Active Learning in Multi-label Classification of Bioacoustic Data

Hannes Kath[1,2], Thiago S. Gouvêa[1], and Daniel Sonntag[1,2]

[1] German Research Center for Artificial Intelligence (DFKI), Oldenburg, Germany
{hannes.kath, thiago.gouvea, daniel.sonntag}@dfki.de
[2] University of Oldenburg, Applied Artificial Intelligence (AAI), Oldenburg, Germany

Abstract. Passive Acoustic Monitoring (PAM) has become a key technology in wildlife monitoring, providing vast amounts of acoustic data. The recording process naturally generates multi-label datasets; however, due to the significant annotation time required, most available datasets use exclusive labels. While active learning (AL) has shown the potential to speed up the annotation process of multi-label PAM data, it lacks standardized performance metrics across experimental setups. We present a novel performance metric for AL, the 'speedup factor', which remains constant across experimental setups. It quantifies the fraction of samples required by an AL strategy compared to random sampling to achieve equivalent model performance. Using two multi-label PAM datasets, we investigate the effects of class sparsity, ceiling performance, number of classes, and different AL strategies on AL performance. Our results show that AL performance is superior on datasets with sparser classes, lower ceiling performance, fewer classes, and when using uncertainty sampling strategies.

Keywords: Passive Acoustic Monitoring · Multi-label data · Active Learning · Transfer Learning · Speedup Factor

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 [21, 22]. 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 [20], and basic research [17]. However, effectively utilizing this vast amount of data for sound event detection poses significant challenges due to the need for annotated data to train machine learning models. The low quality of automatically generated annotations for PAM datasets often requires manual annotation.

Multi-label PAM datasets. Because PAM datasets are typically recorded without controlling the sources of sound events, PAM recordings are naturally multi-label datasets, meaning that multiple labels can be assigned to a single point in time. However, there is a lack of availability of multi-label PAM datasets. The main reasons for this scarcity are that only 21 % of the data examined in the literature is published [1], and that manually labelling a PAM dataset with multi-label annotations can be extremely time-consuming and costly, typically requiring annotation times of 10 to 15 times the duration of the audio [12]. For this reason, the vast majority of PAM datasets are annotated with exclusive labels, neglecting the multi-label aspect. Few research efforts are dedicated to the development of multi-label annotation tools for PAM datasets [8, 12].

Active learning (AL) is a promising research direction for accelerating the annotation and analysis of large multi-label PAM datasets [9]. AL provides an iterative strategic approach to prioritize the most informative samples for annotation. Unlike random sampling of data points, AL algorithms intelligently select samples that accelerate model convergence (see fig. 1). AL strategies can be categorized into prediction-based, data-based and model-based strategies [24]. Prediction-based strategies use model outputs to select the next batch of samples, e.g., uncertainty sampling prioritizes samples close to the decision boundary. Data-based strategies select samples based on the internal structure of the data, e.g., diversity sampling selects batches that cover the entire input space. Model-based strategies select samples based on the change in the model, e.g., prioritizing samples with high model influence regardless of the label. While model-based strategies directly aim to improve model performance, they are rarely used in practice due to the high computational cost of training the model for all samples in all label combinations [14]. Multi-label AL is not a new task, yet this research field has gained popularity only recently [26]. Various approaches have been proposed to the problem of selecting samples while optimizing for multiple classes [13, 16, 26], with the practice of pooling results from established strategies applied to binary classifiers being used as a baseline [14]. The evaluation of AL performance is primarily done by visually comparing the learning curves of the applied sampling strategies [15, 19]. While this approach is feasible for a few and clearly separated curves, it complicates the comparison of AL strategies across datasets, experiments, and projects [15]. Evaluation metrics provide statistical justification by quantifying features of the learning curve [15]. The most popular evaluation metrics used in the literature for (multi-label) AL strategies are accuracy [2, 23] or multi-class accuracy [10], sensitivity and specificity [11, 23], and area under the curve [10, 25] or multi-class area under the receiver operating characteristic curve [5, 11]. These methods offer numerically comparable results under equal experimental setups, but they are limited by their variation with different numbers of training samples, leading to incomparability when different experimental setups are used.

