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

Anonymous submission

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 leveraging pre-trained models from related tasks or datasets to bootstrap the learning process for sound event detection. Furthermore, active learning enables strategic selection of the most informative samples for annotation, effectively reducing the annotation burden and improving model performance. We propose a novel 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 vocalizations, can be effectively used for predicting anurans in PAM data: a linear classifier constructed using these embeddings outperforms the benchmark by a noteworthy 21.7 %. In terms of active learning, our results indicate a superiority over random sampling, although no clear winner emerged 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.

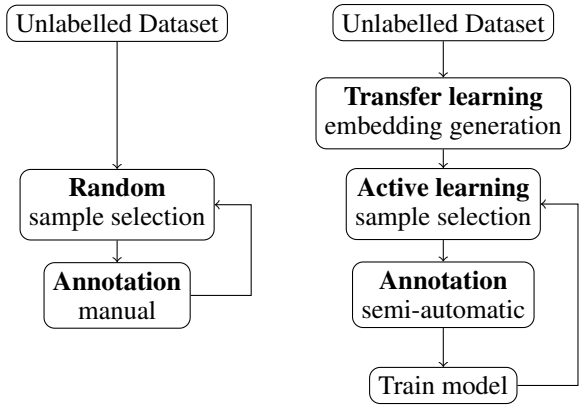

Figure 1: Workflow for annotating passive acoustic monitoring datasets, comparing the conventional approach (left) with the proposed approach (right).

Sound event detection in PAM datasets traditionally requires domain-specific annotated data, 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. To address this issue, we propose a novel approach that leverages transfer learning and active learning to optimize sound event detection in PAM datasets.

Transfer learning has shown 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. Instead of randomly labeling all data points, an active learning algorithm 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

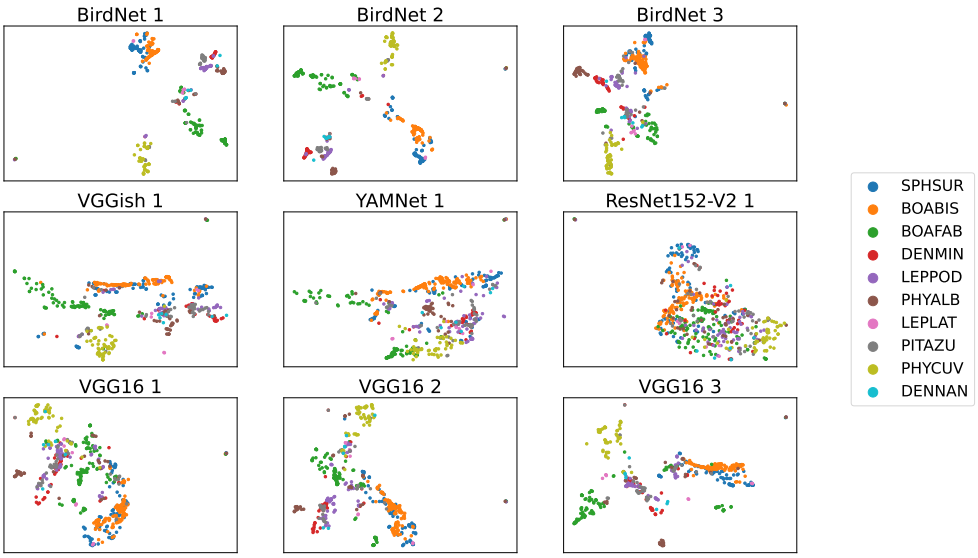

Figure 2: UMAP plots for diffenerent embedding layers of different embedding models for AnuraSet. Colors indicate top 10 classes. For UMAP generation, we selected randomly 5000 samples. In the plot we show only samples that are aligned to exactly one class.

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 workload. 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.

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 bioacoustics datasets. Species identification can be formulated as (multi-label) sound event detection, a machine learning problem for which convolutional neural networks (CNNs) are known to work well (Hershey et al. 2017). While the idea of applying CNNs for species identification in bioacoustics is not new (see Stowell 2022 for a survey), real life applications are often limited by scarcity of annotated data. In fact, deep learning models for species detection in PAM datasets are often trained with focal recordings (e.g., (Kahl et al. 2021)). 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) used foundation CNNs pre-trained on large image (ImageNet) and sound (AudioSet) datasets to solve sound event detection tasks, and found that models pre-trained on the audio domain perform better. Dufourq et al. (2022) applied transfer learning to PAM datasets. They compared 12 different CNN architectures pretrained on ImageNet as feature extractors for single-species detection in PAM datasets, and found that ResNets (101V2, 152V2) (He et al. 2016) performed best, followed by VGG16 (Simonyan and Zisserman 2015); Dufourq et al. did not explore models

Model	Pre-Training	Layer		AnuraSet		Noronha	
		# from last	Size	Micro F1	Macro F1	Micro F1	Macro F1
BirdNet	Bird vocalizations	1	1024	**0.800**	**0.595**	0.693	**0.786**
		2	6144	0.759	0.570	0.682	0.562
		3	4608	0.771	0.588	**0.714**	0.541
VGGish	AudioSet	1	128	0.406	0.300	0.256	0.302
YAMNet	AudioSet	1	1024	0.563	0.412	0.439	0.512
VGG16	ImageNet	1	4096	0.487	0.370	0.068	0.083
		2	4096	0.519	0.403	0.263	0.372
		3	25088	0.729	0.509	0.365	0.515
ResNet152-V2	ImageNet	1	2048	0.154	0.129	0.042	0.051

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 provide micro and macro F1 scores calculated for the test set of AnuraSet and Noronha dataset. Each F1 score represents the average result of multiple independent runs with different random seed values. Due to the constant standard deviation of less than 0.05, it is not included in the table.

pre-trained on audio datasets. Çoban et al. (2020) used VGGish, a VGG variant pre-trained on AudioSet, to detect sound events in a PAM dataset; the event classes were coarse grained (e.g., 'songbird', 'waterbird', and 'insect') as opposed to fine grained (e.g., species identity). McGinn et al. (2023) investigated 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 found 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.

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) used a synthetic dataset built by recombining environmental sounds with urban soundscape background to study how active learning can improve aupon random selection in the context of prototype based classification with models trained with few-shot episodes. Qian et al. (2017) used active learning to improve on the data efficiency of bird species classifiers applied to a museum sound collection (likely focal recordings); their classifiers operated on low level descriptors, which are interpretable feature extractors that might afford lower performance than deep learning methods. Allen et al. (2021) used active learning to detect humpback whale songs (single species) in a very large PAM dataset (187,000 h); they used 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) precluded 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) implemented 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 We took advantage of AnuraSet, a recently released benchmark multi-species PAM dataset comprised of 27 hours of manually created expert annotations for 42 species of anurans (frogs and toads) from two different biomes (Cañas et al. 2023). In addition, we used a novel small, manually annotated portion of a multi-year PAM dataset acquired collected in Fernando de Noronha, Brazil. The selected portion, here referred to as the Noronha set, consists of 75 minutes annotated by an expert for 5 species of oceanic birds. For all experiments, we divided each audio file into 3-second segments referred to as 'samples'. A sample was considered positive for a given event class whenever event occurrence overlapped 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 dataset, we randomly selected a third of the dataset.

