

ADAPTIVE DIVERSITY-UNCERTAINTY ACTIVE LEARNING WITH REDUNDANCY CONTROL FOR BIOACOUSTIC EVENT CLASSIFICATION

Technical Report

Gabriel Dubus^{1}, Hugo Magaldi¹, Anatole Gros-Martial²,*

¹ Eco-Anthropologie, Muséum National d'Histoire Naturelle, UMR7206, CNRS, Paris, France,

² Centre d'Etudes Biologiques de Chizé, UMR 7372, CNRS, La Rochelle Université, France,

* Corresponding author: Gabriel Dubus: gabriel.dubus@mnhn.fr

ABSTRACT

Active learning is a promising framework for reducing annotation costs in large-scale bioacoustic monitoring, where expert labeling is expensive and data distributions are highly heterogeneous across environments. However, existing sample selection strategies often rely on static criteria that do not adapt to the evolving reliability of model predictions during training. This limitation can lead to sub-optimal exploration–exploitation trade-offs and redundant sample selection.

We propose an active learning strategy for multilabel bioacoustic event classification that jointly models predictive uncertainty, embedding-space diversity, and intra-batch redundancy. The method introduces an adaptive weighting scheme that progressively shifts from diversity-driven exploration in high-uncertainty regimes toward uncertainty-driven exploitation as the model becomes more confident, reflecting the increasing reliability of the classifier. To further improve annotation efficiency, a greedy Maximum Marginal Relevance (MMR) procedure is used to enforce diversity among selected samples within each acquisition batch.

We evaluate the proposed approach within the BioDCASE 2026 Task 4 active learning framework on terrestrial (BirdSet) and marine (ATBFL) benchmarks using pretrained audio embeddings and a fixed annotation budget. Experimental results show consistent improvements in learning efficiency and competitive in terms of macro mean Average Precision (mAP) and Area Under the Learning Curve (AULC) across heterogeneous acoustic domains. The gains are particularly pronounced on structured terrestrial soundscapes, while performance remains competitive under noisier marine conditions. These findings demonstrate that adaptive acquisition strategies combining uncertainty estimation, embedding-space diversity, and redundancy-aware batch construction provide an effective and robust solution for annotation-efficient bioacoustic learning.

Index Terms— Active Learning, Bioacoustics, Multilabel Classification, Sample Selection, Uncertainty Sampling, Diversity Sampling

1. INTRODUCTION

Large-scale bioacoustic monitoring relies on passive acoustic systems that continuously generate massive volumes of unlabeled audio data in both terrestrial and marine environments. Although these recordings contain rich ecological information, only a small fraction can be manually annotated due to the high cost and expertise required for reliable labeling [1]. These annotations are

typically required to train machine learning models for acoustic event detection and classification.

This imbalance between data availability and annotation budget raises a central question: *given a large pool of unlabeled acoustic data, how should samples be selected for annotation in order to maximize model performance under a strict labeling budget?*

This work was submitted to the BioDCASE 2026 Task 4 workshop challenge on Active Learning for Bioacoustics, which specifically addresses annotation-efficient learning across terrestrial and marine acoustic domains. The challenge provides a standardized active learning framework, pretrained PerchV2 embeddings, and fixed classifier training procedures, requiring participants to focus exclusively on the design of acquisition strategies under a constrained annotation budget.

Active Learning (AL) provides a principled framework to address this challenge by iteratively selecting the most informative samples for labeling [2]. AL methods aim to optimize acquisition strategies in order to improve model performance while minimizing annotation effort [3]. However, designing effective acquisition functions remains challenging, particularly in bioacoustic settings where signals are multilabel, highly imbalanced, and strongly domain-dependent across recording locations [1, 4].

Within the BioDCASE Task 4 setting, methods are evaluated on both BirdSet [5], a large-scale terrestrial avian benchmark, and ATBFL (Acoustic Trends Blue Fin Library) [6], a marine bioacoustic dataset derived from passive acoustic monitoring datasets in the Southern Ocean. These datasets exhibit substantially different acoustic structures, label densities, and noise characteristics, making cross-domain generalization a central challenge for AL methods.