Research Contributions. We propose a quantitative performance metric for AL termed the speedup factor. The metric represents the fraction of samples required for an AL strategy to achieve the same model performance as when using random sampling. The speedup factor remains constant regardless of the number of training samples, providing a comparable score even across experimental setups. Furthermore, driven by the quest to accelerate the annotation of multi-label PAM datasets, we investigate the performance of AL on a synthetically generated variant of the Watkins Marine Mammal Sound Database (WMMD) [18], as well as on AnuraSet [3], a multi-label PAM dataset of reasonable size. Multi-label datasets often exhibit significant class imbalance, where the amount of samples containing a particular class, termed **positive samples**, varies significantly across classes. We examine the correlation between AL performance and the number of positive samples. Since neural models typically perform best on balanced datasets with 50 % positive samples [6], and random sampling on average selects the same fraction of positive samples as in the unlabelled set, we hypothesize that AL performance decreases as the fraction of positive samples in the dataset increases. Conversely, we expect to see an improvement in performance for sparse datasets where random sampling has difficulty selecting positive samples. The maximum performance, termed **ceiling performance**, of a neural model for a given dataset depends on several factors, including model architecture, class similarity, class sparsity, and feature characteristics. We investigate the correlation between AL performance and ceiling performance. Since classes with high ceiling performance have more room for improvement through AL strategies, while classes with low ceiling performance may already be facing challenges and could benefit from AL, we do not hypothesize a clear trend in this regard. The number of classes in multi-label datasets is not fixed, but open-ended. We investigate the correlation between the performance of AL and the number of classes. Since AL strategies are forced to optimize for multiple goals when the dataset contains more classes, we hypothesize that AL performance decreases as the number of classes increases. Due to the high computational cost associated with model-based AL strategies and the large number of experiments, we limit our analysis to one uncertainty-based method and one diversity-based method. Diversity strategies rely on the internal structure of the data and do not consider information from previously sampled data. As a result, they face challenges such as repeatedly sampling from the same clusters and sampling from clusters irrelevant to the classification task [14]. Uncertainty strategies aim to refine the decision boundary by selecting samples close to it, but they can also face challenges such as getting trapped by sampling from small clusters where the decision boundary is unclear, rather than establishing a reliable decision boundary for the entire space [14]. Because diversity sampling strategies select data regardless of model performance and the classes being optimized for, we hypothesize that uncertainty sampling outperforms diversity sampling. The present study tests the following four hypotheses on two multi-label PAM datasets:

H1 AL performance decreases as the number of positive samples in the dataset increases.
H2 AL performance changes as the ceiling performance changes.
H3 AL performance decreases as the number of classes increases.
H4 AL performance decreases when using diversity sampling strategies instead of uncertainty sampling strategies.

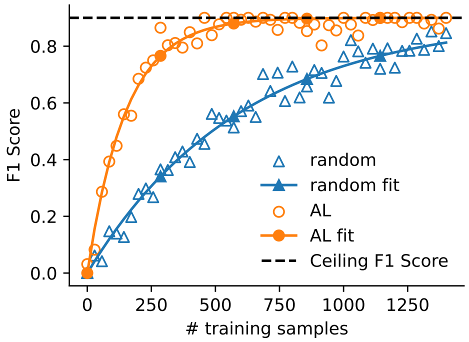

Fig. 1: Schematic representation of the learning curve for a neural model employing both random sampling and active learning (AL). The curve is fitted using the function $P = a\left(1 - e^{-\frac{x}{b}}\right)$, where the ceiling F1 score is $a = 0.9$ and the parameterized learning rates are $b_{random} = 600$ and $b_{AL} = 150$.

2 Methodology

2.1 Active learning validation metric

The learning curves of neural models typically exhibit three prominent features: performance starting at 0 for 0 training samples, a general increase with the addition of training data, and an asymptotic shape with the ceiling performance as the limit. Therefore, the learning curves of neural models can be effectively modeled using the equation

$$P = a\left(1 - e^{-\frac{x}{b}}\right), \tag{1}$$

where P is the model performance, a is the ceiling performance, b is the parameter that represents the learning rate, and x is the number of training samples.

Active learning algorithms are designed to identify and select relevant samples from the unlabelled dataset. As a result, they reduce the number of samples required to achieve a predetermined level of model performance. Figure 1 shows a schematic comparison of learning curves using random sampling and AL. Given the assumption that increasing the number of training samples improves performance, a corresponds to the performance of the model when trained on the entire dataset. The parameter b is calculated from eq. (1) using the least squares method, using the samples obtained from the experimental data. Therefore, a is constant for all sampling strategies, while b varies depending on the sample selection method. We define the speedup factor as an evaluation metric for AL, representing the fraction of samples required to achieve the same level of performance using AL as compared to random sampling. The speedup factor (S) serves as validation metric for assessing the efficiency of active learning. Inverting eq. (1) and assuming that $a_{AL} = a_{random}$, the speedup factor is calculated with eq. (2).

The speedup factor remains constant regardless of the number of training samples; for example, a factor of 0.4 means that the AL strategy requires, on average, 40 % of the amount of data that random sampling requires to achieve the same performance. A lower speedup factor indicates that AL requires fewer data samples to achieve comparable performance, while a speedup factor greater than 1 indicates that random sampling is more efficient than AL.

$$S = \frac{x_{AL}}{x_{random}} = \frac{-\ln\left(1 - \frac{P}{a_{AL}}\right) \cdot b_{AL}}{-\ln\left(1 - \frac{P}{a_{random}}\right) \cdot b_{random}} = \frac{b_{AL}}{b_{random}} \tag{2}$$

2.2 Active learning strategies

To test the AL performance, we opted for simple but interpretable and well-established AL strategies to test hypotheses H1-H4 [14].