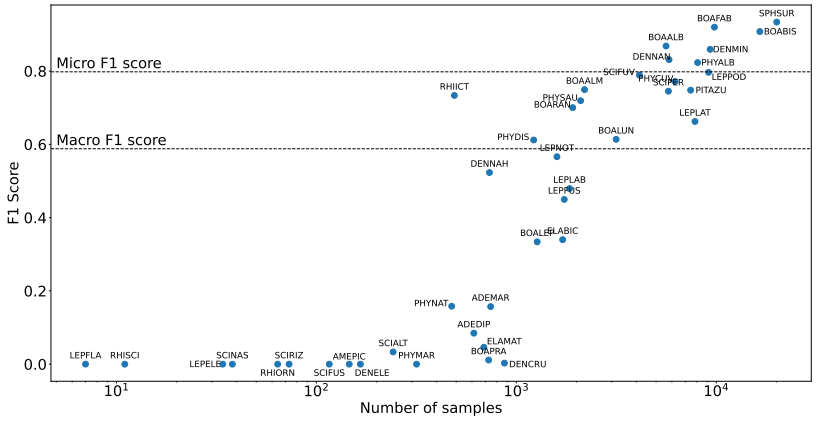

Figure 3: Number of samples for each species in the AnuraSet, and the associated F1 score for a linear classifier (logistic regression) on features extracted with BirdNet.

Transfer Learning We explored 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 of different domains and modalities, with varying degrees of similarity to the target modality (audio) and domain (multiple anuran species in PAM data). Following (Dufourq et al. 2022), we tested ResNet152-V2 (He et al. 2016) and VGG16 (Simonyan and Zisserman 2015); these are CNNs pretrained on ImageNet (Deng et al. 2009), a dataset on the visual modality. VGGish[1], a variant of VGG11A (Simonyan and Zisserman 2015), and YAMNet[2], 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 more directly related to the predictions it makes (Bengio 2009). We evaluated embeddings at different layers within the CNNs (table 1). For VGG16 we investigated the last three layers before the final classification layer ('fc2', 'fc1', and 'flatten'). For ResNet152-V2 we only investigated the last embedding layer ('avg_pool'). Considering our future goal of implementing a real-time pipeline with transfer learning and active learning, we decided 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 for use as feature extractors, for those we only used the last embedding layer. For BirdNet we investigated 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 have not investigated 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).

Low dimensional representations of AnuraSet embeddings in each model/layer combination were generated with UMAP, a neighbor embedding method that seeks to preserve in the lower dimensional representation the same distance between points observed in the high dimensional embedding space (McInnes, Healy, and Melville 2020). UMAP embeddings were computed for a subset of 5000 random samples from the top 10 classes from AnuraSet. Sound event detection performance afforded by each embedding was evaluated with a linear classifier (single fully connected layer). Since samples might contain overlapping calls from different species, we treated it as a multi-label classification problem and used logistic activation and binary cross-entropy loss function. Linear classifiers were trained on frozen embeddings (no fine tuning) for up to 1000 epochs with early stopping based on validation loss (minimum delta of 0.1, a patience of 10 epochs, with reinstatement of best weights). In cases where the embedding models output an array of time points for each input sample, we treated it as a multiple instance learning problem by applying the classifier to each time point, and then pooling predictions with an exponential softmax (Wang, Li, and Metze 2019) (see appendix A.1 for detailed information). Metrics reported are macro F1 score (table and main figures), and macro precision and recall (appendix). Results are reported as mean computed over multiple runs with different random seeds; spread reported in all figures is standard error of the mean (SED).

Active Learning We explored a number of sampling strategies, both uncertainty- and diversity-based, myopic (greedy) and adaptive (batch mode), as well as combinations thereof. Importantly, in all cases 5 % of samples were chosen at random. Uncertainty scores were computed with least confidence, ratio, and entropy methods (see

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

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

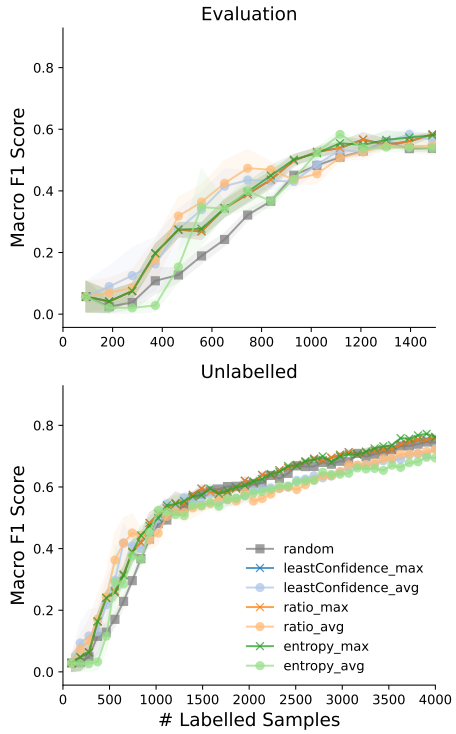

Figure 4: Active learning on AnuraSet with uncertainty-based sampling strategies ('least confidence', 'ratio' and 'entropy') and score aggregation methods ('max' and 'average'). Macro F1 score computed on evaluation data (top), and on the portion of the training data that remains unlabelled (bottom). Mean ± SEM across 5 independent runs.

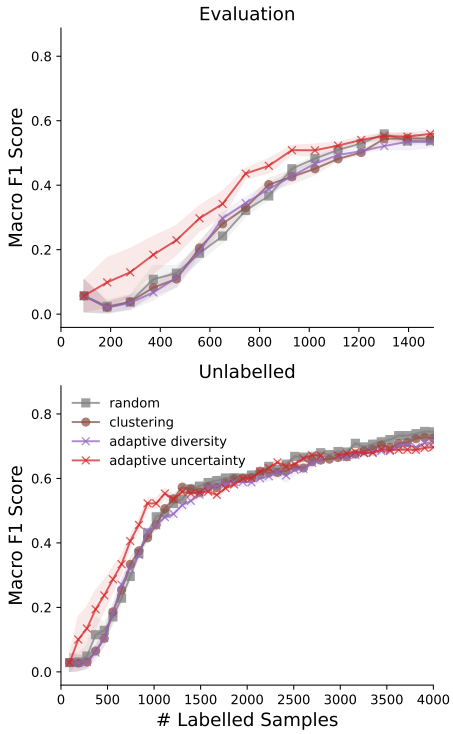

Figure 5: Active learning on AnuraSet with a **diversity**-based sampling strategy ('clustering') and two **adaptive** ones (see text on sections 3 and A.2 for details). Macro F1 score computed on evaluation data (top), and on the portion of the training data that remains unlabelled (bottom). Mean ± SEM across 5 independent runs.

appendix A.2 for details). In multi-label problems, each sample gives rise to an array of uncertainty scores (one per class); we aggregate scores for each sample by taking either the maximum of the average across classes (reported separately). Note that all uncertainty score methods decay monotonically away from 0.5 and thus share the same maximum. As a diversity method we used an established k-means based clustering technique (Monarch 2021) described in detail in appendix A.2. We further investigated adaptive, non-myopic sampling methods, which minimize batch redundancy; we selected one adaptive method for uncertainty sampling and one for diversity sampling. For details on adaptive sampling methods as well as all combinations, we refer to appendix A.2.

Class labels are available for all samples used in this study, and an active learning scenario was 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. Batch sizes represented 1 % of the data for AnuraSet, and 0.1 % for Anuraset. The classifier heads used were identical to those from the transfer learning experiment, always applied to data embeddings with the same selected model (BirdNet-1, see section 4.1).

4 Results

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. 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.

