Machine Learning Approaches for Automated Classification of Southern Resident Killer Whale Echolocation Clicks in the Salish Sea: Lasso Regression or Ensemble Tree-Based Methods?

by

April Houweling

B.Sc., University of British Columbia, 2019

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Sciences
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2025

Declaration of Committee

Name: **April Houweling**

Degree: **Master of Science**

Thesis title: **Machine Learning Approaches for Automated Classification of Southern Resident Killer Whale Echolocation Clicks in the Salish Sea: Lasso Regression or Ensemble Tree-Based Methods?**

Committee: **Chair:** Liangliang Wang
Professor, Statistics and Actuarial Science

Ruth Joy
Co-Supervisor
Assistant Professor, Environmental Science
Adjunct Professor, Statistics and Actuarial Science

Robert Tibshirani
Co-Supervisor
Professor, Department of Statistics
Stanford University
Adjunct Professor, Statistics and Actuarial Science

Owen Ward
Internal Examiner
Assistant Professor, Statistics and Actuarial Science

David Stenning
Committee Member
Assistant Professor, Statistics and Actuarial Science

Abstract

Several commercial shipping routes in the Salish Sea pass through critical habitat of the endangered Southern Resident killer whale (SRKW), a population increasingly threatened by elevated underwater noise levels. To mitigate the impacts of vessel noise on SRKW in a cost-effective manner, automated algorithms must differentiate SRKW from a sympatric population of West Coast Transient (WCT) killer whales. This study utilizes a multi-year dataset from the Boundary Pass Underwater Listening Station. Twenty-one machine learning classifiers were trained and evaluated using logistic regression with varying lasso-penalization or ensemble tree methods. A population-specific killer whale echolocation click detector-classifier is presented that integrates a Teager-Kaiser click detector with an eXtreme Gradient Boosting classifier to predict SRKW or WCT presence using 12 acoustic features (Event Level Precision: 0.83, Recall: 1.00, F_2 Score: 0.96, MCC: 0.85, ROC-AUC: 0.98). The findings of this study support the development of automated vessel alert systems for SRKW in the Salish Sea.

Keywords: EXtreme Gradient Boosting; Pre-trained Lasso; Underwater Acoustics; Model Selection; Feature Selection; Species Protection

Dedication

I dedicate this thesis to David Hannay, Chief Science Officer at JASCO Applied Sciences, whose support provided me with the opportunity to conduct statistical acoustic research on hundreds of thousands of killer whale echolocation clicks recorded in Boundary Pass. The Underwater Listening Station in Boundary Pass has continuously captured high-quality acoustic data since June 2020, creating an unprecedented resource for bioacoustic research. His contributions have been instrumental in making this work possible.

I also dedicate this thesis to my fiancé, Tom Fruin, for his unwavering support, optimism, and encouragement throughout my theoretical statistics coursework at Simon Fraser University. His belief in my abilities and constant motivation have been invaluable as I pursue my passion for studying cetacean acoustics.

Acknowledgements

I extend my deepest gratitude to my supervisors, Dr. Ruth Joy and Dr. Robert Tibshirani, for their invaluable experience, guidance, and support throughout my MSc journey, along with review of my thesis. Pursuing this degree has been a true test of my abilities as a trained biologist and bioacoustician, and Ruth and Rob's encouragement has been a constant source of motivation at every stage of this endeavor.

I am also grateful to the Humans and Algorithms Listening to Orcas (HALLO) team, particularly Dr. Kaitlin Palmer, Dr. Harald Yurk and Dr. Michael Dowd, for their expert knowledge and experience in the statistics and bioacoustic fields.

I extend my thanks to all those who funded this research, specifically JASCO Applied Sciences (JASCO), Mitacs, and Simon Fraser University. My appreciation extends to Transport Canada (TC), the Vancouver Fraser Port Authority (VFPA), and JASCO for providing access to a multi-year acoustic dataset and to JASCO for providing a dedicated workspace at their Saturna Island office. A special thanks to JASCO employees, Dr. Katie Kowarski, Dr. Jennifer Wladichuk, Connor Grooms, Forest Stothart, Dorian Visserot, Ash Hill, Briand Gaudet, and Trent Johnson, for their expert acoustic knowledge, continuous improvements of JASCO's PAMlab software, as well as an existing odontocete click detector-classifier, which have been invaluable to this research.

I would like to acknowledge Saturna Island Marine Research and Education Society (SIMRES), specifically Susie Washington-Smyth and Maureen Welton, for welcoming me to Saturna Island and sharing acoustic data from the SIMRES East Point hydrophone. In addition, I would like to thank Lucy Quale for sharing visual sightings from Spyhopper and Scott and Val Viers for sharing acoustic data from the Orcasound Lab hydrophone on San Juan Island.

A heartfelt thank you to Professor Dr. Rachel Altman for her dedication to her students and her commitment to ensuring thorough understanding of statistical concepts. Her guidance was instrumental in helping me navigate my first year of coursework successfully. Finally, I would like to acknowledge my fellow graduate students, Yulia Kozhevnikova, So Yeon Park, and Roxana Darvishi for their contributions to the development of a preliminary Bayesian averaged classification model and a killer whale echolocation click detector incorporating normalizing flows, stochastic gradient descent, Hamiltonian Monte Carlo, and Gibbs sampling with particle filtering.

Table of Contents

List of Tables

List of Figures

Glossary

10 dB down bandwidth Difference between the 10 dB downwards high and low frequencies (kHz). *pages.* 15, 39

10 dB down high frequency From the peak frequency, the algorithm stepped upward in frequency until the power spectral density was 10 dB below the peak (kHz). *pages.* 15, 38, 39

10 dB down low frequency From the peak frequency, the algorithm stepped downward in frequency until the power spectral density was 10 dB below the peak (kHz). *pages.* 15, 38, 39

3 dB down bandwidth Difference between the 3 dB downwards high and low frequencies (kHz). *pages.* 15, 39

3 dB down high frequency From the peak frequency, the algorithm stepped upward in frequency until the power spectral density was 3 dB below the peak (kHz). *pages.* 15, 38, 39

3 dB down low frequency From the peak frequency, the algorithm stepped downward in frequency until the power spectral density was 3 dB below the peak (kHz). *pages.* 15, 39

broadband Sound that is spread across a wide range of frequencies [41]. *pages.* 4, 5, 14, 48

burst-pulse Discrete isolated series of high repetition rate pulses produced at a relatively constant rate, frequently exceeding 300 pulses per second [41, 4]. *pages.* 3, 14

buzz A rapid increase in the repetition rate of clicks emitted by an actively echolocating animal [41] ($\leq$ 10 ms inter-click interval [73]). *pages.* 4, 14

centroid frequency Frequency dividing a spectrum in halves of equal energy (i.e., frequency at which 50% of the selection energy was distributed above and below this point; kHz) [41]. *pages.* 15, 38, 39

click rate Number of echolocation clicks (detected and classified as a species-level killer whale by the JASCO odontocete click detector) per five-minute WAV file. *pages.* 15, 20, 38, 43, 45, 53, 54

duration End time minus start time of the detection box. *pages.* 15, 16, 38

echolocation clicks A forward projected signal of short duration whose primary function is echo ranging, target detection, or discrimination [41]. *pages.* xiv, 1, 4–10

ecotype A group within a species that is specifically adapted to certain local environmental conditions and foraging techniques . *pages.* 1, 2

event A continuous segment of acoustic recording with killer whale vocalizations, where no more than 30 minutes pass between vocalizations. *pages.* 10, 12, 15, 33, 36

high frequency Frequency at the top of the detection box (kHz). *pages.* 15, 38, 39, 45

low frequency Frequency at the bottom of the detection box with a 5 kHz high pass filter (kHz). *pages.* 15, 38, 39, 45

mask Reduced ability to detect, recognize, or understand sounds of interest because of interference by other sounds [41]. *page.* 5

multi-path Phenomenon when a single signal reaches a receiver in two or more paths.. *page.* 8

narrow-band Sound that is concentrated in a narrow range of frequencies [41]. *page.* 3

odontocete The taxonomic division of cetaceans (whales, dolphins, and porpoises) that comprises the toothed whale. *pages.* xiv, 5, 8–10, 14, 16, 46–49, 51

off-axis When a signal is received from any other direction than directly in front of, or behind, the whale. *pages.* 8, 9, 17

on-axis When a signal is received directly in front of, or behind, the whale. *pages.* 8, 17, 52

peak frequency Centre frequency of the band with the highest spectral density (kHz) [41]. *pages.* 15, 39

phonic lips Air spaces within the nasopharyngeal passage that rapidly open and close thereby producing sound [41]. *page.* 3

pulsed call A series of pulses with high repetition rates that appear as a tonal signal, typically with harmonics . *pages.* 2, 4, 5, 12, 14, 48, 56

signal-to-noise ratio Ratio of signal power to noise power [41], defined as $\mathrm{SNR} = 20 \times \log_{10}\left(\frac{\sqrt{\int_{T_2}^{T_3} s(t)^2\,dt}}{\sqrt{\int_{T_0}^{T_1} n(t)^2\,dt}}\right)$. Here, $s(t)$ is the pressure time-series of the echolocation click, and $n(t)$ is the pressure time-series of the preceding noise window. T_2 and T_3 define the start and end of the echolocation click window, and T_0 and T_1 define the start and end of a noise window of equal duration and bandwidth, offset by 2 ms from the echolocation click window. *page.* 12, 17

Chapter 1

Introduction

The growing number of vessels worldwide has raised underwater ambient noise levels near shipping routes [30, 51, 43]. Cetaceans, a group including whales, dolphins and porpoises, rely almost exclusively on sound for navigation, communication, predator detection, foraging, and reproduction, and are vulnerable to negative impacts of elevated ambient noise levels [21, 69, 39, 40, 64]. Conservation activities related to monitoring underwater noise levels in critical habitat include quantification of vessel noise [22, 37] and real-time detection of marine mammals through visual and acoustic monitoring [11]. Visual sightings provide valuable information on species, population, and occasionally individual presence [12], but they are limited to small areas, daylight hours, good weather conditions, and surface-active animal behaviours. In contrast, passive acoustic monitoring (PAM) via underwater microphones, or hydrophones, is a tool to autonomously monitor the presence of marine mammals and assess their temporal and spatial distributions [20, 45, 46]. PAM relies on detecting signals produced by marine mammals, such as songs, calls, or whistles used between individuals, or echolocation clicks used to forage and navigate. Hydrophone-facilitated PAM operates continuously, recording ‘acoustic data’ that includes animal vocalizations day and night, in all weather conditions. However, distinguishing between acoustically similar populations and species requires both expert knowledge of the groups and classification algorithms trained with observations from each group.