We use 'ratio' as the AL uncertainty sampling strategy [14]. The uncertainty score Φ for a binary classifier is calculated using the equation

$$\Phi_{bi}(y) = \frac{0.5 - |y - 0.5|}{0.5 + |y - 0.5|}, \tag{3}$$

where $y \in [0; 1]$ represents the output of the model prediction. Multi-label classifiers with n species use n binary classifiers, hence n uncertainty scores per sample are computed. To obtain a meaningful single uncertainty score for a sample, we use the maximum of the computed uncertainty scores [9].

Following [14], we implement k-means clustering using the Euclidean distance measure as diversity sampling strategy. 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 [14].

2.3 Experimental Setup

Datasets. In this study we take advantage of AnuraSet, 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 [3]. To the best of our knowledge, this is currently the only freely available multi-label PAM dataset of reasonable size. 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. The authors provide a training/evaluation split for all files. The AnuraSet dataset has a high degree of class imbalance among its classes. We used the evaluation set as defined by the original authors. Additionally, we synthesized another multi-label dataset using underwater sound[3] and inserting events from the WMMD [18] at random

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

positions, including the possibility of overlapping events. We segmented the data into three-second files. From the total duration of 7.8 h of audio with 40 highly unbalanced classes, the evaluation set (1.6 h) includes 20 % of the events from each class.

Pre-processing. Recent studies show that the use of BirdNet [7], a neural model trained on vocal bird recordings, is an effective approach for embedding PAM datasets [4,9]. As a pre-processing step, all three-second files are resampled to 48 kHz and embedded using the penultimate layer of BirdNet 2.4. This layer has an embedding size of 1024, which has shown superior performance [9].

Training. Since the goal of this study is not to construct a state-of-the-art classifier, but to investigate the dependencies of different factors on AL performance, a linear (multi-label) classifier is trained to evaluate the performance of active learning. The resulting architecture consists of a single fully connected layer with an equal number of output nodes as there are species in the dataset. Because the classifier comprises only a single layer without shared weights, it facilitates the assignment of results to different causes. Each output node indicates the presence or absence of a particular 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 BirdNet embeddings (no fine-tuning) for a maximum of 1 000 epochs. Early stopping criteria are based on validation loss, with a minimum delta of 0.1 and a patience of 10 epochs, with reinstatement of the best weights.

Active learning experiments are conducted starting with a random selection of 20 samples. Then, using the sampling strategies random, ratio max, and clustering, 20 samples are added per iteration until a total of 1400 samples (70 iterations) is reached.

Evaluation. All experiments are conducted with 5 random seeds to ensure robustness and reliability of the results. All results are evaluated on the respective evaluation set, ensuring that this set is never used for training purposes. After each sample selection iteration, all selected samples are used to train a linear classifier, which is then evaluated on the evaluation set. Given the highly imbalanced classes, we use the macro F1 score as score aggregation metric for experiments with multi-label datasets to assign equal importance to all species. To compute the speedup factor, we calculate the mean across random seeds of the true positive, false positive and false negative values to compute the mean macro F1 score for the currently used AL sampling strategy and for random sampling. Using the macro F1 score as learning curve, the b-values and the respective speedup factor are computed. If the speedup factor cannot be computed, e.g. if the class is extremely sparse and no positive samples are selected for the training set over all iterations, the results are discarded. To compute the ceiling performance, we use all available data from the training split to train the model and average the result over the independent runs.

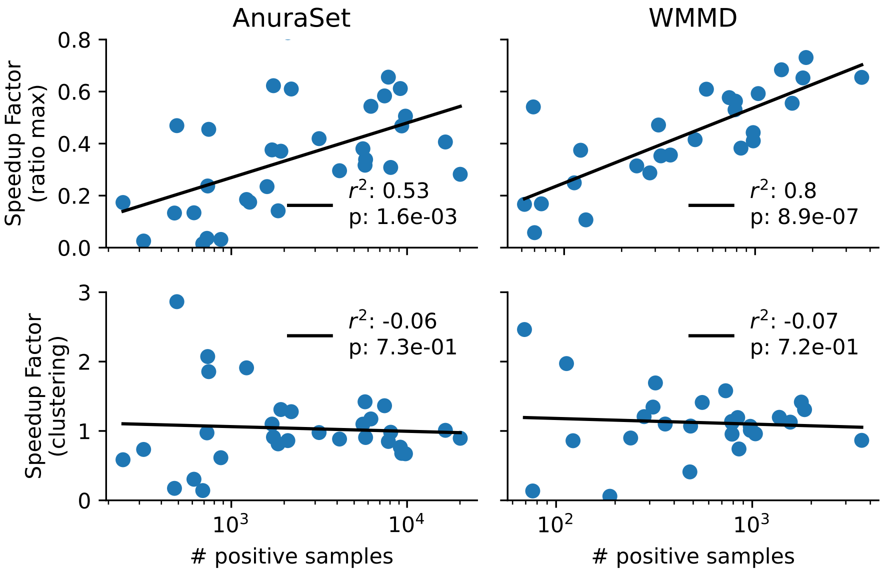

Fig. 2: Speedup factor of the single-label datasets over the number of positive samples of the respective class. AnuraSet (left) and WMMD (right) with the active learning strategies ratio max (top) and clustering (bottom).

