'samples'. A sample is considered positive for a given event class whenever event occurrence overlaps with the sample, even if only partially and briefly. We resample the audio for the models trained in the audio domain to 48 kHz (BirdNET) or 16 kHz (YAMNet and VGGish). Since ResNet152-V2 and VGG16 take images as input, spectrograms were calculated for each sound sample using the native sampling rate of the audio signal, employing a window size of 512 samples and an overlap of 256 samples. Lastly, each spectrogram was resized to the

[3] https://www.youtube.com/watch?v=sCc3UtzZDEo

dimensions required by each of the two convolutional models.

2.2.4. *Training*

To evaluate the performance of sound event detection, a linear multi-label classifier is trained. The resulting architecture consists of a single fully connected layer with one output node per species in the dataset. Each output node indicates the presence or absence of that species and is independent of the other output nodes. A binary cross-entropy loss function and logistic activation are used since we train a multi-label classifier. The classifiers are trained on frozen embeddings (no fine-tuning) for a maximum of 1000 epochs. Early stopping criteria, starting from epoch 50, are based on the validation loss, with a minimum delta of 0.1 and a patience of 10 epochs, with reinstatement of the best weights. VGGish and YAMNet take as input audio segments shorter than the 3-s samples employed in this study. In order to generate a single embedding point for each 3-s sample, we split each sample into shorter segments expected by the models, resulting in an array of time points as output. We then proceeded with a multiple instance learning approach by applying the classifier to each element in the array, and then pooling the output array into a scalar value with the exponential softmax function $\hat{y} = \sum_i y_i \frac{exp(y_i)}{\sum_j exp(y_j)}$ (Wang et al., 2019).

2.3. *Active learning*

The active learning experiments are carried out in the embedding space of the selected transfer learning model (BirdNet-1, see Section 3.1). We explore a range of sampling strategies: uncertainty and diversity based, myopic (greedy) and adaptive (batch mode), and combinations thereof. Fig. 2 provides a schematic overview of the pure families of sampling strategies: Random sampling selects samples arbitrarily, uncertainty sampling targets samples near the decision boundary, and diversity sampling focuses on samples that span the entire data space. In all cases, 5% of the samples are selected at random. 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 training process. To ensure the robustness of the results, we averaged them over 30 independent runs.

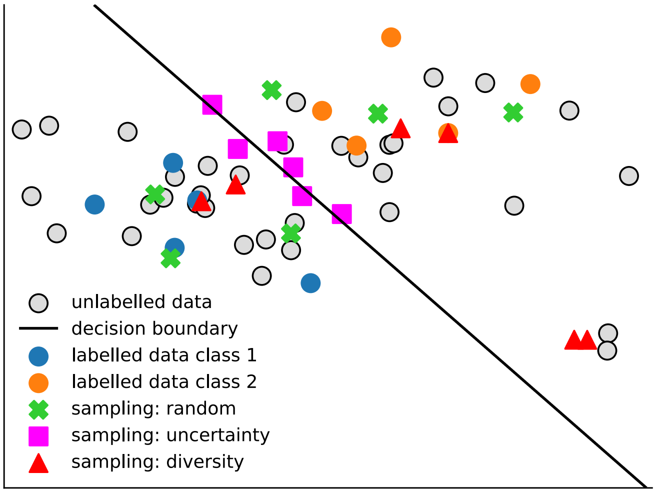


Fig. 2. Schematic representation of the random, uncertainty, and diversity sampling strategies. Random sampling selects arbitrary samples. Uncertainty sampling selects samples based on their proximity to the model's decision boundary, calculated using the existing class 1 and class 2 labels. Diversity sampling uses the internal data structure for sampling purposes, such as clustering the data and selecting samples from each cluster.

Uncertainty sampling strategies compute uncertainty scores for each unlabelled sample and select those with the highest scores. Following Monarch (2021), we implement ‘least confidence’ (Φ_{LC}), ‘ratio’ (Φ_{RC}) and ‘entropy’ (Φ_{EN}). Fig. 3 shows the uncertainty score Φ corresponding to a prediction y for a signle species within a binary model. With n species, and therefore n binary classifiers, n uncertainty scores per sample are computed. Deriving a single uncertainty score per sample involves score aggregation, where we explore the techniques of 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 $[0.5; 1]$ (see Fig. 3). 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 to achieve comprehensive coverage of the data space with the selected samples, ensuring an even distribution and avoiding class imbalance. Unlike uncertainty sampling, diversity sampling selects samples directly based on the structure of the dataset, without relying on model predictions or labels. 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 aim to reduce redundancy within the selected batch of samples during 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 et al., 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.

