

Positive-Unlabelled Active Learning to Curate a Dataset for Orca Resident Interpretation

Bret Nestor^{*†}

bret.nestor@ubc.ca

Translancean Research Foundation
Vancouver, Canada
University of British Columbia
Vancouver, Canada

Jasmine Moore

University of Calgary
Calgary, Canada

Bohan Yao

University of Washington
Seattle, United States

Jasper Kanes

Ocean Networks Canada
Victoria, Canada

Abstract

This work presents the largest curation of Southern Resident Killer Whale (SRKW) acoustic data to date, also containing other marine mammals in their environment. We systematically search all available public archival hydrophone data within the SRKW habitat (over 30 years of audio data). The search consists of a weakly-supervised, positive-unlabelled, active learning strategy to identify all instances of marine mammals. The resulting transformer-based presence or absence classifiers outperform state-of-the-art classifiers on 3 of 4 expert-annotated datasets in terms of accuracy and energy efficiency. The fleet of WHISPER detection models range from 0.58 (0.48-0.67) AUROC with WHISPER-tiny to 0.77 (0.63-0.93) with WHISPER-large-v3. Our multiclass species classifier obtains a top-1 accuracy of 53.2% (11 train classes, 4 test classes) and our ecotype classifier obtains a top-1 accuracy of 33.6% (4 train classes, 5 test classes) on the DCLDE-2026 dataset.

We yield 919 hours of SRKW data, 230 hours of Bigg's orca data, 1374 hours of orca data from unlabelled ecotypes, 1501 hours of humpback data, 88 hours of sea lion data, 246 hours of pacific white-sided dolphin data, and over 784 hours of unspecified marine mammal data. This SRKW dataset is larger than DCLDE-2026, Ocean Networks Canada, and OrcaSound combined. The curated species labels are available under CC-BY 4.0 license, and the corresponding audio data are available under the licenses of the original owners. The comprehensive nature of this dataset makes it suitable for unsupervised machine translation, habitat usage surveys, and conservation endeavours for this critically endangered ecotype.

^{*}Work completed while at the University of Toronto and the University of Washington

[†]Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Under review for KDD '26, Jeju, Korea

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
<https://doi.org/XXXXXXX.XXXXXXX>

CCS Concepts

• **Computing methodologies** → *Neural networks*; **Active learning settings**; **Semi-supervised learning settings**; • **Applied computing** → *Environmental sciences*.

Keywords

Marine Biology, Active Learning, Positive-Unlabelled Learning, Killer Whale, Orca, Bioacoustics

ACM Reference Format:

Bret Nestor, Bohan Yao, Jasmine Moore, and Jasper Kanes. 2026. Positive-Unlabelled Active Learning to Curate a Dataset for Orca Resident Interpretation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Under review for KDD '26)*. ACM, New York, NY, USA, 18 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

The southern resident killer whales (SRKWs) are an ecotype of *Orcinus orca* who inhabit the waters of the northeastern pacific. They are designated as endangered by both the Department of Fisheries and Oceans Canada [44] and the U.S. Fish and Wildlife Service [43]. The International Union for the Conservation of Nature (IUCN) lists orcas' endangered status as "data deficient" because they are comprised of diverse ecotypes [25], though recent work has proposed that Northeast Pacific ecotypes should be classified as distinct species [37]. With this distinction, SRKWs who diverged from other lineages approximately 100,000 years ago, could also receive threatened status under the IUCN [37].

The critical decrease in SRKW population was caused by extensive capture for entertainment, compounded by a failure to rebound due to decreased prey availability owing to dam construction, increased contaminants, and increased noise pollution [30]. They face direct threats from human activity, such as ship strikes [54] and ship noise [63], which reduces their ability to echo-locate prey [61]. They also face threats indirectly caused by human activity, such as a lack of large Chinook salmon prey availability [20, 45]

In the Northeastern Pacific the SRKWs diet constitutes fish, whereas Bigg's orcas are mammal-eating [9]. Due to their choice of prey, resident orcas are more vocal than the mammal-eating Bigg's ecotype [15]. The SRKW socialisation and echolocation make them easier to detect acoustically, as compared to other cetaceans. As

their habitat coincides with the heavily monitored Salish Sea and Northeast Pacific Ocean, we hypothesise that there are ample instances of SRKWs within unsearched archival datasets.

The goal of this work is to collect all archival instances of SRKWs to build a Dataset for Orca Resident Interpretation (DORI). Accomplishing this requires 1) using effective classifiers to discover instances of SRKW data from unlabelled archival data, 2) developing effective species and ecotype labelling techniques, and 3) labelling data.

We develop competitive marine mammal detection classifiers, and species classifiers, resulting in the largest curation of marine mammal acoustic data to date ¹. From a set of more than 30 years (260,000 hours) of archival data, we discover 919 hours of SRKW data.

1.1 Passive Acoustic Monitoring

Passive acoustic monitoring (PAM) involves setting up a recording device in a region of interest, and permitting it to record indiscriminately without human oversight [36]. Data is later analysed to determine the presence of sound-producing species. PAM datasets are unique because they are recorded in-the-wild without confounding research vessel interference [14, 65]. One of the unique challenges of passive acoustic data is that it generates a tremendous amount of data, yet only a small fraction of it involves the species of interest. This results in datasets with extreme class imbalance. Historically, PAM was reviewed by human annotators [62], but algorithms have been developed to preprocess and filter acoustic data at scale, and to automatically detect species presence within these large-scale datasets [19, 27]. Audio is often recorded with a duty-cycle to reduce labelling effort or storage requirements in remote locations. Though duty cycling could potentially cause missed identifications [55] or inaccurate durations [52].

Human-annotation would be infeasible the scale required to search this archive [53]. The most comprehensive approach to date has been OrcaLab, a digitalization of 20,000 hours of acoustic data from OrcaLab, in northern resident killer whale habitat [62]. SRKW acoustic surveys been fragmented. From 2006-2011, several hydrophone deployments along the United States coast detected 131 SRKW events [11]. From 2009-2011, 175 days of monitoring contained SRKWs at Swiftsure Bank [56]. From 2015-2017, 96 days of SRKW events were detected in the winter season in the northern Salish Sea [50]. In 2018, 46 days with SRKW events were detected off of Lime Kiln [52]. Between 2018-2020, 605 days contained SRKW presence at Swiftsure Bank [56]. These studies included duty cycles and intermittent deployment. Cumulatively, SRKWs were detected on 1053 days. These acoustic data are not available publicly. However, during these same periods, there are publicly available (CC-BY-NC-SA 4.0) recordings from OrcaSound, amounting to 122 whale days from 2018-2022.

1.2 Active Learning and Positive-Unlabelled Learning

Active learning is a process that reduces labelling effort by prioritising samples that are informative. Typically this is done by sorting

samples according to their uncertainty, diversity, or model influence [34]. A key aspect of active learning is iterating between model retraining, and labelling unlabelled samples that would maximally benefit the model. When high entropy samples are selected for active learning, the resulting model outperforms a model trained on the same quantity of randomly labelled data [26]. In the binary classification case this corresponds to a logit probability closest to 0.5 [33]. There is a natural tension between exploring within-class diversity and samples far from class distributions. Modern algorithms seek to strike a balance between these, especially when classes are imbalanced [1], heterogeneous [26], or unknown (known as open-set annotation) [40]. Other works seek to be task-agnostic by predicting the model's loss from the covariates in order to sample under-represented and highly influential samples [66]. In contrast to PAM approaches, which duty-cycle the unlabelled set to have an unbiased measure of the data, active learning takes a biased selection of the data to uncover the maximum model performance per labelling effort.

Under class-conditional mislabelling, simple models are noise tolerant [38]. This means that when candidates are mislabelled, the classifier trained on these mislabelled samples can still approximate the performance of a classifier trained on correctly labelled samples. In the most extreme case of positive-unlabelled learning (where noise is only applied to randomly flip some of the positive labels to negatives/unlabelled) the loss function for a weighted regression problem serves as a surrogate for the true underlying loss [29]. While it relies on knowing the class proportions, this is only a single hyperparameter that is tuned in cross validation [38]. The convergence rate of PU learning has been derived to be:

$$\begin{cases} O\left(\frac{V}{n \cdot e_m \cdot h}\right) & h \geq \sqrt{\frac{V}{n \cdot e_m}} \\ O\left(\sqrt{V/(n \cdot e_m)}\right) & h < \sqrt{\frac{V}{n \cdot e_m}} \end{cases}$$

Where V is the Vapnik-Chervonenkis dimension, n is the number of samples, e_m is the propensity of a positive sample to be labelled under the selected completely-at-random (SCAR) assumption, and h is Massart margin that quantifies the difficulty of classifying the data [12]. A lower Massart margin, h , indicates a harder to classify example. Fortunately, in the active learning setting e_m approaches 1 as labelling progresses. This drives the bound towards the standard classification bound. For a difficult to classify problem, or for a large pretrained model, having unlabelled positive samples in your dataset incurs virtually no penalty as compared to a standard training setup when the negative instances vastly outnumber the positive instances.

These bounds show that models with sufficient data can lower their empirical risk on *I.I.D.* test data irrespective of architecture choices. We use transformer-based models that are universal function approximators for sequential data problems [67].

Knowing that performant models can be uncovered using noisy labels is fortunate for passive acoustic monitoring, since we can rely on non-experts. The challenge for machine learning models in this problem setting is that the SCAR assumption used to derive the bounds may be broken by selectively identifying high SNR samples. Systematic archive searches require amateur (noisy) labellers, since the quantity of available archival data grows faster than experts can process them. In a related work investigating the ability of

¹Data available at <https://huggingface.co/collections/DORI-SRKW/dori>

citizen scientists to annotate deep-sea images, the citizen scientists positively identified 50–65% of expert annotated positive images, while maintaining a true positive rate of 36–46% [32]. This enriched sensitivity reduces expert burden so long as it is above the ambient positivity rate of the data. While these capabilities might not transfer from imaging to audio annotation, we hypothesize that amateur listeners can similarly identify high-SNR instances leading to ample labelled marine mammal acoustic events.

We hypothesize that by combining positive-unlabelled learning, amateur labelling, and active-learning, we can recover classifiers that often outperform existing state-of-the-art classifiers. The resulting models should be small enough and efficient enough to search the ever growing archives of SRKW habitat.

2 Methods

A schematic overview of data sources are given in Figure 1.

2.1 Data Sources

We curate data within the region bounded by 40N to 51N latitude, and -132E to -120E longitude, through the end of 2022. We only explore publicly available hydrophone data under permissive data use licenses or agreements.