In this work, we study pool-based AL under a fixed annotation budget using pretrained acoustic embeddings. We propose a unified acquisition strategy that jointly models predictive uncertainty, embedding-space diversity, and batch-level redundancy control. The method introduces an adaptive weighting scheme driven by global model confidence over the unlabeled pool, progressively shifting from diversity-driven exploration in high-uncertainty regimes toward uncertainty-driven exploitation as the model becomes more confident. A greedy Maximum Marginal Relevance (MMR) procedure is used to ensure diversity within each acquisition batch.

Contributions:

- We propose a unified AL strategy combining predictive uncertainty, embedding-space diversity, and batch-level redundancy reduction with a global-confidence adaptive weighting scheme

that dynamically balances exploration and exploitation during training.

- We integrate a greedy Maximum Marginal Relevance (MMR) mechanism to reduce redundancy within acquisition batches.
- We evaluate the method on BirdSet and ATBFL under a fixed annotation budget, demonstrating consistent improvements across heterogeneous bioacoustic domains.

2. MATERIALS AND DATASETS

We evaluate our method on the BioDCASE 2026 Task 4 benchmark, which includes both terrestrial and marine bioacoustic datasets. The task is formulated as a pool-based active learning setting, where only the sample acquisition strategy is learned while the classifier and training pipeline remain fixed.

2.1. Active Learning Framework

Evaluation is conducted using the BaseAL framework. At each active learning cycle, a batch of samples is selected from an unlabeled pool, their labels are automatically revealed (*oracle labeling*), and added to the labeled set. The classifier is then retrained on the updated labeled set. This process is repeated until a fixed annotation budget of 500 samples is reached.

All methods operate on precomputed 5-second audio embeddings extracted using the PerchV2 model. These embeddings serve as fixed representations for training a randomly initialized multilabel classification head, ensuring a controlled comparison of sampling strategies.

2.2. Datasets

The benchmark consists of two complementary bioacoustic datasets designed to evaluate generalization across terrestrial and marine acoustic environments. We use precomputed embeddings extracted using the PerchV2 model [7, 8].

2.2.1. BirdSet (Terrestrial)

BirdSet is a large-scale avian bioacoustic dataset containing recordings from diverse geographic regions [5]. For this task, we use three of the eight validation sets of Birdset: HSN, POW, and UHH. These subsets vary significantly in class cardinality (19,41 and 25 respectively), and label density (average labels per sample are 0.524, 2.833 and 1.058 respectively), making them well suited for evaluating robustness across structured soundscapes.

2.2.2. ATBFL (Marine)

ATBFL is a marine bioacoustic dataset derived from long-term monitoring of Antarctic whale populations [6]. It consists of multiple site-year subsets representing distinct acoustic environments. Compared to BirdSet, ATBFL exhibits higher noise levels, making it a more challenging domain for embedding-based similarity and diversity modeling.

2.3. Evaluation Protocol

All methods are evaluated under a fixed annotation budget of 500 samples per subset. Performance is measured using the Area Un-

der the Learning Curve (AULC) for macro mAP, averaged over 5 independent runs.

3. METHOD

We propose ADU-MMR (Adaptive Diversity–Uncertainty + Maximum Marginal Relevance), an active learning strategy that combines predictive uncertainty, embedding-space diversity, and batch-level redundancy reduction. The method operates on fixed audio embeddings and follows an adaptive exploration–exploitation schedule throughout the annotation process.

3.1. Problem Setting

Let L_t and U_t denote the labeled and unlabeled sets at iteration t . At each active learning step, a batch $S_t \subset U_t$ of size B is selected for annotation. Each sample x_i is represented by a fixed pretrained embedding extracted using PerchV2. The classifier produces a multilabel probability vector $p_i = (p_{i,1}, \dots, p_{i,C})$, where C is the number of classes. After annotation, selected samples are added to the labeled set and the model is updated.

3.2. Uncertainty Estimation

Predictive uncertainty is estimated using the sum of binary entropies computed from the multilabel class probabilities:

$$h_i = - \sum_{c=1}^C [p_{i,c} \log(p_{i,c} + \epsilon) + (1 - p_{i,c}) \log(1 - p_{i,c} + \epsilon)] \quad (1)$$

where ϵ is a small constant introduced for numerical stability. To ensure compatibility with the diversity score, $\tilde{h}_i = \text{Norm}(h_i)$ is defined as uncertainty values linearly normalized to the interval $[0, 1]$ through a min-max normalization over the unlabeled pool at iteration t .