Combined sampling strategies address 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

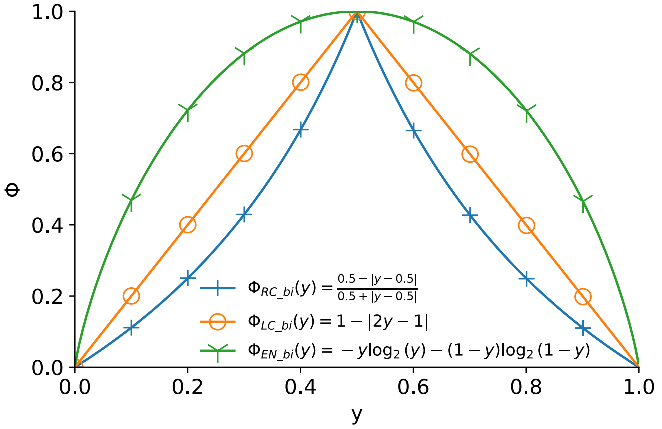


Fig. 3. Uncertainty scores for binary inputs computed with the methods ‘ratio’ (Φ_{RC_bi}), ‘least confidence’ (Φ_{LC_bi}) and ‘entropy’ (Φ_{EN_bi}), given a model prediction y.

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.

2.4. *Computational resources*

The computational analyses[4] including pre-processing steps are conducted using Python programming language version 3.10. Key libraries include librosa for audio processing, scikit-maad for spectrogram computation, and TensorFlow for learning processes. We used BirdNet V2.4,[5] VGGish,[6] YAMNet,[7] VGG16[8] and ResNet152-V2[9] as embedding models for the present study.

We used an Intel® Core™ i7-1165G7 CPU, and 32 GB of RAM for all computations. The pre-processing, especially the computation of the embeddings, requires approximately 100 h of CPU time for all three datasets for all embedding models and layers. The active learning experiments require approximately 6 h of CPU time per random seed for all three datasets. Table 1 shows the CPU time required for all sampling strategies, normalized to random sampling.

Table 1
Required computational time for all active learning (AL) strategies. The computational time is normalized, i.e. it is divided by the time required for random sampling.

AL Family	AL Strategy	$\frac{t_{strategy}}{t_{random}}$
uncertainty	least confidence avg	1.06
	entropy avg	1.23
	ratio avg	1.10
	ratio max	1.09
diversity	clustering	2.04
adaptive	adapt uncertainty	1.98
	adapt diversity	9.98
combination	ratio max + clustering	2.01
	adapt uncert + clustering	3.86
	ratio max + adapt div	12.20
	adapt uncert + adapt div	16.96

3. 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.

3.1. *Transfer learning*

We start by testing different pre-trained models as feature extractors for species-level sound event detection. All performance metrics are computed on the held-out evaluation sets described in section 2.1.

To gain intuition on the potential of each embedding model, we generate low-dimensional neighbor visualizations for high-dimensional embeddings of samples using UMAP, a neighbor embedding method that aims to preserve the distances between points observed in the high-dimensional embedding space within the low-dimensional representation (McInnes et al., 2020). The visualization in Fig. 4 shows that the BirdNet embeddings exhibit 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 multi-label classifiers on embeddings derived from the AnuraSet (frequent, common, rare and all), Noronha set, and Watkins datasets, using all pre-trained models. The quantitative results presented in Table 2 are largely consistent with the intuitions afforded by the neighborhood embedding visualizations, with BirdNet performing best, followed by intermediate layers of VGG16, albeit with much lower dimensionality.

Overall, we find that BirdNet-1 performes best as a feature extractor for multi-label classification for the utilized PAM datasets. The analysis of the frequent, common and rare parts of AnuraSet shows that this result is independent of the number of positive samples. Fig. 5 shows the single class F1 score for samples embedded with BirdNet-1 for each of the 42 classes of 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.

3.2. *Active learning*

We investigate the effect of active learning by emulating the annotation of the common partition of the AnuraSet, the Noronha set and the Watkins dataset. Due to the superior performance of the transfer