Ocean Networks Canada. The Ocean Networks Canada archival repository was the largest source investigated, spanning 60057 days of hydrophone deployments. Due to storage considerations, we only retained 3143 days of high likelihood hydrophone data for the active learning simulation. These data are distributed under a CC-BY license.

SanctSound. SanctSound involved temporary hydrophone deployments along the outer Washington coast between 2018 and 2022 [21]. We analysed 1788 days of data. Due to storage considerations, we only retained 455 days of the highest probability samples. Machine learning detections of Orcas were provided with annotated start and end time of acoustic bouts. Ecotype determinations were not included. Humpbacks were labelled by day-level presence, rather than in bouts.

Ocean Observatories Initiative. Ocean Observatories initiative operates hydrophone nodes near the continental shelf and Yaquina Bay [24]. These data have not been thoroughly explored for marine mammal events. The archives contain 10370 days of hydrophone deployments within the specified range, usable under a custom terms².

OrcaSound. OrcaSound is a community collective of self hosted hydrophone nodes and volunteer annotators [46]. Their data contain human and machine detections through various deployments in the southern Salish Sea. We include 58 days of data from 4 locations from 2018–2022. These data were selected because they coincided with nearby sightings, so they represent an enriched prevalence.

DCLDE 2026 dataset. This dataset is curated specifically for the 2026 Biennial Conference and Workshop on Detection, Classification, Localization, and Density Estimation of Marine Mammals

using Passive Acoustics (DCLDE) [47]. There are 35 hours of marine mammal audio, including 12.2 hours of killer whale data, and 4.1 hours of SRKW data. It contains several orca ecotypes for ecotype classification problems. Some of the contributed data overlaps with the data from the previously listed sources. For this reason, we intentionally omit OrcaSound and Ocean Networks Canada data when this dataset is used as an external test set.

DEEPAL Fieldwork. This dataset is collected using a towed hydrophone array, with the intention of classifying audio clips as containing, or not containing orcas [7, 8]. The dataset requires a data use agreement to access.³

FGPD & PSGCH 20203 datasets. These datasets are entirely sourced from Ocean Networks Canada, and retain the CC-BY 4.0 license for the data. The FGPD set features 365 5-minute segments of data from a hydrophone node at hahuuli [41] (lat 48.8138, lon -125.2746, depth 95m), and PSGCH contains 31 5-minute segments from August 2023 in the Salish Sea [42] (lat 49.0395, lon -123.4252, depth 298). The segments were collected by taking the maximum scored sample for a given day, using logits drawn from several models. This means that the segment-level prevalence of marine mammals is over-representative, but the day-level prevalence of marine mammals is representative. The labels are 1, if a marine mammal is present, and a 0 otherwise (including snapping shrimp, fish grunts). The marine mammals in FGPD include humpbacks, Bigg’s transients, NRKW, SRKW, and sea lions. PSGCH contains humpbacks and Bigg’s transients. The introduced test sets are chronologically disjoint from our training set, and are labelled by an expert bioacoustician.

Other hydrophone operators are also in the region, such as the British Columbia Hydrophone Network, Raincoast Conservation Foundation, and the Lime Kiln State Park Lighthouse. These sites were not included because they did not have comprehensive coverage, permissive data, or archival data in the study period.

Sightings data. We obtained records from the Pacific Whale Watch Association (under a limited data use agreement), Orca Networks’ archival data [39], iNaturalist [17, 23], and prior literature [58] to build a thorough profile of sightings. In total, there were 92745 unique sightings, of which 855 were reported within 1 km of a hydrophone.

2.2 Preprocessing

Audio files were converted from their original format to a lossless flac encoding for storage purposes using ffmpeg (<https://ffmpeg.org/>). If the audio files were longer than 5 minutes, they were also separated into 5-minute long segments at this step. Each segment which was 5-minutes long or less was resampled to 32 kHz (to maintain a Nyquist frequency of 16 kHz) and processed through a high-pass filter at 1 kHz. The default whisper-tiny parameters were used for mel spectrogram preprocessing (a *chunk_length* of 30, a *feature_size* of 80, and a *hop_length* of 160). 15-second long segments of each file were fed through the 39 million parameter

²<https://oceanobservatories.org/how-to-use-acknowledge-and-cite-data/>

³access instructions at <https://lme.tf.fau.de/dataset/deepal-fieldwork-data-2017-2018-dlfd/>

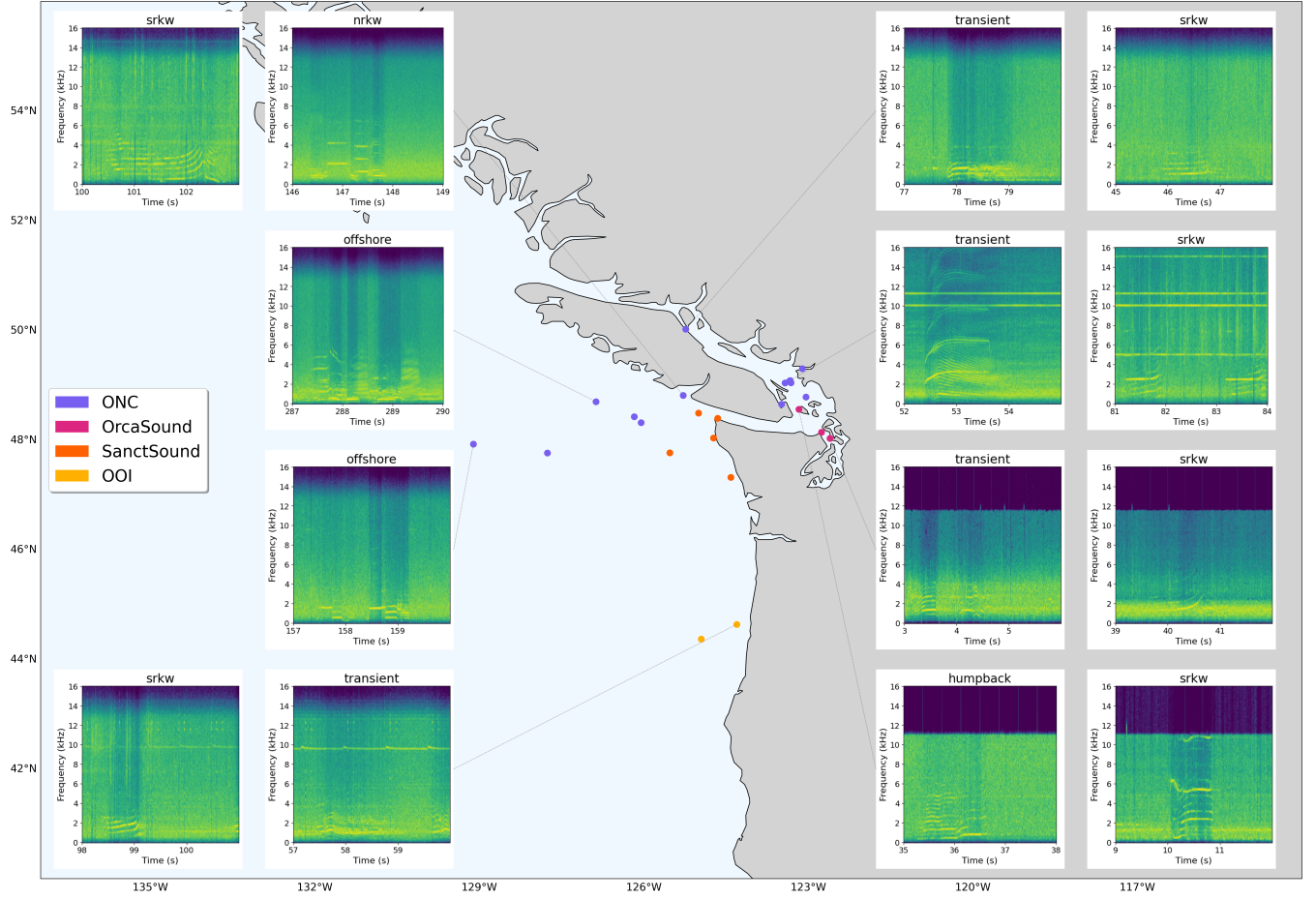


Figure 1: Hydrophone deployment sites where data are collected from.

whisper-tiny encoder [51] to obtain a feature set with the dimensions of *audio_chunks*, *hidden_steps*, *hidden_size*. Layer normalization was applied to the outputs at this stage [3]. A mean-pooling was applied across the *hidden_steps* dimension. All dimensions were averaged until we obtained a single vector per file with the dimension: *hidden_size*. We refer to these per-file vector representations as embeddings throughout the rest of the paper.

The median embedding required 0.000045 kWh of electricity to process [13]. The consumption of all embeddings was 153.37 kWh, which was conducted in regions with >90% of electricity sourced from nuclear, hydroelectric, wind, or solar. Less than 10 kg of CO₂ was produced while generating the embeddings, which was offset through the Great Bear Forest Carbon Project (<https://www.carbonzero.ca>).

2.3 Weakly Supervised Learning

The embeddings were given to a standard logistic regression classifier implemented in scikit-learn with an L-2 penalty [49]. We also train a classifier using SGD and a hinge-loss. The purpose of this hinge loss is to evaluate a model training scheme that removes the bias caused by deleting unlikely samples. Positive labels included

all "seed" labels, which were ONC labelled data, and all human labelled positive instances at the time of training. Since the positive instances were so rare, samples were randomly selected from the unlabelled data to serve as negative instances during training. The class priors were not adjusted to accommodate the likelihood of the negative instances of being positive. We leave this to future work, as we can adjust the priors based on location, bathymetry, time-of-year, and time-of-day, to inform the priors for each sample. In some active learning strategies definitive negatives (those already labelled by a human) supplemented the negative class. Data were split by time, with audio files recorded before 2021 being relegated to training, those recorded in 2021 used as validation, and those in 2022 as a held-out test set.

2.4 Active Learning

We followed several different active learning strategies. This included **Positive only** sampling where at each step only the top *k* samples were labelled. Files that were not labelled positive remained in the unlabelled group in case a low SNR sample was missed. Next, we applied a standard **entropy-based** labelling strategy [26]. The top-*k* samples sorted in descending order by entropy were labelled.

Samples which were labelled as negatives were kept as guaranteed negatives for future model training iterations. We employ a **loss-estimate** strategy as described in prior work [66]. The labelling strategy was similar to that of the entropy strategy, except that samples were sorted by descending order of predicted loss. Two hybrid approaches were conducted, the first being a mixture of the positive and entropy sampling approaches. At each iteration, half of the labelling budget was applied to the positive labelling strategy and half to the entropy labelling strategy. We refer to this approach as **mixed** active learning. The second hybrid approach, **alternating**, would alternate the labelling strategy between the positive labelling and entropy labelling strategies. For these experiments, 500 5-minute long samples were labelled at each labelling step.

