Active learning: Human-in-the-Loop Strategies to Efficiently Analyse Big Acoustic Datasets

Joel Hoefs, NPSC3000
Supervised by Dr Paul Nguyen Hong Duc,
Dr Evgeny Sidenko, Prof. Christine Erbe
CMST Dept. Curtin University

October 11, 2024

Abstract

This paper explores methods of sound event detection and classification for terrestrial life in long-duration audio, applying and comparing state-of-the-art active learning methods to minimize labelling efforts for passive acoustic monitoring tasks. We take the most robust case of discrete signature classification by segmentation against a vast, diverse dataset from Mauritius and find that most traditional clustering and diversity-based sampling methods are intractable. Existing deep active learning research does not prioritize efficiency at this scale interrupting the annotation workflow significantly. In order for classification of precise, rare bird calls, we design a lightweight classification-by-segmentation pipeline and propose a novel method for information diversity sampling on model embeddings with uncertainty. This allowed the model to correctly adjust to subtle features that discriminate the target classes in a large diverse dataset. In this case, the proposed method outperformed baselines achieving 0.88% of the potential performance accessing just 10% labelled data.

1 Introduction

Eco-acoustics is a new area of research concerning the relationship between an environment's soundscape and ecological health via surveillance efforts, often involving vast amounts of passive acoustic data for sound event detection and classification. Investigating the long duration temporal patterns, richness and diversity of animal sounds (*biophony*), natural sounds (*geophony*) and human sounds (*anthrophony*) involves the use of robust detection and classification algorithms that summarise terabytes of acoustic data.

Conventionally a domain expert wisely chooses sound samples , observes, listens to, and annotates signals of interest in order to produce meaningful labelled datasets all while minimizing subjectivity and human error. Now with the overwhelming data

surplus generated by modern Passive Acoustic Monitoring (PAM), the data greed of deep models far out-asks manual annotation efforts (Ren et al., 2021; Kholghi et al., 2018) and thus more efficient methods are called for.

Large annotation tasks may be assisted through interactive state space representations of audio (Mosqueira-Rey et al., 2022), revealing patterns through semi-supervised clusters (Phillips et al., 2018; S. Zhao et al., 2020) and other heuristic visual aids that guide the discovery process speeding up the annotation bottleneck. Supervised learning systems trained on these annotations may continuously monitor, stream analytics and adapt with ecosystem changes via on-line training paradigms. The resulting habitual mappings of species is then used to infer ecosystem stability, impact of human noise pollution and understand population dynamics with greater precision (Bateman & Uzal, 2022). Yet for classification to be robust, adaptive, and verifiable, the full data collective must be taken advantage of.

Active Learning (AL) is an extension of supervised learning, where the sampling order is optimized singling out the most informative, diverse and representative unlabelled instances to be annotated. This accelerates classifier accuracy and reduces the manual annotation effort required to reach baselines for quality model performance. Traditional AL focuses on smaller scale one-by-one sampling, a harmful process for deep models designed for batch learning leading to over-fitting and impractical training time. The batched-inputs, overconfident soft-max response and long training/evaluation/query times in deep learning has prompted the attention of new deep AL strategies to adapt (Ren et al., 2021). Thus, a human-in-the-loop deep AL approach is proposed to minimize the annotation effort, known as the *labelling budget*, whilst maintaining satisfactory classification metrics for supervised learning of bird call signatures for robust PAM. This method will be applied to a fully unlabeled, raw PAM dataset of 4300hrs against 4 discrete bird-call signatures that are not clusterable. The specificity, rarity, and subjectivity of the chosen signals provides an opportunity to improve these methods for nuanced and practical applications.

2 Existing Methods

Sound event detection and classification pipelines in PAM are mostly un-standardized, with the choice of method instead being adapted for situational data and target signals. However, there is proven superiority of Convolutions Neural Network (CNN) implementations for noisy and dynamic soundscapes owing to their ability to summarize and explain high level features invariant to spectral context (EmreÇakır et al., 2017). CNN's possess a degree of robustness that drastically outperforms traditional energy-based algorithms as demonstrated by Allen et al. (2021).

Ryazanov et al. (2021), Fischer et al. (2023) and Boiński et al. (2022) used the Residual Neural Network (ResNet) architecture for classification by segmentation of spectral-temporal bounding boxes in marine and environmental soundscapes, generalizing a small acoustic datasets with variable noise conditions surprising well. Fischer et al. (2023) elaborates the effectiveness of a Bayesian Resnet-32 model with AL through disagreement. Bayesian models are able to output precise uncertainty

measures without model variance but are less durable for larger datasets (Sener & Savarese, 2018).

(Shishkin et al., 2021) demonstrates state-of-the-art performance of Medoid Active Learning (MAL) on pretrained embeddings of Mel-Frequency Cepstrum Coefficients (MFCC) against a deep Bayesian classifier. A drastic improvement is shown over the previously most efficient method with MAL and semi supervised active learning with a support vector machine (Shuyang, Heittola, & Virtanen, 2017). Later, Shishkin et al. improves the MAL-embeddings method with Gaussian Dense active learning on raw MFCC's, likely being the most effective AL method for the Urban8k benchmark to this date. However, These methods strictly classify pre-segmented sound clips as weak labels. Practical studies of soundscape biophony are naturally un-segmented, where the occurrence of sound events or target signals is separate task onto itself via segmentation-by-classification methods producing strong labels.