3.3. Embedding-Based Diversity

To encourage exploration of underrepresented regions of the embedding space, we compute the distance between each unlabeled sample and its nearest labeled neighbor. For each sample $x_i \in U_t$:

$$d_i = \min_{x_j \in L_t} \|x_i - x_j\|_2^2. \quad (2)$$

The resulting distances are normalized to obtain $\tilde{d}_i$. When no labeled samples are available, all samples are assigned a diversity score of 1.

3.4. Adaptive Exploration

The relevance score of each sample is defined as:

$$s_i = \alpha_t \tilde{h}_i + (1 - \alpha_t) \tilde{d}_i. \quad (3)$$

The adaptive coefficient α_t is driven by the global uncertainty of the model over the unlabeled pool. We compute:

$$H_t = \frac{1}{|U_t|} \sum_{x_i \in U_t} \tilde{h}_i. \quad (4)$$

The coefficient is then defined as:

$$\alpha_t = 0.5 \cdot \left(\frac{\max(0, \tau - H_t)}{\tau} \right)^2, \quad (5)$$

This formulation reduces reliance on uncertainty in early stages when predictions are unreliable, and gradually increases its influence as the model becomes more confident. The threshold τ has been set empirically at $\tau = 0.05$ to ensure that uncertainty is only exploited once the model reaches a sufficiently low-entropy regime, where predictions become more reliable and uncertainty estimates more informative.

3.5. Batch Selection via Maximum Marginal Relevance

Selecting the top- B samples according to s_i can lead to redundant acquisitions. To mitigate this issue, we employ a greedy Maximum Marginal Relevance (MMR) procedure. This formulation follows the classical Maximum Marginal Relevance framework of Carbonell and Goldstein [9]. Embeddings are ℓ_2 -normalized and pairwise cosine similarities are computed. At each step, the candidate maximizing:

$$\text{MMR}(x_i) = s_i - \lambda \max_{x_j \in S_t} \cos(x_i, x_j) \quad (6)$$

where S_t denotes the set of samples selected so far within the current greedy construction of the acquisition batch at iteration t . The procedure is initialized with $S_t = \emptyset$ and proceeds by sequentially adding the sample maximizing the MMR criterion until B samples have been selected.

The MMR trade-off parameter was set to $\lambda = 0.3$. This value was selected empirically based on preliminary experiments, providing a good balance between score s_i maximization and redundancy reduction across the considered datasets.

The first selected sample corresponds to the highest acquisition score. Subsequent selections are penalized according to their similarity with previously selected samples, promoting diversity within the acquisition batch.

Algorithm 1 Adaptive Diversity–Uncertainty Active Learning

```

1: for each iteration  $t$  do
2:   Compute normalized uncertainty scores  $\tilde{h}_i$ 
3:   Compute normalized diversity scores  $\tilde{d}_i$ 
4:   Compute  $\alpha_t$ 
5:   Compute acquisition scores  $s_i = \alpha_t \tilde{h}_i + (1 - \alpha_t) \tilde{d}_i$ 
6:   Initialize  $S_t \leftarrow \emptyset$ 
7:   while  $|S_t| < B$  do
8:     Compute MMR scores  $\text{MMR}(x_i) = s_i - \lambda \max_{x_j \in S_t} \cos(x_i, x_j)$ 
9:     Select the highest-scoring sample
10:    Add it to  $S_t$ 
11:  end while
12:  Query annotations and update  $L_t$ 
13: end for
```

4. EXPERIMENTS

We evaluate ADU-MMR on the BirdSet (terrestrial) and ATBFL (marine) bioacoustic datasets using pretrained PerchV2 embeddings. Both datasets exhibit strong class imbalance and multilabel

Method	AULC (mAP macro)	Best mAP
ADU-MMR	0.505	0.590
CoreSet	0.479	0.583
Margin Sampling	0.422	0.532
Random	0.405	0.482

Table 1: Overall performance averaged across BirdSet and ATBFL datasets over 5 runs per dataset, then averaged across datasets. The proposed method outperforms baselines in both AULC and mAP.

Dataset	ADU-MMR	CoreSet	Margin	Random
AULC (mAP macro)				
ATBFL	0.454	0.462	0.441	0.464
HSN	0.632	0.556	0.444	0.382
POW	0.482	0.476	0.461	0.438
UHH	0.443	0.421	0.341	0.337
Best mAP				
ATBFL	0.484	0.489	0.483	0.494
HSN	0.716	0.696	0.600	0.477
POW	0.617	0.608	0.607	0.559
UHH	0.543	0.541	0.440	0.399

Table 2: Per-dataset performance of ADU-MMR and baselines in terms of AULC (macro mAP) and best achieved mAP. Best values per row are highlighted in bold.

annotations, making them suitable benchmarks for active learning under realistic acoustic conditions.

