PB-MFS: MULTI-FUNNEL SELECTION WITH PARETO-BALANCED UNCERTAINTY AND DIVERSITY FOR BIOACOUSTIC ACTIVE LEARNING

Technical Report

Haoyu Wang[1,2,*], *Han Dou*[3,*], *Hui Chen*[3], *Yushan Du*[1], *Liushi Tu*[3], *Ningning Pan*[1], *Hao Zhang*[3], *Jingyao Wu*[4], *Gang Li*[2], *Gongping Huang*[3,†]

[1] Southwestern University of Finance and Economics, China
[2] MiLM Plus, Xiaomi Inc.
[3] IASP Lab, Wuhan University, China
[4] Massachusetts Institute of Technology, USA

ABSTRACT

Active learning mitigates annotation costs in bioacoustic classification; however, effective batch selection remains challenging for multi-label audio data. Conventional uncertainty-based strategies frequently sample redundant instances, whereas diversity-based approaches tend to overlook informative samples crucial for optimizing the current classifier. To address these limitations, we propose Pareto-Balanced Multi-Funnel Selection (PB-MFS), a novel batch active learning strategy that decouples candidate filtering from final batch selection. Specifically, PB-MFS initially constructs a survivor set via multiple complementary acquisition funnels. Subsequently, it determines the optimal annotation batch by harmonizing uncertainty and diversity through a Pareto optimization objective. The proposed method integrates diversity-driven cold-start selection, multi-funnel survivor voting, and an uncertainty-diversity trade-off leveraging pretrained audio embeddings. Comprehensive evaluations on the four BioDCASE 2026 Task 4 subsets (HSN, POW, UHH, and ATBFL) demonstrate that PB-MFS achieves an aggregated mean AULC of 0.510. Furthermore, it outperforms the strongest baseline, CoreSet, by an absolute margin of 0.050, delivering a 10.9% relative performance gain.

***Index Terms*—** active learning, bioacoustics, multi-label classification, uncertainty sampling, diversity sampling

1. INTRODUCTION

Bioacoustic monitoring has become a vital tool for large-scale studies of biodiversity, ecosystem dynamics, and species activities. Although modern recorders enable massive audio collection with minimal human intervention, reliable annotation remains costly and time-consuming [1, 2]. This bottleneck is especially pronounced in multi-label bioacoustic classification, where one audio segment often contains multiple overlapping acoustic events. Thus, reducing annotation effort while preserving model performance is essential for practical bioacoustic learning systems.

Active learning addresses this issue by iteratively selecting small batches of unlabeled instances for annotation and retraining the model on the augmented labeled dataset [3]. The key challenge is to identify instances with maximum information utility under constrained annotation budgets. In batch active learning, this difficulty increases because acquired samples must be individually informative and collectively non-redundant.

For multi-label bioacoustic classification, a single acquisition criterion is often insufficient. Uncertainty sampling tends to prioritize ambiguous but redundant instances [4], whereas diversity-based selection broadly explores the embedding space but can overlook informative samples crucial for refining the current classifier [5]. Therefore, prior batch active learning methods combine uncertainty and diversity metrics to reduce redundancy [6]. Following this principle, we aggregate multiple acquisition signals before applying a final uncertainty-diversity selection rule.

We propose Pareto-Balanced Multi-Funnel Selection (PB-MFS), a novel batch active learning strategy for multi-label bioacoustic classification. PB-MFS decouples candidate filtering from final batch selection. It initializes the labeled dataset with a PCA-based farthest-first cold-start mechanism, then constructs a compact survivor set through multiple filtering funnels driven by complementary acquisition criteria. Finally, it selects the annotation batch using a Pareto-balanced objective, prioritizing samples with both high uncertainty and substantial diversity.