The Northeast Pacific hosts numerous shipping routes that are frequented year-round by many species of marine mammals, including three acoustically distinct ecotypes of killer whales (*Orcinus orca*): fish-eating Residents, marine mammal-eating Transients or Bigg’s, and shark-eating Offshores [8, 25, 53, 58]. These ecotypes are further divided into populations. The Resident ecotype consists of the Northern Resident (NRKW) and Southern Resident killer whale (SRKW) populations [24], while the Transient ecotype consists of the West Coast Transient (WCT) killer whale population, which is further divided into the Inner Coast Transient (ICT) and Outer Coast Transient (OCT) sub-populations [67, 18] (Figure 1.1). The SRKW, NRKW, and WCT populations are further divided into socially, acoustically, or genetically distinct taxonomic groups (sub-populations, clans, subclans, pods, matrilines, etc.). The Offshore killer whale (OKW) is the least studied of the northeastern Pacific ecotypes and if or how these animals differentiate into populations, clans, etc. is unknown [26]. In this study, we focus on the SRKW and WCT [1] populations (Figure 1.1) that inhabit SRKW critical habitat in the Salish Sea. NRKW typically range farther north, while OKW inhabit a range much farther offshore and thus were not considered in this study.

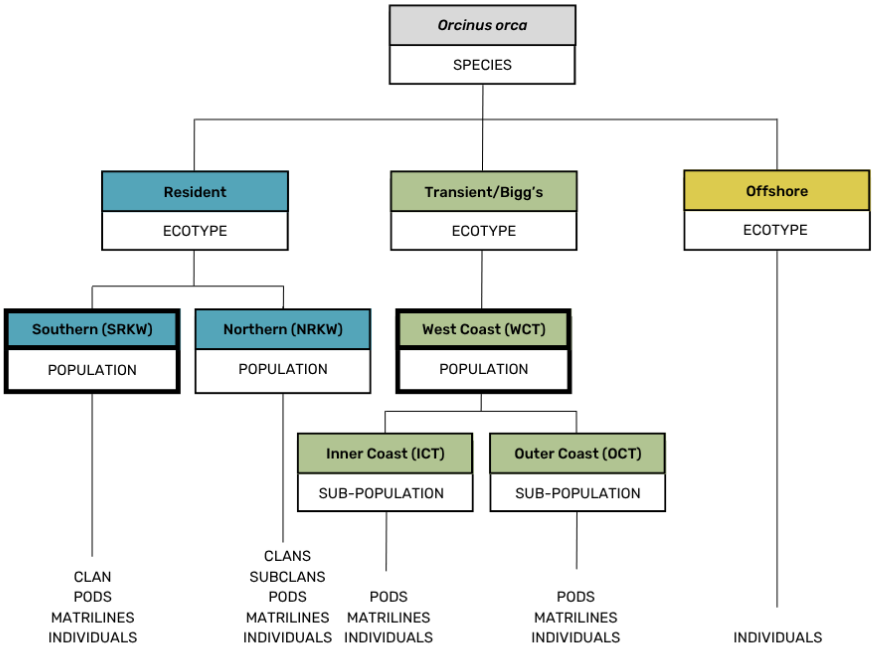

Figure 1.1: A hierarchical taxonomic classification of observed *Orcinus orca* ecotypes, populations (the Southern Resident (SRKW), Northern Resident (NRKW), and West Coast Transient (WCT) killer whale), and sub-populations (the Inner-Coast Transient (ICT) and Outer-Coast (OCT) killer whale) in the northeast Pacific based on geographical, genetic, ecological, and acoustic distinctions.

[1] The WCT population studied likely only includes individuals from the ICT sub-population, yet this distinction requires OCT pulsed call experts or photo identification of all vocalizing individuals.

SRKW are listed as endangered under Canada's Species at Risk Act (SARA) [47] and the United States' Endangered Species Act (ESA) [54], with only 73 individuals remaining and their population trajectory continuing to decline [72]. WCT killer whales are listed as a species of 'Special Concern' under SARA, but are not listed under ESA. This classification in Canada indicates that the population is not currently considered endangered or threatened, but may become so if factors that threaten its survival and recovery are not addressed. Under SARA, federally designated 'critical habitat' for killer whales receives legally mandated safeguards against its destruction, including threats like excessive underwater vessel noise. Given these vulnerabilities and legal protections, proactive management of vessel traffic in the Salish Sea is essential. Voluntary slowdowns by commercial vessels have shown to be effective in reducing underwater noise levels [42, 68] and the Vancouver Fraser Port Authority (VFPA) has expressed interest in implementing automated acoustic detection systems for SRKW, enabling vessels to slow down or reroute in real-time when the whales are detected. This dynamic approach would help mitigate acoustic disturbances associated with vessel traffic.

Killer whales rely heavily on sound as vision is limited in underwater environments. Underwater sound has a very low rate of attenuation relative to in-air sound propagation, thus it can travel longer distances and with less distortion, allowing marine mammals to detect and interpret acoustic cues more effectively than visual ones. Consequently, killer whales have evolved specialized auditory and vocal anatomy to adapt to their environment, along with an extremely rapid auditory nervous system response [1]. Their phonic lips, specialized structures used to generate sound, and highly developed inner ears enable them to produce and perceive a wide range of sounds for navigation, communication, and social interactions [49]. Killer whales produce three types of acoustic signals: Echolocation clicks, pulsed signals, and tonal whistles (Figure 1.2). Tonal whistles are continuous narrow-band frequency modulated tones with few to no harmonic components, while pulsed signals, including pulsed calls and burst-pulses, are generated as a series of pulses with high repetition rates and appear tonal [23]. Both pulsed signals and tonal whistles are primarily used for

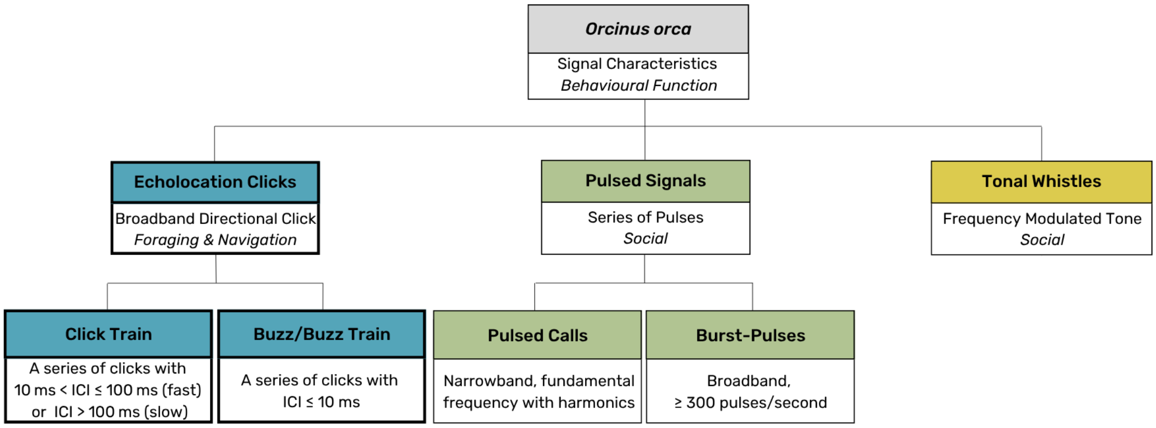

Figure 1.2: A hierarchical classification of *Orcinus orca* vocalizations based on signal characteristics and behavioral function. Vocalizations are divided into three categories: Echolocation clicks, pulsed signals, and tonal whistles. Narrower categories are defined by inter-click interval (ICI), pulse rates, and frequency bandwidth.

socializing and communication [24]. Echolocation clicks, which are broadband short-duration directional signals, are primarily used for echolocation during navigation and prey detection [23, 1]. Echolocation clicks can be produced as single isolated clicks or in structured sequences called trains. These trains can be evaluated based on inter-click-interval (ICI) as slow (ICI $>$ 100 ms), fast (10 ms $<$ ICI $\leq$ 100 ms), or buzz trains (ICI $\leq$ 10 ms) [4, 73, 41]. Buzzes (i.e., buzz trains) are an important indicator of foraging activity [73, 64], and due to their short ICI, buzzes may be observed as tonal depending on the spectrogram settings (e.g., time resolution and Fast Fourier Transform (FFT) size). Additionally, terminal buzzes may indicate prey capture attempts, as they often occur in the final approach phase of hunting [33]. Figure 1.3 displays a visual waveform and spectrogram representation of each type of killer whale vocalization. Understanding these signals and their potential use is essential for real-time acoustic monitoring and conservation efforts in the Salish Sea. In summary, broadband signals, such as echolocation clicks, contain a wide range of frequencies, allowing for high-resolution detection of objects, while lower frequency sounds, such as pulsed calls, occupy a narrower frequency range and are likely more suited for long-range communication.