4.1 Transfer Learning

We started 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 generated low dimensional neighbor embedding visualizations for high dimensional embeddings of samples from the top 10 classes of AnuraSet (figure 2). The BirdNet embeddings show a clear separation between class clusters, with more pronounced differentiation in layers closer to the final layer. VGGish and YAMNet

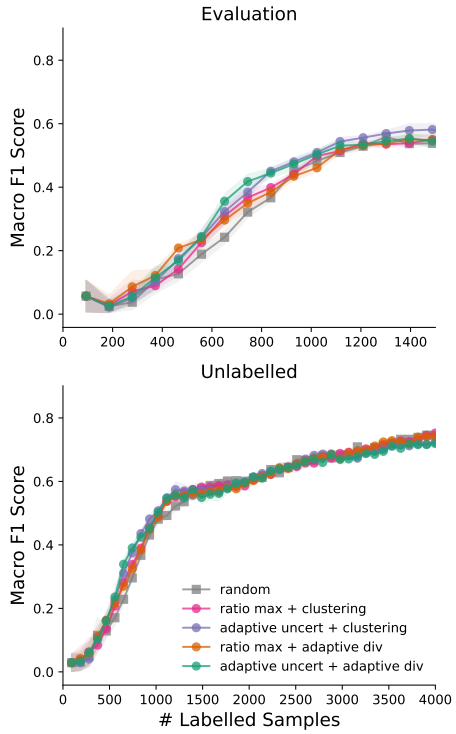

Figure 6: Active learning on AnuraSet with **mixed** diversity- and uncertainty-based sampling strategies (see text on sections 3 and A.2 for details). Macro F1 score computed on evaluation data (top), and on the portion of the training data that remains unlabelled (bottom). Mean ± SEM across 5 independent runs.

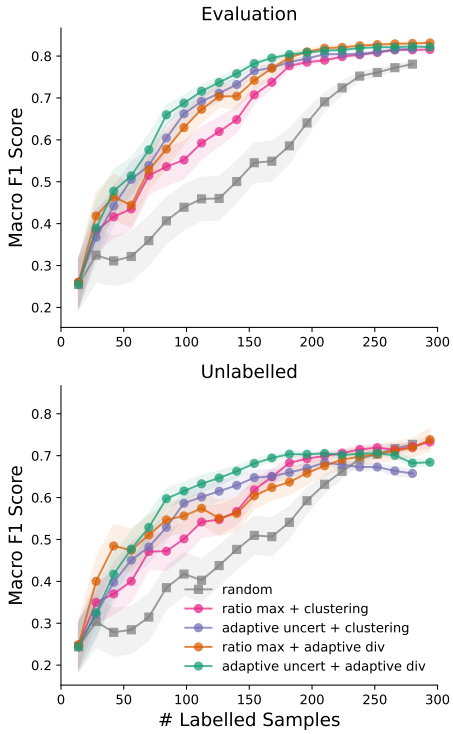

Figure 7: Active learning on the **Noronha** dataset with the same mixed diversity- and uncertainty-based sampling strategies as in figure 6. Macro F1 score computed on evaluation data (top), and on the portion of the training data that remains unlabelled (bottom). Mean ± SEM across 30 independent runs.

show effective cluster separation for only a subset of clusters, while ResNet152-V2 embeddings appeared as a continuum, salt-and-pepper pattern in the low dimensional representation. Cluster separation was visible for VGG16, with more apparent separation for layers further away for the top.

Next, we trained linear classifiers on embeddings of both AnuraSet and Noronha datasets with all pre-trained models. These quantitative results largely match the intuitions afforded by neighbor embedding visualizations, with BirdNet-1 performing best, followed by intermediate layers of VGG16 (albeit with a much smaller dimensionality, see table 1).

Overall, we found that BirdNet-1 performed best as a feature extractor for multi-label classification for the PAM datasets AnuraSet and Noronha dataset, resulting in the best macro F1 scores in both cases. Figure 3 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 macro F1 score and class size, and consequently a wide gap between macro and micro F1 scores. BirdNet-1 was used as a feature extractor for all active learning experiments.

4.2 Active Learning

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 constructed evaluation sets that reflected the distribution of the original training data. For the AnuraSet, we used the identical evaluation set used by the original authors (Cañas et al. 2023). For the Noronha dataset, we randomly selected a third of the data to form the evaluation set. As illustrated in figure 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 full dataset and the remaining unlabelled subset. Consequently, we will present results specific to this remaining unlabelled subset. In all active learning experiments, we used the embeddings generated by BirdNet 1, which showed the best overall performance. Due to the significant imbalance of classes in AnuraSet and the Noronha dataset, we used the macro F1 score as the evaluation metric, and provide the corresponding macro precision and macro recall values in appendix B. As a baseline, all figures show the performance of random

sampling. This subsection focuses on presenting the results for the AnuraSet, as the AnuraSet contains a larger amount of annotated data and includes more species than the Noronha dataset. We provide all further results for the AnuraSet and the Noronha dataset in appendix B.

We investigated the uncertainty sampling strategies 'least confidence', 'ratio' and 'entropy' with the score aggregation methods 'max' and 'average' ('avg'). Figure 4 shows the F1 score. For the evaluation set, all uncertainty sampling strategies show superior performance compared to random sampling, reaching convergence just below an F1 score of 0.6. No uncertainty sampling strategy clearly outperforms the others. Looking at precision and recall in figure 8, we observe a rapid convergence of precision, slightly above 0.8. On the other hand, recall does not show any convergence and remains significantly lower than precision at around 0.2. The remaining unlabelled subset shows similar F1 score behaviour. Differences include the lack of convergence and the superior performance of the 'max' score aggregation method over the 'average' method. This contrast is vividly illustrated in Figure 8, where the 'max' score aggregation consistently produces higher recall values.

We further investigated the diversity sampling strategy of 'clustering' and explored two adaptive approaches – one for uncertainty and the other for diversity. The results of the F1 score are shown in figure 5. For the evaluation set, the adaptive uncertainty method clearly outperforms the others, which perform similarly to random sampling. Figure 9 show similar results for precision and recall as the uncertainty sampling strategies, demonstrating the superior performance of the adaptive uncertainty method for recall. The remaining unlabelled subset shows a comparable pattern in F1 score behaviour, highlighting the improved performance of the adaptive uncertainty sampling strategy in terms of F1 score, attributed to its superior recall performance.

Figure 6 shows the F1 score for the combined sampling strategies described in appendix A.2. We chose the 'ratio max' uncertainty sampling strategy for the combination due to the superior performance of the 'max' versions in certain scenarios and the simplicity of calculating the ratio. In the evaluation set, all strategies perform slightly better than random sampling, with no single method emerging as the clear best. Figure 10 illustrates that random sampling achieved a higher accuracy compared to the combined strategies, albeit with a lower recall. The remaining unlabelled subset shows similar F1 scores for all combined sampling strategies and random sampling in figure 6. This similarity is also reflected in precision and recall, as shown in figure 10.

For the Noronha dataset, the results of the combined sampling strategies are very different, as shown in figure 7. For the evaluation dataset, all strategies clearly outperform random sampling, with the best being the combination of the adaptive uncertainty method and the adaptive diversity method. As shown in figure 13, this phenomenon can be attributed to the significantly higher precision of the combined sampling strategies.

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 last embedding layer of BirdNet, a CNN trained on focal recordings of bird vocalizations (Kahl et al. 2021), furnishes the most effective feature space. Notably, a 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 pioneers this intersection. In our exploration of uncertainty-based sampling strategies, originally designed for multi-class classification, we note 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 is 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 dataset, 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 raises 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 doesn't encompass the optimization for accuracy metrics, exemplified by F1 Score. Instead, our primary goal lies in identifying efficient strategies that synergize 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.

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.

6 Acknowledgments

Omitted in anonymous version.

7 Data and Code Availability

AnuraSet is a public dataset (Cañas et al. 2023). The Noronha dataset will be made publicly available in the near future.

References

Allen, A. N.; Harvey, M.; Harrell, L.; Jansen, A.; Merkens, K. P.; Wall, C. C.; Cattiau, J.; and Oleson, E. M. 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. 0 citations (Semantic Scholar/DOI) [2023-08-12].