3 Experiments

The following experiments investigate the influence of the number of positive samples, the ceiling performance, the number of classes, and the active learning strategy on the performance of active learning, measured by the speedup factor.

3.1 Single-label data

To isolate the variables of interest, we begin by evaluating hypotheses H1, H2 and H4 on single-label data. Single-label datasets are generated by iteratively selecting one class from the dataset.

To test hypothesis H1, fig. 2 shows the speedup factor for the single-class datasets of AnuraSet and WMMD over the total number of positive samples for the respective class used. Using ratio max, fig. 2 (top) shows a strong positive correlation between the speedup factor and the number of positive samples. The speedup factor remains consistently below 1. Conversely, using clustering, fig. 2 (bottom) shows no significant correlation. The speedup factor is around 1.

Figure 6 provides a detailed overview of the results for AnuraSet, showing the learning curves for both a rare and a frequent species over the AL sampling strategies ratio max and clustering. As can be seen in the figures, clustering does not improve model performance compared to random sampling, while ratio max improves model performance, showing higher improvement for rare classes.

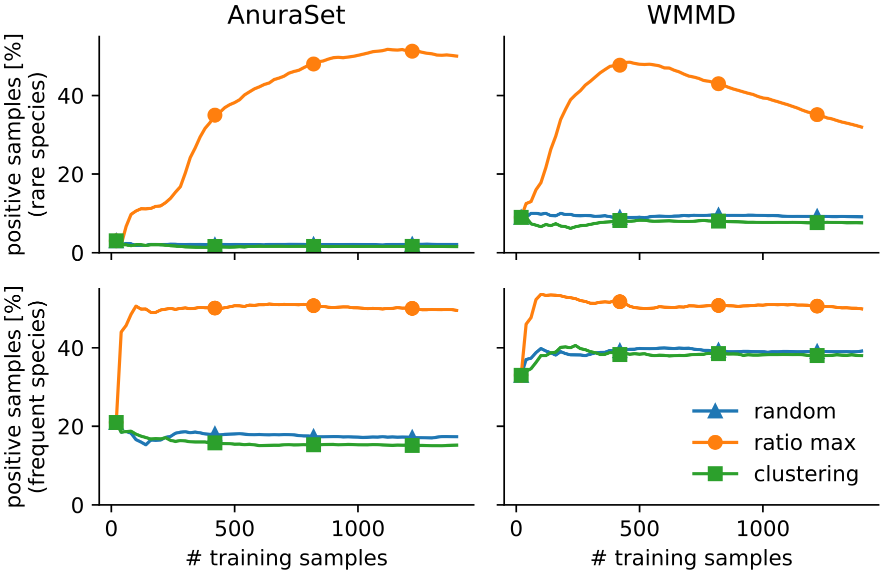

Fig. 3: Fraction of positive samples in the training set selected by the sampling strategies random, ratio max and clustering over the total amount of training samples for two single-label datasets from AnuraSet: one with a low number of positive samples in the dataset (left, class ELABIC, 1705 positive samples (1.8 %)) and the other with a high number of positive samples in the dataset (right, class BOABIS, 16524 positive samples (17.4 %)).

Figure 3 (left) shows for AnuraSet for all 3 sampling strategies the fraction of positive samples in the training set over the total amount of training data for a rare class and a frequent class. The proportion of positive samples for the rare class in the entire AnuraSet is 1.8 % and for the frequent class 17.4 %. After 70 iterations, the number of positive samples in the training set for the ratio max sampling strategy is 701 (50.0 %) for the rare dataset and 693 (49.5 %) for the frequent dataset. In contrast, for random sampling and clustering, the proportion remains approximately at the default level. Figure 3 (right) shows similar trends for the WMMD dataset, with the difference that for the rare species the fraction for ratio max decreases after a certain number of iterations.

These results indicate that ratio max is effective as an AL strategy for PAM datasets, while clustering does not provide significant benefits. Therefore in the following we present and discuss the results of the AL strategy ratio max. Corresponding figures for clustering are provided in the appendix.