Clustering algorithms have also shown promising results classifying weak labels in harmony with AL. Kholghi et al. (2018) implemented hierarchy clustering followed by K-means on acoustic index vectors with AL. Phillips et al. (2018) reused this method to describing temporal patterns of clusters, heavily reducing data storage with acoustic indexes from 5.7TB to just 1.14 Million descriptive vectors. Hilasaca et al. (2021) used this method for insects, birds and frogs from spectrograms with high performance accompanied by MAL and contour sampling. These clustering implementations benefit from helpful visualizations of state-space representations that assist the manual annotation process facilitating cooperative learning paradigms.

S. Zhao et al. (2020) implemented a binary sound event detector as a pre-trained gated-CNN with a bi-GRU head on log-mel spectrogams. It is shown that Mismatch-first farthest traversal (MFFS) sampling outperforms uncertainty based techniques for event detection achieving acceptable performance with 2% annotated data. Shuyang et al. (2018) also demonstrated the strength of MFFS on the UrbanSound8K dataset using a SVM.

Strong label classification of long-duration audio requires automated sound event detection, which is either a separate task to classification or combined into a single deep model. Mesaros et al. (2021) provides an excellent summary of segmentation-by-classification methods, introducing the Convolution Recurrant Neural Network (CRNN) with frequency-restricted pooling in the convolutions layers, preserving temporal context for the recurrent layers that can then trace the frame-wise time-steps of class occurrences as a binarized matrix. Mesaros et al. covers new methods of teacher-student learning paradigms for tagging and boundary detection respectively. Z. Zhao et al. (2017) instead implemented A Gaussian mixture model for event-level segmentation with a multi class SVM for classification. This method however relies on classifying the majority of signals that occur and may be problematic for rarer targets.

Parcerisas et al. (2024) repurposed a pre-trained 'You Only Learn Once' (YOLO) model for bounded box spectrogram segmentation and a SVM for classification.

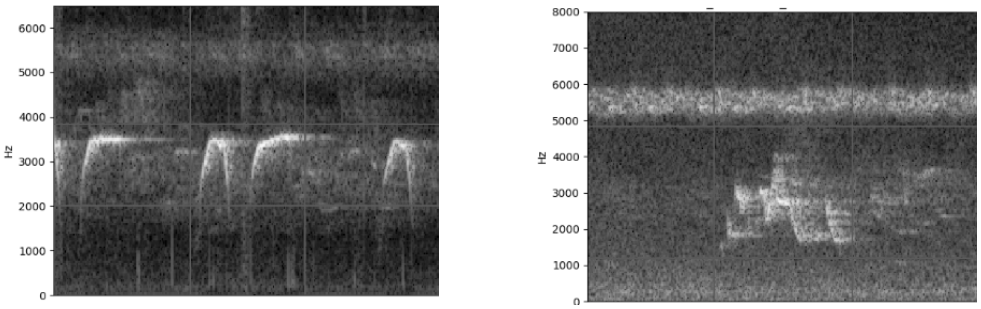
Venkatesh et al. (2022) introduced an efficient YOLO-like algorithm for a strong label regression problem of 0.3 second chunks rather than frames. Venkatesh et al.'s 'You Only Hear Once' (YOHO) model was designed for efficiency and scalability drastically reducing the parameter load and training time, a crucial part for active learning in PAM. Luo et al. (2021) showed a capsule neural network (CapsNN) that outperforms YOHO for polymorphic soundscapes. CapsNN architectures can effectively represent overlapping acoustic objects and their characteristics such as rotation, hue and shift in a hierarchical capsule routing mechanism, where CNN's otherwise loose this information through pooling layers. This makes CapsNN excellent at distinguishing target signals and its variants in rain, wind and from great distances.

3 Method

3.1 Data

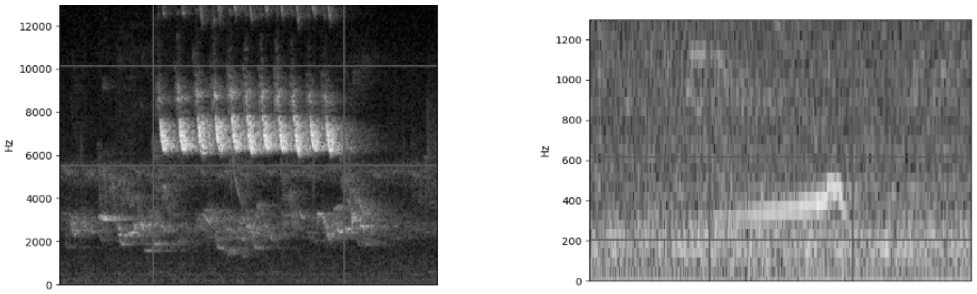
Our data was gathered from a Song Meter SM4 Acoustic Recorder site in the Ebony Forest reserve, Mauritius. Recordings are 10-minutes each and sampled at 96khz with a 60% duty cycle from February to November in 2023 approximating 4300 recording hours. The Ebony Forest reserve has a highly diverse ecosystem and abundance of endangered species, making it perfect for species recognition as annotation efforts and deep learning techniques have yet to be applied here.