The primary contributions of this work are threefold. First, we introduce a multi-funnel candidate filtering strategy that aggregates complementary acquisition criteria to optimize bioacoustic active learning. Second, we propose a Pareto-balanced batch selection objective that explicitly harmonizes uncertainty and diversity during final sample selection. Third, we systematically evaluate the proposed strategy within the BaseAL active learning framework, demonstrating consistent superiority over standard baselines under stringent annotation budgets.

2. RELATED WORK

Active learning reduces annotation cost by selecting informative unlabeled samples under a limited budget. Settles [3] summarizes the main paradigms, noting that pool-based active learning suits large unlabeled collections. This setting is relevant to bioacoustic monitoring, where abundant passive recordings require expert effort for species identification, event verification, and multi-label annotation. Prior active learning applications in long-duration environmental audio [7] and wildlife passive acoustic monitoring [8] demonstrate that sample selection is critical under limited bud-

* Equal contribution.
† Corresponding author: gongpinghuang@whu.edu.cn

gets. Classical active learning strategies include uncertainty-based, diversity-based, and representative selection. Uncertainty sampling, originating from Lewis and Gale [4], includes least-confidence and margin-based methods but remains unreliable during cold start. Diversity methods like CoreSet [5] and farthest-first traversal reduce batch redundancy but may favor outliers. Representative methods like TypiClust [9] combine clustering and typicality to select samples but may under-cover rare events. Hybrid active learning combines multiple signals to address these limitations. BADGE [6] jointly models uncertainty and diversity in gradient embeddings. In bioacoustics, hybrid strategies prove useful as data often contain sparse events, repeated soundscapes, long-tailed classes, and complex noise. Rauch et al. [10] highlight domain transfer, annotation scarcity, and environmental variation in avian bioacoustics, while Zhang and Virtanen [11] combine committee disagreement with farthest traversal for bioacoustic sound event detection. PB-MFS follows this hybrid direction for multi-label bioacoustic classification. It combines PCA-based farthest-first cold start, multi-funnel survivor voting for candidate filtering, and Pareto-balanced uncertainty-diversity selection for final batch construction. This design targets cold-start instability, batch redundancy, and rare-event under-coverage under limited annotation budgets.

3. METHODS

3.1. Problem Setup

Let $\mathcal{U}_t$ and $\mathcal{L}_t$ denote the unlabeled pool and the labeled dataset, respectively, at active learning cycle t. During each cycle, the sampling mechanism selects an annotation batch $\mathcal{B}_t \subset \mathcal{U}_t$ of size $|\mathcal{B}_t| = k_t$, where k_t represents the annotation budget for the current cycle. Upon selection, an oracle provides the ground-truth multi-label annotations, denoted as $y_i = [y_{i1}, \dots, y_{iC}]^\top \in \{0,1\}^C$, for each instance $i \in \mathcal{B}_t$. The datasets are subsequently updated such that $\mathcal{L}_{t+1} = \mathcal{L}_t \cup \mathcal{B}_t$ and $\mathcal{U}_{t+1} = \mathcal{U}_t \setminus \mathcal{B}_t$, followed by retraining the classifier on the augmented labeled dataset $\mathcal{L}_{t+1}$. Each audio segment i is represented by a fixed embedding vector z_i, which is extracted from the provided audio representation. Once at least one labeled batch becomes available, the classifier outputs multi-label prediction probabilities $p_i = [p_{i1}, \dots, p_{iC}]^\top \in [0,1]^C$, where C denotes the total number of target classes. The objective is to select instances that optimally improve the performance of the classifier under a constrained annotation budget, while avoiding the inclusion of redundant samples within a single batch.

3.2. Overview of PB-MFS

As illustrated in Figure 1, we propose PB-MFS, a batch active learning strategy tailored for multi-label bioacoustic classification. PB-MFS applies distinct acquisition rules contingent upon the availability of labeled examples. During the initial cycle, when model predictions are either unavailable or unreliable, the method employs a diversity-based cold-start mechanism. In subsequent cycles, PB-MFS distills the expansive unlabeled pool into a compact survivor set via multiple filtering funnels before executing the final batch selection based on a Pareto-balanced uncertainty-diversity rule. This design decouples coarse candidate screening from fine-grained batch construction, thereby mitigating the reliance of the acquisition process on any single scoring metric.