Killer whale vocalizations are diverse and vary between populations and sub-populations. Early research identified that SRKW and WCT killer whales produce distinct stable pulsed call repertoires. These vocalizations are unique to each group and can be used for reliable

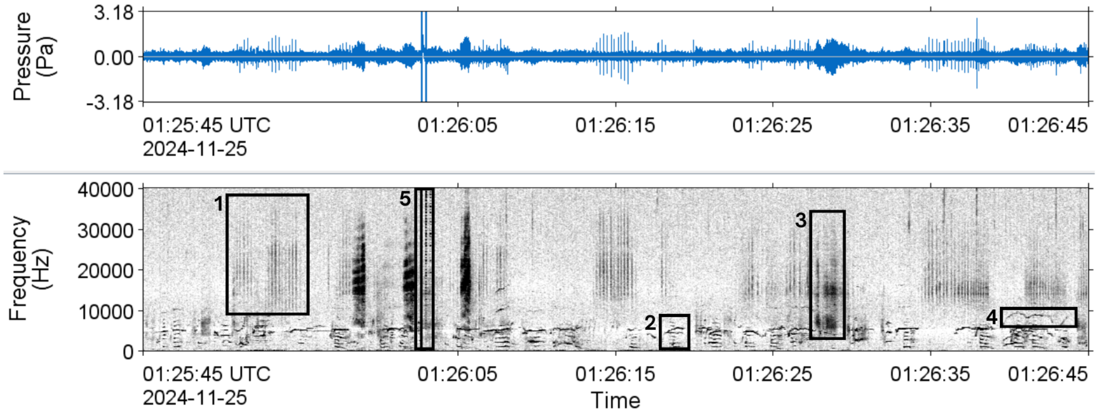

Figure 1.3: Waveform (top) and spectrogram (bottom) of Southern Resident killer whale (SRKW) echolocation clicks (1), pulsed call (2), buzz (3), and whistle (4), and the Underwater Listening Station (ULS) Acoustic Doppler Current Profiler (ADCP) ping (5) recorded on November 25, 2024 (UTC; 2 Hz frequency resolution, 0.125 s time window, 0.03125 s time step, Hanning window, normalized across time).

classification [23]. More recently, research has shown ICT and OCT sub-populations have subtle distinct pulsed call dialects [58]. These pulsed calls allow trained bioacousticians to differentiate killer whale populations and sub-populations based on their vocal signatures [23]. Despite these known distinctions, current deep learning detector-classifiers often struggle to differentiate killer whale pulsed calls from overlapping or confounding signals, especially evolving humpback whale (*Megaptera novaeangliae*) [31, 32] songs which can occupy similar frequency ranges to those of killer whale pulsed calls, and tonal or whistling vessel noise generated by ship propellers [27, 28, 44, 7]. Classification becomes even more challenging in the presence of multiple calling individuals within a pod, where overlapping vocalizations can obscure individual calls. Additionally, low-frequency vessel noise can acoustically mask killer whale signals, making these vocalizations harder to detect.

Killer whale echolocation clicks have been studied less than pulsed calls. However, given their non-tonal, short-duration, and broadband acoustic characteristics, they may be particularly well-suited for automated detection and classification. Their broadband nature enhances their detectability, especially when a high pass filter is applied to remove low-frequency vessel noise. The frequent use of echolocation clicks by SRKW for navigation and foraging also makes them strong candidates for acoustic detection and classification in ongoing monitoring and mitigation efforts [38]. Moreover, the risk of misclassification is limited to a small number of classes as few echolocating odontocetes inhabit the Salish Sea.

Potential confusion is likely to occur mainly with WCT, while the relatively rare Pacific white-sided dolphin (*Lagenorhynchus obliquidens*) and NRKW are of minimal concern due to their infrequent occurrence. SRKW echolocation clicks are reported to have higher peak frequencies, occur in longer echolocation click trains, and are used more frequently than WCT echolocation clicks, potentially reflecting their differences in prey types and hunting strategies [5, 23, 48, 19]. Despite this broader understanding, research on WCT echolocation clicks is often limited by insufficient acoustic data for robust analysis or training of detector-classifiers [23, 48, 50]. However, a long-term acoustic dataset from critical SRKW habitat in the Salish Sea, encompassing the Strait of Georgia, the Strait of Juan de Fuca and Puget Sound (Figure 1.4), provides a valuable resource, containing hundreds of thousands of killer whale echolocation clicks documented over multiple seasons and years. This dataset is critical as reliable machine learning models need to be trained on data that ideally represents the full range of the animals' targeted call type, and those vocalizations should remain relatively static over time [60].

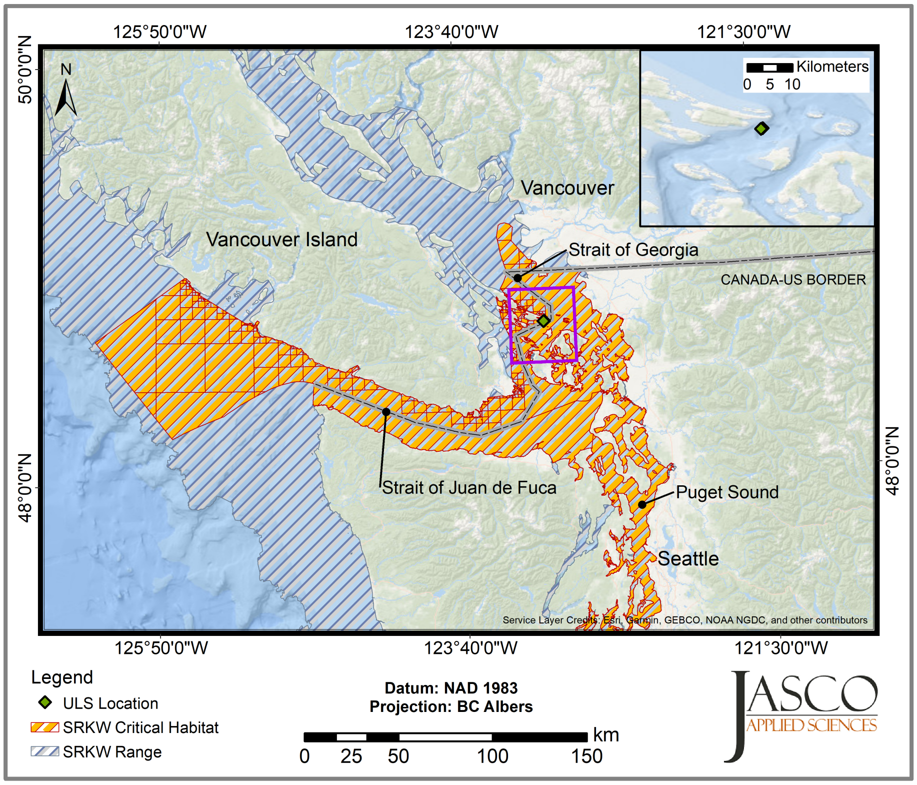

Figure 1.4: Map of the Underwater Listening Station (ULS) in the Salish Sea with Southern Resident killer whale (SRKW) critical habitat and range.

The Boundary Pass Underwater Listening Station (ULS) is a state-of-the-art PAM station equipped with hydrophone arrays and advanced signal processing technology, designed to capture high-resolution sounds from marine mammals and vessel traffic. The ULS (Frames A and B) was installed in June 2020 between Saturna and Waldron Islands, sponsored by Transport Canada (TC) and the VFPA, and is operated by JASCO Applied Sciences (JASCO; Figure 2.2). This station records at an ultra-high sample rate (512,000 samples per second), capturing the fine temporal details of echolocation clicks even at high frequencies. The dataset also includes recordings of a busy shipping lane where noise level fluctuates significantly. The ULS is located in 'critical habitat' for SRKW and is used frequently by the WCT killer whales, providing a rich dataset to investigate acoustic differences between the two populations, and to develop automated classifiers capable of distinguishing these populations in near real-time.