The classification targets were chosen to appropriately test preciseness and ambiguity, a deliberately different task from baseline datasets like Urban-8k and YouTube-8M. Instead of recognising long-term patterns produces by the species, we focus on discrete bird call signatures.



(a) '*nr* Syllable' (b) 'Triangular Flourish'

Figure 1: Spectrograms of the chosen target audio signatures (Part 1)



(a) 'Fast Trill' (b) 'Upsweep'

Figure 2: Spectrograms of the chosen target audio signatures (Part 2)

The 3khz 'nr syllable' in figure 1a forming an 'nr'-like shape on the spectrogram is that of a Mascarene Paradise-Flycatche *(Terpsiphone bourbonnensis)* and is chosen for an easy to detect and repetitive signal. The 3khz 'Traingle Flourish' in figure 1b is likely produced by the Mauritius Grey White-eye *(Zosterops mauritianus)* and is chosen for its highly ambiguous, inconsistent shape which is easily miss-classified in the mid-range introducing annotation subjectivity. The 6khz 'Trill' sound produced by the Mauritius Kestrel in figure 2a *(Falco punctatus)* and the 500hz 'Upsweep' produced by the Pink Pigeon in figure 2b (*Nesoenas Mayeri*) are chosen to force the classifier to predict over the full frequency width. Annotation lengths range from 0.3 (smallest upsweeps) to 2 seconds (longest trills).

3.2 Initial Annotation

Deep-AL queries require an initial 'warm-start' annotated-set before they are capable of producing meaningful model-based uncertainty measures. The most efficient method to expedite this process is semi-supervised MAL (Shishkin et al., 2021, 2024; Shuyang et al., 2017; Hilasaca et al., 2021). This, whilst highly efficient for sound-clip tagging, introduces some over-diversity bias in the initial dataset (Shishkin et al., 2021) and requires the data to be n-clusterable, unlike in some PAM where the noise class overlaps all data and target classes represent $\approx 1\%$ of samples. Instead we used a normalize cross-correlation between our template signals and every other recording in the dataset (Briechle & Hanebeck, 2001).

$$NCC(k) = \frac{\sum_{n=0}^{N-1}(T_n - \bar{T})(X_{n+k} - \bar{X})}{\sqrt{\sum_{n=0}^{N-1}(T_n - \bar{T})^2 \cdot \sum_{n=0}^{N-1}(X_{n+k} - \bar{X})^2}}$$

T and X denote the template and recording spectrograms respectively, where $NCC(k)$ computes the normalized correlation coefficient of T in X lagged at frame k. 354 Recordings where randomly selected of which 123 with significant correlation coefficient peaks were chosen for annotation. Annotation was done using temporal-spectral bounding boxes with the Sonic Visualiser tool (Cannam et al., 2010) totalling 1980 initial annotations.

Label	N-annotations	Mean-Mid-Freq (hz)	Mean-duration (s)
Fast Trill	465	7153	1.111
nr Syllable	969	3050	0.520
Triangular Flourish	230	2754	0.707
Upsweep	317	404	0.835

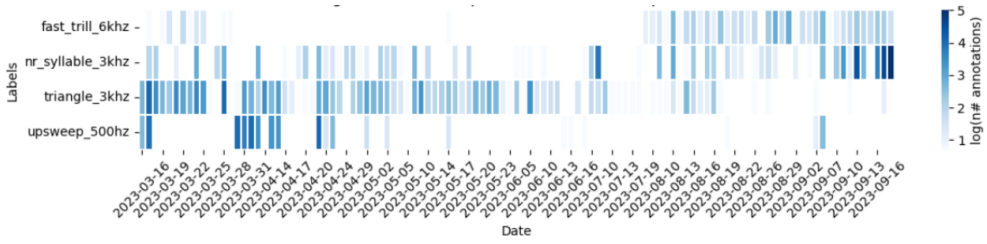


Figure 3: Logarithmic heat-map of annotations per recording date

3.3 Data Representation

Due to the extremely large size of our dataset, minimizing training and evaluation time is crucial. The pre-trained embedding architectures YAM-net (Drossos et al., 2020) and VGG-ish (Hershey et al., 2017) did not capture enough nuance to properly discriminate the classes due to their precise shapes. Instead, 40-band log-mel spectrograms were computed with a 2048 sample window length and 512 overlap. Spectrograms are then segmented into 2.5 second long chunks with a 0.5 second overlap with each chunk representing a single 'sample' or 'instance' of the dataset. Each instance is allocated a label based off whether a significant enough proportion of its corresponding annotation occurs within the chunk.

3.4 Model

The Resnet-16 architecture is adopted, due to its efficiency and ability to maintain contextual information through is residual layers. More robust architectures like Resnet-50 and CRNN-variants had too much of a computational burden for AL queries against the >250,000 spectrogram dataset, significantly impacting the annotation workflow by hours. We avoid the use of Bayesian CNN's and accompanying query methods as they do not scale well to large datasets due to batch sampling (Sener & Savarese, 2018).