We simulate the learning process with 5 randomly seeded models. We demonstrate the experiments using the labels produced in this work, and in a second setting where 30% of the positive labels are randomly flipped to negative for each new batch of 500 labels.

2.5 Labelling

Data were labelled using both the available spectrogram and audio. The labelling strategy used to curate the dataset most closely resembled the "alternating" approach. In addition to purely algorithmic searches, heuristics were applied to try to capture more samples; Every file before and after a marine mammal detection was manually searched, regardless of its likelihood. Hydrophone array deployments were ensured to have the same label for each device to reduce the number of false negatives. Due to data storage capacity, low-probability samples from ONC were periodically deleted and removed from disk throughout the labelling protocol. For this reason, our embeddings are slightly over-representative of the SRKW prevalence.

Species labels were seeded by using previous ONC annotations, and Discovery of Sound in the Sea examples [64]. Ecotype determinations were made by labelling calls according to those recorded in prior literature [16, 59]. When available, sightings data were used to clarify species and ecotype labels.

2.6 Benchmark Models

As a benchmark, we re-implement the PAMguard Whistle&Moan detector and **LDA classifier** [18, 19]. We also adapt the **ROCCA random forest classifier** [5] as a detection model. Audio data are loaded and resampled to 32 kHz. Echolocation clicks are removed using a sliding exponential filter:

$$signal_{low} = \frac{1}{\left(1 + \frac{signal - \text{mean}(signal)}{\gamma * \text{std}(signal)}\right)^p}$$

where p and γ are hyperparameters. A spectrogram is created from the transformed audio data and converted into decibels. Spectral noise is removed by subtracting a sliding median filter over the spectrogram, introducing a kernel size hyperparameter, κ . The decaying running-average is subtracted to remove tonal noise, introducing an exponential decay hyperparameter, α . Subsequently, Gaussian smoothing is applied to the spectrogram. The spectrogram is then binarised by thresholding it at a specific decibel value, β . Pixels are grouped into regions by connecting groups of neighbouring pixels. Regions with fewer than min_length pixels and fewer than

min_count pixels are omitted. The remaining binary map should highlight only the detections of cetacean whistles and moans. The hyperparameters $\gamma, p, \alpha, \kappa, \beta, min_length$, and min_count are tuned using a random hyperparameter search. In addition, an optional high-pass filter at 1 kHz is added to the hyperparameter space to remove low frequency noise.

While this detection filter does not constitute a machine learning algorithm, the PAMGuard software [18, 19] also employs a latent discriminant analysis (LDA) classifier and an implementation of the random forest classifier highlighted in prior work [5]. For the LDA classifier, features are procured from the binarised spectrogram by cycling through the regions. If a region is of sufficiently long duration, it is split into slices. A second order polynomial is fit to the binary pixels in each slice. The zeroth, first, and second coefficients of the polynomials are collected for each slice. As a result, for any spectrogram, there is a distribution of these coefficients. Subsequently, the mean, standard deviation, and skew are calculated for the distribution, resulting in nine features per audio file. Though the original purpose of this classifier was to classify various species, we use it also as an enriched presence or absence classifier. LDA classifiers are then hyperparameter tuned for the detection filter hyperparameters as well as slice length.

We employed a similar approach for the ROCCA random forest algorithm [5]. In the original implementation, ROCCA relies upon human-generated whistle contours to create classifications, which is infeasible on large datasets. Therefore, we opted to use the contours generated by the PAMGuard whistle & moan detector. In some cases, this could generate over 10,000 contours per 5-minute audio file. For each contour, 57 attributes of the signal were extracted (i.e., starting frequency, slope, harmonics) [5]. Training was conducted using weakly supervised labels (i.e., propagating the label from the entire spectrogram to each contour). For hyperparameter tuning on the validation set, all contours in the spectrogram were ensembled such that the label receiving the most votes became the classification for the spectrogram.

For both of these techniques, it is important to note that the algorithm does not have a consistent computational cost per audio file, but rather depend on the number of contours produced in the whistle & moan detector.

We also use the **Animal-Spot** library as a benchmark [7, 8]. This model is a ResNet-based classifier [22] that classifies fixed spectrogram segments, as is common in passive acoustic monitoring [27]. We train the 11.69K parameter ResNet18 on the DEEPAL fieldwork dataset's training split and hyperparameter tune using the validation split. This dataset has two classes, "noise" and "orca". To make the model more competitive in speed and efficiency, we enable mixed precision during inference, and modify the data-loader to load 5-minute increments in parallel and split them into the appropriately formatted spectrograms.

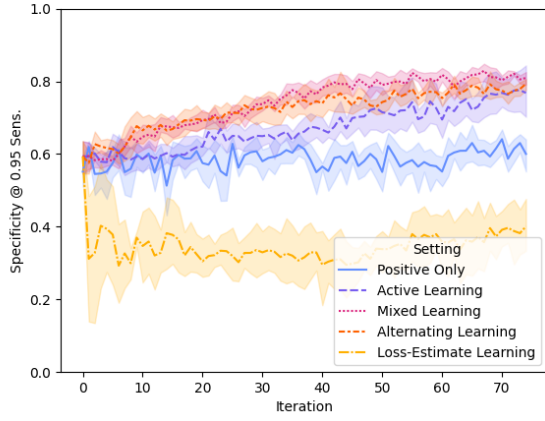
3 Results

3.1 Active Learning

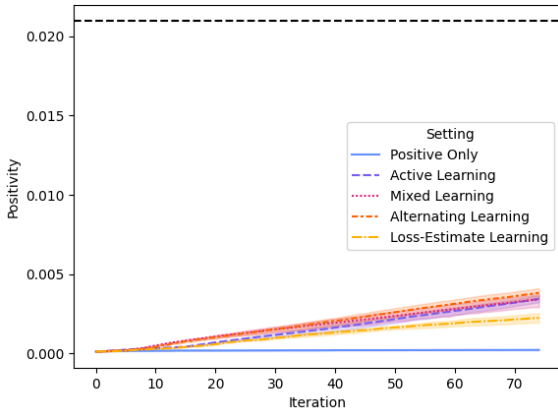
The positive-unlabelled active learning experiments are demonstrated in Figures 2 & 3 for the positive-unlabelled setting on the disjoint testing dataset (2022). There are two objectives of this approach; one is to develop high-performing classifiers to maximise

the likelihood of labelling positive instances, and the other objective is to label all possible positive instances.

First, we seek to evaluate the classifier performance across labelling iterations. When trying to construct the best possible classifier (as evaluated with validation and test performance as a function of labelling iterations), the mixed labelling strategies are preferred (Figure 2a). This is echoed in the second objective where the mixed labelling strategies are slightly more efficient at discovering data (Figure 2b).



(a) Specificity at 95% Sensitivity vs labelling iterations.



(b) Positivity rate vs. active learning iteration. The dashed line represents the positivity rate for the entire embedding dataset.

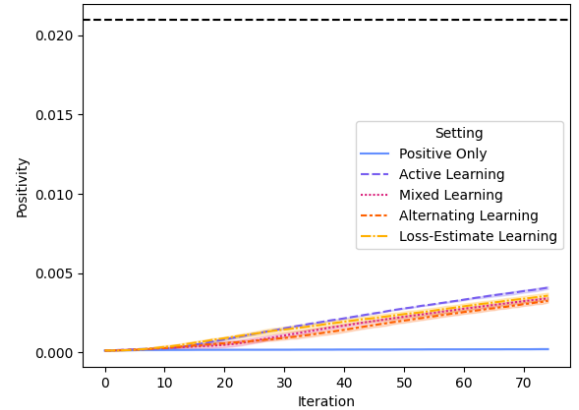
Figure 2: Active learning without noisy labels. Mixing positive and active learning helps discover samples with no penalty compared to entropy-only sampling.

Bioacousticians may be wary of trusting models trained using non-expert labelled data. We simulate scenarios where varying percentages of true positives are missed. Anecdotally, these may be low SNR cases where there is surrounding vessel noise or distant marine mammals. Under these scenarios, we surprisingly observe

that the models using noisy data and a mixed or alternating labelling approach *outperform* the clean data (Figure 3a). This exploration comes at an expense. We observe that under the noisy-labelling scenario entropy-based active learning discovers positive instances faster than the other approaches (Figure 3b).



(a) Specificity at 95% Sensitivity vs labelling iterations.



(b) Positivity rate vs. active learning iteration. The dashed line represents the positivity rate for the entire embedding dataset.

Figure 3: Active learning with label noise. At each iteration 30% of positively labelled samples are assigned flipped to negatives. In this setting, entropy-only sampling slightly outperforms the mixed strategies.

3.2 Benchmarking

Next, we compare the capabilities of our classifiers to the pre-existing techniques. We find that compared to PAMGuard Whistle & Moan detector + LDA [19], ROCCA [5], and ANIMAL-SPOT [7, 8] the fine-tuned whisper is competitive across several held out test sets (Table 1). While Animal-Spot attains highest performance on

the domain it is trained on, it is unsurprising that the whisper-model, fine-tuned with the most data generalizes well to in-domain and out-of-domain datasets. As model capacity grows, and the number of trainable parameters increases, so did the AUROC on the held-out test sets.

The whisper and ResNet models exceeded real-time detection (Table 2). The whisper model was more energy efficient during inference regardless of the hardware. Animal-Spot and whisper-tiny were competitive in processing time on GPU. In 1 day of consumer GPU processing, almost an entire year of 32 kHz PAM data can be processed on a single GPU.

3.3 Citizen Science Labelling Evaluation

We compare the performance of non-expert labellers with that of an expert bioacoustician on the most likely audio files across 365 days at one site, and 31 days at another (396 samples total). The concordance between amateur and the expert labeller, as measured by Cohen’s kappa, was 0.714, 0.580, and 0.229. The new instances of marine mammal data discovered in the paper have been labelled by the amateur labeller with a Cohen’s Kappa of 0.714.

3.4 Species and Ecotype Classification

Species labels are available for 55,872 of 65,580 files. 392 are provided by ONC, 31,118 are provided by SanctSound [21], 22,567 are labelled by a human annotator in this project, and 1926 are pseudo-labelled by propagating labels to the files occurring immediately before or after a labelled event. These labels cover humpbacks (1501 h), orcas (2484 h), pacific white sided dolphins (246 h), sea lions (88 h), minke whales (0.3 h), fin whales (166 h), sperm whales (10 h), gray whales (0.2 h), or multiple species (20 h). These labels were also acquired using a combination of active learning and prioritising labelling likely orca files. During training, we omit the SanctSound labels, as they feature imprecise start and end times. A simple multiclass classifier is trained atop of the whisper-tiny embeddings. We evaluate the model on the DCLDE-2026 dataset (excluding ONC and OrcaSound contributions), which include humpback whales, orcas, and other dolphins. We map the pacific white-sided dolphin class to the other dolphin class in DCLDE-2026. The species-classifier obtains a top-1 classification accuracy of 53.2% (11 train classes, 4 test classes).