Campos, I. B.; Fewster, R.; Truskinger, A.; Towsey, M.; Roe, P.; Vasques Filho, D.; Lee, W.; and Gaskett, A. 2021. Assessing the potential of acoustic indices for protected area monitoring in the Serra do Cipó National Park, Brazil. *Ecological Indicators*, 120: 106953. 6 citations (Semantic Scholar/DOI) [2023-08-12].

Cañas, J. S.; Toro-Gómez, M. P.; Sugai, L. S. M.; Restrepo, H. D. B.; Rudas, J.; Bautista, B. P.; Toledo, L. F.; Dena, S.; Domingos, A. H. R.; de Souza, F. L.; Neckel-Oliveira, S.; da Rosa, A.; Carvalho-Rocha, V.; Bernardy, J. V.; Sugai, J. L. M. M.; Santos, C. E. d.; Bastos, R. P.; Llusia, D.; and Ulloa, J. S. 2023. AnuraSet: A dataset for benchmarking Neotropical anuran calls identification in passive acoustic monitoring. 0 citations (Semantic Scholar/arXiv) [2023-08-12] 0 citations (Semantic Scholar/DOI) [2023-08-12] arXiv:2307.06860 [cs, eess].

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, 248–255. IEEE Computer Society.

Dufourq, E.; Batist, C.; Foquet, R.; and Durbach, I. 2022. Passive acoustic monitoring of animal populations with transfer learning. *Ecological Informatics*, 70: 101688. 12 citations (Semantic Scholar/DOI) [2023-08-12].

Gemmeke, J. F.; Ellis, D. P. W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 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. 1977 citations (Semantic Scholar/DOI) [2023-08-12] ISSN: 2379-190X.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Identity Mappings in Deep Residual Networks. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, 630–645. Cham: Springer International Publishing. ISBN 978-3-319-46493-0. 8224 citations (Semantic Scholar/DOI) [2023-08-12].

Hershey, S.; Chaudhuri, S.; Ellis, D. P. W.; Gemmeke, J. F.; Jansen, A.; Moore, R. C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; Slaney, M.; Weiss, R. J.; and Wilson, K. 2017. CNN Architectures for Large-Scale Audio Classification. 1746 citations (Semantic Scholar/arXiv) [2023-08-12] arXiv:1609.09430 [cs, stat].

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. 9994 citations (Semantic Scholar/arXiv) [2023-08-12] arXiv:1704.04861 [cs].

Kahl, S.; Wood, C. M.; Eibl, M.; and Klinck, H. 2021. BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61: 101236. 129 citations (Semantic Scholar/DOI) [2023-08-12].

Konyushkova, K.; Sznitman, R.; and Fua, P. 2017. Learning Active Learning from Data. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 4225–4235.

Lostanlen, V.; Salamon, J.; Cartwright, M.; McFee, B.; Farnsworth, A.; Kelling, S.; and Bello, J. P. 2019. Per-Channel Energy Normalization: Why and How. *IEEE Signal Processing Letters*, 26(1): 39–43. Conference Name: IEEE Signal Processing Letters.

McGinn, K.; Kahl, S.; Peery, M. Z.; Klinck, H.; and Wood, C. M. 2023. Feature embeddings from the BirdNET algorithm provide insights into avian ecology. *Ecological Informatics*, 74: 101995. 3 citations (Semantic Scholar/DOI) [2023-08-12].

McInnes, L.; Healy, J.; and Melville, J. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv:1802.03426 [cs, stat].

Monarch, R. M. 2021. *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-centered AI*. Simon and Schuster. ISBN 978-1-61729-674-1. Google-Books-ID: W2U0EAAAQBAJ.

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. 34 citations (Semantic Scholar/DOI) [2023-08-12].

Ross, S. R. P.-J.; O'Connell, D. P.; Deichmann, J. L.; Desjonquères, C.; Gasc, A.; Phillips, J. N.; Sethi, S. S.; Wood, C. M.; and Burivalova, Z. 2023. Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions. *Functional Ecology*, 37(4): 959–975. 5 citations (Semantic Scholar/DOI) [2023-08-12] _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1365-2435.14275.

Ryazanov, I.; Nylund, A. T.; Basu, D.; Hassellöv, I.-M.; and Schliep, A. 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. 6 citations (Semantic Scholar/DOI) [2023-08-12] Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Sethi, S. S.; Bick, A.; Ewers, R. M.; Klinck, H.; Ramesh, V.; Tuanmu, M.-N.; and Coomes, D. A. 2023. Limits to the accurate and generalizable use of soundscapes to monitor biodiversity. *Nature Ecology & Evolution*, 1–6. 0 citations (Semantic Scholar/DOI) [2023-08-12] Publisher: Nature Publishing Group.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. 9999 citations (Semantic Scholar/arXiv) [2023-08-03] arXiv:1409.1556 [cs] version: 6.

Stowell, D. 2022. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10: e13152. 69 citations (Semantic Scholar/DOI) [2023-08-12].

Sueur, J.; Farina, A.; Gasc, A.; Pieretti, N.; and Pavoine, S. 2014. Acoustic Indices for Biodiversity Assessment and Landscape Investigation. *Acta Acustica united with Acustica*, 100(4): 772–781. 300 citations (Semantic Scholar/DOI) [2023-08-12] 252 citations (Crossref) [2022-10-19].

Sueur, J.; Pavoine, S.; Hamerlynck, O.; and Duvail, S. 2008. Rapid Acoustic Survey for Biodiversity Appraisal. *PLOS ONE*, 3(12): e4065. 485 citations (Semantic Scholar/DOI) [2023-08-12] 357 citations (Crossref) [2022-10-19] Publisher: Public Library of Science.

Sugai, L. S. M.; and Llusia, D. 2019. Bioacoustic time capsules: Using acoustic monitoring to document biodiversity. *Ecological Indicators*, 99: 149–152. 34 citations (Semantic Scholar/DOI) [2023-08-12].

Sugai, L. S. M.; Silva, T. S. F.; Ribeiro, J. W.; and Llusia, D. 2019. Terrestrial Passive Acoustic Monitoring: Review and Perspectives. *BioScience*, 69(1): 15–25. 215 citations (Semantic Scholar/DOI) [2023-08-12].

Tian, J.; Mithun, N.; Seymour, Z.; Chiu, H.-P.; and Kira, Z. 2022. Striking the Right Balance: Recall Loss for Semantic Segmentation. ArXiv:2106.14917 [cs] version: 2.

Tsalera, E.; Papadakis, A.; and Samarakou, M. 2021. Comparison of Pre-Trained CNNs for Audio Classification Using Transfer Learning. *Journal of Sensor and Actuator Networks*, 10(4): 72. 17 citations (Semantic Scholar/DOI) [2023-08-12] Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

Wang, Y.; Cartwright, M.; and Bello, J. P. 2022. Active Few-Shot Learning for Sound Event Detection. In *Interspeech 2022*, 1551–1555. ISCA. 2 citations (Semantic Scholar/DOI) [2023-08-12].

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 *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 31–35. 146 citations (Semantic Scholar/DOI) [2023-08-12] ISSN: 2379-190X.

Çoban, E. B.; Pir, D.; So, R.; and Mandel, M. I. 2020. Transfer Learning from Youtube Soundtracks to Tag Arctic Ecoacoustic Recordings. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 726–730. Barcelona, Spain: IEEE. ISBN 978-1-5090-6631-5. 4 citations (Semantic Scholar/DOI) [2023-08-12].