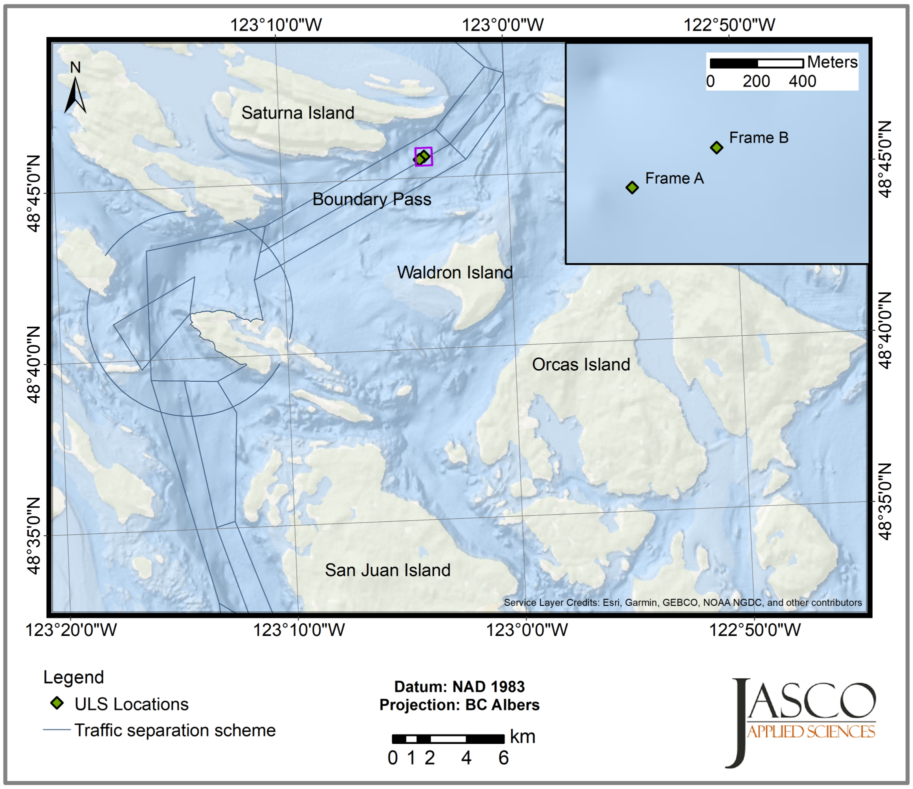

Figure 1.5: Map of the location of the Underwater Listening Station (ULS) frames in Boundary Pass.

Although echolocation clicks are well suited for detection and classification, analyzing click signals within acoustic datasets presents several challenges, particularly due to other impulsive sounds and multi-path propagation [35]. This occurs when an acoustic signal reflects off surfaces like the seafloor, ocean surface, or air sacs within an odontocete's noise production organs [36], causing multiple versions of the same signal to reach the receiver (hydrophone). Multi-path propagation can distort signals by leading to inverted or overlapping echoes that affect both the temporal and spectral properties. Specifically, the received sound level of a killer whale echolocation click (the magnitude per frequency recorded) often exhibits variability due to factors such as source directionality (e.g., whale orientation as on-axis or off-axis relative to the hydrophone), source distance (e.g., animal movement relative to the hydrophone), and environmental conditions (e.g., water temperature, salinity, depth, current, boundaries, background noise, etc.). Additionally, killer whale echolocation clicks differ in their energy distribution depending on behavioural contexts, such as navigation or foraging [49]. This variability highlights the need for robust classifier design and training that accounts for differences in echolocation click usage, source location and directionality, and variations in the underwater acoustic environment. Doing so ensures that the model learns the true acoustic properties of each echolocation click rather than noise artifacts present in the recordings [59].

Despite these challenges, echolocation clicks are a strong candidate for automated detection and classification due to their distinct acoustic characteristics. Echolocation click characteristics such as spectral energy ratios, peak frequencies, and zero-crossing characteristics have been successfully used in previous studies to classify other species of odontocete echolocation clicks. For example, random forest models trained on third-octave level (TOL) energy ratios between 16–25 kHz and 25–40 kHz successfully distinguish beluga whale (*Delphinapterus leucas*) and narwhal (*Monodon monoceros*) echolocation clicks [74]. Similarly, Risso's dolphin (*Grampus griseus*) and Pacific white-sided dolphin echolocation clicks could be distinguished by comparing unique spectral peak and notch frequency patterns [61]. Beaked whale species could be differentiated by their echolocation signals using peak frequency, centre frequency, -10 dB bandwidth, and inter-pulse interval [6] or using a Teager-Kaiser energy detector that isolates echolocation clicks from ambient noise through nonlinear energy tracking, with an additional classifier that computed zero-crossing characteristics of each echolocation click [46, 45]. This thesis builds on previous methods used to classify species-specific echolocation clicks and investigates a conservation-oriented and an arguably more precise classification problem: developing an interpretable automated detector-classifier to classify echolocation clicks from two populations (SRKW and WCT) within the same species.

Wavelet-based smoothing techniques were initially explored to detect killer whale echolocation clicks and improve feature extraction by reducing variability in off-axis signals. While wavelets are effective for isolating time-frequency features in noisy environments, their performance was limited in the spectral domain by the low inter-population variability and high intra-population variability of recorded killer whale echolocation clicks, specifically those recorded off-axis. Deep learning approaches, such as convolutional neural networks (CNNs), were also explored for their ability to efficiently identify complex patterns in large datasets. However, deep neural networks often require expert tuning and optimization, and their lack of interpretability further complicates validation within an ecological context. Additionally, their high computational complexity and resource demands make them less suitable for real-time monitoring applications, where speed, transparency, and reliability are critical. As a result, we built multiple interpretable machine learning models that include several acoustic echolocation click characteristics (hereafter referred to as features), trained on a large dataset collected using an existing JASCO Teager-Kaiser odontocete click detector-classifier (hereafter referred to as a click detector), to classify SRKW and WCT echolocation clicks in near real-time. Using a detector to select echolocation clicks, as opposed to manual annotation where analysts must identify and draw a box around individual signals within recordings, substantially increases the amount of observations available for model training. Logistic regression with least absolute shrinkage and selection operator (lasso) penalization [65], univariate-guided sparse regression (UniLasso) [13], and pre-trained lasso [17] methods were utilized for their interpretability and feature selection capabilities that prevent overfitting. We explored random forest [9] and eXtreme gradient boosting (XGBoost) [14] to capture nonlinear relationships and optimize classification performance through hyperparameter fine-tuning.

This thesis proceeds by describing the acoustic data collection within the next *Observations and Data Sources* chapter (Chapter 2). Chapter 2 introduces the Boundary Pass acoustic echolocation click repository collected using JASCO's odontocete click detector, and explains the process of selecting killer whale echolocation clicks, assigning population labels, and extracting acoustic features from each echolocation click. Chapter 3 outlines the *Methods* of building and training 21 supervised classification machine learning models: 13 logistic regression models with variants of lasso penalization and eight ensemble-based models, including random forest and XGBoost variants. Chapter 4 summarizes the *Results* of each classifier's feature selection or weighting and calculated performance metrics on a test dataset, examining whether elapsed time or prediction level (echolocation click, file or event) influences classification outcomes. This chapter also describes the final automated population-specific killer whale detector-classifier algorithm and evaluates it's test performance on two other locations. Finally in Chapter 5, the *Discussion* explores the implications of this work, its limitations, and future directions for automated conservation-orientated acoustic classification research and implementation.

Chapter 2

Observations and Data Sources

2.1 Acoustic Data

2.1.1 Training and Test Data Source

Since June 2020, underwater sound has been continuously recorded at the cabled Underwater Listening Station (ULS) deployed and maintained by JASCO in Boundary Pass. The ULS consists of two sub-sea moorings, spaced approximately 300 m apart at a depth of 192 m, positioned between the international shipping lanes. Each mooring contains 4-element hydrophone arrays mounted on tetrahedral frames (Frames A and B) deployed using a cable-lay vessel and a remotely operated vehicle (Figure 2.1). Both frames are connected to shore via a fiber-optic and electrical cable spanning 2.7 km along the seabed.

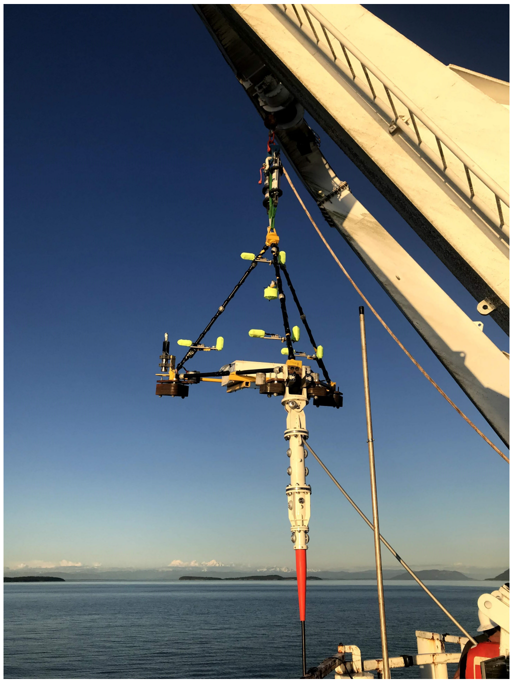

Figure 2.1: Deployment of an Underwater Listening Station (ULS) frame in Boundary Pass. Copyright of JASCO Applied Sciences.

Each frame of the ULS is equipped with eight hydrophones: four GTI M36-V35 omnidirectional hydrophones (GeoSpectrum Technologies Inc.; −165 ± 3 dB re 1 V/μPa sensitivity) and four HTI-99-HF hydrophones (High Tech Inc.; −165 dB re 1 V/μPa sensitivity). To minimize noise from water flow over the acoustic transducers, each hydrophone is protected by a cage with a shroud covering. The ULS records sound at a sampling rate of 512,000 samples per second (10 Hz to 256 kHz bandwidth) with 24-bit resolution per channel, using JASCO OceanObserver^{TM} data acquisition systems (www.jasco.com/oceanobserver).