Our dataset contains ecotype annotations for 2484 hours of orca data. This spans SRKW (919 hours), Bigg’s transients (230 hours), NRKW (3 hours), offshore (15 hours). SRKWs were labelled by pulsed calls. If the calls featured an S01, S03, S13, S16, S18, S19, S33, S36, S42, or S44, then the file was determined to be an SRKW [16]. Bigg’s were determined by T04 or T07 calls. NRKWs were more rare in the dataset, so there was no systematic labelling technique. Offshores were primarily determined by ruling out other ecotypes at deep water nodes. Pseudo-labels were generated for files neighbouring ecotype identifications, even if the pseudo-labelled file only contained echolocation or whistles. Using these 4 classes, the ecotype-classifier obtains a top-1 accuracy of 33.6% on the DCLDE-2026 dataset (4 train classes, 5 test classes). Our dataset does not include Southern Alaskan resident orcas.

3.5 SRKW Encounters and Habitat Usage

We compare the quantity of SRKW encounters with other sources of reporting defined by days with SRKW sightings (Table 3). While sightings have produced more events than this dataset, expanding the openly available hydrophone network to regions like Swiftsure Bank may enable this scalable approach to match those of sightings networks [56].

When comparing our observations to those in prior literature, we observe 170 additional whale-days [58]. This means that at least 20.8% of SRKW whale-days were missed during the study’s time period within the study’s coordinates [58]. 68.9% of these missed events occurred during daylight (solar altitude $>0^\circ$) 23.3% occurred at night (solar altitude $\leq -12^\circ$) with the remainder occurring at twilight. Over 60% of the missed detections happened between July–September, with the missed events being evenly distributed across the years. Across our entire dataset, 47.2% of vocalizations occurred during daylight, 38.8% at night, and the remainder during twilight.

We also compare our observations to another prior work [60]. This work used 6 data sources, which involved contributed reports from hydrophones and sightings networks. We find roughly a tenth of the number of occurrences in the study region, likely due to a lack of hydrophone coverage. Despite this, the temporal trend of annual SRKW whale-days has a Pearson correlation of 69.1%. Interestingly, the Bigg’s population had a correlation coefficient of -25.6%. One explanation could be that SRKWs are relatively noisy, so they are likely to be detected on the hydrophone each time they pass [6, 37]. Bigg’s on the other hand, are silent while hunting [6, 37].

4 Discussion

There are some easy tangible steps to produce better classifiers. In all of our experiments, we only use the whisper-tiny model with frozen weights during fine-tuning. This allows us to pre-compute and store all of the embeddings. Using a large whisper architecture or fine-tuning the entire model would likely improve performance. This model is designed with end-uses in mind. Bioacousticians may need to run these algorithms them aboard ships during field work, or the embeddings may be computed locally on the node to reduce the quantity of data that must be streamed.

The algorithms we develop are specifically designed to detect pulsed calls and whistles within orca communication frequencies. By having a sampling rate of only 32 kHz, we exclude some dolphin species, while having a high-pass filter at 1 kHz we lose sensitivity for blue, sei, fin, humpback, gray, and north pacific right whales. Researchers interested in detecting these species should modify these parameters accordingly.

Human labour availability is an important aspect of this work. Appropriate prior distributions are required to effectively conduct positive-unlabelled learning. Using traditional marine biologist surveys may reduce labelling burden when expanding to new species or new locations. Overall, the quantity of findings in these large archives using machine learning leaves *more* work for marine biologists to do, rather than less. In addition, we observe that three of the amateur labellers had positive-valued Cohen’s Kappa scores. We cite theoretical literature, and demonstrate empirically that the penalties to noisily labelled data are inconsequential. Datasets can feasibly be labelled by non-experts

Table 1: Cetacean detection AUROC on several held out test sets. "LR" means logistic regression and "SGD" indicates stochastic gradient descent with a hinge loss on frozen embeddings. "FT" indicates that all weights were unfrozen and updated. * a maximum of 1000 contours are subsampled.

Model	Deepal ComParE [7]	DCLDE-2026 [47] excl. ONC & OS	FGPD 2023	PSGCH.H3 2023/08
PAMGuard LDA ^a	0.520	0.559	0.490	0.390
ROCCA ^b [5]	0.467	0.533 ^c	0.582	0.548
ANIMAL-SPOT [8]	0.648	0.466	0.529	0.348
Whisper-tiny [51] + LR	0.480	0.669	0.542	0.624
Whisper-tiny [51] + SGD	0.600	0.671	0.644	0.852
Whisper-tiny [51] + FT	0.568	0.786	0.678	0.957
Whisper-base [51] + FT	0.732	0.749	0.599	0.786
Whisper-large-v3 [51] + FT	0.716	0.779	0.632	0.933
Wav2Vec 2 [4] + FT	0.604	0.655	0.568	0.657

^aAlgorithm implemented in python to operate across platforms. Hyper-parameter tuning was conducted to outperform the PAMGuard defaults.

^bAlgorithm implemented in python to operate across platforms. Hyper-parameter tuning was conducted to outperform the PAMGuard defaults.

^cGenerated contours were randomly subsampled so that only 1000 contours were selected

Table 2: Computational efficiency of presence/absence classifiers. All benchmarks are conducted on the FGPD and PSGCH.H3. Except for LDA and ROCCA, all settings used 8 concurrent workers. "+S" means that the contours found are subsampled to increase the file processing throughput.

System	Model	Processing Time (audio:compute) ↑	Power Efficiency (day audio/kWh) ↑	RAM (MB) ↓	VRAM (MB) ↓
CPU	PAMGuard LDA [19]	55.34	46.80	6130	-
	ROCCA [5, 19]	0.494	0.06	16368	-
	ROCCA ^a [5, 19]	20.23	17.29	5771	-
	Animal-Spot [8]	25.02	7.41	6512	-
	Whisper-tiny [51]	122.09	26.51	2237	-
	Whisper-base [51]	23.38	4.95	3388	-
	Whisper-large-v3 [51]	4.42	1.20	7394	-
	Wav2Vec 2 [4]	17.31	3.67	2917	-
RTX 2060	Animal-Spot [8]	280.73	21.23	4995	3491
	Whisper-tiny [51]	273.57	106.81	4003	154
	Whisper-base [51]	137.83	54.66	3888	717
	Whisper-large-v3 [51]	23.56	9.42	4228	4077
	Wav2Vec 2 [4]	78.73	32.04	3896	1077

^aGenerated contours were randomly subsampled so that only 1000 contours were selected

and still contribute meaningfully to science. Promising projects such as bioacoustic or satellite surveys for marine mammals [10] may benefit from citizen science labelling initiatives (for example, <https://www.wwf.org.uk/learn/walrus-from-space>).

Unsupervised machine translation [31] has reignited interest in translating non-human animals. We note that this dataset is approximately the same size as the English libraspeech corpus [48]. Unlike audiobooks, our dataset contains varying SNRs caused by ship noise, diarization challenges [35], SRKW dialects [16], and contributions from nearly 100 individual orcas throughout the years. Nonetheless, we suspect that recent works investigating whale linguistics using this dataset may find similar properties with SRKWs.

For example, humpback vocalizations follow the same statistical structure as human languages [2], sperm whales embed context in their vocalizations [57], and bottlenose dolphins will respond to whistles that individually identify them [28]. The openly licensed dataset we curated should serve as a useful tool for investigating animal linguistics.

We welcome contributions from the community to add species, ecotype, pod, and individual identifications for this dataset through versioned merge requests.

Table 3: Recorded southern resident killer whale events across a variety of sources through 2022. An "event" is defined as a day in which SRKWs are spotted. * 203.6 hours of confirmed events with the remainder being interpolated from the median event duration.

Source	Events	Audio (hrs)
Shields (2018-2022) [58]	647	-
PWWA (2018-2022)	317	-
Orca Network	1995	-
iNaturalist [23]	14	-
OrcaSound	122	214.6*
ONC annotations	10	3.1
DCLDE [47]	87	4.1
DORI (ours)	560	919

5 Conclusion

We present the dataset for orca resident interpretation (DORI). This dataset is a curation of over 30 years of data resulting in 5298 hours of marine mammals and 919 hours of SRKW audio. We propose a novel active learning strategy that outperforms both entropy-based labelling and loss estimate labelling strategies. The newly-discovered quantity of data enables us to fine tune state of the art classifiers that outperform existing tools in terms of accuracy, speed, and energy efficiency. This work serves as a template for citizen science on vast uncurated datasets. These data are available for unbiased observations of SRKW behaviour, policy and planning activities for conservation, and to translate the SRKW dialect to human languages.

6 Ethics, Fairness, and Limitations

This paper focuses only on publicly accessible data through Ocean Networks Canada (CC-BY 4.0), OrcaSound (CC-BY-NC-SA 4.0), Ocean Observatories Initiative (publicly accessible under custom data usage agreement), and SanctSound [21]. Our resulting curated dataset is publicly available, which reduces barriers to entry for marine biologists. Many of the hydrophones which data are collected from in Canada are located on Coast Salish waters off the coast of British Columbia. The coastal territories that coincide with the area of this study are mostly unceded and ancestral land of the Coast Salish. Hydrophone network operators, such as ONC use CARE principles for indigenous governance (<https://www.gida-global.org/whoware>) in order to align with the UN's Declaration of Rights for Indigenous Peoples. While the source data is managed by partners, we respect indigenous rights to ownership, control, access, and possession, especially in unceded territorial waters. In some nations, whales are regarded as kin, and should be considered to have personhood and the afforded rights.

We intentionally omit data acquired from live-capture events and aquariums.

The Whisper models are distributed under the Apache-2.0 License. Our fine-tuned model weights and code are distributed under a BigScience Open RAIL-M License which restricts unethical use

cases. These restrictions prohibit use for harming others, discriminating against others, impersonating others, or to be used in law enforcement circumstances, amongst others.

Data collected in this study contains strong sampling bias, as hydrophones are placed in regions which are determined to be optimal and feasible. This may present as an over-abundance of marine mammal data compared to other locations within SRKW habitat. Passive acoustic monitoring overcomes some sampling bias from sightings and sea-based surveys by monitoring at night and during inclement weather.