A Detailed Methods

A.1 Transfer Learning

Exponential softmax function

$$y = \frac{\sum_i y_i \exp(y_i)}{\sum_i \exp(y_i)} \tag{1}$$

The recording-level probability, calculated using the exponential softmax function, is predominantly influenced by the higher frame-level probabilities, similar to max-pooling. However, frames with lower probabilities are also given the opportunity to receive an error signal, increasing the influence of backpropagation during the training process (Wang, Li, and Metze 2019). For this reason, we chose the exponential softmax function for pooling over different time points.

A.2 Active Learning

The following describes the active learning methods used in this paper in detail.

Symbol	Meaning
n	Number of classes
x	Input sample
y^*	Label with highest probability
Θ	Model parameters
$\Phi_{**}(x)$	Uncertainty score derived from method $**$
$P_\Theta(y\|x)$	Probability of y, given x, using Θ

Table 2: Mathematical symbols

Random sampling
Random sampling serves as a baseline for active learning techniques and describes the random sample selection. It can also be used as an easy diversity sampling method.

Uncertainty sampling
Uncertainty sampling strategies calculate an uncertainty score for each unlabelled sample. The samples with the highest uncertainty scores are selected. As we are addressing a multi-label problem with n classes, our approach involves training n binary classifiers. As a result of this strategy, each classifier generates n uncertainty scores per sample. In the following we describe the score computation of each uncertainty sampling method used in this paper.

Least confidence sampling computes the margin between the label with the highest probability and 1, scaled to [0, 1].

$$\Phi_{LC}(x) = (1 - P_\Theta(y^*|x)) \cdot \frac{n}{n-1} \tag{2}$$

For binary classifiers, the equation is simplified to

$$\Phi_{LC\_bi}(x) = (1 - P_\Theta(y^*|x)) \cdot \frac{2}{2-1} \tag{3}$$
$$= 2 - 2 \cdot P_\Theta(y^*|x) \tag{4}$$
$$= 2 - 2 \cdot (0.5 + |y^* - 0.5|) \tag{5}$$
$$= 1 - |2y^* - 1| \tag{6}$$

Ratio sampling computes the ratio between the two most confident predictions, scaled to [0,1].

$$\Phi_{RC}(x) = \frac{P_\Theta(y_2^*|x)}{P_\Theta(y^*|x)} \tag{7}$$

For binary classifiers, the equation is simplified to

$$\Phi_{RC\_bi}(x) = \frac{1 - P_\Theta(y^*|x)}{P_\Theta(y^*|x)} \tag{8}$$
$$= \frac{1}{P_\Theta(y^*|x)} - 1 \tag{9}$$
$$= \frac{1}{0.5 + |y - 0.5|} - 1 \tag{10}$$

Entropy sampling computes the entropy of the sample predictions, scaled to [0, 1].

$$\Phi_{EN}(x) = -\frac{\sum_y P_\Theta(y|x) \cdot \log_2(P_\Theta(y|x))}{\log_2(n)} \tag{11}$$

For binary classifiers, the equation is simplified to

$$\Phi_{EN\_bi}(x) = -P_\Theta(y|x) \cdot \log_2(P_\Theta(y|x)) \tag{12}$$
$$-(1 - P_\Theta(y|x)) \cdot \log_2(1 - P_\Theta(y|x)) \tag{13}$$
$$= -y \cdot \log_2(y) - (1-y) \cdot \log_2(1-y) \tag{14}$$

Uncertainty score combination methods distil a single uncertainty score for each sample from the computed n uncertainty scores of the n different classes. We explored two well-established options: computing the average uncertainty score across all classes and selecting the maximum uncertainty score across all classes (Monarch 2021, chapter 3). $\Phi_{LC\_bi}(x)$, $\Phi_{RC\_bi}(x)$ and $\Phi_{EN\_bi}(x)$ are all strictly monotonically increasing in the range [0;0.5] and strictly monotonically decreasing in the range [0.5;1], leading to

$$\forall X \subseteq \{x_i | x_i \in [0,1]\} : \arg\max(\Phi_{LC\_bi}(X)) \tag{15}$$
$$= \arg\max(\Phi_{RC\_bi}(X)) \tag{16}$$
$$= \arg\max(\Phi_{EN\_bi}(X)). \tag{17}$$

We therefore provide only one uncertainty score using the maximum score selection approach. We chose $\Phi_{RC\_bi}(x)$.

Diversity sampling
Diversity sampling strategies aim to ensure coverage of the entire input space. Unlike random sampling, which may disproportionately select samples from an over-represented class, diversity sampling techniques select samples evenly distributed across the input space, avoiding redundancy and increasing representativeness (Monarch 2021, chapter 4). Unlike uncertainty sampling strategies, the diversity sampling strategies we used directly select samples without the need for class-specific score calculations. As a result, there's no need for a method of combining scores. In the following, we describe clustering, the diversity sampling strategy used in this paper.

Clustering involves using all unlabelled samples and subjecting them to a clustering algorithm. In our implementation, we chose the k-means clustering algorithm using Euclidean distance. Within each cluster, our approach involved selecting specific samples: 1 centroid (the sample with the smallest distance to the cluster centre), 1 outlier (the sample furthest from the nearest cluster centre), and 3 randomly selected samples. The number of clusters is determined inversely; for example, if our goal is to annotate 100 samples at a rate of 5 samples per cluster, we use k-means clustering with 20 clusters, following the approach in (Monarch 2021, chapter 3).

Adaptive active learning
Adaptive active learning addresses the problem of selecting multiple samples during each iteration, a scenario where the selected samples may have significant redundancy between them (Monarch 2021, chapter 5). During a single annotation cycle (as shown in Figure 1), adaptive active learning strategies iteratively select a reduced number of samples. In particular, all previously selected samples are included in the process of selecting subsequent samples. Consequently, this approach reduces the redundancy of newly labelled samples within a single iteration. Our investigation included one adaptive uncertainty sampling strategy and one adaptive diversity sampling strategy.

Adaptive uncertainty sampling uses the trained model to classify the validation set and labelling data as 'correct' for accurately labelled instances and 'incorrect' for inaccurately labelled instances. The last layer of the original model is replaced by a single node, which is retrained using the generated labels. The unlabelled subset is fed into the trained model, and the samples most likely to be labelled as 'incorrect' are selected. These selected samples are then added to the validation set, labelled 'correct'. Now, the next iteration of sample selection starts. The model is retrained using the updated validation set and the subsequent small subset of unlabelled data is selected (Konyushkova, Sznitman, and Fua 2017). In our implementation, we use 5 rounds of sample selection when using adaptive uncertainty sampling.

Adaptive diversity sampling works towards reducing the distribution gap between the training data and the unlabeled data. The validation set, distributed in the same way as the training set, is labelled 'validation'. Meanwhile, the unlabelled subset is labelled 'unlabelled'. The last layer of the original model is replaced by a single node, which is retrained using the generated labels. The unlabelled subset is fed into the trained model, and the samples most likely to be labelled as 'unlabelled' are selected. These selected samples are then added to the validation set, labelled 'validation'. Now, the next iteration of sample selection starts. The model is retrained using the updated validation and unlabelled sets and the subsequent small subset of unlabelled data is selected. In our implementation, we use 5 rounds of sample selection when using adaptive diversity sampling.

Combining active learning strategies
Combining active learning strategies aims to overcome the limitations of pure strategies (Monarch 2021, chapter 5). While uncertainty sampling strategies select samples close to decision boundaries, they often introduce redundancy. Conversely, diversity sampling strategies aim to cover the entire input space, but may miss critical regions, such as samples near decision boundaries. Therefore, we investigated combination methods that combine an uncertainty sampling strategy with a diversity sampling strategy. We investigated two combination methods for sample selection strategies.