Since January 2022, JASCO bioacoustic experts have manually logged detailed notes on daily killer whale acoustic events in Boundary Pass using Excel (v2501) and JASCO's PortListen software, which runs a high-sensitivity automated killer whale contour detector similar to that described in Moloney et al. [52] and Dewey et al. [20]. The detailed notes include manual population identification (based on killer whale pulsed calls), event timestamps, acoustic detection counts, recording quality (signal-to-noise ratio (SNR), vessel noise, etc.), and call types present. Audio files were stored in uncompressed Waveform Audio Value (WAV) format, a common audio file format for raw sound recordings. For this study, 1401 five-minute audio files (117.8 hours) recorded between 2022 and 2025 were selected, each containing manually confirmed echolocation click signals from SRKW or WCT killer whales. Although on at least one occasion SRKW and WCT killer whale populations were simultaneously recorded on the ULS, none of these events were included in this study; such co-occurrence events are rare as SRKW and WCT typically display avoidance behaviour. Where available, manual acoustic classifications were validated with killer whale sightings data from Spyhopper (Table A.1 in Appendix A). To build a robust acoustic classifier, we compiled a diverse training and test dataset by selecting multiple hydrophone channels from both ULS frames, which include varying killer whale (echolocation click source) orientations and distances, environmental conditions, and vessel presence (i.e., ambient noise levels).

2.1.2 Additional Test Data Sources

Two acoustic test datasets from other locations in the Salish Sea and with differing hydrophone set ups were provided by community partners, Saturna Island Research and Marine Education Society (SIMRES) and Orcasound. These locations include: (1) EP01: SIMRES listening station at East Point on Saturna Island, and (2) OS-J-NB: Orcasound listening station at Orcasound Lab on San Juan Island (Figure 2.2). Each dataset ($\sim$ 120 minutes) was comprised of acoustic files containing SRKW echolocation clicks, manually classified based on the presence of known SRKW pulsed calls. EP01 and OS-J-NB hydrophones were located at shallower depths and both have lower sampling rates than the Boundary Pass ULS. Table 2.1 provides a summary of the acoustic recordings, including the date, duration, location, and depth of each deployment, as well as the number of WAV files and the hydrophone type.

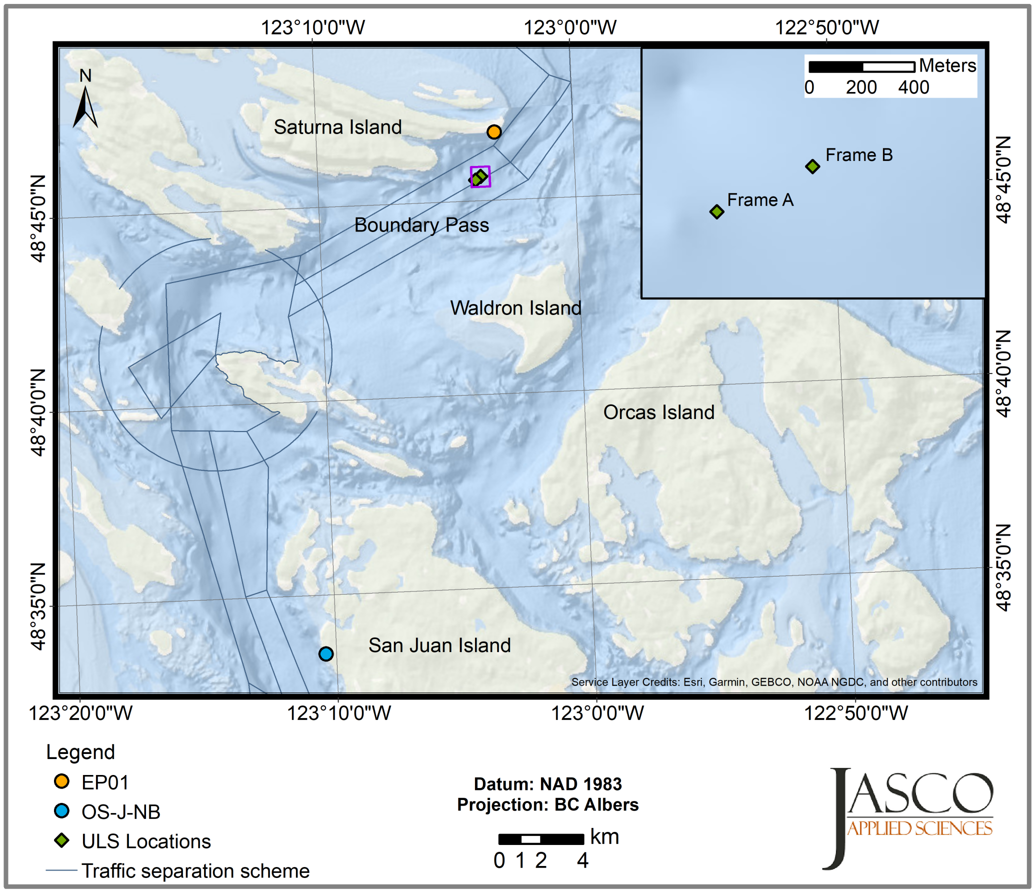

Figure 2.2: Map of the location of the Underwater Listening Station frames in Boundary Pass, the Saturna Island Marine Research and Education Society (SIMRES) underwater listening station (EP01) at East Point, Saturna Island, and the Orcasound underwater listening station (OS-J-NB) at Orcasound Lab, San Juan Island in the Salish Sea.

Table 2.1: Summary of additional acoustic test datasets from other locations in the Salish Sea (sps: samples per second; kHz: kilohertz).

Date of Acoustic Recording	Location (depth)	Hydrophone (sampling rate)
Dec 29, 2024 (28 WAV files)	East Point (20 m)	icListen HF (128,000 sps/64 kHz)
Mar 6, 2025 (27 WAV files)	Orcasound Lab (8 m)	Aquarian (48,000 sps/24 kHz)

2.2 Echolocation Click Observations and Feature Extraction

Echolocation click selection and acoustic feature extraction followed three steps:

1. Multiple five-minute WAV files surrounding known timestamps of killer whale echolocation click activity (SRKW or WCT, as classified by expert analysts based on pulsed calls) were retrieved from archived acoustic data stored on hard drives. Files were visually and aurally inspected using JASCO's PAMlab (v11.4.2) software to verify echolocation click presence. A curated repository of selected WAV files was uploaded to JASCO's data-warehouse, ensuring proper deployment information (hydrophone calibration) was linked to each WAV file for accurate acoustic measurements.

2. JASCO's automated odontocete click detector was applied to the repository (Figure 2.3). This detector selected individual killer whale echolocation clicks using the following steps: (1) a high pass filter of 5 kHz was applied to remove any vessel noise, (2) a Teager-Kaiser energy detector identified possible click events, (3) zero-crossing characteristics of the detection were extracted, (4) the detection characteristics were compared to a killer whale-specific zero-crossing echolocation click characteristic template, and (5) the detection was classified as a species-level killer whale echolocation click if under a Mahalanobis distance threshold [45, 46].

3. Across all detected species-level killer whale echolocation clicks, four false positives were omitted. Individual echolocation clicks from buzzes were not omitted as this was not feasible in this study. The click detector detected very few false positives, so it was not advantageous to try and adjust it. Additionally, it is possible that a small number of individual broadband pulses from burst-pulses were also included in the dataset, as they are acoustically similar to buzzes and are most reliably distinguished based on the behavioural context in which they occur. For each echolocation click, 12 acoustic features were calculated and extracted using JASCO's PAMlab (v11.4.2), JASCO's Ark software (v2.6.3) and R (v4.4.1). Echolocation clicks were manually labeled by population (SRKW = 1, WCT = 0) based on pulsed call types and visual sightings. Only features from individual echolocation clicks were analyzed, not echolo-

cation click trains (e.g., ICI) due to the low sensitivity of the click detector (i.e., not all echolocation clicks were detected).

