Title

Authors: Kaitlin Palmer, April Houweling, Jennifer Wladichuk, James Pilkington, Ruth Joy

# Abstract

# Introduction

# Methods

## Training Data

Annotated acoustic data for this study were obtained from the publicly available 2024 DCLDE annotated dataset (cite XXX, available at XXX). This dataset comprises acoustic annotations identifying communities and ecotypes of killer whales, specifically northern residents (NRKW), southern residents (SRKW), transients (TKW), and offshore killer whales (OKW). The annotation file additionally includes labels for humpback whales, undetermined biological sounds (e.g., dolphin whistles and other cetacean vocalizations), and abiotic noise sources, notably vessel noise. The full dataset covers approximately 13 years of annotations provided by ten organizations and recorded at 31 locations ranging from Oregon, USA through Alaska USA with the majority of the sites located within the Strait of Georgia and Salish Sea.

Initially, we attempted to build classifiers using all annotations that had confidently identified ecotypes, resulting in large multi-class training sets exceeding 30,000 annotations. However, these initial classifiers were neither sufficiently accurate nor computationally practical, prompting a more focused approach leveraging existing expert knowledge of killer whale call types.

The granularity of label detail within this dataset varies significantly. While most killer whale annotations specify the ecotype, others are broadly classified simply as killer whale, or not. Among the 86,823 confirmed killer whale annotations, only 5,032 are categorically labeled according to call type from call catalogues—comprising 3,419 SRKWs, 1,374 NRKWs, 105 TKWs, and 134 OKWs. Catalogues for transient and offshore killer whales, housed at the Canadian Department of Fisheries and Oceans, remain unpublished. Despite the limitations this poses for cross-validation of call types, we have included these labels to capture a broader diversity of vocalizations than would be possible without such categorization, recognizing that these categories may be subject to future revision.

We therefore constructed a targeted, balanced training dataset emphasizing high-quality, confidently annotated pulsed calls with an emphasis on ensuring call types, where known, were robustly included in the dataset. First, annotations were rigorously filtered to include only killer whale vocalizations assigned with high confidence to specific ecotypes (Southern Resident [SRKW], Transient [TKW], and Offshore [OKW]). Uncertain or ambiguous call annotations (indicated by question marks or uncommon vocalizations such as whistles, buzzes, and rasps) were excluded. We then identified pulsed calls that contained a call type annotation within each ecotype. Several of the annotations contained multiple call types indicating that multiple animals were present in the same annotation. For these cases, calls were annotations were duplicated and such that it was represented in each call category. Some killer whale call types are further delineated into subtypes. Sub-type was not considered in building the dataset. For call types with fewer than 100 available examples, we applied data augmentation by randomly shifting the temporal center of each annotated segment within ±50% of its original duration.

Following call-type-level balancing, additional vocalizations annotated only at the ecotype level were randomly sampled to ensure equal representation (~4600 examples) across each ecotype. Lastly, two supplementary classes were established: humpback whales, included due to known acoustic confusion with killer whale calls, and a diverse background class representing abiotic noises and undetermined biological sounds. These classes were created using stratified sampling across providers to maintain ecological representativeness. The resulting balanced and curated dataset thus optimized classifier performance while reflecting realistic behavioural and environmental diversity.

## Evaluation Data

Classifier performance was assessed using a fully independent evaluation dataset collected as part of the Malahat deployment (JASCO Applied Sciences) from four distinct hydrophone stations located within the Salish Sea, spanning latitudes 48.50° to 48.78° N and longitudes 123.21° to 124.54° W. Recordings were obtained using Autonomous Multichannel Acoustic Recorders (AMAR) deployed at depths ranging from 74 to 237 meters, sampling at a rate of 64 kHz. The evaluation dataset comprises annotations from acoustic data collected continuously between October 2015 and February 2017.

Annotations used for model validation included pulsed calls of Southern Resident (SRKW, total annotations = 11,135) and Bigg’s (Transient, total annotations = 2,491) killer whales, along with humpback whale (HW, total annotations = 1,380), abiotic (vessel and environmental noise, total annotations = 5,947), and undetermined biological sounds (total annotations = 3). Offshore killer whales (OKW), northern resident killer whales (NRKW), and other rare classes were not detected in this evaluation dataset. Notably, annotations in the Malahat dataset were not comprehensive, meaning that not all killer whale calls within the recorded files were annotated. Therefore, it was not feasible to apply the classifier exhaustively to all recorded audio data, limiting validation strictly to the annotated subset. The uneven distribution across ecotypes and supplementary classes reflects typical ecological conditions and underscores the realism of the performance evaluation. Dataset selection was run in R version 4.3.3.

## Model Selection and Training

We utilized the BirdNET Analyzer 1.3.1 with model 2.4 (Kahl et al. 2021, v2.4) for this analysis. BirdNET is an open-source convolutional neural network (CNN) framework specifically developed for rapid deployment in bioacoustic applications. BirdNET has been widely validated in terrestrial contexts, achieving high accuracy in bird species detection and classification tasks, even under challenging acoustic conditions (Kahl et al. 2021). Importantly, BirdNET has proven adaptable to marine bioacoustic applications, successfully supporting classification tasks involving marine mammals and underwater acoustic signals (Kahl et al. 2021, Ghani et al. 2023). The selection of BirdNET was motivated by its flexibility, ease of use, built-in optimization tools for automated hyperparameter tuning, and demonstrated effectiveness in real-world ecological monitoring scenarios, ensuring rapid and reliable implementation for time critical conservation applications.