Filtering involves pre-selecting samples using one active learning strategy and then selecting samples from within that subset using another active learning strategy (Monarch 2021, chapter 5). We used this method for 'combi: ratio max / clustering'. Using the uncertainty sampling strategy 'ratio sampling', we selected 50 % of the unlabelled data. We then applied the proposed clustering algorithm to this subset and finalised the sample selection.

Hybrid sampling combines the results of two different sampling methods, allowing each method to contribute independently to the selection of part of the total sample. We used this method for 'combi: adaptive uncertainty / clustering', 'combi: ratio max / adaptive diversity' and 'combi: adaptive uncertainty / adaptive diversity'. In all cases, each of the combination methods selected 50 % of the total number of samples.

B Supplementary Figures

Figures 8 to 10 show precision and recall computed on AnuraSet.

Figures 11 to 13 show F1 score, precision and recall computed on the Fernando de Noronha dataset.

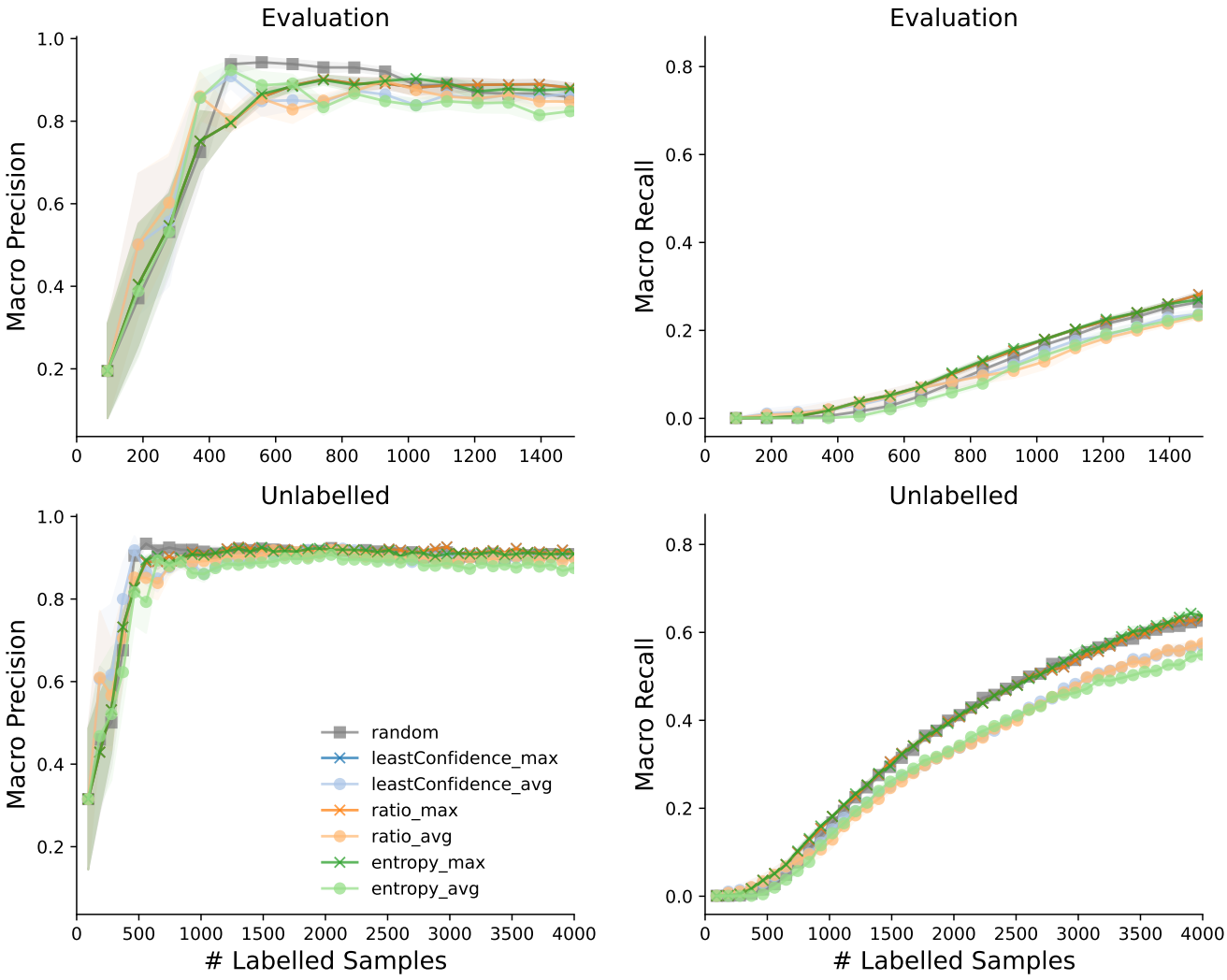

Figure 8: Precision and recall of active learning on AnuraSet with mixed diversity- and uncertainty-based sampling strategies (compare with figure 4). Metrics computed on evaluation data (top), and on the portion of the training data that remains unlabelled (bottom). Mean ± SEM across 5 independent runs.