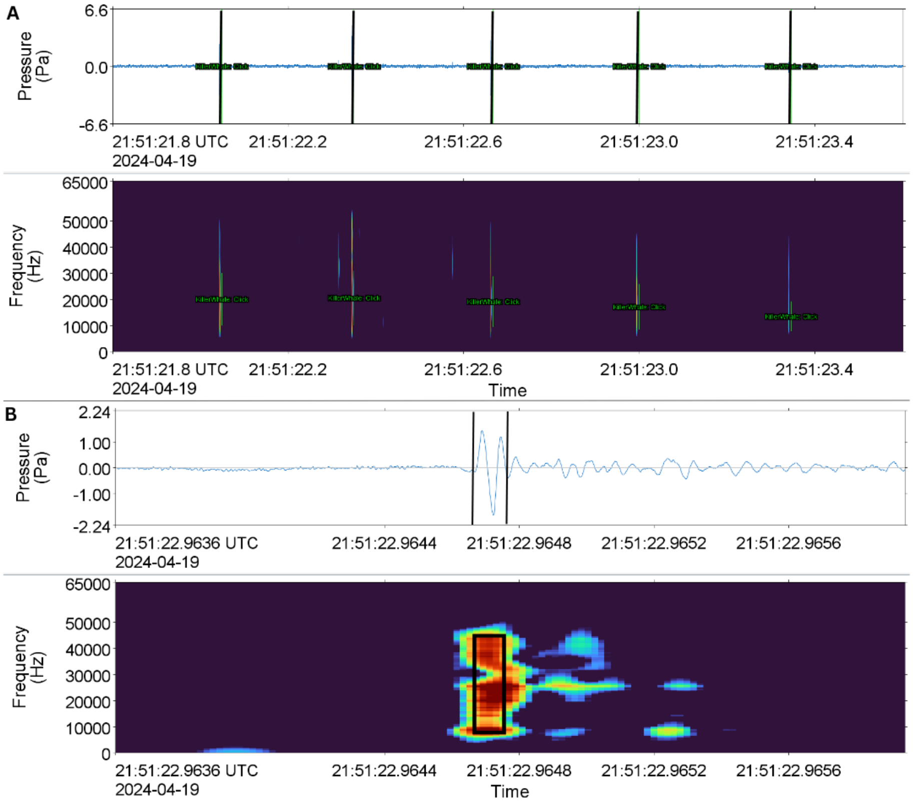

Figure 2.3: A) Zoom Out: Waveform (top) and spectrogram (bottom) of five consecutive Southern Resident killer whale (SRKW) echolocation clicks on April 19, 2024 (UTC; 128 Hz frequency resolution, 0.001 s time window, 0.0005 s time step, Hanning window, normalized across time). Black vertical lines (and superimposed text) represent the outputs of the JASCO automated odontocete click detector; B) Zoom In: Waveform (top) and spectrogram (bottom) of a single SRKW echolocation click (512 Hz frequency resolution, 0.000266 s time window, 0.00002 s time step, Hanning window, normalized across time). The black box represents an automated killer whale echolocation click detection.

A total of 205,737 observations of individual echolocation clicks (SRKW: 188,636; WCT: 17,101) were labeled, comprising 24 events with SRKW and 26 events with WCT echolocation clicks (Table A.1). Twelve acoustic features were calculated for each killer whale echolocation click (Table 2.2). Eleven echolocation click features were extracted using JASCO's PAMlab (v11.4.2): *duration*, *high frequency*, *low frequency*, *centroid frequency*, *peak frequency*, *10 dB down bandwidth*, *3 dB down bandwidth*, *10 dB down high frequency*, *3 dB down high frequency*, *10 dB down low frequency*, *3 dB down low frequency*. The final extracted feature, *click rate*, was computed using R (v4.4.1) to incorporate a feature that quantified echolocation click use. Details of how each acoustic feature was calculated are included in the Glossary.

Table 2.2: Acoustic features and units calculated for each Southern Resident killer whale (SRKW) and West Coast Transient (WCT) echolocation click. The Glossary provides definitions for each feature.

Acoustic Echolocation Click Feature	**Measurement Unit**
Click Rate	(echolocation clicks per 5-minute file)
Duration	microseconds (µs)
High Frequency	kilohertz (kHz)
Low Frequency	kilohertz (kHz)
Centroid Frequency	kilohertz (kHz)
Peak Frequency	kilohertz (kHz)
10 dB Down Bandwidth	kilohertz (kHz)
3 dB Down Bandwidth	kilohertz (kHz)
10 dB Down Low Frequency	kilohertz (kHz)
3 dB Down Low Frequency	kilohertz (kHz)
10 dB Down High Frequency	kilohertz (kHz)
3 dB Down High Frequency	kilohertz (kHz)

2.2.1 Refined Echolocation Click Selection

To account for any selection bias from pre-determined settings built into the JASCO odontocete click detector, detection boundaries were expanded and then refocused on the echolocation click based on energy. First, detections were systematically expanded in time and frequency. In time, detections were expanded to a 0.00256 s duration [6, 63], centered on the absolute maximum of the detection's waveform. In frequency, detections were expanded upward and downward by 5% of the detection frequency bandwidth. Echolocation clicks were selected to retain 95% of the total expanded selection energy. Acoustic echolocation click features were then calculated using the final selection criteria, except for echolocation click *duration* which was calculated using Teager-Kaiser energy from the initial detection. As a result, echolocation click durations may be slightly shorter and not fully comparable to other methods that use 95% energy.

2.2.2 Signal-To-Noise Ratio (SNR) Binning

To test whether high signal-to-noise ratio (SNR) improved classification performance, an SNR-based binning approach was implemented. Selected echolocation clicks were categorized into three SNR quantile bins: *Low* (0%–33.3%), *Medium* (33.3%–66.6%), and *High* (66.6%–100%). A high SNR can correspond to a close source (i.e., killer whale) or an on-axis echolocation click produced in a low ambient noise level, while a medium or low SNR could result from high ambient noise levels (e.g., due to a close passing vessel), a distant source, or an off-axis echolocation click. Machine learning models were built separately for all three SNR bins and compared to models trained on *all* SNR levels, with the exception of the pre-trained lasso model, which is able to incorporate *all* SNR levels through pre-training and then fine-tune to each SNR bin separately. The SNR binning method enables the exploration of whether high SNR echolocation clicks improve the predictive ability of the classifier.

2.3 Environmental Data

For each echolocation click, we characterized oceanographic conditions including water temperature, salinity, mid-water column current, and wind speed using the Workhorse Quartermaster Acoustic Doppler Current Profiler (ADCP) mounted on ULS Frame A and Environment Canada's weather station on Saturna Island. This data is critical to ensure environmental variability is present in the training set. Temperature and salinity affect the speed of sound and the attenuation coefficient, with warmer temperatures and higher salinity typically increasing sound speed. Mid-water currents influence the movement of sound-reflecting layers, potentially altering acoustic scattering and signal coherence, and wind speed, a key driver of surface-generated noise, can impact background ambient noise levels and the detectability of acoustic signals. Exploratory data analysis confirmed that models were trained across a range of oceanographic conditions (Figure 2.4). This approach ensures that the models capture the intrinsic features of echolocation clicks rather than artifacts introduced by environmental variability.

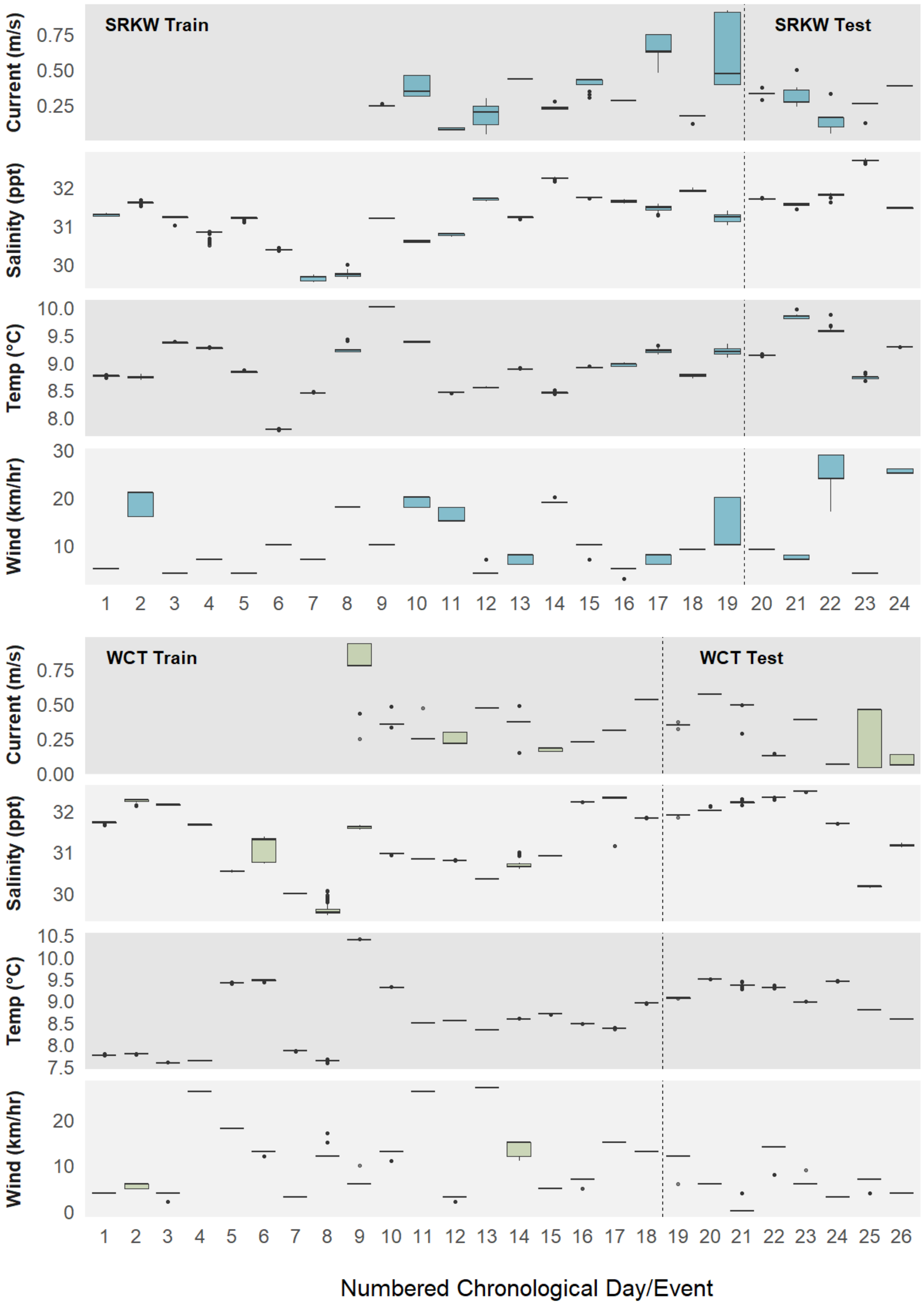

Figure 2.4: Box plots of environmental conditions (water temperature (°C), salinity (ppt), wind speed (km/hr), and mid-water column current speed (m/s)) from the corresponding 24 Southern Resident killer whale events (SRKW; blue) and 26 West Coast Transient events (WCT; green). Box plots represent environmental conditions when echolocation clicks were recorded by the ULS. Each event is identified by a chronological event number. The dashed vertical line separates the training data (left) from test data (right). Mid-water column current speed data (top panel) was not collected prior to October 2023.