3.3. PB-MFS Acquisition Strategy

3.3.1. Diversity-Based Cold Start

As illustrated in the upper panel of Figure 1, the cold-start phase constructs the initial batch via PCA-based farthest-first traversal. In the first cycle, $\mathcal{L}_t$ is empty, so uncertainty-based metrics are unavailable. PB-MFS initializes selection using a diversity-driven strategy in a 256-dimensional PCA space. Naive farthest-first traversal is sensitive to outliers, as the initial seed may lie at an extreme point. To mitigate this, candidates are ranked by centroid distance and an adaptive skip is applied to the high-distance tail. The initial seed is chosen by the largest local distance gap within a restricted window, and remaining samples are selected via farthest-first traversal. This ensures coverage of the acoustic embedding space before model-based acquisition.

3.3.2. Multi-Funnel Candidate Filtering

As illustrated in the middle panel of Figure 1, the multi-stage filtering process refines the unlabeled pool into a compact set of surviving instances. Following the cold-start phase, PB-MFS incorporates a candidate refinement procedure to avoid the execution of the final greedy selector across the entire pool of unlabeled instances. Initially, the method constructs an approximate candidate pool $\mathcal{P}_t \subseteq \mathcal{U}_t$, which comprises all available instances. Subsequently, multiple filtering funnels are applied to $\mathcal{P}_t$. Each funnel constitutes an ordered sequence of acquisition criteria. This ordering is critical, as each criterion exclusively evaluates the instances that survive the preceding filter. PB-MFS employs five distinct families of criteria. To ensure the reproducibility of the initial filtering process, the scores corresponding to these criteria are formally defined as follows:

$$
\begin{aligned}
q_{\mathrm{CF}}(i) &= \min_{j \in \mathcal{L}_t} \|z_i - z_j\|_2^2, \\
q_{\mathrm{MM}}(i) &= 1 - \frac{1}{0.5C} \sum_{c=1}^{C} |p_{ic} - 0.5|, \\
q_{\mathrm{LCB}}(i) &= \sum_{c=1}^{C} p_{ic} \frac{1}{n_c + 1}, \\
q_{\mathrm{NC}}(i) &= \frac{1}{C} \sum_{c=1}^{C} |p_{ic} - \bar{y}_{ic}^{(K)}|, \\
q_{\mathrm{LCI}}(i) &= \left| \sum_{c=1}^{C} p_{ic} - \frac{1}{|\mathcal{L}_t|} \sum_{j \in \mathcal{L}_t} \sum_{c=1}^{C} y_{jc} \right|.
\end{aligned}
$$

Here, $n_c = \sum_{j \in \mathcal{L}_t} y_{jc}$ denotes the number of labeled examples associated with class c, and $\bar{y}_{ic}^{(K)}$ represents the average label value for class c computed from the K nearest labeled neighbors of candidate i. Specifically, the CF criterion favors candidates located far from existing labeled examples; the MM criterion prioritizes instances with highly uncertain predictions; the LCB criterion focuses on underrepresented classes; the NC criterion selects candidates whose predictions diverge from those of the local labeled neighborhoods; and the LCI criterion targets candidates that exhibit a label cardinality different from that of the current labeled dataset. For composite criteria, such as the combination of NC and LCI, PB-MFS normalizes each constituent score to the range $[0,1]$ and computes the average of these normalized values.