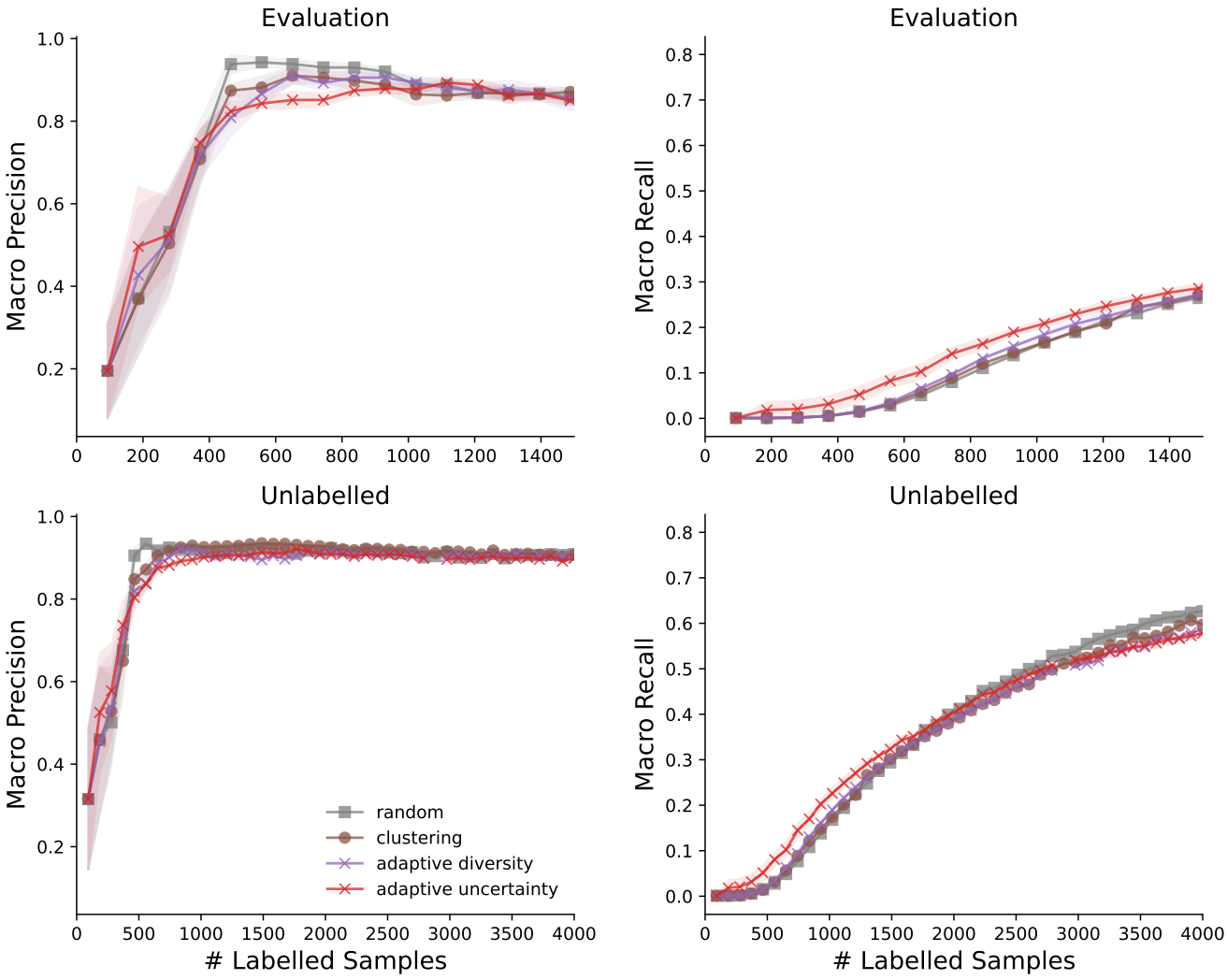

Figure 9: Precision and recall of active learning on AnuraSet with a diversity-based sampling strategy ('clustering') and two adaptive ones (compare with figure 5). Metrics computed on evaluation data (top), and on the portion of the training data that remains unlabelled (bottom). Mean ± SEM across 5 independent runs.

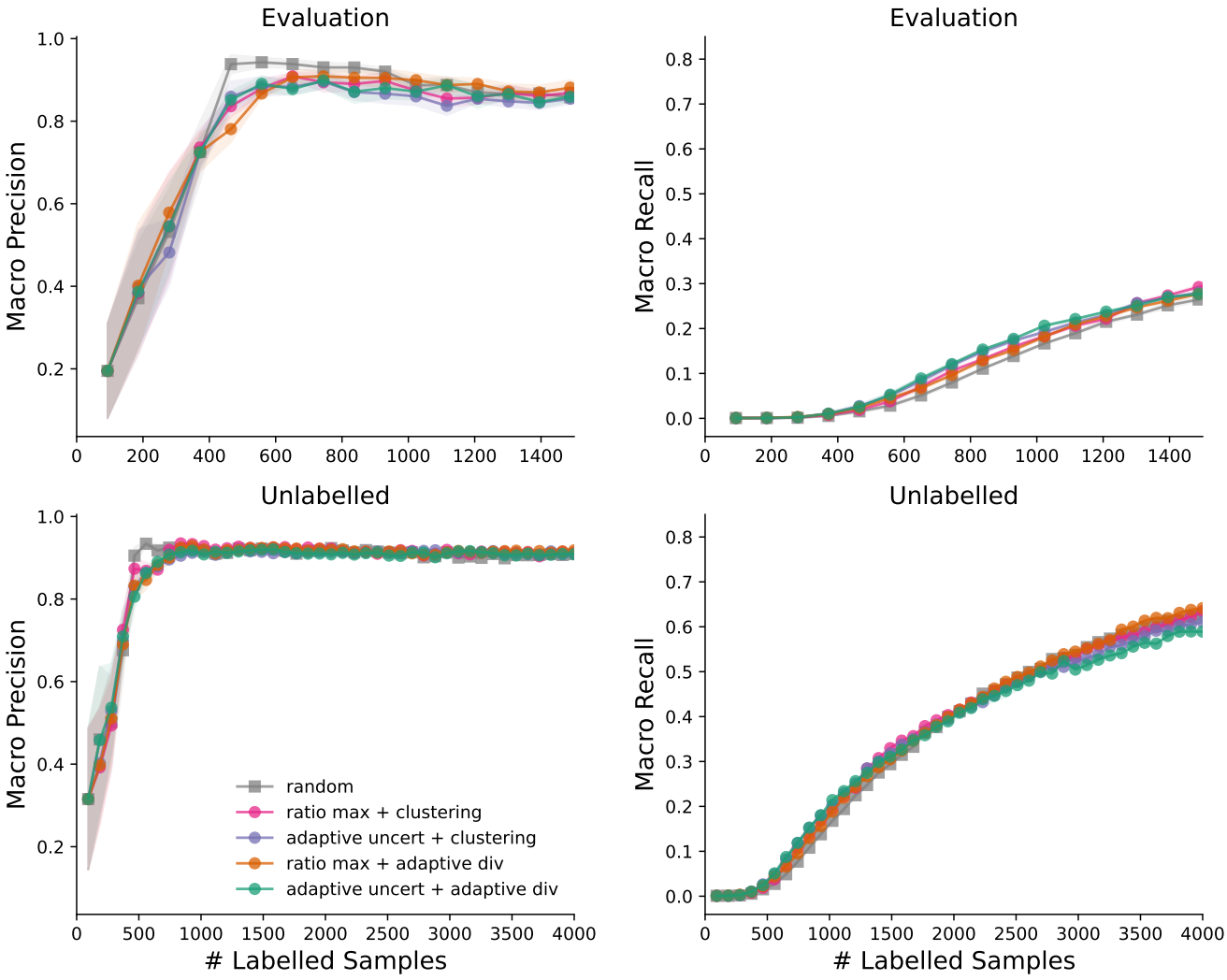

Figure 10: Precision and recall of active learning on AnuraSet with mixed diversity- and uncertainty-based sampling strategies (compare with figure 6). Metrics computed on evaluation data (top), and on the portion of the training data that remains unlabelled (bottom). Mean ± SEM across 5 independent runs.

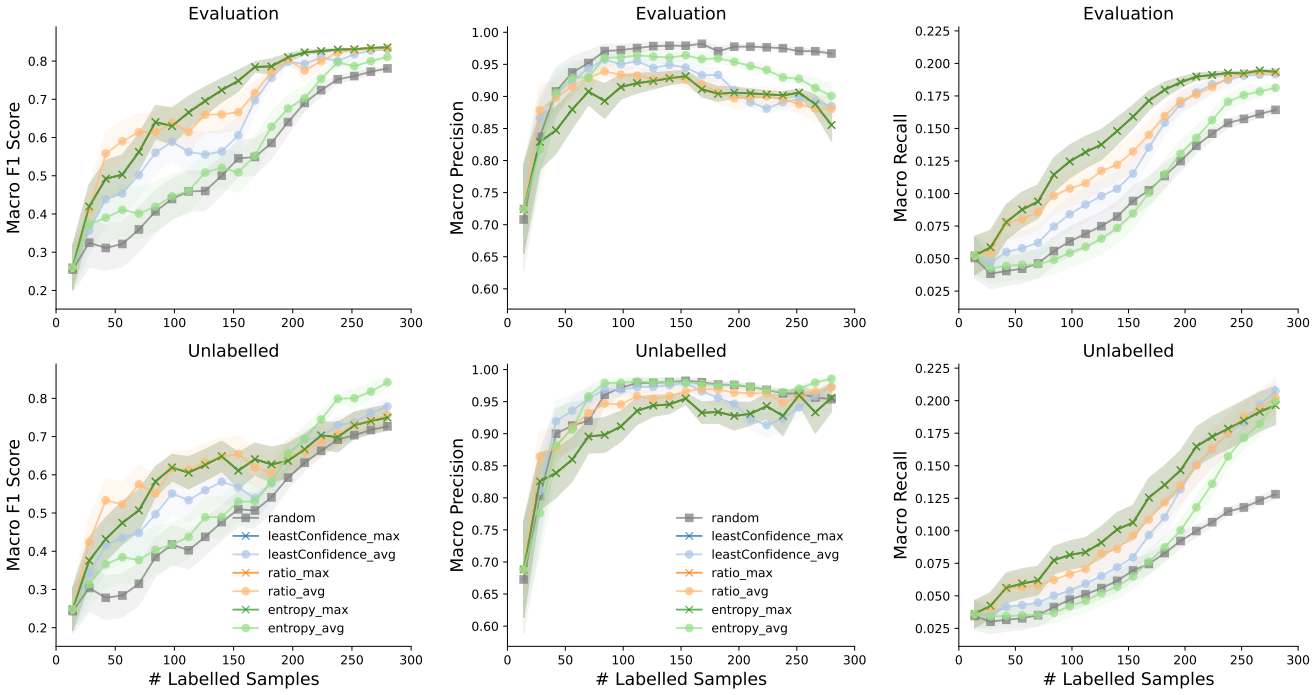

Figure 11: F1 score, precision and recall of active learning on the Fernando de Noronha dataset with uncertainty-based sampling strategies (compare with figures 4 and 8). All maximum-aggregation methods select the same samples and are thus overlaid. Metrics computed on evaluation data (top), and on the portion of the training data that remains unlabelled (bottom). Mean $\pm$ SEM across 30 independent runs.