Chapter 3

Methods

To differentiate SRKW from WCT killer whale populations in Boundary Pass, we trained supervised machine learning models using 12 acoustic features extracted from each echolocation click. We developed and trained 21 binary classification models, focusing primarily on two modeling approaches: supervised logistic regression with lasso regularization and ensemble-based classification trees. Within the logistic regression framework, we implemented three lasso-penalization methods: standard lasso, UniLasso, and pre-trained lasso. Logistic regression with lasso regularization is particularly well-suited for our classification problem as it performs feature selection by automatically penalizing the most irrelevant features for distinguishing populations, and thus reduces the risk of overfitting to the training data. This is especially valuable as the 12 acoustic echolocation click features contain correlated and possibly irrelevant features. Within the classification tree framework, we employed random forest and XGBoost as our ensemble methods. Random forest and XGBoost are well suited for capturing complex, non-linear relationships, which may aid in accurately identifying subtle differences in echolocation click features between killer whale populations. Together, lasso and tree-based approaches provide a balance between interpretability and predictive performance, making them sensible candidates for analyzing acoustic data without using (black-box) neural network-based methods.

Feature selection was initiated by a manual step informed by expert acoustic knowledge (Table 2.2). This step omitted the selection of amplitude-based features (including selecting SNR as a feature) as no distance to the echolocation click source (killer whale) was available. Exploratory analysis demonstrated that SRKW typically travel closer to the ULS and thus have higher amplitude echolocation clicks than WCT, and although incorporating features that leverage this would improve predictive performance in Boundary Pass, this model would not generalize well to other locations. Additionally, misclassification would occur if the killer whale populations changed their travel patterns.

In summary, the 12 acoustic features selected include click rate (number of echolocation clicks detected per five-minute file), echolocation click duration, and ten frequency-based features (Table 2.2). For the lasso models, subsequent feature selection was performed through penalization, which refined the initial set of 12 acoustic features by removing highly correlated variables and retaining those with the strongest predictive signal. Classification tree methods did not perform feature selection, rather weighted features by importance. In the following sections, we provide a detailed overview of each model, including the general form, cross-validation procedures, hyperparameter settings, and SNR bin used to train and test each model.

3.1 Logistic Regression and Lasso Penalization Models

3.1.1 Overview

To begin we fit logistic regression models to our data. Logistic regression offers low flexibility, as it assumes a linear relationship between the binary response and the echolocation click features through the logit link function. The linear assumption reduces the variance of predictions, $\hat{y}$, by preventing overfitting on the training data. However, its low flexibility comes at the cost of introduced bias when the true relationship is likely more complex. For our dataset collected from the ULS, model assumptions were generally satisfied (Appendix B), although several acoustic features were highly correlated. We addressed the issue of high-dimensional collinearity by incorporating lasso penalization into our model fitting process.

We trained a total of 13 logistic regression models, each a variant of a standard lasso [65], a UniLasso [13], or a pre-trained lasso-penalized model [17] (Figure 3.1). These models integrated feature selection with varying levels of sparsity, employed different cross-validation strategies, and utilized SNR-based binning to investigate whether high SNR echolocation clicks improve predictive performance.

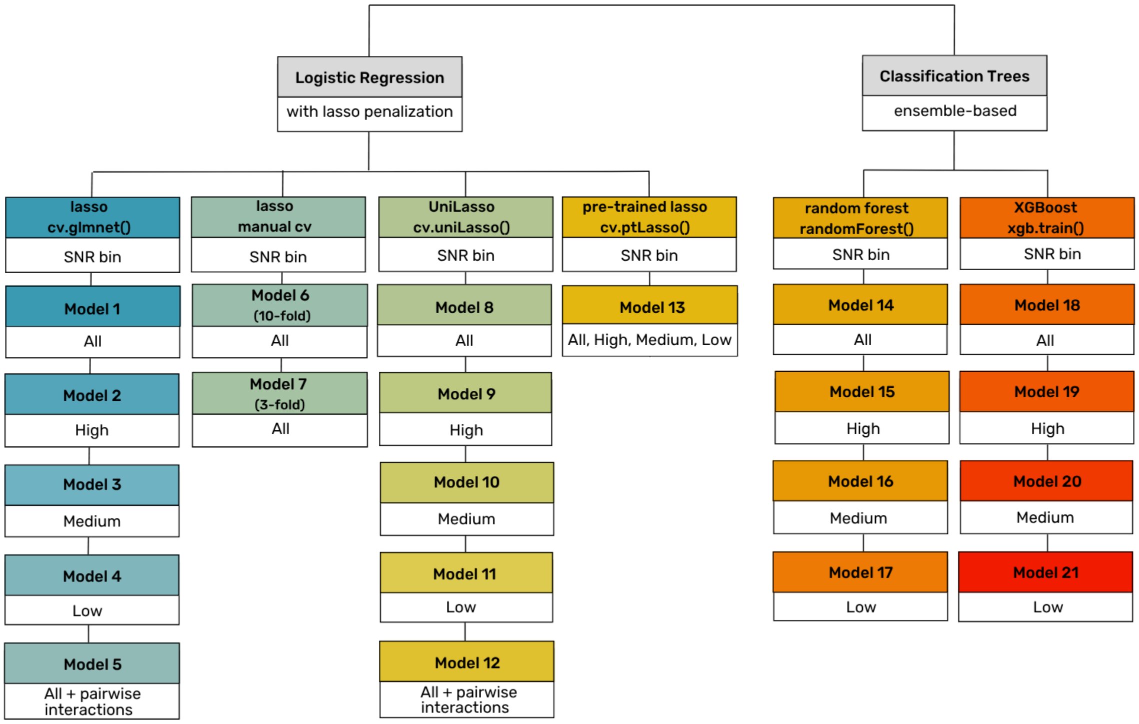

Figure 3.1: Flowchart illustrating all models fit in this study, categorized by modeling approach (logistic regression with lasso penalization vs. ensemble-based classification trees), method (lasso, univariate-guided lasso (UniLasso), pre-trained lasso, random forest, eXtreme gradient boosting (XGBoost)), x-fold cross-validation function or manual approach (`cv.glmnet`, 10-fold manual, 3-fold manual, `cv.uniLasso`, `cv.ptLasso`), inclusion of pairwise interactions, and signal-to-noise ratio (SNR) bins (*All*, *High*, *Medium*, *Low*). Each model variant is uniquely numbered and coloured for clarity.

General Form

The binary response variable, Y_i, is assumed to follow a Bernoulli distribution:

$$Y_i \sim \text{Bernoulli}(\pi_i(x)),$$

where π_i represents the probability that observation, Y_i, is a SRKW. The response variable is defined as:

$$Y_i = \begin{cases} 0, & \text{if WCT (West Coast Transient killer whale)} \\ 1, & \text{if SRKW (Southern Resident killer whale)} \end{cases}$$

Assuming independence across observations, the logistic regression model is specified as:

$$\operatorname{logit}(\pi_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \tag{3.1}$$

where x_{ij} denotes the jth acoustic feature for the ith echolocation click, and p represents the total number of features (here, $p = 12$ when including all features).

Pairwise Interactions

To account for potential interactions between features, equation (3.1) was expanded to include all pairwise interactions. We first centered each predictor by subtracting its mean:

$$\tilde{x}_{ij} = x_{ij} - \bar{x}_j,$$

where $\bar{x}_j$ is the mean of the jth predictor over all observations. Next, the product of the jth and kth centered features is added to equation (3.1):

$$\operatorname{logit}(\pi_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{1 \leq j < k \leq p} \gamma_{jk} \left(\tilde{x}_{ij} \tilde{x}_{ik} \right), \tag{3.2}$$

where γ_{jk} represents the coefficient associated with the interaction between the jth and kth features. This model allows us to capture both the main effects of the 12 acoustic features and the additional influence due to the interactions between the 66 pairs of unique features, which potentially decreases bias as we increase the number of features. A pairwise model (equation 3.2) was fit and the classification performance was compared to all other models after implementing standard lasso and UniLasso penalization.

Penalization

The lasso, UniLasso, and pre-trained lasso are regularization techniques used in regression models to prevent overfitting and improve interpretability by shrinking or setting coefficients to zero. All methods add a penalty to the loss function, but they differ in how they apply this penalty or whether they incorporate pre-training. Each lasso penalization method is described in more detail in the following subsections.