The annotation budget is fixed to 500 samples with a batch size of 25, resulting in 20 active learning iterations. All experiments are repeated over 5 independent runs, and results are reported as mean $\pm$ standard deviation.

We compare the proposed method against the following baselines: Random sampling, Margin-based uncertainty sampling, CoreSet selection [10] and TypiClust (diversity-aware baseline).

4.1. Results

Table 1 reports AULC and mAP scores on BirdSet and ATBFL for ADU-MMR and baselines.

Averaged across datasets, ADU-MMR achieves the best performance with 0.594 mAP and 0.505 AULC, outperforming all competing methods. These results indicate that the proposed strategy provides a consistent trade-off between exploration and exploitation across heterogeneous acoustic domains.

Table 2 reports per-dataset results of ADU-MMR and baselines. On BirdSet, ADU-MMR consistently improves performance across all subsets. On HSN, it significantly outperforms CoreSet (0.632 vs 0.556 AULC), indicating strong gains in both uncertainty- and diversity-driven regimes. Similar improvements are observed on POW, where it achieves 0.482 AULC, outperforming both CoreSet and random sampling. On UHH, gains are smaller but consistent, with ADU-MMR achieving the best AULC and mAP. On ATBFL, all methods show comparable performance, Random achieving best scores.

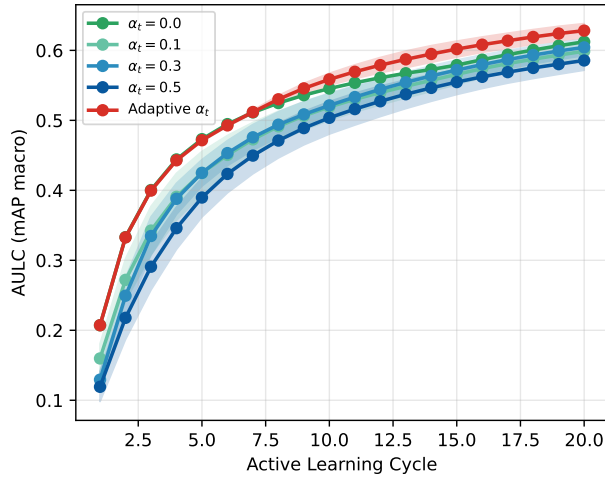


Figure 1: Evolution of AULC across active learning cycles for different fixed values of α (0, 0.1, 0.3, 0.5) as well as the proposed adaptive schedule, results for BirdSet-HSN dataset. The adaptive schedule consistently outperforms all fixed-weight configurations, particularly during early active learning cycles.

4.2. Effect of the adaptive weighting scheme

We analyze the impact of the adaptive weighting parameter α_t on the active learning performance. Figure 1 reports the evolution of AULC across active learning cycles for different fixed values of α (0, 0.1, 0.3, 0.5) as well as the proposed adaptive schedule, computed with the example of the BirdSet-HSN dataset.

Overall, the adaptive strategy consistently outperforms all fixed weighting configurations across active learning cycles. The largest performance gap is observed in the early stages, where high values of α underperform due to unreliable uncertainty estimates from a limited labeled set. In this regime, diversity-based selection is more effective than uncertainty-driven sampling.

At the beginning of the active learning process, the adaptive strategy behaves similarly to the $\alpha = 0.0$ configuration, as it assigns a low value to α_t and prioritizes diversity. As training progresses, the two strategies diverge: the fixed $\alpha = 0.0$ configuration saturates, while the adaptive approach continues to improve. This suggests that uncertainty estimates become increasingly reliable as the model is exposed to more labeled data, making uncertainty-based acquisition progressively more beneficial.

These results highlight the importance of dynamically balancing uncertainty and diversity rather than relying on a fixed weighting scheme.

5. DISCUSSION

The results highlight a clear dependency of active learning performance on dataset structure and embedding quality. On BirdSet, the proposed method consistently outperforms all baselines, indicating that pretrained embeddings provide a meaningful geometric structure that can be effectively exploited by both uncertainty and diversity-based acquisition strategies.