To test hypothesis H2, fig. 4 (left) shows the ceiling performance for the single-class datasets over the total number of positive samples for the respective class used. A strong positive correlation between the speedup factor and the number of positive samples is recognisable. Observing a strong correlation between the ceiling performance and the number of positive samples (fig. 4 (left))

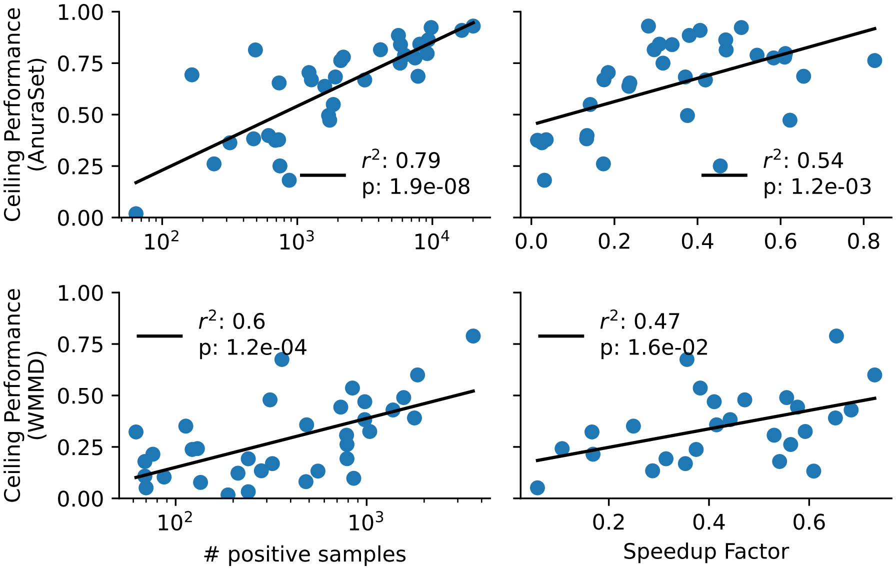

Fig. 4: Ceiling performance of the single-label datasets of AnuraSet (top) and WMMD (bottom) over the number of positive samples of the respective class (left) and the speedup factor (right) for the active learning strategy ratio max.

and a strong correlation between the speedup factor and the number of positive samples (fig. 2 (top)), we also observe a strong positive correlation of the speedup factor and the ceiling performance (fig. 4 (right)).

3.2 Multi-label data

To test hypothesis H3, we create multi-label datasets.

To isolate the variable of interest (number of classes), we first create datasets that include classes with similar amounts of positive samples, ceiling performances and speedup factors. Figure 5 (top) shows a strong positive correlation between the speedup factor and the number of classes. Table 1a shows the fraction of positive samples and table 1b the F1 score for each species of AnuraSet after 25 iterations, over the number of classes for which the AL strategy ratio max was optimized. Table 1a shows a noticeable tendency for the proportion of positive samples to decrease as the number of classes increases, with the most significant decrease observed when optimizing for 2 classes instead of 1. Figure 7 (left) illustrates that this trend holds not only for the specific value of 25 iterations, but also more generally over all iterations. Table 1b shows that the F1 score for each class decreases as the number of classes increases. Figure 7 (right) shows similar results to the WMMD dataset, but with the difference that the proportion of selected positive samples begins to decrease after several iterations.

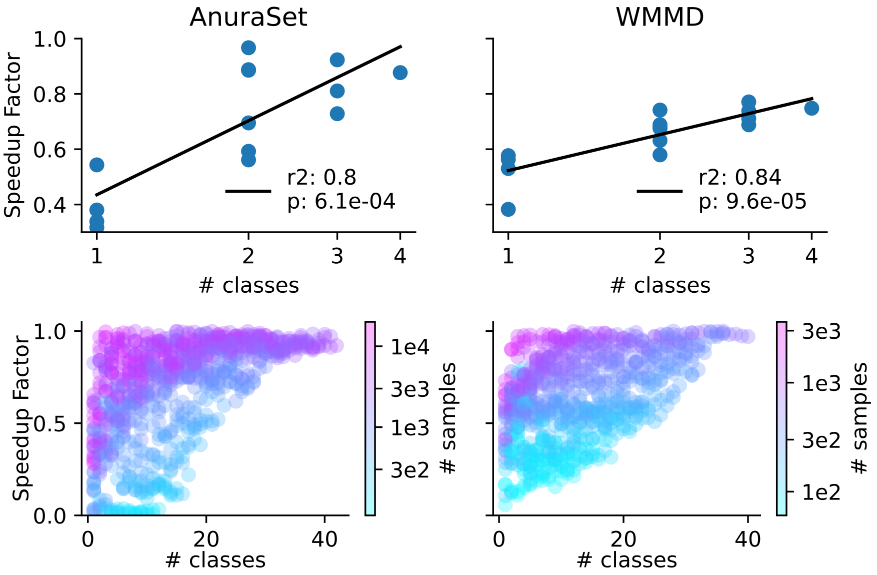

Fig. 5: Active learning (ratio max) performance over the used number of classes for multi-label datasets generated with AnuraSet (left) and WMMD (right). Top row: 4 classes with similar properties (left: classes: BOAALB, SCIPER, DENNAN, PHYCUV, # positive samples: 5917 ± 309, ceiling performance: $\mu = 0.82$, $\sigma^2 = 0.003$, speedup factor: $\mu = 0.46$, $\sigma^2 = 0.012$; right: classes: FalseKillerWhale, SpinnerDolphin, NorthernRightWhale, BowheadWhale, # positive samples: 787 ± 56, ceiling performance: $\mu = 0.43$, $\sigma^2 = 0.105$, speedup factor: $\mu = 0.51$, $\sigma^2 = 0.006$). All combinations are used. Bottom row: All classes. For each number of classes, all combinations with neighboring numbers of positive samples are used. The mean number of positive samples is color-coded.