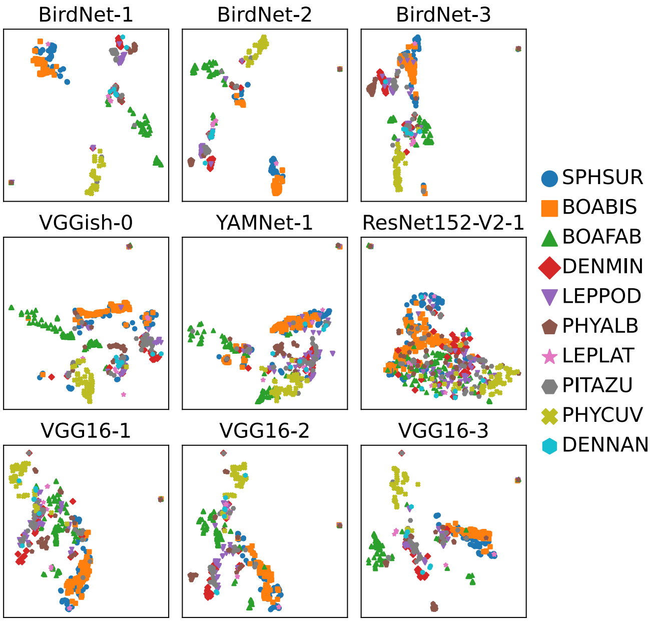


Fig. 4. UMAP plots for different embedding layers of different embedding models for AnuraSet. For UMAP generation, we randomly select 5000 samples and discard all samples that are aligned to more than one class. Colors and shapes indicate the 10 classes with the highest occurrence frequency. Layers are numbered according to their distance from the classification layer, e.g. ‘BirdNet-1’ is the last layer before the classification layer of the BirdNet model.

[4] https://github.com/HKathman/pam_annotation_experiments

[5] https://github.com/kahst/BirdNET-Analyzer/tree/main/checkpoints/V2.4

[6] https://tfhub.dev/google/vggish/1

[7] https://tfhub.dev/google/yamnet/1

[8] tensorflow.keras.applications.vgg16.VGG16(weights=’imagenet’).

[9] tensorflow.keras.applications.resnet_v2.ResNet152V2 (weights=’imagenet’).

Table 2
Size and performance of embedding layers from different transfer learning models. The layers are labelled in reverse order, with layer 1 being the last layer before the classification layer. We analysed the frequent, common and rare part as well as the whole dataset of AnuraSet, the Noronha set and the Watkins dataset. We provide micro (Mic) and macro (Mac) F1 scores calculated for the evaluation set. Each score represents the average result of 30 runs.

				AnuraSet											
		Layer		Frequent		Common		Rare		All		Noronha set		Watkins	
Model	Pre-Training	#	Size	Mic F1	Mac F1	Mic F1	Mac F1	Mic F1	Mac F1	Mic F1	Mac F1	Mic F1	Mac F1	Mic F1	Mac F1
BirdNet	Bird vocalisations	1	1024	**0.901**	**0.888**	**0.791**	**0.789**	0.487	0.406	**0.797**	**0.588**	**0.747**	0.610	**0.393**	**0.378**
		2	6144	0.870	0.858	0.764	0.763	0.495	0.451	0.752	0.575	0.729	0.541	0.353	0.340
		3	4608	0.869	0.855	0.766	0.764	**0.516**	**0.453**	0.765	0.578	0.726	0.517	0.356	0.347
VGGish	AudioSet	0	128	0.612	0.567	0.271	0.227	0.005	0.034	0.409	0.313	0.553	0.641	0.145	0.155
YAMNet	AudioSet	1	1024	0.748	0.702	0.479	0.440	0.077	0.127	0.560	0.406	0.632	**0.728**	0.128	0.128
VGG16	ImageNet	1	4096	0.690	0.594	0.394	0.393	0.032	0.100	0.492	0.374	0.335	0.396	0.088	0.120
		2	4096	0.696	0.608	0.443	0.449	0.059	0.146	0.504	0.388	0.324	0.382	0.146	0.165
		3	25,088	0.856	0.829	0.707	0.692	0.370	0.337	0.726	0.513	0.502	0.581	0.294	0.280
ResNet152-V2	ImageNet	1	2048	0.695	0.619	0.050	0.066	0.001	0.007	0.159	0.128	0.145	0.177	0.033	0.070

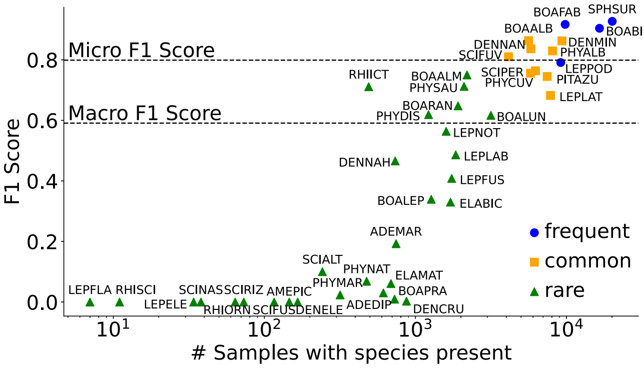


Fig. 5. Transfer learning applied to AnuraSet using features extracted from the last layer before the classification layer of BirdNet. A linear classifier (logistic regression) is used. The resulting F1 score for each species is plotted against the number of samples containing that species. Frequent, common and rare species are defined according to (Cañas et al., 2023).

learning results of the BirdNet-1 embedding, we use this embedding as the feature extractor for all subsequent active learning experiments.