The model is implemented with the Tensorflow library and compiled with focal-loss using the default focusing parameter $\gamma = 2$.

$$\mathcal{L}_{focal} = -\sum_{i=1}^{i=n}(1-p_c)^{\gamma} log_2(p_c)$$

Additionally, It was found that using an explicit noise class did in fact increase performance (Z. Zhao et al., 2017), thus the model is fitted with 5 soft-max classes.

3.5 Active Learning Methods

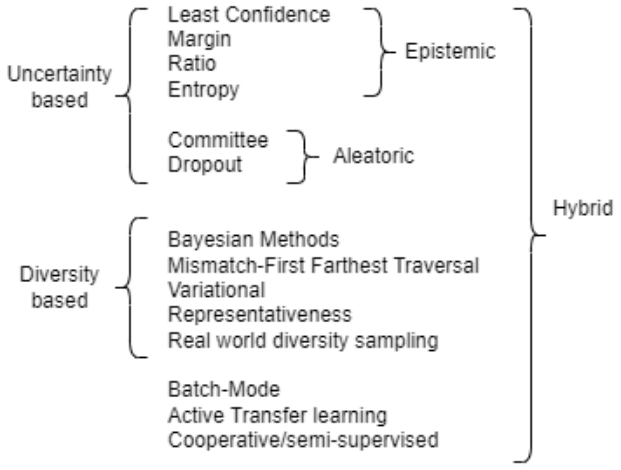


Figure 4: Summarized deep active learning methods

In the AL cycle, the model is first fitted to the currently labelled dataset and makes a prediction of the unlabeled data. A *query method*, or *acquisition function* ϕ accesses the unlabelled set and, with help from the predictions, optimizes a batch of unlabelled instances to be trained on. The primary goal of ϕ is to produce samples near the decision boundary (uncertain) far away from each other in the labelled set (diverse) and able to explain as many neighbouring instances in the unlabelled set (dense). A sample's degree of uncertainty may not be captured through the model if it's a statistical outlier (aleatoric), which is instead exposed through inter-model disagreement or dropout. An emipirical study by (Sener & Savarese, 2018) shows that uncertainty methods provide little value for deep-CNN's compare to diversity.

We find that the overwhelmingly large dataset makes all point-to-point distances metrics inefficient and often impossible, significantly restricting the choice of ϕ. In order to calculate diversity, a randomly chosen batch of samples is taken and diversities are computed among them via pairwise euclidean distances. Alternatives to this method are explored later in this section.

1.) *Least Confidence Sampling* (LC) is a model-based uncertainly method that determines the informativeness of a sample through the greatest soft-max output for labels $y \in Y$ per instance $x \in X$. Samples with the lowest, greatest prediction are queried:

$$\phi(X) = \arg\min_{x \in X} \left\{ \arg\max_{y \in Y} P(y|x) \right\} \tag{1}$$

2.) *Entropy Sampling*, or *max-entropy* is another model-based uncertainty method

that measures chaotic/indecisive predictions through the soft-max entropy. It is defined as follows:

$$\phi(X) = \arg\max_{x \in X} En(x) = \arg\max_{x} \left\{ -\sum_{y \in Y} P(y|x) log P(y|x) \right\} \tag{2}$$

LC and Entropy sampling brings forth informative samples but neglects their level of diversity, density and aleatoric uncertainty.

3.) *Information Density* (IDen) incorporates a similarity measure to maximise sampling representativeness within the unlabeled dataset X_U, balanced with uncertainty which here is computed with entropy:

$$Sim(x, \hat{x}) = \frac{1}{1 + D(x, \hat{x})} \tag{3}$$

$$\phi(X) = \arg\max_{x \in X} \left\{ \left[\frac{1}{X_U} \sum_{\hat{x} \in X_U} Sim(x, \hat{x}) \right] * En(x) \right\} \tag{4}$$

Where D computes the euclidean distance.

4.) *Information Diversity Sampling* (IDiv) maximises sampling diversity within the labeled dataset X_L balanced against the same uncertainly measure in (4):

$$\phi(X) = \arg\max_{x \in X} \left\{ \left[1 - \frac{1}{X_L} \sum_{\hat{x} \in X_L} Sim(x, \hat{x}) \right] * En(x) \right\} \tag{5}$$

5.) *Adaptive Information Diversity sampling.* Due to the imprecise diversity measures of the above method, a variation of IDiv is proposed incorporating a buffered selection of uncertain samples S. The selected S is refined based on samples that are internally diverse to each other, guaranteeing uncertainty and diversity within the batch itself.

$$S = \{s_1, s_2 ..\} = \arg\max_{x} En(x) \tag{6}$$

$$\phi(X) = \arg\max_{x \in X} \left\{ 1 - \frac{1}{S} \sum_{\hat{x} \in S} Sim(x, \hat{x}) \right\} \tag{7}$$

6.) *Embedding Information Diversity sampling*, inspired by embedding gradient sampling, implements the above method on the 64 embedding layer output from the selected set rather than the samples itself. This is designed to properly discriminate between the subtle feature-variation between our targets signals and focus on equally sampling classes and class variations.