7 GenAI Disclosure

Generative AI was used in the preparation of some code. It was also used to broaden the literature search beyond what was found by traditional indexed database searches. Generative AI was used in neither the writing of the paper, nor in the creation of citations, which were downloaded as bibtex or RIS format from source.

Acknowledgments

This project was funded by the UW Allen School Postdoc Research Award.

Advanced computing resources are provided by the Digital Research Alliance of Canada (Alliance), the organization responsible for digital research infrastructure in Canada, and ACENET, the regional partner in Atlantic Canada. ACENET is funded by the Alliance, the provinces of New Brunswick, Newfoundland & Labrador, Nova Scotia and Prince Edward Island, as well as the Atlantic Canada Opportunities Agency.

This material is based upon work supported by the National Science Foundation under Cooperative Agreement No. 1743430 (which supports the OOI) or other relevant NSF award number.

We would like to thank Mark Hamilton for providing the advice to use layer normalisation for the embeddings.

References

- [1] Umang Aggarwal, Adrian Popescu, and Céline Hudelot. 2020. Active Learning for Imbalanced Datasets. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1417–1426. doi:10.1109/WACV45572.2020.9093475
- [2] Inbal Arnon, Simon Kirby, Jenny A. Allen, Claire Garrigue, Emma L. Carroll, and Ellen C. Garland. 2025. Whale song shows language-like statistical structure. *Science* 387, 6734 (2025), 649–653. arXiv:https://www.science.org/doi/pdf/10.1126/science.adq7055 doi:10.1126/science.adq7055
- [3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
- [4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 12449–12460. https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9ba3227870bb6d7f07-Paper.pdf
- [5] Yvonne Barkley, Julie Nicola Oswald, James V. Carretta, Shannon Rankin, Alexis Rudd, and Marc O. Lammers. [2011]. Comparison of real-time and post-cruise acoustic species identification of dolphin whistles using ROCCA (Real-time Odontocete Call Classification Algorithm). https://repository.library.noaa.gov/view/noaa/3797 Technical Memorandum.
- [6] LANCE G. BARRETT-LENNARD, JOHN K.B. FORD, and KATHY A. HEISE. 1996. The mixed blessing of echolocation: differences in sonar use by fish-eating and mammal-eating killer whales. *Animal Behaviour* 51, 3 (1996), 553–565. doi:10.1006/anbe.1996.0059
- [7] Christian Bergler, Hendrik Schröter, Rachael Xi Cheng, Volker Barth, Michael Weber, Elmar Nöth, Heribert Hofer, and Andreas Maier. 2019. ORCA-SPOT: An Automatic Killer Whale Sound Detection Toolkit Using Deep Learning. *Scientific Reports* 9, 1 (29 Jul 2019), 10997. doi:10.1038/s41598-019-47335-w
- [8] Christian Bergler, Simeon Q. Smeele, Stephen A. Tyndel, Alexander Barnhill, Sara T. Ortiz, Ammie K. Kalan, Rachael Xi Cheng, Signe Brinklov, Anna N. Osiecka, Jakob Tougaard, Freja Jakobsen, Magnus Wahlberg, Elmar Nöth, Andreas Maier, and Barbara C. Klump. 2022. ANIMAL-SPOT enables animal-independent signal detection and classification using deep learning. *Scientific Reports* 12, 1 (19 Dec 2022), 21966. doi:10.1038/s41598-022-26429-y
- [9] Michael Andrew Bigg. 1982. An assessment of killer whale (*Orcinus orca*) stocks off Vancouver Island, British Columbia. *REP. INT. WHALING COMM* 32 (1982), 655–666. https://api.semanticscholar.org/CorpusID:222306729
- [10] Alex Borowicz, Hieu Le, Grant Humphries, Georg Nehls, Caroline Höschle, Vladislav Kosarev, and Heather J. Lynch. 2019. Aerial-trained deep learning networks for surveying cetaceans from satellite imagery. *PLOS ONE* 14, 10 (10 2019), 1–15. doi:10.1371/journal.pone.0212532
- [11] M. Bradley Hanson, Candice K. Emmons, Eric J. Ward, Jeffrey A. Nystuen, and Marc O. Lammers. 2013. Assessing the coastal occurrence of endangered killer whales using autonomous passive acoustic recorders. *The Journal of the Acoustical Society of America* 134, 5 (11 2013), 3486–3495. arXiv:https://pubs.aip.org/asa/jasa/article-pdf/134/5/3486/13325241/3486_1_online.pdf doi:10.1121/1.4821206
- [12] Olivier Coudray, Christine Keribin, Pascal Massart, and Patrick Pamphile. 2023. Risk Bounds for Positive-Unlabeled Learning Under the Selected At Random Assumption. *Journal of Machine Learning Research* 24, 107 (2023), 1–31. http://jmlr.org/papers/v24/22-067.html
- [13] Benoit Courty, Victor Schmidt, Sasha Luccioni, Goyal-Kamal, MarionCoutarel, Boris Feld, Jérémy Lecourt, LiamConnell, Amine Saboni, Inimaz, supatomic, Mathilde Léval, Luis Blanche, Alexis Cruveiller, ouminasara, Franklin Zhao, Aditya Joshi, Alexis Bogroff, Hugues de Lavoreille, Niko Laskaris, Edoardo Abati, Douglas Blank, Ziyao Wang, Armin Catovic, Marc Alencon, Michal Stęchly, Christian Bauer, Lucas Otávio N. de Araújo, JPW, and MinervaBooks. 2024. *mlco2/codecarbon: v2.4.1*. doi:10.5281/zenodo.11171501
- [14] Lusseau D, Bain DE, Williams R, and Smith JC. 2009. Vessel traffic disrupts the foraging behavior of southern resident killer whales *Orcinus orca*. *Endangered Species Research* 6, 3 (2009), 211–221. doi:10.3354/esr00154
- [15] Volker B. Deecke, John K.B. Ford, and Peter J.B. Slater. 2005. The vocal behaviour of mammal-eating killer whales: communicating with costly calls. *Animal Behaviour* 69, 2 (2005), 395–405. doi:10.1016/j.anbehav.2004.04.014
- [16] John Kenneth Baker Ford. 1984. *Call traditions and dialects of killer whales (Orcinus orca) in British Columbia*. Ph.D. Dissertation. University of British Columbia. doi:10.14288/1.0096602
- [17] GBIF.Org User. 2024. Occurrence Download. doi:10.15468/DL.ERW6Z5
- [18] Douglas Gillespie, Marjolaine Caillat, Jonathan Gordon, and Paul White. 2013. Automatic detection and classification of odontocete whistles. *The Journal of the Acoustical Society of America* 134, 3 (09 2013), 2427–2437. arXiv:https://pubs.aip.org/asa/jasa/article-pdf/134/3/2427/15307906/2427_1_online.pdf doi:10.1121/1.4816555
- [19] Douglas Gillespie, David K. Mellinger, Jonathan Gordon, David McLaren, Paul Redmond, Ronald McHugh, Philip Trinder, Xiao-Yan Deng, and Aaron Thode. 2009. PAMGUARD: Semiautomated, open source software for real-time acoustic detection and localization of cetaceans. *The Journal of the Acoustical Society of America* 125 (04 2009), 2547–2547. doi:10.1121/1.4808713
- [20] M. Bradley Hanson, Candice K. Emmons, Michael J. Ford, Meredith Everett, Kim Parsons, Linda K. Park, Jennifer Hempelmann, Donald M. Van Doornik, Gregory S. Schorr, Jeffrey K. Jacobsen, Mark F. Sears, Maya S. Sears, John G. Sneva, Robin W. Baird, and Lynne Barre. 2021. Endangered predators and endangered prey: Seasonal diet of Southern Resident killer whales. *PLOS ONE* 16, 3 (03 2021), 1–27. doi:10.1371/journal.pone.0247031
- [21] Leila Hatch, Jack Barkowski, Simone Baumann-Pickering, Carrie Wall Bell, Benjamin Best, Jennifer Brown, Annamaria DeAngelis, John Joseph, Ella Kim, Danielle Kitchen, Annebelle C. M. Kok, Anke Kügler, Anurag Kumar, Marc O. Lammers, Tetyana Margolina, Megan McKenna, Karlina Merckens, Lindsey Peavey Reeves, Timothy J. Rowell, Mandy Shoemaker, Jenni A. Stanley, Alison Stimpert, Brandon Southall, Brian Stone, Mitchell Tartt, Sofie Van Parijs, and Eden J. Zang. 2024. *SanctSound: Building Data Systems for Sound Decisions*. Springer International Publishing, Cham, 1547–1557. doi:10.1007/978-3-031-50256-9_58
- [22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. doi:10.1109/CVPR.2016.90
- [23] iNaturalist. [n.d.]. iNaturalist. https://www.inaturalist.org
- [24] NSF Ocean Observatories Initiative. 2023. Broadband Acoustic Receiver (Hydrophone) (CE02SHBP-LJ01D) data from Oregon Shelf Cabled Benthic Experiment Package at Coastal Endurance from 2015-01-01 to 2018-01-01. https://dataexplorer.oceanobservatories.org/ooi/array/CE/subsite/CE02SHBP/instrument_type/HYDBB/data
- [25] IUCN. [n.d.]. IUCN Red List of Threatened Species. https://www.iucnredlist.org/
- [26] Ajay J. Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. 2009. Multi-class active learning for image classification. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2372–2379. doi:10.1109/CVPR.2009.5206627
- [27] Stefan Kahl, Connor M. Wood, Maximilian Eibl, and Holger Klinck. 2021. BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics* 61 (2021), 101236. doi:10.1016/j.ecoinf.2021.101236
- [28] Stephanie L. King and Vincent M. Janik. 2013. Bottlenose dolphins can use learned vocal labels to address each other. *Proceedings of the National Academy of Sciences* 110, 32 (2013), 13216–13221. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1304459110 doi:10.1073/pnas.1304459110
- [29] Ryuichi Kiryo, Gang Niu, Marthinus C du Plessis, and Masashi Sugiyama. 2017. Positive-Unlabeled Learning with Non-Negative Risk Estimator. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/7ece53cf90577442771720a370c3c723-Paper.pdf
- [30] Robert C. Lacy, Rob Williams, Erin Ashe, Kenneth C. Balcomb III, Lauren J. N. Brent, Christopher W. Clark, Darren P. Croft, Deborah A. Giles, Misty MacDuffee, and Paul C. Paquet. 2017. Evaluating anthropogenic threats to endangered killer whales to inform effective recovery plans. *Scientific Reports* 7, 1 (26 Oct 2017), 14119. doi:10.1038/s41598-017-14471-0
- [31] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*. https://openreview.net/forum?id=H196sainb
- [32] Daniel Langenkämper, Erik Simon-Lledó, Brett Hosking, Daniel O. B. Jones, and Tim W. Nattkemper. 2019. On the impact of Citizen Science-derived data quality on deep learning based classification in marine images. *PLOS ONE* 14, 6 (06 2019), 1–16. doi:10.1371/journal.pone.0218086
- [33] David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Dublin, Ireland) (SIGIR '94)*. Springer-Verlag, Berlin, Heidelberg, 3–12.
- [34] Dongyuan Li, Zhen Wang, Yankai Chen, Renhe Jiang, Weiping Ding, and Manabu Okumura. 2025. A Survey on Deep Active Learning: Recent Advances and New Frontiers. *IEEE Transactions on Neural Networks and Learning Systems* 36, 4 (2025), 5879–5899. doi:10.1109/TNNLS.2024.3396463
- [35] Louis Mahon, Benjamin Hoffman, Logan S. James, Maddie Cusimano, Masato Hagiwara, Sarah C. Woolley, Felix Effenberger, Sara Keen, Jen-Yu Liu, and Olivier Pietquin. 2025. Robust detection of overlapping bioacoustic sound events. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems workshop: AI for non-human animal communication*. https://openreview.net/forum?id=b7x4CtWynM
- [36] DAVID K. MELLINGER, KATHLEEN M. STAFFORD, SUE E. MOORE, ROBERT P. DZIAK, and HARU MATSUMOTO. 2007. An Overview of Fixed Passive Acoustic Observation Methods for Cetaceans. *Oceanography* 20, 4 (2007), 36–45. http://www.jstor.org/stable/24860138
- [37] Phillip A. Morin, Morgan L. McCarthy, Charissa W. Fung, John W. Durban, Kim M. Parsons, William F. Perrin, Barbara L. Taylor, Thomas A. Jefferson, and Frederick I. Archer. 2024. Revised taxonomy of eastern North