Finally, we used all classes for our analysis. To minimize the variance within the selected datasets regarding the number of positive samples, we first sorted all classes based on their respective number of positive samples. Subsequently, we generated datasets containing 1 to 40 (42) classes. For n classes, we utilized all combinations of classes with neighboring numbers of positive samples, resulting in $40\ (42) - n + 1$ combinations. Figure 5 (bottom) shows the speedup factor over the number of classes with all used combinations for the sampling strategy ratio max. The average number of positive samples across all classes used for each dataset is color-coded. There is a strong positive correlation between the number of classes and the speedup factor, where the speedup factor tends to be lower with fewer classes and approaches 1 with more classes. There is also a clear positive correlation between the number of positive samples in the dataset and the speedup factor, with the speedup factor being lower for datasets with fewer positive samples and higher for datasets with more positive samples. While for some combinations the speedup factor exceeds 1, indicating that random

Table 1: Fraction of positive samples and F1 score for each species for created datasets with 1 to 4 classes from AnuraSet after 25 iterations (500 training samples) for the sampling strategy ratio max. For datasets with 2 and 3 classes, each class combination is used, resulting in three combinations per class. The results are then averaged to obtain the final results.

(a) positive samples [%]

	# **classes**			
Species	1	2	3	4
PHYCUV	52.7	30.0	25.1	20.0
BOAALB	47.7	37.6	18.1	19.9
DENNAN	47.8	28.9	27.6	9.9
SCIPER	56.8	21.0	19.7	26.7

(b) F1 Score

	# **classes**			
Species	1	2	3	4
PHYCUV	0.59	0.48	0.41	0.39
BOAALB	0.87	0.75	0.52	0.61
DENNAN	0.75	0.57	0.54	0.25
SCIPER	0.73	0.39	0.36	0.48

sampling outperforms AL, for the majority of combinations it remains below 1, indicating that AL is more efficient in those cases.

4 Discussion

In this study, we introduced a quantitative performance measure for AL, the speedup factor, which is independent of the number of training samples and states the fraction of samples needed to reach the same level of performance using AL compared to random sampling. Motivated by the desire to develop a methodology capable of annotating and interpreting large volumes of multi-label PAM data, we investigate the influence of several factors on AL performance. These factors include the number of positive samples, the ceiling performance, the number of classes, and the AL sampling strategy.

H1. We hypothesised, that AL performance decreases as the number of positive samples in the dataset increases. Our results shown in figs. 2 and 6 support this hypothesis for single-class datasets, showing a strong positive correlation between the number of positive samples and the speedup factor. Figure 3 shows that random sampling maintains an equivalent proportion of positive samples in the training set as in the original unlabelled dataset. By selecting samples close to the decision boundary, the AL method ratio max is trained to select approximately 50 % positive samples. Since a model tends to perform better with a more balanced dataset, the effect of AL is more pronounced for sparse classes where the fraction of positive samples selected by random sampling is very small. Figure 5 (bottom) shows the same trend for multi-label datasets, where the color indicates that the speedup factor decreases for sparser multi-label datasets. Using the sampling strategy ratio max on the smaller dataset WMMD with sparse classes results in a noticeable decrease in the number of unlabelled positive samples over iterations, resulting in a decreasing proportion of positive samples within the training set, as shown in fig. 3 (top right) and fig. 7 (right).

H2. We hypothesised, that AL performance changes as the ceiling performance changes. While fig. 4 (right) shows a decrease in AL performance with increasing ceiling performance, fig. 4 (left) also shows a strong correlation between the number of positive samples and ceiling performance. Therefore, we conclude that the ceiling performance does not directly influence the AL performance, but rather the number of positive samples has a strong influence on both the AL performance and the ceiling performance.

H3. We hypothesised, that AL performance decreases as the number of classes increases. Holding the variables of positive samples, ceiling performance, and speedup factors constant for the single-label datasets, fig. 5 (top) illustrates that the performance of AL decreases as the number of classes increases. As shown in fig. 7, this phenomenon occurs because the selected fraction of positive samples decreases when optimizing for more classes, resulting also in lower model performance (see table 1b). Figure 5 (bottom) shows consistent results even when considering different numbers of positive samples, ceiling performances, and speedup factors, asymptotically approaching a speedup factor of 1 as the number of classes increases.

H4. We hypothesised, that AL performance decreases when using diversity sampling strategies instead of uncertainty sampling strategies. Figures 2 and 6 show that for single-label datasets, ratio max significantly outperforms random sampling, while clustering exhibit an average speedup factor of 1, indicating a learning curve close to random sampling. Figure 3 reveals that the number of selected positive samples using clustering is even slightly lower than that for random sampling. In addition, we conducted all experiments using diversity sampling. The results indicate that the correlation between the speedup factor and both maximum performance and the number of classes suggests no performance advantage when using clustering compared to random sampling.