Birdnet requires all data to be sampled at 48khz and represents audio as a single channel mel spectrogram created with Fast Fourier Transform (FFT) window size of 10.7 ms (512 samples) and an overlap of 25%, representing a 8 ms time step but limits the maximum frequency range to 15khz. This range is consistent with the frequency range of many cetaceans including killer whales. For all audio files in the training and evaluation data we re-sampled to 48khz, downsampling higher sample rate audio and as well as interpolating data recorded at lower sample rates. Filtering and processing was done using the load function in librosa (McFee et al. 2025) and setting the target sample rate to 48khz.

**V2.4,**

## Model Evaluation

Models were evaluated using three complementary approaches. First, model predictions were assessed using precision-recall curves generated for each target class in a one-versus-all framework wherein only one class was applied to each validated segment. This approach assumes that only one of the selected classes (or background) is present, regardless of confidence score.

However, models output allow for multiple species to be present in the same segment. Moreover BirdNET confidence scores are class-specific; equivalent scores may lead to disparate outcomes in important metrics like precision, recall, and true positive rates (Wood and Kahl 2024). Thus, for the second evaluation technique we employed logistic regression analysis to relate the classifier's confidence scores to their probability of correct classification following the recommendations of Wood and Kahl 2024. From the logistic analysis we derived the confidence scores corresponding to a 90% correct classification probability (P90) for each target class in the evaluation dataset (humpback whales, SRKW, and TKW). P90 scores were then used in the third evaluation technique that built confusion matrices based. In this analysis the model output across all classes was retained for every segment in the validation set.

on P90 thresholds were used to construct confusion matrices by classifying predictions at or above these thresholds as positive detections for the corresponding class, while predictions below these thresholds were classified as background. Recall values—defined as the proportion of correctly classified instances relative to all ground-truth instances at the specified P90 thresholds—were calculated and annotated along the diagonal of each confusion matrix. This visualization approach enables practitioners to select appropriate thresholds balancing acceptable recall with classification accuracy according to management or monitoring needs. All analysis was done in Python 3.9.21

# Results

Through the process of building the training data, 26 SRKW, nine transient, and five offshore killer whale calls were identified. There was considerable variation in the identification of calls types across the dataset. The most common SRKW calls were S04 and S01 with 864 and 684 calls identified and least common were the S31, S32 and S41 with one, five and five examples. There were overall fewer transient calls with associated call types with a total of 112 in total. There were 131 total offshore killer whales with call types with S17 having 110 calls and S02 having just one example (Table 1).

Table 1 Calls identified in the initial training dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SRKW | | Transient | | Offshore | |
| Call Id | Count | Call Id | Count | Call Id | Count |
| S01 | 684 | T01 | 27 | OFF02 | 1 |
| S02 | 214 | T02 | 21 | OFF07 | 16 |
| S03 | 66 | T03 | 7 | OFF17 | 110 |
| S04 | 864 | T04 | 19 | OFF19 | 4 |
| S05 | 119 | T07 | 11 | OFF30 | 3 |
| S06 | 32 | T08 | 5 |  |  |
| S07 | 24 | T11 | 5 |  |  |
| S08 | 21 | T12 | 11 |  |  |
| S10 | 153 | T13 | 6 |  |  |
| S12 | 23 |  |  |  |  |
| S13 | 8 |  |  |  |  |
| S14 | 3 |  |  |  |  |
| S16 | 301 |  |  |  |  |
| S17 | 121 |  |  |  |  |
| S18 | 67 |  |  |  |  |
| S19 | 243 |  |  |  |  |
| S22 | 6 |  |  |  |  |
| S31 | 5 |  |  |  |  |
| S32 | 1 |  |  |  |  |
| S33 | 17 |  |  |  |  |
| S36 | 165 |  |  |  |  |
| S37 | 36 |  |  |  |  |
| S40 | 10 |  |  |  |  |
| S41 | 5 |  |  |  |  |
| S42 | 95 |  |  |  |  |
| S44 | 213 |  |  |  |  |

There were sufficient SRKW calls annotated to call level to include in the classification task and as such no SRKW annotations without call types were added to the final dataset. Only two of the datasets within the larger DCLDE data provided call type for transient killer whales. As such, 3700 TKW calls not identified to call type were added into the dataset. This include data from the Canada Department of Fisheries and Oceans Cetacean Research Program (DFO CRP; n = 2247), Canada Department of Fisheries and Whale and Dolphin Listening Program (DFO WLDP; n= 265), and JASCO/Vancouver Frasier Port Authority (n =9), Ocean Networks Canada (ONC; n =103), Scripps Institute of Oceanography (SIO; n=701), and the University of Alaska Fairbanks (UAF; n = 375). Notably there is considerably les information known about many of these datasets and the transient community to which the animals belong is not always known. Offshore annotations added to the offshore call type data were from the DFO CRP (n =1622), ONC (n=546), SIO (n=522) and UAF (n=103).

The resulting BirdNet model contained 2048 hidden units, used 0.33% dropout with a batchsize of 64. The learning rate was 0.005 and cropping was not used as 3 sec segments were pre-processed for the validation. Upsampling mode was linear, and the ration was 0. The model used mixup but did not use label smoothing.

A graph of a number of different colored lines

AI-generated content may be incorrect. A graph of a diagram

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

Figure 1 Classifier confusion matrix with P90 threshold valuds for southern resident, transient, and humpback whales. No offshore killer whales were available for evaluation.

# Discussion

**Precision Recall, one vs many**

**Probabilistic thresholds**

**Confusion Matrix**