7.) *Uncertain Core-Set Sampling.* Sener and Savarese (2018) defined a formal upper-bound for AL-loss (Core-set loss) in which the problem of minimizing is equivalent to the NP-Hard k-centers problem. A greedy-k-centers algorithm is adopted to

perform core-set sampling, applied to an uncertain large batch of X_L and a random large batch from X.

For the above methods, we use ranked-batch mode active learning (Cardoso et al., 2017), where arbitrate query sizes are drawn from the unlabelled set and ranked, allowing this batch to be easily refined with new queries that can be ran simultaneously with annotation. In order to compare methods, the model is fitted with an initial labelling budget of 3%, and after each query, the model is evaluated, reset and retrained on the newly expanded dataset until over-fitting is detected through validation loss. All experiments were done with arbitrary query sizes on an Amazon Web Services c4.2xlarge EC2 instance.

In order to compare methods, mean-Average-Precision (mAP) is used due to the imbalanced and rare classes. Defined as the mean area under the precision-recall curve or average precision (AP) across all classes. AP is interpolated from all precision and recall metrics from an arbitrary amount of classification thresholds, here 200 thresholds is used. Precision P, the positive prediction accuracy, and recall R, the true label accuracy, are defined in terms the True Positives TP, False Negatives FN and False Positives FP for each class.

$$P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN} \tag{8}$$

$$AP = \sum_{n=1}^{200} (R_n - R_{n-1}) P_n \tag{9}$$

$$mAP = \frac{1}{n} \sum_{k=1}^{k=n} AP_k \tag{10}$$

4 Results

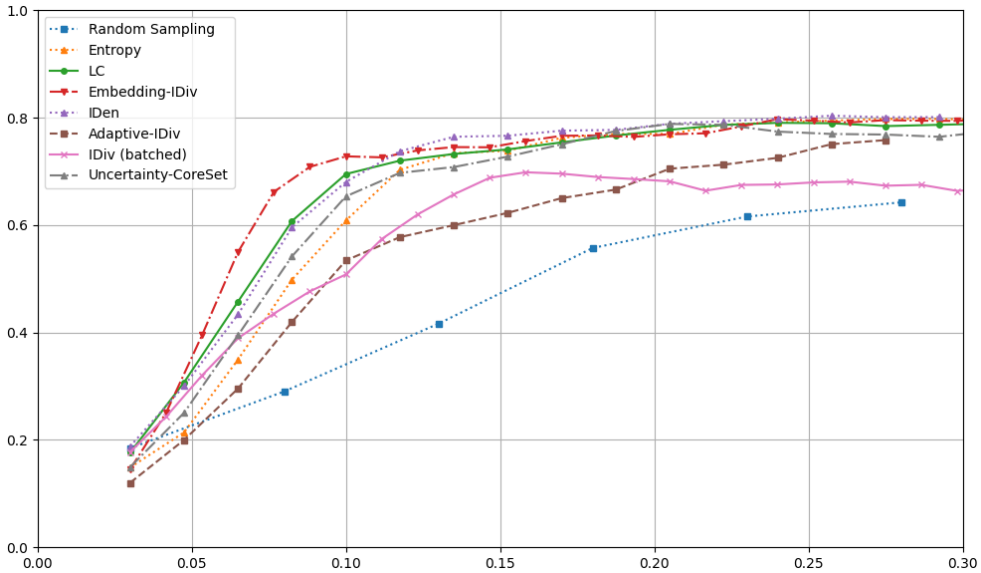
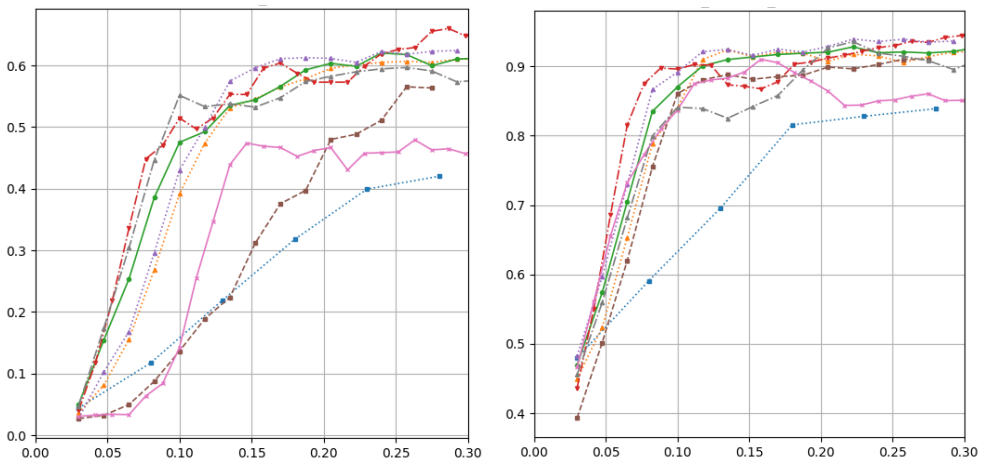