Regarding the limitations of our elaboration, while the speedup factor effectively captures the AL performance, it could be better adapted to the learning curves by allowing a shift along the x-axis (see fig. 6). For this study, we limited our analysis to two well-established, computationally tractable AL strategies in the domain of PAM datasets. Future research include investigating these claims for multi-label datasets from other domains, such as image, text, and tabular data, as well as exploring more sophisticated AL strategies.

Acknowledgements

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

References

1. Baker, E., Vincent, S.: A deafening silence: a lack of data and reproducibility in published bioacoustics research? Biodiversity Data Journal **7**, e36783 (2019). https://doi.org/10.3897/BDJ.7.e36783
2. Boney, R., Ilin, A.: Semi-supervised and active few-shot learning with prototypical networks. arXiv preprint arXiv:1711.10856 (2017)
3. Cañas, J., Toro-Gómez, M., Sugai, L., et al.: A dataset for benchmarking neotropical anuran calls identification in passive acoustic monitoring. Scientific Data **10**(1), 771 (2023). https://doi.org/10.1038/s41597-023-02666-2
4. Ghani, B., Denton, T., Kahl, S., Klinck, H.: Global birdsong embeddings enable superior transfer learning for bioacoustic classification. Scientific Reports **13**(1), 22876 (2023)
5. Hand, D.J., Till, R.J.: A simple generalisation of the area under the roc curve for multiple class classification problems. Machine learning **45**, 171–186 (2001)
6. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering **21**(9), 1263–1284 (2009). https://doi.org/10.1109/TKDE.2008.239
7. Kahl, S., Wood, C.M., Eibl, M., Klinck, H.: Birdnet: A deep learning solution for avian diversity monitoring. Ecol. Informatics **61**, 101236 (2021). https://doi.org/10.1016/j.ecoinf.2021.101236
8. Kath, H., Gouvêa, T., Sonntag, D.: A Human-in-the-Loop Tool for Annotating Passive Acoustic Monitoring Datasets. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. pp. 7140–7144. Macau, SAR China (2023). https://doi.org/10.24963/ijcai.2023/835
9. Kath, H., Serafini, P.P., Campos, I.B., Gouvêa, T.S., Sonntag, D.: Leveraging Transfer Learning and Active Learning for Sound Event Detection in Passive Acoustic Monitoring of Wildlife. In: 3rd Annual AAAI Workshop on AI to Accelerate Science and Engineering (AI2ASE) (2024)
10. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(3), 453–465 (2014). https://doi.org/10.1109/TPAMI.2013.140
11. Liu, W., Zhang, H., Ding, Z., Liu, Q., Zhu, C.: A comprehensive active learning method for multiclass imbalanced data streams with concept drift. Knowledge-Based Systems **215**, 106778 (2021). https://doi.org/https://doi.org/10.1016/j.knosys.2021.106778
12. Lüers, B., Serafini, P.P., Campos, I.B., Gouvêa, T.S., Sonntag, D.: 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 (2024)

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

13. Möllenbrok, L., Sumbul, G., Demir, B.: Deep active learning for multi-label classification of remote sensing images. IEEE Geosci. Remote. Sens. Lett. **20**, 1–5 (2023). https://doi.org/10.1109/LGRS.2023.3305647
14. Monarch, R.: Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-centered AI. Simon and Schuster (Jul 2021)
15. Pupo, O.G.R., Altalhi, A.H., Ventura, S.: Statistical comparisons of active learning strategies over multiple datasets. Knowl. Based Syst. **145**, 274–288 (2018). https://doi.org/10.1016/J.KNOSYS.2018.01.033
16. Reichart, R., Tomanek, K., Hahn, U., Rappoport, A.: Multi-task active learning for linguistic annotations. In: McKeown, K.R., Moore, J.D., Teufel, S., Allan, J., Furui, S. (eds.) ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA. pp. 861–869. The Association for Computer Linguistics (2008)
17. Ross, S., O'Connell, D., Deichmann, J., et al.: Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions. Functional Ecology **37**(4), 959–975 (2023). https://doi.org/10.1111/1365-2435.14275
18. Sayigh, L., Daher, M., Allen, J., Gordon, H., Joyce, K., Stuhlmann, C., Tyack, P.: The watkins marine mammal sound database: An online, freely accessible resource. In: Proceedings of Meetings on Acoustics. vol. 27, p. 040013 (01 2016). https://doi.org/10.1121/2.0000358
19. Settles, B.: Active Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers (2012). https://doi.org/10.2200/S00429ED1V01Y201207AIM018
20. Sueur, J., Pavoine, S., Hamerlynck, O., Duvail, S.: Rapid Acoustic Survey for Biodiversity Appraisal. PLOS ONE **3**(12), e4065 (2008). https://doi.org/10.1371/journal.pone.0004065
21. Sugai, L., Llusia, D.: Bioacoustic time capsules: Using acoustic monitoring to document biodiversity. Ecological Indicators **99**, 149–152 (2019). https://doi.org/10.1016/j.ecolind.2018.12.021
22. Sugai, L., Silva, T., Ribeiro, J., Llusia, D.: Terrestrial Passive Acoustic Monitoring: Review and Perspectives. BioScience **69**(1), 15–25 (2019). https://doi.org/10.1093/biosci/biy147
23. Tharwat, A., Schenck, W.: Balancing exploration and exploitation: A novel active learner for imbalanced data. Knowledge-Based Systems **210**, 106500 (2020). https://doi.org/https://doi.org/10.1016/j.knosys.2020.106500
24. Tharwat, A., Schenck, W.: A survey on active learning: State-of-the-art, practical challenges and research directions. Mathematics **11**(4) (2023). https://doi.org/10.3390/math11040820
25. Vasilakes, J., Rizvi, R., Melton, G.B., Pakhomov, S., Zhang, R.: Evaluating active learning methods for annotating semantic predications. JAMIA Open **1**(2), 275–282 (06 2018)
26. Wu, J., Sheng, V.S., Zhang, J., Li, H., Dadakova, T., Swisher, C.L., Cui, Z., Zhao, P.: Multi-label active learning algorithms for image classification: Overview and future promise. ACM Comput. Surv. **53**(2) (mar 2020). https://doi.org/10.1145/3379504