From a machine learning perspective, the two objectives outlined in the beginning of Section 3 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 report results for the held-out evaluation sets described in Section 2.1. As illustrated in Fig. 1, the process of annotating an entire dataset using active 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.

To gain intuition on the potential of active learning, we compare the detection rate over time using random sampling and active learning. The 'Ground Truth' curve of Fig. 6 shows the cumulative occurrences of the class DENMIN (AnuraSet) throughout the day, aggregated across all samples recorded at each respective time. We present the number of positive detections using different numbers of training samples. The initial training set of 20 samples is identical for both sampling strategies, resulting in identical curves for random sampling and active learning. The F1 score is 0.01, and the curve remains constant over time. Gradually adding training samples increases the F1 score and reshapes the curve to resemble the 'Ground Truth'. With a small number of training samples (100 or 200), the training set selected by active learning yields significantly more occurrence detections and a higher F1 score. This effect is less pronounced when using 1000 training samples.

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

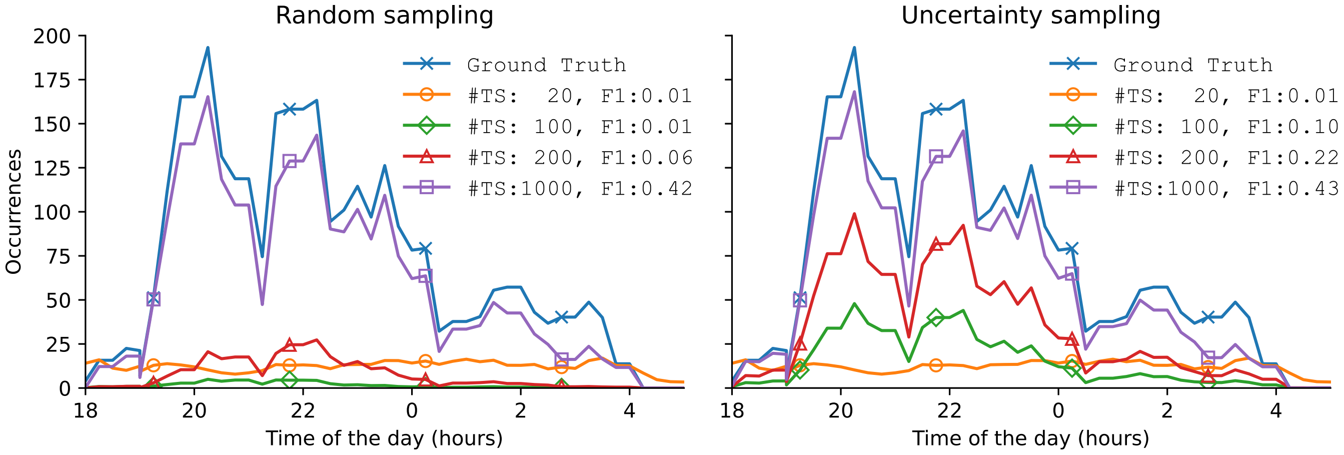


Fig. 6. Cumulative occurrence detections for a single class of AnuraSet (DENMIN) over time of day, aggregated across all samples recorded at each respective time. Shown are the ground truth and the number of positive detections using 20, 100, 200 and 1000 training samples (#TS) with the corresponding F1 score for the sampling methods random (left) and uncertainty (right). A moving average filter with a window size of 4 is applied to to each curve.

'average' ('avg'). The top row of Fig. 7 shows the results for the evaluation set of the three datasets. The score aggregation method 'max' consistently outperforms 'average' and surpasses random sampling.

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 for the evaluation set are shown in the center row of Fig. 7. For the common partition of AnuraSet and the Noronha set, the adaptive uncertainty method shows a slight performance advantage over other methods, with all methods outperforming random sampling. For the Watkins dataset, none of the strategies clearly outperform random sampling.

The bottom row of Fig. 7 shows the F1 score for the evaluation set 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 all datasets, all combinations used outperform random sampling, with ratio max + adaptive diversity being the best by a small margin.

Evaluating the results of the unlabelled datasets leads to comparable conclusions (see fig. S1).

Looking at precision and recall values of the uncertainty methods in Fig. 8, we observe a rapid convergence of precision for all methods. On the other hand, recall does not show any convergence and remains significantly lower than precision. 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. The comprehensive results presented in Figs. S2 to S4 yield consistent conclusions across all active learning methods for both the evaluation and unlabelled sets.