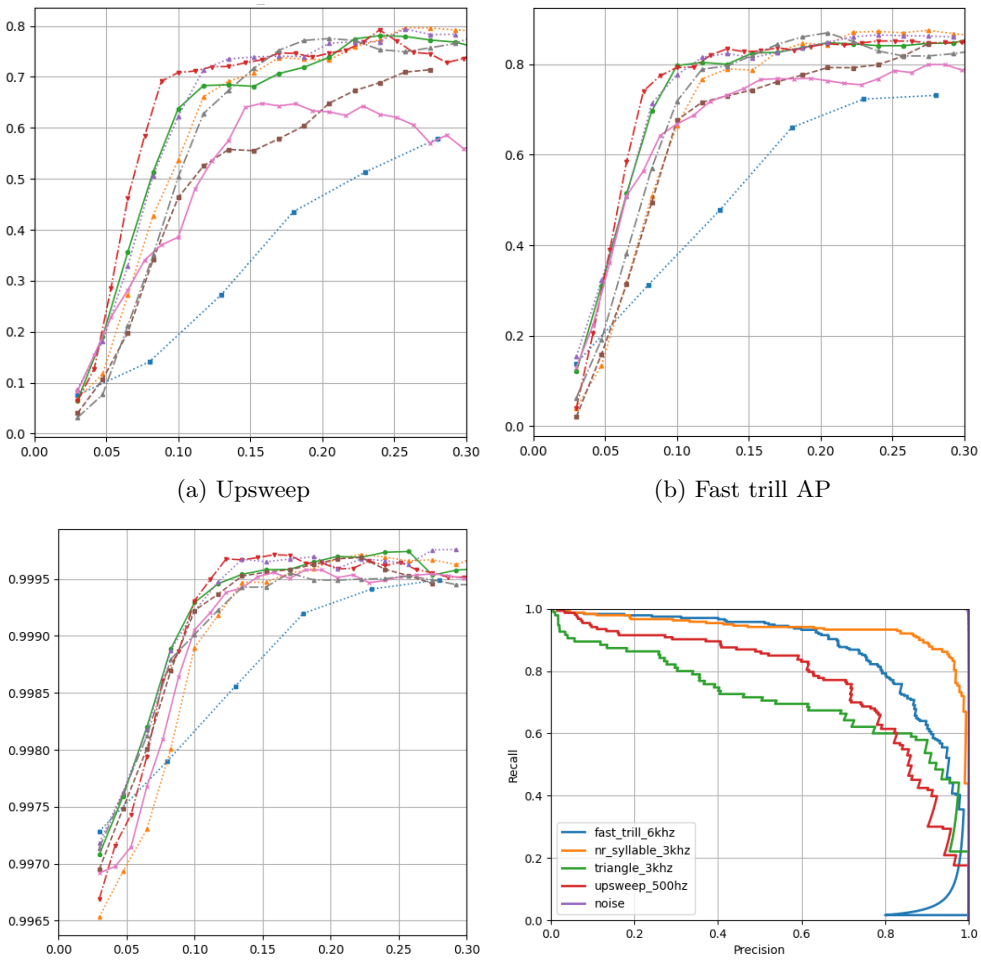
Figure 5: mAP vs. labelling budget for all query methods up to 30% labelled data.



(a) Triangular flourish

(b) *nr* Syllable

Figure 6: Classwise AP vs labelling budget for the 8 query methods (Part 1)



(a) Upsweep

(b) Fast trill AP

(c) Noise

(d) PR-curve for 100% labelling budget

Figure 7: Remainder classwise AP vs labelling budget for the 8 query methods and the 200 thresholds precision-recall curve for the model trained on the full dataset.

Experiments for each method took 3 hours and 25 minutes on average with the most costly queries (IDen, IDiv) taking 157.89 ms per sample, unlikely to slow down annotations for the $\approx 76,000$ chunks that made up the 30% labelling budget.

All AL methods outperformed random sampling significantly, and as expected, the ambiguous 'Triangular Flourish' class in Figure 6a acted as the main discriminator between methods, specifically the traditional diversity methods; IDiv and adaptive-IDiv. The 'Upsweep' class also had this effect in Figure 7a, likely due to its low resolution and low energy variations. The Embedding-IDiv method produced the best results with a 5-10% labelling budget and was the quickest to adapt to these difficult classes. Embedding-IDiv had the highest mAP (0.81) at the last iteration,

whereas the model trained on all the data had an mAP of 0.84. In the 0.15% to 30% range, all uncertainty methods performed roughly the same, possible indicating exhaustion of diverse samples within the uncertainty buffer.

5 Discussion

The perfomance of the uncertainty-based methods contradicts expectations for CNN's against large datasets (Sener & Savarese, 2018). This is inherently due to the the fine feature discrimination required to sample out particular shapes rather than summarize clear and distinct sound motifs. Additionally, the extremely rare class occurrences required AL methods to automatically over-sample and distinguish these classes from the heavily diverse noise class, rendering traditional diversity techniques less effective. The subjective class 'Triangular Flourish' and it's many variations required far more sampling than its less-ambiguous peers. Raw diversity-based methods, IDiv and Adaptive-IDiv, failed to recognize this class as informative resulting in their overall failure. Inferring diversity via the embedding space allowed the model to correctly adjust, performing best overall achieving 0.96% the mAP of a model trained on the full dataset with a 30% labelled budget, and that of 0.88% with just a 10% labelling budget. This gives promising results for variational-AL and variational-adversarial-AL methods for this case due to their similar embedding-like inferencing.