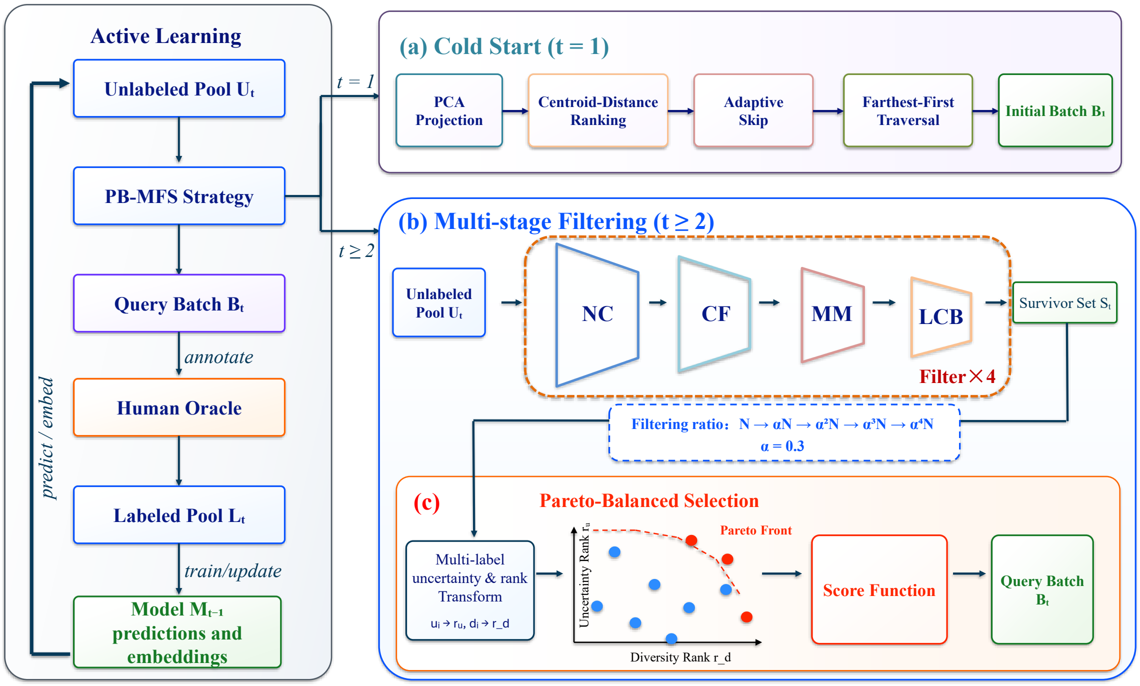

Figure 1: Overview of the proposed Pareto-Balanced Multi-Funnel Selection (PB-MFS) framework. The system consists of three core modules: (a) **Cold Start** ($t = 1$), which initializes the active learning cycle by extracting a diverse initial batch; (b) **Multi-stage Filtering** ($t \geq 2$), which employs a cascade of specialized filters to progressively prune the unlabeled pool into a high-quality survivor set; and (c) **Pareto-Balanced Selection**, which constructs the final query batch by systematically unifying multi-label uncertainty and diversity ranks.

Within each funnel, PB-MFS retains the top-ranking candidates evaluated under the current criterion and discards the remainder. Given a current survivor set of size N, the subsequent set retains $\max(4k_t, \mathrm{round}(\alpha N))$ candidates, where α denotes the funnel retention ratio. This lower bound prevents early criteria from aggressively over-pruning the candidate pool. The current implementation instantiates four distinct funnels derived from a base criterion sequence. In the reported configuration, the base sequence is defined as $\{\mathrm{NC} \to \mathrm{CF} \to \mathrm{MM} \to \mathrm{LCB}\}$. The four funnel variants modify this base structure by replacing NC with NC+LCI in two configurations, and replacing LCB with LCI in the remaining two, thereby generating complementary perspectives on neighborhood disagreement and label-cardinality behavior. After applying all funnels, the resulting survivor lists are aggregated via a voting mechanism. Let $\mathcal{S}_t^{(f)}$ denote the survivor list associated with funnel f; the vote count for candidate i is computed as:

$$v_i = \sum_f \mathbf{1}\{i \in \mathcal{S}_t^{(f)}\}.$$

Candidates that survive a greater number of funnels receive higher priority, and the instances accumulating the highest vote counts constitute the final survivor set $\mathcal{S}_t$. The size of this survivor set is constrained to at least $\max(4k_t, 100)$ candidates, provided that sufficient examples remain available.