4. Discussion

This research investigates the combination of transfer learning and active learning to efficiently support and accelerate the detection of sound events in large, multi-label PAM datasets. In our study, we use AnuraSet, an expert-annotated multi-label PAM soundscape dataset unparalleled in terms of number of label classes (42 species of frogs and toads) and duration of annotated segments. To generalise our findings, we also use the Noronha set, a smaller, unpublished, multi-label dataset expert-annotated for seabirds, and Watkins, which we created synthetically by placing sound events from the Watkins Marine Mammal Sound Database (Sayigh et al., 2016) into a noisy background, resulting in a multi-label dataset. By exploring the applicability of different embedding layers of embedding models, each with varying degrees of similarity to the target modality of the acoustic data and the domain of passive acoustic monitoring, we observe that the penultimate layer of BirdNet (Kahl et al., 2021), a CNN trained on data most closely related to both the target modality and the domain, yields the best performance as a feature embedding model. Using BirdNet-1 as the embedding model, active learning sample selection strategies significantly reduce the number of samples required for annotation to achieve model convergence.

Automatic sound event detection for multi-label PAM datasets requires large training sets of multi-label PAM data. While PAM data inherently consist of multi-label annotations, the considerable time required for annotation (Lüers et al., 2024) explains why most PAM datasets are published with single or small sets of label classes. The proposed approach for accelerating the generation of multi-label PAM datasets differs from conventional workflows by applying feature extraction techniques on the dataset and subsequently selecting the most informative samples to speed up model convergence. While using a pipeline of transfer learning and active learning has shown to reduce annotation time in various domains such as labelling images from camera traps (Norouzzadeh et al., 2021) and PAM data (van Osta et al., 2023), our study presents the first systematic investigation of different embedding models and active learning strategies for annotating PAM

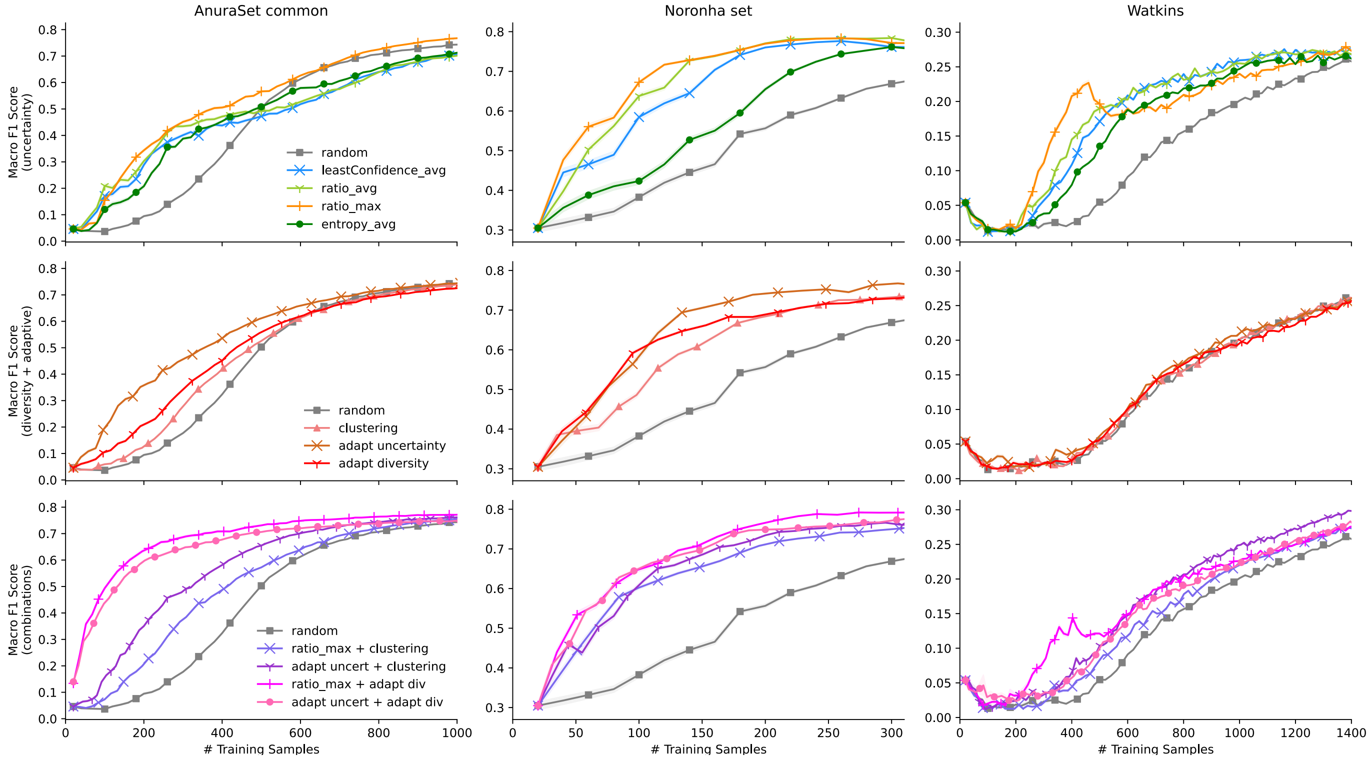


Fig. 7. Active learning on the common partition of AnuraSet, Noronha set, and Watkins using the embeddings of BirdNet-1. Macro F1 score computed on evaluation data. Mean ± SEM across 30 independent runs. *Top:* uncertainty-based sampling strategies ('least confidence', 'ratio' and 'entropy') and score 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.