Aleatoric uncertainty methods, such as drop and committee-based AL were avoided due to the lack of classifiers that would efficiently produce meaningful comparisons without using a second deep CNN. As shown with the traditional diversity methods here, data-based outliers would be less valuable for such low class occurrences in an already diverse soundscape.

After considering the most robust case for bird-call recognition in PAM, it is found that traditional diversity-based techniques do not suffice but can provide value at the embedding level. Pre-trained acoustic embedding frames may not properly capture these rare and discrete bird-call signatures. Light-weight sliding-window CNN models, such as ResNet-16, can however appropriately recognize these shapes efficiently on log-mel spectrograms. The combination of lightweight classification, training and queries with ranked batch-mode AL allows for large annotation workflows to go uninterrupted achieving maximal performance with minimal human resources.

These findings have informed how modern AL methods are applied to datasets of this scale against discrete acoustic signatures. As it stands, further research needs to done on AL methods optimized for larger, complex datasets where existing options like MAL and MFFS become obsolete. Additionally, more diverse baseline datasets would promote the curation of standardized methods in sound event detection, similar to those found in computer vision and natural language processing.

Professional and Technical Development

The above sections form the project write-up that will likely not be published due to incomplete methods and a lack significant additions to the field (in my personal opinion). Due to this fact, it does not follow any specific journal formatting style (default latex article). I've added this section to finish off the requirements under the standard report marking rubric for non-article reports just in case.

Initially the project was designed to incorporate AL-backed annotation software utilizing the most efficient methods found here, this however was not completed due to excessive pivots, issues and restarting on my part which pushed the schedule back immensely. Due to these push backs, the vast majority of the projects work did not culminate here as all the aforementioned experiments and methods were desperately formed in the last month. This experience has informed my career prospects greatly, and I have hard-learnt many valuable lessons in research, software and acoustics.

For technical development, I have invested significant time into learning formal python data science project workflows, large data pipeline curation, efficient ML on said pipelines, acoustic signal processing, cloud computing and more. Specifically: numpy, pandas, Tensorflow, librosa, soundfile, modAL, scikit-learn, scikit-maad, scipy.signal, matplotlib, tkinter, environment handling, AWS and bash scripting. I have consulted supervisors extensively about producing code that fits their expectations and needs. Whilst no software was made, the package code is designed for re-purposing in existing annotation workflows within the department. As such, the modality, documentation and format standards for the code base where heavily communicated in order to ensure that the means of the project target real-world bottlenecks and can be adapted to improve existing methods for acousticians. For example, visual heuristics and design choices for low-effort annotation applications, code re-usability and verify-ability with multiple datasets, ensuring the code is decoupled and able to be repurposed for future students and researchers if they wish to expand. This allowed me to write and test software integrated in a client feedback loop, a crucial skill in all software development prospects.

Multidisciplinary skills were mostly self-taught through various internet resources such as papers, articles, documentation, blogs, videos and lectures. Among these, there were 2 book readings:

- '*Human-in-the-loop machine learning: a state of the art* (Mosqueira-Rey et al., 2022)'. Recommended by my supervisor, this was essentially a dictionary for AL methods and their many variations. Throughout the project I used this book as a guide for all AL related tasks.
- '*Probabilistic Machine Learning: An Introduction (P. Murphy, 2022)*'. Despite the name, is a very in-depth look at both probabilistic models and discrete models from a probabilistic perspective. I only ending up reading 30% of this book,

but was enough for me to grasp Bayesian models and their applications. This was my initial research focus but the approach was scrapped due to unforeseen constraints.

The research performed here is easily expandable for variational AL, large data optimized AL methods, a software implementation and automated systems for intelligent annotations of terrestrial life in the general case.

(GitHub)
(Summary Table)

References

Allen, A. N., Harvey, M., Harrell, L., Jansen, A., Merkens, K. P., Wall, C. C., ... Oleson, E. M. (2021, March). A Convolutional Neural Network for Automated Detection of Humpback Whale Song in a Diverse, Long-Term Passive Acoustic Dataset. *Frontiers in Marine Science*, *8*. doi: 10.3389/fmars.2021.607321

Bateman, J., & Uzal, A. (2022, September). The relationship between the Acoustic Complexity Index and avian species richness and diversity: A review. *Bioacoustics*, *31*(5), 614–627. doi: 10.1080/09524622.2021.2010598

Boiński, T. M., Szymański, J., & Krauzewicz, A. (2022, January). Active Learning Based on Crowdsourced Data. *Applied Sciences*, *12*(1), 409. doi: 10.3390/app12010409

Briechle, K., & Hanebeck, U. D. (2001). Template matching using fast normalized cross correlation. In D. P. Casasent & T.-H. Chao (Eds.), *Optical pattern recognition xii* (Vol. 4387, pp. 95–102). SPIE / International Society for Optics and Photonics. Retrieved from `https://doi.org/10.1117/12.421129` doi: 10.1117/12.421129

Cannam, C., Landone, C., & Sandler, M. (2010, October). Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the ACM multimedia 2010 international conference* (pp. 1467–1468). Firenze, Italy.