- Pacific killer whales (*Orcinus orca*): Biggs et al. 2019; and resident ecotypes deserve species status. *Royal Society Open Science* 11, 3 (2024), 231368. [arXiv:https://royalsocietypublishing.org/doi/pdf/10.1098/rsos.231368](https://royalsocietypublishing.org/doi/pdf/10.1098/rsos.231368)
- [38] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. 2013. Learning with Noisy Labels. In *Advances in Neural Information Processing Systems*, C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (Eds.), Vol. 26. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2013/file/3871bd64012152bfb53fd04b401193f-Paper.pdf
- [39] Orca Network. [n. d.]. Recent Whale Sightings. <https://www.orcanetwork.org/recent-sightings>. Accessed: 2023-10-14.
- [40] Kun-Peng Ning, Xun Zhao, Yu Li, and Sheng-Jun Huang. 2022. Active Learning for Open-Set Annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 41–49.
- [41] Ocean Networks Canada Society. 2020. Folger Deep Hydrophone Deployed 2020-03-08. doi:10.34943/8aa1dc24-2e86-45ad-a46b-6ed1c9a3e0df
- [42] Ocean Networks Canada Society. 2023. Strait of Georgia Central Hydrophone Deployed 2023-03-17. doi:10.34943/b169fbab-3287-44f0-b5d6-a27cb2d6b1a0
- [43] National Oceanic and Department of Commerce Atmospheric Administration National Marine Fisheries Service. [n. d.]. Enumeration of endangered marine and anadromous species. 50 CFR 224.101. <https://www.ecfr.gov/current/title-50/part-224/section-224.101>
- [44] Government of Canada. 2023. Killer Whale (*Orcinus orca*), Northeast Pacific southern resident population. <https://species-registry.canada.ca/index-en.html#/species/699-5>
- [45] K. B. Oke, C. P. Cunningham, P. A. H. Westley, M. L. Baskett, S. M. Carlson, J. Clark, A. P. Hendry, V. A. Karatayev, N. W. Kendall, J. Kibele, H. K. Kindsvater, K. M. Kobayashi, B. Lewis, S. Munch, J. D. Reynolds, G. K. Vick, and E. P. Palkovacs. 2020. Recent declines in salmon body size impact ecosystems and fisheries. *Nature Communications* 11, 1 (19 Aug 2020), 4155. doi:10.1038/s41467-020-17726-z
- [46] OrcaSound. [n. d.]. Projects. <https://www.orcasound.net/portfolio/>. Accessed: 2023-10-14.
- [47] K. J. Palmer, Emma Cummings, Michael G. Dowd, Kait Frasier, Fabio Frazao, Alex Harris, April Houweling, Jasper Kanes, Oliver S. Kirsebom, Holger Klinck, Holly LeBlond, Lauren Laturnus, Craig Matkin, Olivia Murphy, Hannah Myers, Dan Olsen, Caitlin O'Neill, Bruno Padovese, James Pilkington, Lucy Quayle, Amalis Riera Vuibert, Krista Trounce, Svein Vagle, Scott Veirs, Val Veirs, Jen Wladichuk, Jason Wood, Tina Yack, Harald Yurk, and Ruth Joy. 2025. A Public Dataset of Annotated *Orcinus orca* Acoustic Signals for Detection and Ecotype Classification. *Scientific Data* 12, 1 (03 Jul 2025), 1137. doi:10.1038/s41597-025-05281-5
- [48] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5206–5210. doi:10.1109/ICASSP.2015.7178964
- [49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
- [50] James F. Pilkington, Eva H. Stredulinsky, Katherine Gavrilchuk, Sheila J. Thornton, John K. B. Ford, and Thomas Doniol-Valcroze. 2023. Patterns of winter occurrence of three sympatric killer whale populations off eastern Vancouver Island, Canada, based on passive acoustic monitoring. *Frontiers in Marine Science* Volume 10 - 2023 (2023). doi:10.3389/fmars.2023.1204908
- [51] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.), PMLR, 28492–28518. <https://proceedings.mlr.press/v202/radford23a.html>
- [52] Zoe R. Rand, Jason D. Wood, and Julie N. Oswald. 2022. Effects of duty cycles on passive acoustic monitoring of southern resident killer whale (*Orcinus orca*) occurrence and behavior. *The Journal of the Acoustical Society of America* 151, 3 (03 2022), 1651–1660. [arXiv:https://pubs.aip.org/asa/jasa/article-pdf/151/3/1651/16523645/1651_1_online.pdf](https://pubs.aip.org/asa/jasa/article-pdf/151/3/1651/16523645/1651_1_online.pdf) doi:10.1121/10.0009752
- [53] Zoe R. Rand, Jason D. Wood, and Julie N. Oswald. 2022. Effects of duty cycles on passive acoustic monitoring of southern resident killer whale (*Orcinus orca*) occurrence and behavior. *The Journal of the Acoustical Society of America* 151, 3 (03 2022), 1651–1660. [arXiv:https://pubs.aip.org/asa/jasa/article-pdf/151/3/1651/16523645/1651_1_online.pdf](https://pubs.aip.org/asa/jasa/article-pdf/151/3/1651/16523645/1651_1_online.pdf) doi:10.1121/10.0009752
- [54] Stephen Raverty, Judy St. Leger, Dawn P. Noren, Kathy Burek Huntington, David S. Rotstein, Frances M. D. Gulland, John K. B. Ford, M. Bradley Hanson, Dyanna M. Lambourn, Jessie Huggins, Martha A. Delaney, Lisa Spaven, Teri Rowles, Lynne Barre, Paul Cottrell, Graeme Ellis, Tracey Goldstein, Karen Terio, Debbie Duffield, Jim Rice, and Joseph K. Gaydos. 2020. Pathology findings and correlation with body condition index in stranded killer whales (*Orcinus orca*) in the northeastern Pacific and Hawaii from 2004 to 2013. *PLOS ONE* 15, 12 (12 2020), 1–31. doi:10.1371/journal.pone.0242505
- [55] Amalis Riera, John K. Ford, and N. Ross Chapman. 2013. Effects of different analysis techniques and recording duty cycles on passive acoustic monitoring of killer whales. *The Journal of the Acoustical Society of America* 134, 3 (09 2013), 2393–2404. [arXiv:https://pubs.aip.org/asa/jasa/article-pdf/134/3/2393/15307055/2393_1_online.pdf](https://pubs.aip.org/asa/jasa/article-pdf/134/3/2393/15307055/2393_1_online.pdf) doi:10.1121/1.4816552
- [56] A. Riera, J. F. Pilkington, J. K. B. Ford, E. H. Stredulinsky, and N. R. Chapman. 2019. Passive acoustic monitoring off Vancouver Island reveals extensive use by at-risk resident killer whale (*Orcinus orca*) populations. *Endangered Species Research* 39 (2019), 221–234. doi:10.3354/esr00966
- [57] Pratyusha Sharma, Shane Gero, Roger Payne, David F. Gruber, Daniela Rus, Antonio Torralba, and Jacob Andreas. 2024. Contextual and combinatorial structure in sperm whale vocalisations. *Nature Communications* 15, 1 (07 May 2024), 3617. doi:10.1038/s41467-024-47221-8
- [58] Monika W Shields. 2023. 2018–2022 Southern Resident killer whale presence in the Salish Sea: continued shifts in habitat usage. *PeerJ* 11 (2023), e15635. doi:10.7717/peerj.15635
- [59] Marie Souhaut and Monika W Shields. 2021. Stereotyped whistles in southern resident killer whales. *PeerJ* 9, e12085 (Aug. 2021), e12085.
- [60] Joshua D. Stewart, Jane Cogan, John W. Durban, Holly Fearnbach, David K. Ellifrit, Mark Malleson, Melisa Pinnow, and Kenneth C. Balcomb. 2023. Traditional summer habitat use by Southern Resident killer whales in the Salish Sea is linked to Fraser River Chinook salmon returns. *Marine Mammal Science* 39, 3 (2023), 858–875. [arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/mms.13012](https://onlinelibrary.wiley.com/doi/pdf/10.1111/mms.13012) doi:10.1111/mms.13012
- [61] Jennifer B. Tennessen, Marla M. Holt, Brianna M. Wright, M. Bradley Hanson, Candice K. Emmons, Deborah A. Giles, Jeffrey T. Hogan, Sheila J. Thornton, and Volker B. Deecke. 2024. Males miss and females forgo: Auditory masking from vessel noise impairs foraging efficiency and success in killer whales. *Global Change Biology* 30, 9 (2024), e17490. [arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/gcb.17490](https://onlinelibrary.wiley.com/doi/pdf/10.1111/gcb.17490) doi:10.1111/gcb.17490 e17490 GCB-24-0481.R1.
- [62] George Tzanetakis, Mathieu Lagrange, Paul Spong, and Helena Symonds. 2007. Orcheive: Digitizing and Analyzing Orca Vocalizations. In *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)* (Pittsburgh, Pennsylvania) (RIA/O'07). LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, Paris, FRA, 56–63.
- [63] Svein Vagle, Rianna Burnham, Pramod Thupaki, Christine Konrad, Scott Toews, and Sheila J. Thornton. 2021. Vessel presence and acoustic environment within Southern Resident killer whale (*Orcinus orca*) critical habitat in the Salish Sea and Swiftsure Bank area. Canadian Science Advisory Secretariat (CSAS).
- [64] Kathleen J. Vigness-Raposa, Gail Scowcroft, James H. Miller, and Darlene Ketten. 2012. Discovery of Sound in the Sea: An Online Resource. In *The Effects of Noise on Aquatic Life*, Arthur N. Popper and Anthony Hawkins (Eds.). Springer New York, New York, NY, 135–138.
- [65] Rob Williams, David Lusseau, and Philip S. Hammond. 2006. Estimating relative energetic costs of human disturbance to killer whales (*Orcinus orca*). *Biological Conservation* 133, 3 (2006), 301–311. doi:10.1016/j.biocon.2006.06.010
- [66] Donggeun Yoo and In So Kweon. 2019. Learning Loss for Active Learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 93–102. doi:10.1109/CVPR.2019.00018
- [67] Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. 2020. Are Transformers universal approximators of sequence-to-sequence functions?. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=ByxRM0Ntrv>