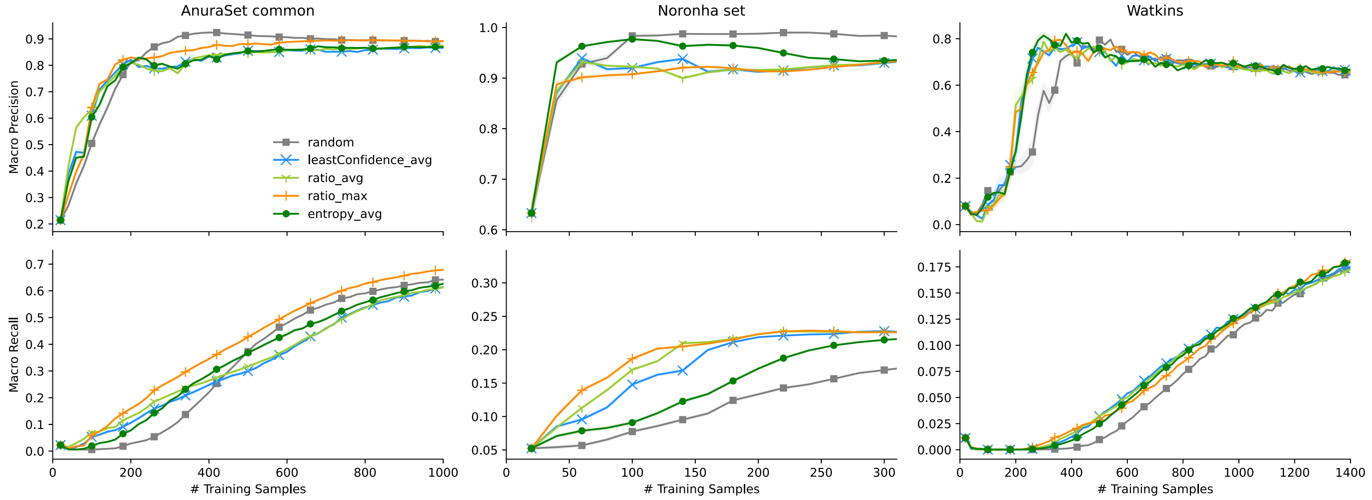


Fig. 8. Precision and recall values for the uncertainty sampling methods of the common partition of AnuraSet, Noronha set, and Watkins dataset.

datasets.

When evaluating embedding models across domains, we observe that higher similarity to the target domain and modality corresponds to improved performance, extending the findings of Ghani et al. (2023) to multi-label PAM datasets. Layers further away from the classification layer tend to capture more abstract features from the data on which they were trained (Bengio, 2009). This phenomenon might explain the performance gain with increasing distance for embedding models that are less related to the target domain and modality (see Table 2, VGG16). Conversely, embedding models that are closely related to the target domain and modality show improved performance by minimising the distance (see Table 2, BirdNet). Overall, we identify the penultimate layer of BirdNet as the most informative for creating embeddings from PAM data. 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 analysis using BirdNet-1 embeddings demonstrates the superiority of active learning over random sampling for multi-label PAM datasets. Most previous active learning efforts in sound event detection for PAM (Allen et al., 2021; Qian et al., 2017) have not utilized features extracted with transfer learning. We build upon the research conducted by van Osta et al. (2023), which implemented active learning on PAM data embedded using a ResNet variant, seemingly pre-trained on ImageNet. First, we compare various embedding models and demonstrate that models trained on data more closely related to PAM than to ImageNet yield superior performance. Second, we provide the first structured analysis of several standard active learning strategies for annotating PAM datasets. Fig. 6 illustrates that the performance of the active learning model exceeds that of the random sampling model in the first iterations. While our qualitative evaluation of active learning methods consistently shows superior performance compared to random sampling, no single method clearly outperforms all others. 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. We find the score aggregation ‘max’ superior to ‘average’ for uncertainty methods.