Cardoso, T. N. C., Silva, R. M., Canuto, S., Moro, M. M., & Gonçalves, M. A. (2017, February). Ranked batch-mode active learning. *Information Sciences*, *379*, 313–337. doi: 10.1016/j.ins.2016.10.037

Drossos, K., Mimilakis, S. I., Gharib, S., Li, Y., & Virtanen, T. (2020, February). *Sound Event Detection with Depthwise Separable and Dilated Convolutions* (No. arXiv:2002.00476). arXiv.

EmreÇakır, Adavanne, S., Parascandolo, G., Drossos, K., & Virtanen, T. (2017, March). *Convolutional Recurrent Neural Networks for Bird Audio Detection* (No. arXiv:1703.02317). arXiv.

Fischer, J., Orescanin, M., Leary, P., & Smith, K. B. (2023, July). Active Bayesian Deep Learning With Vector Sensor for Passive Sonar Sensing of the Ocean. *IEEE Journal of Oceanic Engineering*, *48*(3), 837–852. doi: 10.1109/JOE.2023.3252624

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., ... Wilson, K. (2017, January). *CNN Architectures for Large-Scale Audio Classification* (No. arXiv:1609.09430). arXiv. doi: 10.48550/arXiv.1609.09430

Hilasaca, L. H., Ribeiro, M. C., & Minghim, R. (2021, July). Visual Active Learning for Labeling: A Case for Soundscape Ecology Data. *Information*, *12*(7), 265. doi: 10.3390/info12070265

Kholghi, M., Phillips, Y., Towsey, M., Sitbon, L., & Roe, P. (2018). Active learning for classifying long-duration audio recordings of the environment. *Methods in Ecology and Evolution*, *9*(9), 1948–1958. doi: 10.1111/2041-210X.13042

Luo, L., Zhang, L., Wang, M., Liu, Z., Liu, X., He, R., & Jin, Y. (2021). A system for the detection of polyphonic sound on a university campus based on CapsNet-

RNN. *IEEE access : practical innovations, open solutions*, *9*, 147900–147913. doi: 10.1109/ACCESS.2021.3123970

Mesaros, A., Heittola, T., Virtanen, T., & Plumbley, M. D. (2021, September). Sound Event Detection: A Tutorial. *IEEE Signal Processing Magazine*, *38*(5), 67–83. doi: 10.1109/MSP.2021.3090678

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2022, August). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, *56*. doi: 10.1007/s10462-022-10246-w

Parcerisas, C., Schall, E., te Velde, K., Botteldooren, D., Devos, P., & Debusschere, E. (2024, April). Machine learning for efficient segregation and labeling of potential biological sounds in long-term underwater recordings. *Frontiers in Remote Sensing*, *5*. doi: 10.3389/frsen.2024.1390687

Phillips, Y. F., Towsey, M., & Roe, P. (2018, March). Revealing the ecological content of long-duration audio-recordings of the environment through clustering and visualisation. *PLoS ONE*, *13*(3), e0193345. doi: 10.1371/journal.pone.0193345

P. Murphy, K. (2022). *Probabilistic Machine Learning: An introduction*. MIT Press.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., ... Wang, X. (2021, December). *A Survey of Deep Active Learning* (No. arXiv:2009.00236). arXiv.

Ryazanov, I., Nylund, A. T., Basu, D., Hassellöv, I.-M., & Schliep, A. (2021, February). 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. doi: 10.3390/jmse9020169

Sener, O., & Savarese, S. (2018, June). *Active Learning for Convolutional Neural Networks: A Core-Set Approach* (No. arXiv:1708.00489). arXiv.

Shishkin, S., Hollosi, D., Doclo, S., & Goetze, S. (2021). active learning for sound event classification using monte-carlo dropout and pann embeddings.

Shishkin, S., Hollosi, D., Goetze, S., & Doclo, S. (2024, April). Active Learning for Sound Event Classification Using Bayesian Neural Networks with Gaussian Variational Posterior. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 896–900). Seoul, Korea, Republic of: IEEE. doi: 10.1109/ICASSP48485.2024.10446970

Shuyang, Z., Heittola, T., & Virtanen, T. (2017, March). Active learning for sound event classification by clustering unlabeled data. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 751–755). New Orleans, LA: IEEE. doi: 10.1109/ICASSP.2017.7952256

Shuyang, Z., Heittola, T., & Virtanen, T. (2018, September). An Active Learning Method Using Clustering and Committee-Based Sample Selection for Sound Event Classification. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)* (pp. 116–120). doi: 10.1109/IWAENC.2018.8521336

Venkatesh, S., Moffat, D., & Miranda, E. R. (2022, March). You Only Hear Once: A YOLO-like Algorithm for Audio Segmentation and Sound Event Detection. *Applied Sciences*, *12*(7), 3293. doi: 10.3390/app12073293

Zhao, S., Heittola, T., & Virtanen, T. (2020, September). *Active Learning for Sound Event Detection* (No. arXiv:2002.05033). arXiv.

Zhao, Z., Zhang, S.-h., Xu, Z.-y., Bellisario, K., Dai, N.-h., Omrani, H., & Pijanowski, B. C. (2017, May). Automated bird acoustic event detection and robust species classification. *Ecological Informatics*, *39*, 99–108. doi: 10.1016/j.ecoinf.2017.04.003