3.3.3. Pareto-Balanced Batch Selection

As illustrated in the lower panel of Figure 1, the final annotation batch is selected from $\mathcal{S}_t$ by jointly optimizing a Pareto-balanced uncertainty–diversity objective[12]. For each candidate i, the multi-label uncertainty score is formulated as[13]:

$$u_i = \frac{1}{C} \sum_{c=1}^{C} 4 p_{ic}(1 - p_{ic}),$$

which reaches its maximum when the prediction probabilities approach 0.5. Prior to ranking, PB-MFS applies an uncertainty floor ϵ and an exponent parameter γ, yielding the transformed score $\tilde{u}_i = \max(\epsilon, u_i)^\gamma$. The floor prevents negligible uncertainty values from distorting the ranking, while the exponent regulates the intensity of the uncertainty weighting. These transformed scores are then converted into normalized ranks $r_u(i) \in [0, 1]$. To evaluate diversity, the embeddings of the candidates in $\mathcal{S}_t$ are projected into a 64-dimensional PCA space and subsequently normalized. The selection process operates in a greedy manner, initializing the batch with the candidate that attains the highest uncertainty rank. At each iteration, PB-MFS computes the squared distance from each remaining candidate to the nearest selected sample, converting these distances into normalized diversity ranks $r_d(i) \in [0, 1]$. Subsequently, PB-MFS selects the candidate that maximizes the following objective function:

$$a(i) = \min(r_u(i), r_d(i)) + \eta \sqrt{r_u(i) r_d(i)},$$

where η serves as a balancing hyperparameter. The minimum term penalizes candidates that excel under only a single criterion, whereas the geometric mean term rewards instances that demonstrate both high uncertainty and substantial diversity. This rank-based formulation remains robust against variations in the score scales across different datasets and active learning cycles. Furthermore, it prevents the final batch from being dominated by either redundant uncertain samples or diverse but uninformative instances.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

We evaluate the performance of PB-MFS on the BioDCASE 2026 Task 4 active learning benchmark [14], which encompasses both the terrestrial and marine bioacoustic domains. For the terrestrial domain, we utilize three BirdSet development subsets (HSN, POW, and UHH) [15], whereas the ATBFL subset is employed for the marine domain [16]. All datasets provide precomputed `perch_v2` embeddings[17] and strictly adhere to the official BaseAL format. The experiments follow the official evaluation protocol, wherein a maximum annotation budget of $M = 500$ labeled samples per subset is allocated across 16 active learning cycles. Following the data acquisition in each cycle, the classification head is retrained for 10 epochs. Each experimental configuration is evaluated across five independent runs to ensure statistical robustness. We compare PB-MFS against the official BaseAL baselines: random sampling, margin sampling, CoreSet, and TypiClust. To guarantee a fair comparison, all configurations share an identical model architecture and training pipeline. The primary evaluation metric is the area under the learning curve (AULC), which quantifies the label efficiency across the entire data acquisition process. For a given cycle involving C target classes, the macro mean average precision (mAP) is defined as:

$$\mathrm{mAP}_{\mathrm{macro}} = \frac{1}{C}\sum_{c=1}^{C}\mathrm{AP}_c.$$

Subsequently, the AULC is computed as the normalized area under this performance curve:

$$\mathrm{AULC} = \frac{1}{M}\sum_{t=1}^{T}(n_t - n_{t-1})\,\mathrm{mAP}_{\mathrm{macro}}(n_t),$$

where n_t denotes the cumulative number of labeled samples at cycle t, with the initial state defined as $n_0 = 0$.