When examining the precision and recall values for the active learning methods, we observe a rapid saturation of precision and low recall values. 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 emphasising specificity to false positives. In our methodology, a similar approach could be realised by adjusting learning to penalise 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 optimisation for accuracy metrics, exemplified by F1 Score. Instead, our primary goal was to identify efficient strategies that synergise transfer learning and active learning. To potentially elevate accuracy, strategies such as applying Per-Channel Energy Normalization (PCEN) (Lostanlen et al., 2019), refining spectrogram feature engineering (Dufourq et al., 2022), or employing transfer learning with fine-tuning could be explored.

The present elaboration lays the theoretical foundation for several future research directions. Although studies have shown a correlation between the performance of embedding models and the degree of similarity to the target modality and domain, there is currently no pre-trained model specifically trained on multi-label PAM data. Training such a model is future work that requires extensive training data, especially multi-label annotated PAM data. Implementing a tool based on the presented workflow could streamline the process of collecting such data.

In the field, PAM data is recorded either continuously for permanent monitoring or temporarily to create a new dataset. Continuous monitoring results in a consistent data distribution, which we analyse using the evaluation set. In contrast, temporary monitoring results in a changing data distribution, which we analyse using the unlabelled set. Starting from a completely unlabelled stream or pool of audio data, the combination of transfer learning and active learning provides the theoretical basis for creating an annotation tool that can save annotation time in two ways: Continuous biodiversity monitoring requires automated detection of sound events due to the large amount of incoming data. A tool based on the proposed workflow would require fewer annotated samples to achieve the same model performance, thus reducing annotation time. In addition, when annotating a complete PAM dataset, the proposed workflow can accelerate the process by suggesting labels based on previously sampled data. The user can then adapt these suggestions instead of creating all labels from scratch, further saving annotation time.

CRediT authorship contribution statement

Hannes Kath: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Conceptualization. **Patricia P. Serafini:** Data curation. **Ivan B. Campos:** Resources, Writing – review & editing, Data curation. **Thiago S. Gouvêa:** Conceptualization, Project administration, Writing – review & editing, Writing – original draft, Supervision, Formal analysis. **Daniel Sonntag:** Project administration, Writing – review &

editing, Funding acquisition.

Acknowledgements

This research is part of the Computational Sustainability & Technology Project Field (https://cst.dfki.de/), 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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2024.102710.

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. Front. Mar. Sci. 8.

Bengio, Y., 2009. Learning deep architectures for AI. Found. Trends Mach. Learn. 2 (1), 1–127.

Bicudo, T., Llusia, D., Anciães, M., Gil, D., 2023. Poor performance of acoustic indices as proxies for bird diversity in a fragmented Amazonian landscape. Eco. Inform. 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. Ecol. Indic. 120, 106953.

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

Çoban, E., Pir, D., So, R., 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), pp. 726–730.

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), pp. 248–255.

Dufourq, E., Batist, C., Foquet, R., Durbach, I., 2022. Passive acoustic monitoring of animal populations with transfer learning. Eco. Inform. 70, 101688.

Florentin, J., Dutoit, T., Verlinden, O., 2020. Detection and identification of European woodpeckers with deep convolutional neural networks. Eco. Inform. 55, 101023 (25 citations (Semantic Scholar/DOI) [2024-04-30).

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), pp. 776–780.

Ghani, B., Denton, T., Kahl, S., Klinck, H., 2023. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. Sci. Rep. 13 (1), 22876.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks. In: Computer Vision – ECCV 2016, Lecture Notes in Computer Science, pp. 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), pp. 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., Sonntag, D., 2023. EdgeAL: an edge estimation based active learning approach for OCT segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI, 2023, pp. 79–89.

Kahl, S., Wood, C., Eibl, M., Klinck, H., 2021. BirdNET: a deep learning solution for avian diversity monitoring. Eco. Inform. 61, 101236.

Kholghi, M., Phillips, Y., Towsey, M., Sitbon, L., Roe, P., 2018. Active learning for classifying long-duration audio recordings of the environment. Methods Ecol. Evol. 9 (9), 1948–1958, 9 citations (Semantic Scholar/DOI) [2024-04-30]_eprint. https://doi.org/10.1111/2041-210X.13042.

Konyushkova, K., Sznitman, R., Fua, P., 2017. Learning active learning from data. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, pp. 4225–4235.

Lauha, P., Somervuo, P., Lehikoinen, P., Geres, L., Richter, T., Seibold, S., Ovaskainen, O., 2022. Domain-specific neural networks improve automated bird sound recognition already with small amount of local data. Methods Ecol. Evol. 13 (12), 2799–2810, 11 citations (Semantic Scholar/DOI) [2024-04-30] _eprint. https://doi.org/10.1111/2041-210X.14003.