A Appendix

Received 9 February 2026

Split	Lat.	Lon.	Date	Train	Valid.	Test
ONC-BACAX	48.32	-126.05	2009-2011	✓		
ONC-BIIP	49.30	-123.11	2021	✓		
ONC-CBCH.H1	47.76	-127.73	2020-2021	✓		
ONC-CBYIP	69.11	-105.06	2014-2016	✓		
ONC-CQS64.H1	48.70	-126.87	2017-2020	✓		
ONC-CRIP	50.02	-125.24	2016-2021	✓		
ONC-ECHO1.H1	49.04	-123.32	2016	✓		
ONC-ECHO1.H2	49.04	-123.32	2016	✓		
ONC-ECHO1.H3	49.04	-123.32	2016	✓		
ONC-ECHO2.H1	49.04	-123.32	2014-2016	✓		
ONC-ECHO2.H2	49.04	-123.32	2018	✓		
ONC-ECHO2.H3	49.04	-123.32	2014-2015	✓		
ONC-ECHO2.H4	49.04	-123.32	2014-2016	✓		
ONC-ECHO3.H1	49.04	-123.32	2020-2021	✓		
ONC-ECHO3.H2	49.04	-123.32	2020-2021	✓		
ONC-ECHO3.H3	49.04	-123.32	2020-2021	✓		
ONC-ECHO3.H4	49.04	-123.32	2020-2021	✓		
ONC-FAE	49.08	-123.34	2018	✓		
ONC-FGPD	48.81	-125.28	2009-2021	✓		
ONC-KEMO	47.92	-129.11	2016-2018	✓		
ONC-LSBBL	49.08	-123.34	2016-2017	✓		
ONC-LSHA.H1	49.08	-123.34	2014	✓		
ONC-LSHA.H2	49.08	-123.34	2014-2015	✓		
ONC-LSHA.H3	49.08	-123.34	2014-2015	✓		
ONC-LSHA.H4	49.08	-123.34	2014	✓		
ONC-LSHA.H5	49.08	-123.34	2014	✓		
ONC-NC27.H1	47.76	-127.76	2014-2015	✓		
ONC-NCBC	48.43	-126.18	2010-2011	✓		
ONC-PVIP	48.66	-123.49	2019-2021	✓		
ONC-SCVIP	49.04	-123.43	2015-2021	✓		
ONC-SEHA.H5	49.04	-123.32	2014-2015	✓		
ONC-SEVIP	49.04	-123.32	2019-2020	✓		
ONC-SGC.H2	49.04	-123.426	2017-2018	✓		
ONC-SGE.H1	49.04	-123.32	2008-2015	✓		
ONC-SGE.H2	49.04	-123.32	2011-2012	✓		
ONC-USDDL	49.08	-123.33	2021	✓		
ONC-USSLP	49.085	-123.330	2018-2019	✓		
OOI-CE04OSBP	44.37	-125.00	2018-2021	✓		
OOI-CE02SHBP	44.64	-124.31	2015-2016	✓		
OOI-RS01SLBS	44.52	-125.39	2016-2020	✓		
OOI-RS01SBPS	44.53	-125.39	2016-2021	✓		
SanctSound-OC01	48.39	-124.65	2019-2020	✓		
SanctSound-OC02	48.49	-125.00	2019-2021	✓		
SanctSound-OC03	47.32	-124.41	2020-2021	✓		
SanctSound-OC04	48.04	-124.73	2019-2020	✓		
SanctSound-NRS	47.76	-125.52	2014-2020	✓		

Table 4: Training splits by location and time

Split	Lat.	Lon.	Date	Train	Valid.	Test
ONC-CBCH.H1	47.76	-127.73	2022		✓	
ONC-CRIP	50.02	-125.24	2022		✓	
ONC-ECHO3.H1	49.04	-123.32	2022		✓	
ONC-ECHO3.H2	49.04	-123.32	2022		✓	
ONC-ECHO3.H3	49.04	-123.32	2022		✓	
ONC-ECHO3.H4	49.04	-123.32	2022		✓	
ONC-FGPD	48.81	-125.28	2022		✓	
ONC-PVIP	48.66	-123.49	2022		✓	
ONC-SCVIP	49.04	-123.43	2022		✓	
ONC-USDDL	49.08	-123.33	2022		✓	
OOI-CE04OSBP	44.37	-125.00	2022		✓	
SanctSound-OC02	48.49	-125.00	2022		✓	
ONC-FGPD			2023			✓
ONC-PSGCH			2023/08			✓
DCLDE2026-DFO-CRP						✓
DCLDE2026-DFO-WDLP	48.72	-125.1	2022			✓
DCLDE2026-DFO-WDLP	49.2	-123.3	2021			✓
DCLDE2026-DFO-WDLP	48.94	-123.4	2021			✓
DCLDE2026-DFO-WDLP	48.8	-123.4	2021			✓
DCLDE2026-Scripps	47.4	-124.7	2008-2012			✓
DCLDE2026-Scripps	47.5	-125.4	2011-2013			✓
DCLDE2026-SIMRES						✓
DCLDE2026-SMRU						✓
DCLDE2026-UAF	60.3	-147.0				✓
DCLDE2026-UAF	59.9	-151.8				✓
DCLDE2026-UAF	60.2	-147.8				✓
DCLDE2026-UAF	59.7	-149.5				✓
DCLDE2026-VFPA	-	-				✓
DeepAI			2017-2018			✓

Table 5: Validation and test splits by location and time

Table 6: Cetacean presence/absence classification specificity at 95% sensitivity on several held out test sets. Expressed as a percentage.

Model	Deepal ComParE [7]	DCLDE-2026 [47] excl. ONC & OS	FGPD 2023	PSGCH.H3 2023/08
LDA ^a	5.26	3.52	2.44	9.52
ROCCA ^b [5]	5.39	3.42 ^c	7.32	19.05
ANIMAL-SPOT [8]	19.48	3.77	2.44	0
Whisper-tiny [51] + LR	4.99	9.07	10.24	14.29
Whisper-tiny [51] + SGD	10.55	4.16	12.20	38.10
Whisper-tiny [51] + FT	7.28	18.6	9.76	76.19
Whisper-base [51] + FT	18.47	4.16	6.83	28.57
Whisper-large-v3 [51] + FT	15.55	8.86	12.68	85.71
Wav2Vec 2 [4] + FT	8.88	4.41	7.80	19.05

^aAlgorithm implemented in python to operate across platforms. Hyper-parameter tuning was conducted to outperform the PAMGuard defaults.^bAlgorithm implemented in python to operate across platforms. Hyper-parameter tuning was conducted to outperform the PAMGuard defaults.^cGenerated contours were randomly subsampled so that only 1000 contours were selected

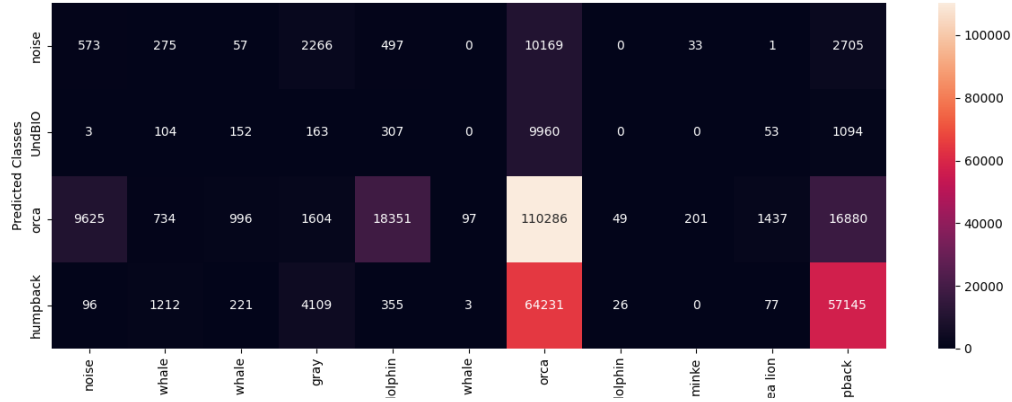
Table 7: Cetacean presence/absence classification AUPRC several held out test sets.

Model	Deepal ComParE [7]	DCLDE-2026 [47] excl. ONC & OS	FGPD 2023	PSGCH.H3 2023/08
LDA ^a	0.205	0.791	0.437	0.353
ROCCA ^b [5]	0.186	0.777 ^c	0.472	0.375
ANIMAL-SPOT [8]	0.379	0.765	0.450	0.282
Whisper-tiny [51] + LR	0.210	0.846	0.432	0.495
Whisper-tiny [51] + SGD	0.277	0.850	0.586	0.761
Whisper-tiny [51] + FT	0.232	0.923	0.666	0.919
Whisper-base [51] + FT	0.421	0.913	0.589	0.659
Whisper-large-v3 [51] + FT	0.367	0.922	0.626	0.807
Wav2Vec 2 [4] + FT	0.309	0.871	0.517	0.566

^aAlgorithm implemented in python to operate across platforms. Hyper-parameter tuning was conducted to outperform the PAMGuard defaults.