A Appendix - Supplementary figures

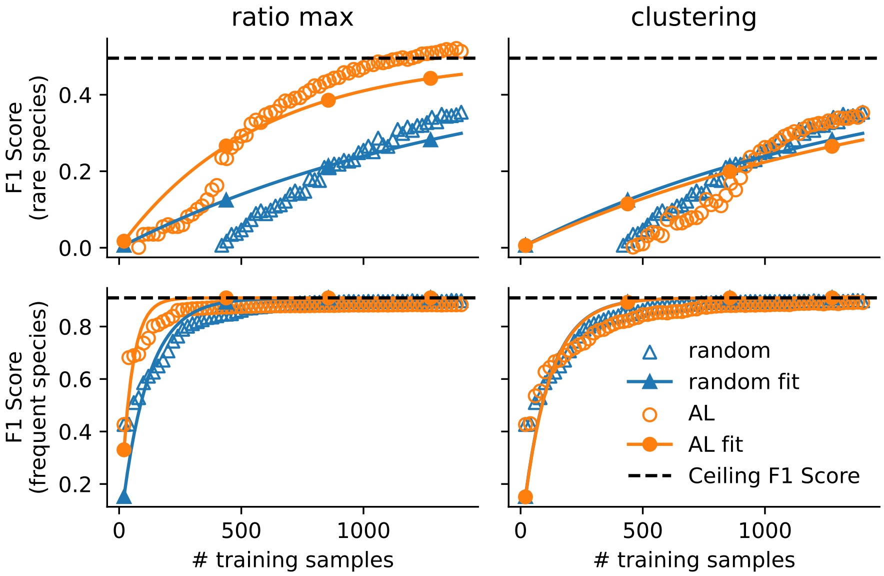

Fig. 6: AnuraSet: Learning curves for two single-label datasets. Positive samples: 1.8 %, (class ELABIC, top row); 17.4 %, (class BOABIS, bottom row). Speedup factors for active learning (AL) strategies: ratio max (left col): 0.3 (ELABIC), 0.5 (BOABIS); clustering (right col): 1.0 (both classes).

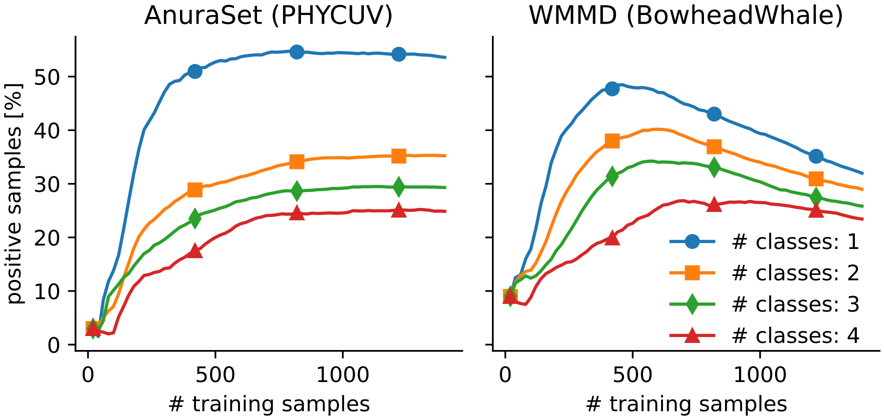

Fig. 7: Fraction of positive samples in the training set selected by the sampling strategy ratio max over the total amount of training samples for multi-label datasets for one class. For 2 and 3 classes, where 3 combinations are possible, the curves are averaged over these combinations.