LeBien, J., Zhong, M., Campos-Cerqueira, M., Velev, J.P., Dodhia, R., Ferres, J.L., Aide, T. M., 2020. A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. Eco. Inform. 59, 101113 (82 citations (Semantic Scholar/DOI) [2024-04-30).

Lostanlen, V., Salamon, J., Cartwright, o., 2019. Per-Channel energy normalization: why and how. IEEE Sign. Proc. Lett. 26 (1), 39–43.

Lüers, B., Serafini, P.P., Campos, I.B., Gouvêa, T.S., Sonntag, D., 2024. BirdNET-annotator: AI-assisted strong labelling of bird sound datasets. In: 3rd Annual AAAI Workshop on AI to Accelerate Science and Engineering (AI2ASE). Vancouver, Canada.

Machado, R.B., Aguiar, L., Jones, G., 2017. Do acoustic indices reflect the characteristics of bird communities in the savannas of Central Brazil? Landsc. Urban Plan. 162, 36–43 (66 citations (Semantic Scholar/DOI) [2024-04-30).

McGinn, K., Kahl, S., Peery, M., et al., 2023. Feature embeddings from the BirdNET algorithm provide insights into avian ecology. Eco. Inform. 74, 101995.

McInnes, L., Healy, J., 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.).

Nolasco, I., Singh, S., Morfi, V., Lostanlen, V., Strandburg-Peshkin, A., Vidaña-Vila, E., Gill, L., Pamuła, H., Whitehead, H., Kiskin, I., Jensen, F.H., Morford, J., Emmerson, M.G., Versace, E., Grout, E., Liu, H., Ghani, B., Stowell, D., 2023. Learning to detect an animal sound from five examples. Eco. Inform. 77, 102258 (12 citations (Semantic Scholar/DOI) [2024-04-30).

Norouzzadeh, M.S., Morris, D., Beery, S., Joshi, N., Jojic, N., Clune, J., 2021. A deep active learning system for species identification and counting in camera trap images. Methods Ecol. Evol. 12 (1), 150–161, 118 citations (Semantic Scholar/DOI) [2024-04-30] _eprint. https://doi.org/10.1111/2041-210X.13504.

Qian, K., Zhang, Z., Baird, A., Schuller, B., 2017. Active learning for bird sound classification via a kernel-based extreme learning machine. J. Acoust. Soc. Am. 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. Funct. Ecol. 37 (4), 959–975.

Sayigh, L., Daher, M., Allen, J., Gordon, H., Joyce, K., Stuhlmann, C., Tyack, P., 2016. The Watkins marine mammal sound database: an online, freely accessible resource. In: Proceedings of Meetings on Acoustics, 27, p. 040013.

Sethi, S., Bick, A., Ewers, R., et al., 2023. Limits to the accurate and generalizable use of soundscapes to monitor biodiversity. Nat. Ecol. Evol. 1–6.

Simonyan, K., 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., Pavoine, S., Hamerlynck, O., Duvail, S., 2008. Rapid acoustic survey for biodiversity appraisal. PLoS One 3 (12), e4065.

Sueur, J., Farina, A., Gasc, A., et al., 2014. Acoustic indices for biodiversity assessment and landscape investigation. Acta Acust. Acust. 100 (4), 772–781.

Sugai, L., Llusia, D., 2019. Bioacoustic time capsules: using acoustic monitoring to document biodiversity. Ecol. Indic. 99, 149–152.

Sugai, L., Silva, T., Ribeiro, J., Llusia, D., 2019. Terrestrial passive acoustic monitoring: review and perspectives. BioScience 69 (1), 15–25.

Swaminathan, B., Jagadeesh, M., Vairavasundaram, S., 2024. Multi-label classification for acoustic bird species detection using transfer learning approach. Eco. Inform. 80, 102471 (3 citations (Semantic Scholar/DOI) [2024-04-30).

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), pp. 5063–5069.

Tsalera, E., Papadakis, A., Samarakou, M., 2021. Comparison of pre-trained CNNs for audio classification using transfer learning. J. Sens. Actuator Netw. 10 (4), 72.

van Osta, J.M., Dreis, B., Meyer, E., Grogan, L.F., Castley, J.G., 2023. An active learning framework and assessment of inter-annotator agreement facilitate automated recogniser development for vocalisations of a rare species, the southern black-throated finch (*Poephila cincta cincta*). Eco. Inform. 77, 102233 (2 citations (Semantic Scholar/DOI) [2024-04-30).

Wang, Y., Li, J., 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), pp. 31–35.

Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M., 2020. Generalizing from a few examples: a survey on few-shot learning. ACM Comput. Surv. 53 (3), 63, 1–63:34.

Wang, Y., Cartwright, M., Bello, J., 2022. Active few-shot learning for sound event detection. In: Interspeech, 2022, pp. 1551–1555.