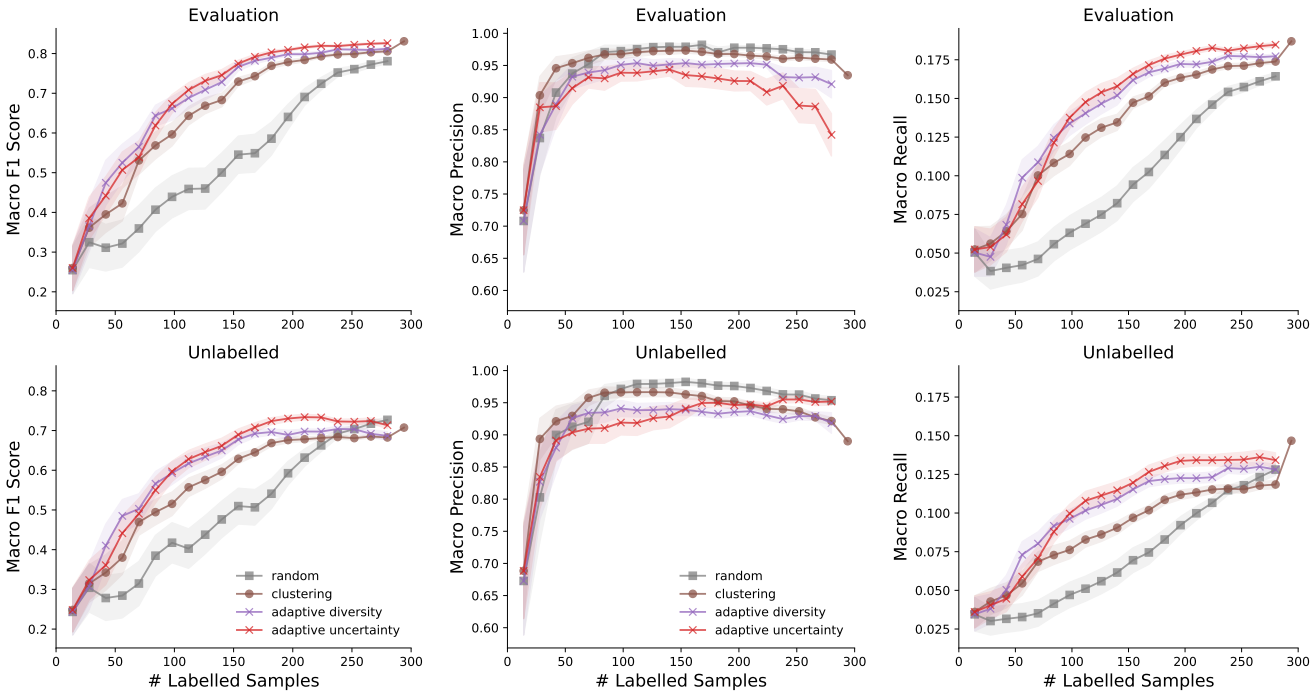

Figure 12: F1 score, precision and recall of active learning on the Fernando de Noronha dataset with a diversity-based sampling strategy ('clustering') and two adaptive ones (compare with figures 5 and 9). Metrics computed on evaluation data (top), and on the portion of the training data that remains unlabelled (bottom). Mean $\pm$ SEM across 30 independent runs.

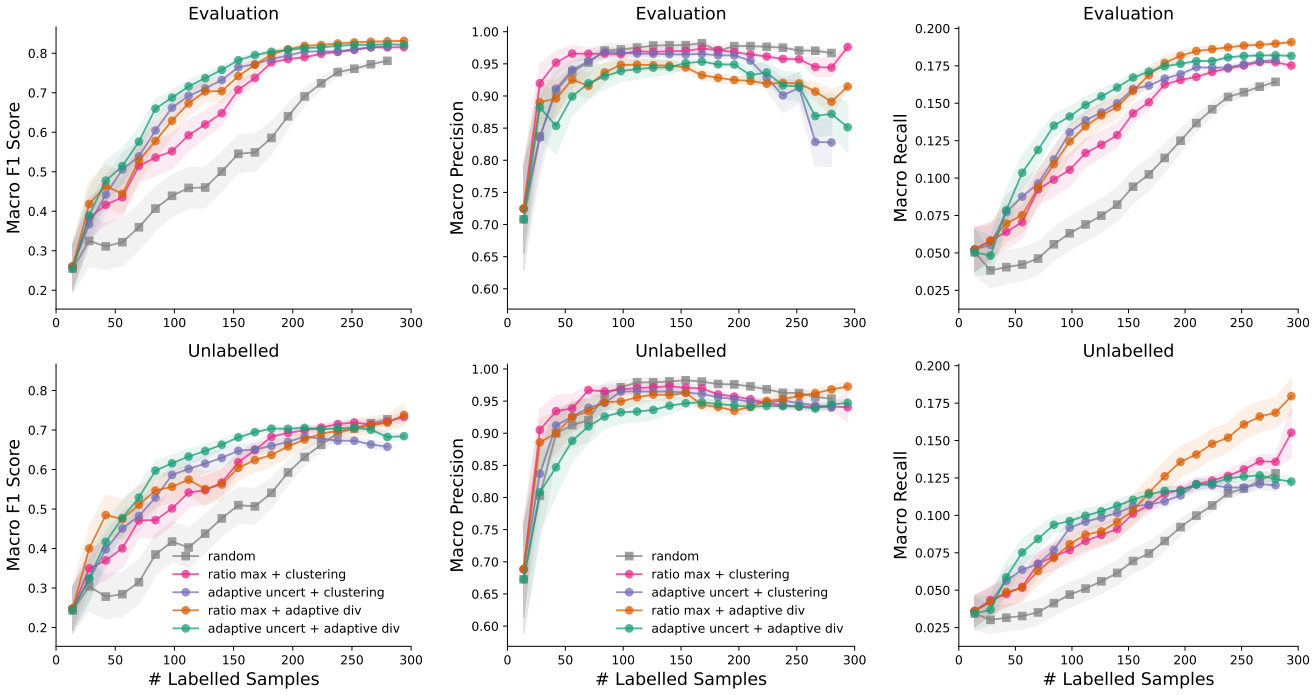

Figure 13: F1 score, precision and recall of active learning on the Fernando de Noronha dataset with mixed diversity- and uncertainty-based sampling strategies (compare with figures 6 and 10). Metrics computed on evaluation data (top), and on the portion of the training data that remains unlabelled (bottom). Panels on the left column repeat figure 7. Mean $\pm$ SEM across 30 independent runs.