In contrast, results on ATBFL are more limited, with all methods

exhibiting similar performance. This suggests that sample selection strategies have a reduced impact in this setting, where performance gains from additional labeled data remain marginal across iterations.

ATBFL is a marine bioacoustic dataset composed of low-frequency recordings of large cetaceans in the Southern Ocean. Several target signals occur at very low frequencies (e.g., blue whale Z-calls around 25 Hz and fin whale calls around 20 Hz). However, the pretrained PerchV2 embeddings may not fully capture such low-frequency acoustic content, which could limit their representational adequacy in this domain. This likely contributes to the limited performance gains observed across active learning strategies on ATBFL. In such cases, when the underlying representation space does not adequately separate relevant acoustic events, improvements from sample selection strategies may be inherently constrained.

We further analyze the role of the adaptive weighting parameter α_t . Early in the active learning process, low values of α_t are beneficial due to unreliable uncertainty estimates from a classifier trained on limited labeled data. In this regime, diversity-based selection provides more stable exploration of the embedding space. As labeling progresses, uncertainty estimates become more reliable, allowing the method to progressively shift toward uncertainty-driven acquisition [11].

6. CONCLUSION

The proposed method combines uncertainty estimation, embedding-based diversity, and batch-level redundancy reduction within a unified active learning framework. It dynamically balances exploration and exploitation through an adaptive weighting scheme and enforces diversity via greedy MMR-based selection.

Experiments on BirdSet and ATBFL show consistent improvements over baselines, with particularly strong gains in structured terrestrial environments. Overall, the results demonstrate the effectiveness of adaptive acquisition strategies for multilabel bioacoustic active learning under strict annotation budgets.

7. ACKNOWLEDGMENT

The author thanks the BioDCASE 2026 Task 4 organizers for providing the BaseAL framework and curated datasets.

8. REFERENCES

- [1] D. Stowell, “Computational bioacoustics with deep learning: A review and roadmap,” *PeerJ*, vol. 10, p. e13152, Mar. 2022.
- [2] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang, “A Survey of Deep Active Learning,” *ACM Computing Surveys*, vol. 54, no. 9, pp. 1–40, Dec. 2022.
- [3] A. Tharwat and W. Schenck, “A Survey on Active Learning: State-of-the-Art, Practical Challenges and Research Directions,” *Mathematics*, vol. 11, no. 4, p. 820, Feb. 2023.
- [4] B. McEwen, K. Soltero, S. Gutschmidt, A. Bainbridge-Smith, J. Atlas, and R. Green, “Active few-shot learning for rare bioacoustic feature annotation,” *Ecological Informatics*, vol. 82, p. 102734, Sept. 2024.

- [5] L. Rauch, R. Schwinger, M. Wirth, R. Heinrich, D. Huseljic, M. Herde, J. Lange, S. Kahl, B. Sick, S. Tomforde, and C. Scholz, “BirdSet: A Large-Scale Dataset for Audio Classification in Avian Bioacoustics,” May 2025.
- [6] B. S. Miller, The IWC-SORP/SOOS Acoustic Trends Working Group, B. S. Miller, K. M. Stafford, I. Van Opzeeland, D. Harris, F. Samaran, A. Širović, S. Buchan, K. Findlay, N. Balcazar, S. Nieukirk, E. C. Leroy, M. Aulich, F. W. Shabangu, R. P. Dziak, W. S. Lee, and J. K. Hong, “An open access dataset for developing automated detectors of Antarctic baleen whale sounds and performance evaluation of two commonly used detectors,” *Scientific Reports*, vol. 11, no. 1, p. 806, Jan. 2021.
- [7] L. Rauch, M. Herde, and B. McEwen, “BioDCASE 2026 Task 4: BirdSet Dataset,” Mar. 2026.
- [8] R. Kurinchi-Vendhan, S. Zhang, and B. McEwen, “BioDCASE 2026 Task 4: ATBFL Dataset,” Mar. 2026.
- [9] J. Carbonell and J. Stewart, “The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries,” *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, June 1999.
- [10] O. Sener and S. Savarese, “Active Learning for Convolutional Neural Networks: A Core-Set Approach,” June 2018.
- [11] P. Soviany, R. T. Ionescu, P. Rota, and N. Sebe, “Curriculum learning: A survey,” *International Journal of Computer Vision*, vol. 130, no. 6, pp. 1526–1565, 2022.