4.2. Main Results and Generalization

Table 1 presents the results of the comparative evaluation. Notably, PB-MFS achieves the best overall performance, securing an overall AULC of 0.510 and outperforming the strongest baseline, CoreSet, by an absolute margin of 0.050. Among the baseline methods, a clear trade-off between label efficiency and computational overhead is evident. Random and margin sampling exhibit minimal computational demands but yield suboptimal label efficiency, indicating that uncertainty alone is insufficient for multi-label bioacoustic scenarios. Conversely, structure-aware methods, such as CoreSet and TypiClust, improve label efficiency but incur substantial computational costs during the sampling phase. PB-MFS effectively resolves this trade-off, delivering superior label efficiency while requiring merely 1.15 seconds per sampling step. Beyond the aggregated metrics, PB-MFS demonstrates robust cross-domain generalization. Compared to the strongest baselines, PB-MFS consistently improves the AULC across all four subsets. Specifically, the proposed method elevates the AULC from 0.503 to 0.623 on the terrestrial HSN subset. Similar steady improvements are observed on the POW (from 0.447 to 0.492) and UHH (from 0.395 to 0.454) subsets, as well as on the marine ATBFL dataset (from 0.456 to 0.469). This verifies that the proposed acquisition strategy generalizes effectively beyond the characteristics of any single dataset.

Table 1: Baseline comparison under the BaseAL evaluation protocol. The evaluation metrics include AULC (mAP macro), sampling wall-time per step, and cumulative annotation cost. Best results are highlighted in **bold** and second-best are <u>underlined</u>.

Method	AULC (mAP macro) ↑	Wall-time (s) ↓	Ann. Cost ↓
Random	0.390	**0.00131**	<u>966.85</u>
Margin	0.399	<u>0.00307</u>	1267.75
CoreSet	<u>0.460</u>	5.04364	1019.05
TypiClust	0.423	5.90751	**958.10**
PB-MFS	**0.510**	1.15090	1146.55

4.3. Ablation Results

To isolate the contributions of individual components, Table 2 presents the results of an ablation study conducted on PB-MFS. Replacing the adaptive PCA-based cold start with a standard initialization leads to the most severe performance degradation ($\Delta\mathrm{AULC} = -0.026$). This confirms that robust initial data coverage is imperative for the success of subsequent active learning cycles. Furthermore, substituting the Pareto-balanced final selection with a single-objective selector yields only a marginal decrease in performance ($\Delta\mathrm{AULC} = -0.002$). This outcome indicates that the upstream multi-funnel filtering accounts for the majority of the efficiency gains, whereas the Pareto optimization primarily serves as a fine-tuning mechanism for batch construction.

Table 2: Ablation study on the components of PB-MFS. Arrows indicate whether higher (↑) or lower (↓) values are optimal. The best results are highlighted in **bold**, and the second-best results are <u>underlined</u>.

Method	AULC ↑	Wall-time (s) ↓	Ann. Cost ↓	Δ AULC
PB-MFS	**0.510**	0.5930	<u>1145.60</u>	**0.000**
PB-MFS w/o Cold Start	0.483	**0.4210**	1163.20	-0.026
PB-MFS w/o Pareto	<u>0.508</u>	<u>0.5891</u>	1150.15	<u>-0.002</u>

5. CONCLUSION AND LIMITATIONS

In this paper, we present PB-MFS, a multi-stage active learning strategy tailored for bioacoustic classification under limited annotation budgets. The proposed method incorporates a diversity-based cold start, multi-funnel survivor voting, and Pareto-balanced uncertainty-diversity selection. This design provides a systematic approach to fusing complementary acquisition signals while mitigating redundancy during batch construction. Extensive evaluations demonstrate that PB-MFS achieves state-of-the-art performance, exhibiting robust cross-domain generalization and superior label efficiency. While the framework relies on the quality of pretrained audio embeddings and introduces minor computational overhead with additional hyperparameters, future work will address these factors by streamlining configurations and enhancing representation robustness, besides evaluating the system beyond the BaseAL protocol and BioDCASE Task 4 datasets.