^bAlgorithm implemented in python to operate across platforms. Hyper-parameter tuning was conducted to outperform the PAMGuard defaults.

^cGenerated contours were randomly subsampled so that only 1000 contours were selected

**Figure 4: Confusion matrix for the actual classes from DCLDE 2026 vs. the whisper-tiny species classifications.****Figure 5: Confusion matrix for the actual ecotype classes from DCLDE 2026 vs. the whisper-tiny ecotype classifications.**

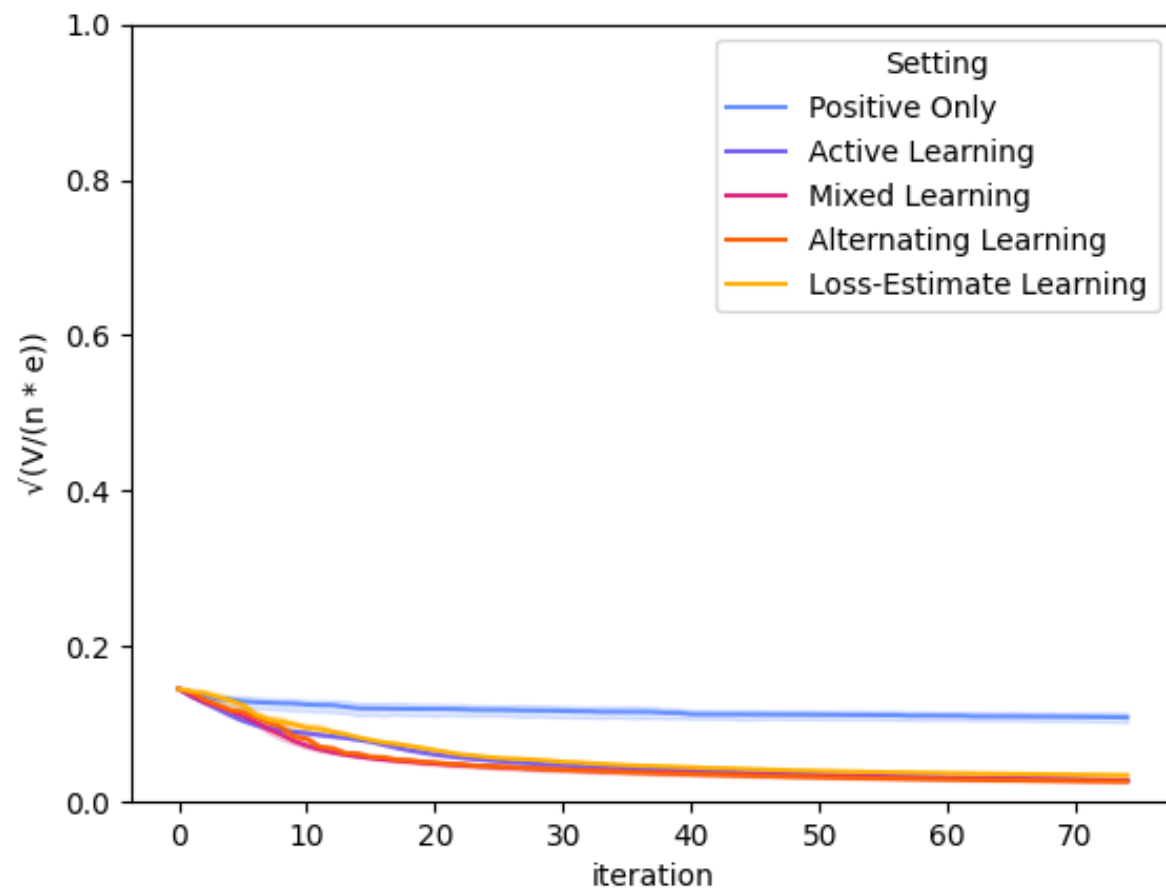
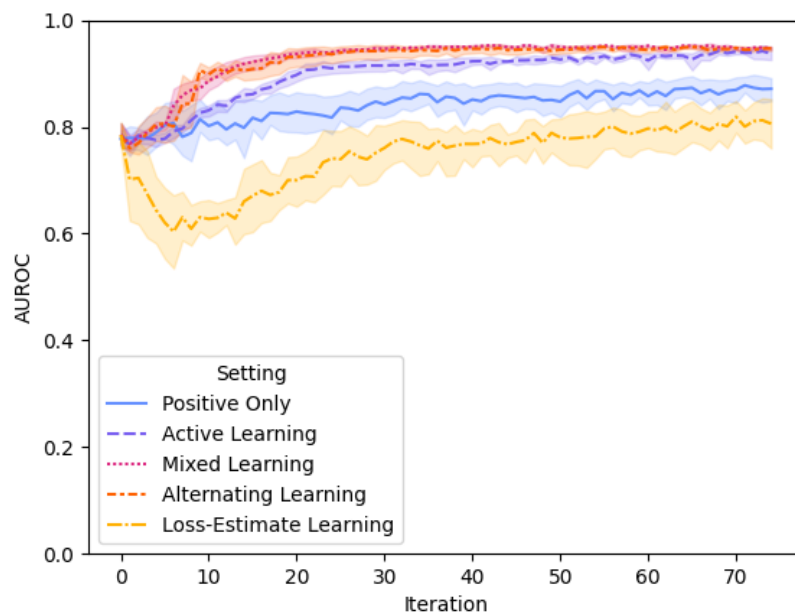
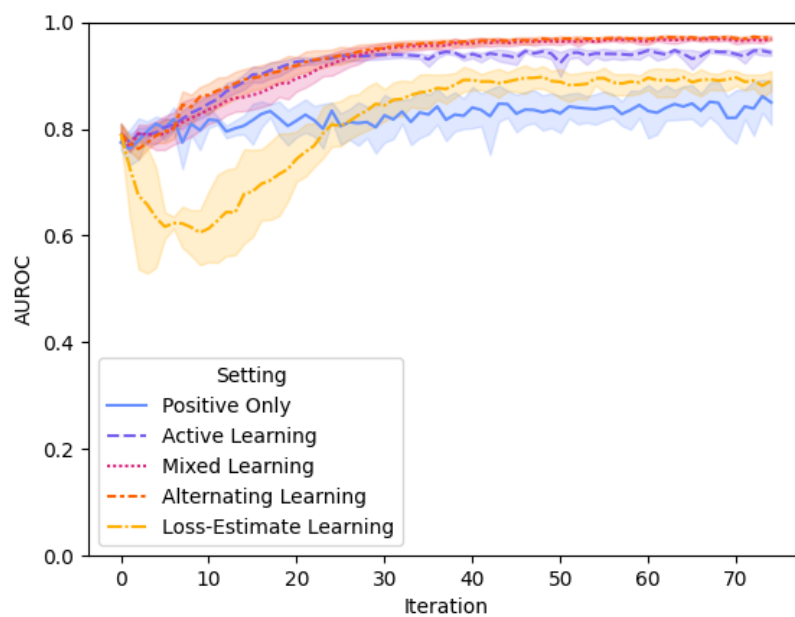


Figure 6: The criterion to determine which convergence regime is applicable.

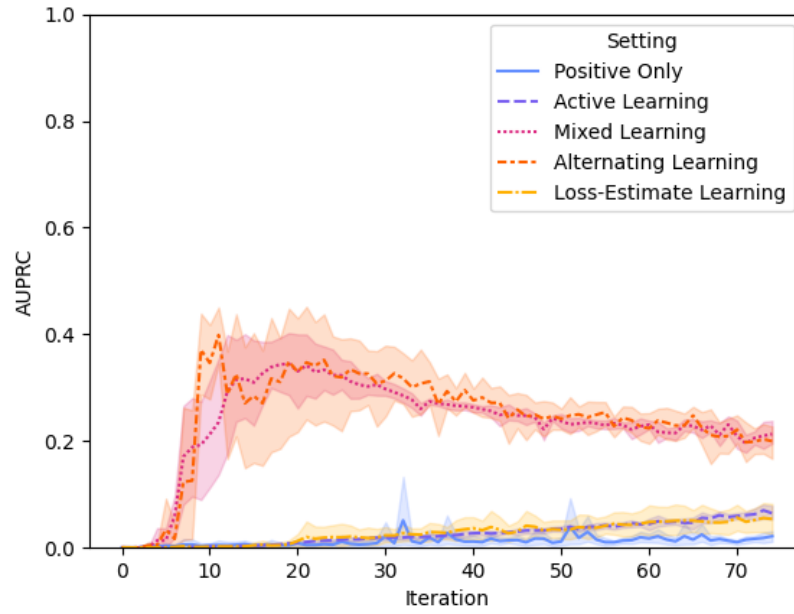


(a) AUROC vs active learning iterations with no label noise

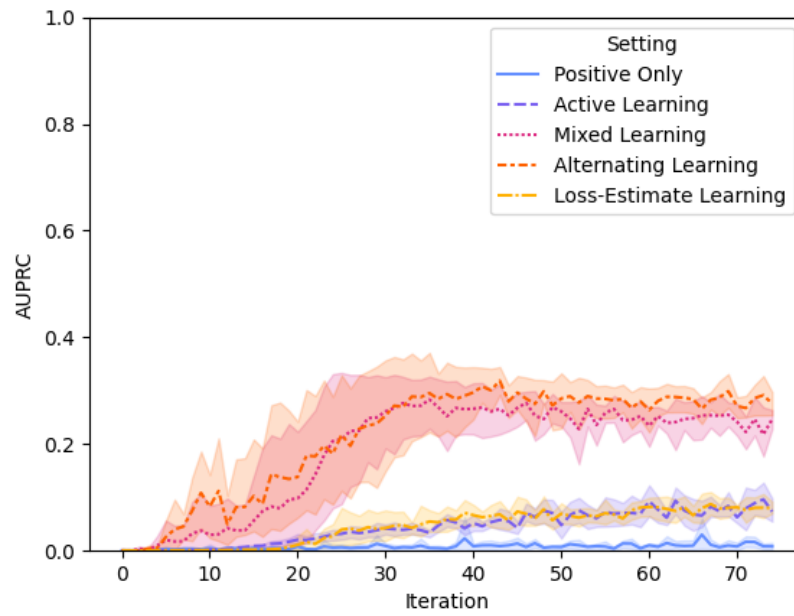


(b) AUROC vs. active learning iteration with label noise.

Figure 7: AUROC across active learning iterations for a variety of settings



(a) AUPRC vs active learning iterations with no label noise



(b) AUPRC vs. active learning iteration with label noise.

Figure 8: AUPRC across active learning iterations for a variety of settings