6. REFERENCES

[1] D. Stowell, "Computational bioacoustics with deep learning: A review and roadmap," *PeerJ*, vol. 10, p. e13152, 2022. [Online]. Available: https://doi.org/10.7717/peerj.13152

[2] L. Rauch, R. Schwinger, M. Wirth, R. Heinrich, J. Lange, S. Kahl, B. Sick, S. Tomforde, and C. Scholz, "Birdset: A multi-task benchmark for classification in avian bioacoustics," *CoRR*, vol. abs/2403.10380, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2403.10380

[3] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009. [Online]. Available: http://burrsettles.com/pub/settles.activelearning.pdf

[4] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *SIGIR '94*. Springer London, 1994, pp. 3–12. [Online]. Available: https://doi.org/10.1007/978-1-4471-2099-5_1

[5] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *6th International Conference on Learning Representations, ICLR 2018*. OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=H1aIuk-RW

[6] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, "Deep batch active learning by diverse, uncertain gradient lower bounds," in *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=ryghZJBKPS

[7] M. Kholghi, Y. Phillips, M. Towsey, L. Sitbon, and P. Roe, "Active learning for classifying long-duration audio recordings of the environment," *Methods in Ecology and Evolution*, vol. 9, no. 9, pp. 1948–1958, 2018. [Online]. Available: https://doi.org/10.1111/2041-210X.13042

[8] H. Kath, P. P. Serafini, I. B. Campos, T. S. Gouvea, and D. Sonntag, "Leveraging transfer learning and active learning for data annotation in passive acoustic monitoring of wildlife," *Ecological Informatics*, vol. 82, p. 102710, 2024. [Online]. Available: https://doi.org/10.1016/j.ecoinf.2024.102710

[9] G. Hacohen, A. Dekel, and D. Weinshall, "Active learning on a budget: Opposite strategies suit high and low budgets," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 2022, pp. 8175–8195. [Online]. Available: https://proceedings.mlr.press/v162/hacohen22a.html

[10] L. Rauch, D. Huseljic, M. Wirth, J. Decke, B. Sick, and C. Scholz, "Towards deep active learning in avian bioacoustics," *CoRR*, vol. abs/2406.18621, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2406.18621

[11] S. Zhang and T. Virtanen, "Hybrid disagreement-diversity active learning for bioacoustic sound event detection," in *Proceedings of the 33rd European Signal Processing Conference*, 2025. [Online]. Available: https://doi.org/10.23919/EUSIPCO63237.2025.11226784

[12] K. e. a. Deb, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE Transactions on Evolutionary Computation*, 2002.

[13] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining*, vol. 3, no. 3, pp. 1–13, 2007. [Online]. Available: https://doi.org/10.4018/jdwm.2007070101

[14] B. McEwen and S. Zhang, "BaseAL: Active learning baseline," 2026. [Online]. Available: https://github.com/BenMcEwen1/BaseAL

[15] L. Rauch, M. Herde, and B. McEwen, "Biodcase 2026 task 4: Birdset dataset," Mar. 2026. [Online]. Available: https://doi.org/10.5281/zenodo.19191603

[16] R. Kurinchi-Vendhan, S. Zhang, and B. McEwen, "Biodcase 2026 task 4: Atbfl dataset," 2026. [Online]. Available: https://doi.org/10.5281/zenodo.19133112

[17] B. van Merriënboer, V. Dumoulin, J. Hamer, L. Harrell, A. Burns, and T. Denton, "Perch 2.0: The bittern lesson for bioacoustics," *arXiv preprint arXiv:2508.04665*, 2025. [Online]. Available: https://arxiv.org/abs/2508.04665