3.1.2 Lasso Penalization

Lasso-penalized logistic regression [65] was employed to identify a set of acoustic features to classify the binary response (WCT or SRKW). The objective function for the lasso penalization logistic regression method combines the negative log-likelihood with its ℓ_1 regularization term on the coefficients. The ℓ_1 penalty encourages sparsity by shrinking some coefficients to zero, thereby removing redundant or uninformative features. This property not only reduces overfitting but also facilitates feature selection by retaining only those acoustic features that contribute to the classification of killer whale populations. In the context of logistic regression, the penalized model is fit by minimizing the objective function

$$\hat{\beta}_{\lambda}^{\text{lasso}} = \arg\min_{\beta} \left\{ -2\mathcal{L}(\hat{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{3.1}$$

where $\mathcal{L}(\hat{\beta})$ is the log-likelihood of the fitted model, and $-2\mathcal{L}(\hat{\beta})$ is the key component of the deviance which serves as the loss function in the logistic regression. The binomial -2 log likelihood is defined as:

$$-2\mathcal{L}(\hat{\beta}) = -2\sum_{i=1}^{n} \Big[y_i \log(\hat{\pi}_i) + (1 - y_i)\log(1 - \hat{\pi}_i) \Big], \tag{3.2}$$

where

$$\hat{\pi}_i = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}}$$

is the predicted probability that observation, i, corresponds to a SRKW ($1 - \hat{\pi}_i$ is the predicted probability that $Y_i = 0$). The lasso-penalized logistic regression then estimates the coefficients by minimizing the following objective function:

$$\hat{\beta}_{\lambda}^{\text{lasso}} = \arg\min_{\beta} \left\{ -2\sum_{i=1}^{n} \Big[y_i \log(\hat{\pi}_i) + (1 - y_i)\log(1 - \hat{\pi}_i) \Big] + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{3.3}$$

where $\lambda \sum_{j=1}^{p} |\beta_j|$ is the ℓ_1 lasso penalty, which contains λ, the regularization parameter controlling the strength of coefficient sparsity (e.g., if $\lambda >> 0$ then the coefficients β_j shrink toward zero). The first term in equation (3.3) is equation (3.2), and the second term is the ℓ_1 penalty encouraging sparsity.

To account for inherent class imbalance prior to model fitting, class weights were computed on the training set so misclassification of the minority killer whale class (WCT) was penalized more heavily. This was implemented through the `weight` argument available in `glmnet` package [29] in the R language. Specifically, sample weights, w_i, were defined as:

$$w_i = \begin{cases} \dfrac{N_1}{N_0}, & \text{if } Y_i = 0 \\ 1, & \text{if } Y_i = 1 \end{cases}$$

where N_0 and N_1 are the counts of class 0 (WCT) and class 1 (SRKW) in the training set, respectively.

Next, the seven lasso-penalized logistic regression models were fit using one of the three different cross-validation methods:

1. **10-Fold Cross-Validation (Figure 3.1; Models 1–5):**

   Five lasso-penalized models were fit using the `cv.glmnet` function in the in `glmnet` package, which standardizes all features prior to model fitting on the training data and calculates the cross-validation error using the area under the Receiver Operator Characteristic Curve (ROC-AUC). The estimated (optimal) regularization parameter, $\hat{\lambda}$, was selected as the value minimizing the cross-validation error across all 10 folds. A final lasso model was then refit on the entire training set using this optimal $\hat{\lambda}$, nonzero features selected, and predictions were generated on the test set.

2. **Manual 10-Fold Cross-Validation Based on File Date (Figure 3.1; Model 6):**

   In addition to `cv.glmnet`, a manual 10-fold cross-validation was performed by standardizing all features prior to model fitting and using ‘file date’ as a grouping variable to prevent including the same date in the training and validation set, as it may result in too small of a regularization parameter, $\hat{\lambda}$, and increase test error. Unique file dates in the training set were randomly assigned to 10 folds, ensuring that all observations from a given date remained in the same fold. For each fold:

- A lasso model was fit over a common grid of λ values.
- Predictions were obtained on the held-out fold and the ROC-AUC was computed.

The average ROC-AUC across folds was used to select the best $\hat{\lambda}$. A final lasso model was then refit on the entire training set using this optimal $\hat{\lambda}$, and predictions were generated on the test set.

3. **Manual 3-Fold Cross-Validation Based on File Date and Future Date Prediction (Figure 3.1; Model 7):**

   Additionally, a 3-fold cross-validation was performed using 'file date' as the grouping variable. Unique file dates in the training set were partitioned into three folds such that all observations from a given date remained in the same fold and the training fold always consisted of earlier dates than the validation fold, thus always predicting forward in time, resembling a real-time model. The same procedure for selecting the regularization parameter, $\hat{\lambda}$, was followed as in the manual 10-fold cross-validation.

For all lasso-penalized models, except the model including pairwise interactions (Model 5; Figure 3.1), a threshold of 0.65 was used to convert predicted probabilities into binary labels, aiming to capture all SRKW calls while minimizing the costly misclassification of WCT as SRKW. A higher threshold of 0.99 was needed to produce the lowest test prediction error for the pairwise model (Model 5; Figure 3.1).

3.1.3 UniLasso Penalization

Building on the standard lasso framework, which applies a single ℓ_1 penalty to all coefficients, we introduce a variant called univariate-guided sparse lasso regression (UniLasso) [13] (Model 8–12; Figure 3.1). This variant imposes an additional constraint on the coefficient estimates by fitting $p = 12$ separate univariate regressions and ensuring that the signs are preserved for the full repression with all $p = 12$ features. This enables increased sparsity. The features are scaled by their univariate estimates and then a non-negative lasso is applied. This preserves the signs of the univariate estimates and gives a boost to features that are stronger on their own. UniLasso integrates marginal (univariate) information into a coherent multivariate framework that preserves the signs of the univariate coefficients and leverages their magnitude, thus enhancing both model stability and interpretability. The UniLasso algorithm is described in Appendix C. The additional constraint is defined as:

$$\operatorname{sign}(\hat{\beta}_j) = \operatorname{sign}(\hat{\beta}_j^{\text{uni}}) \quad \forall\ j.$$

Here, $\hat{\beta}_j^{\text{uni}}$ denotes the univariate estimates. This additional constraint prevents sign flips that might obscure the interpretability of the relationship between acoustic features and the binary response variable.

We implemented five UniLasso models using the `cv.uniLasso` function in the `uniLasso` package [66] in the R language. Similar to the standard lasso models, class weights were incorporated to address class imbalance (where SRKW >> WCT observations) through the `weight` argument available in `cv.uniLasso`, and a threshold of 0.65 was used to convert predicted probabilities into binary labels. A higher threshold of 0.99 was needed to produce the lowest test prediction error for the UniLasso pairwise model (Model 12; Figure 3.1), similar to the standard lasso pairwise model.

3.1.4 Pre-trained Lasso Penalization

To further enhance the standard lasso and UniLasso approaches, we implemented a pre-trained lasso modeling approach using the `cv.ptLasso` function in the `ptLasso` package [16] in the R language (Model 13; Figure 3.1). Pre-trained lasso [17] leverages the pre-training paradigm from neural networks and applies it to sparser, more interpretable regression models. This method first fits an overall model to the entire training set and then fits three separate fine-tuned models for each SNR bin. In this setup, the overall model captures group specific information common across all echolocation clicks, while the individual models capture information specific to each SNR bin. General acoustic echolocation click features are learned from all SNR bin echolocation clicks, and then fine-tuned using smaller datasets for each SNR bin, with an offset and penalty factor to distinguish between SRKW and WCT.

In the first stage (pre-training), a lasso-penalized logistic regression model is fit to the full training set to estimate coefficients (feature weights). In the second stage (fine-tuning), three individual pre-trained models are fit, one per SNR bin. Here, information from the overall model, including predictions and selected features, is passed to the individual models. The amount of transferred information is controlled by a tuning parameter, α, which is selected using 10-fold cross-validation. The general procedure is summarized by the following steps:

1. Fit a single (overall) lasso model to the training set, using `cv.glmnet`. Select the model (weight vector) $\hat{\beta}_0$ along the λ path that minimizes the cross-validation error using `lambda.min`.

2. Fix $\alpha \in [0, 1]$. Define the **offset** and **penalty factor**:

   - Define **offset** as $(1 - \alpha) \cdot (X_k\hat{\beta}_0 + \hat{\mu}_0)$, where $\hat{\mu}_0$ is the estimated overall mean.
   - Let S be the support set of $\hat{\beta}_0$. Define the penalty factor **pf** as

   $$\mathbf{pf}_j = I(j \in S) + \frac{1}{\alpha} \cdot I(j \notin S).$$

3. For each class $k \in \{1, \ldots, K\}$, (in our case, $K = 2$), fit an individual model using `cv.glmnet`, applying the specified **offset** and **penalty factor**. Use these models for prediction within each group.

As with the lasso and UniLasso models, class weights were incorporated to address class imbalance, leveraging the `weight` argument available in `cv.glmnet`, which is internally called by `cv.ptLasso`. A prediction threshold of 0.5 was applied to minimize test error, whereby predicted probabilities greater than 0.5 were classified as SRKW.

3.2 Classification Ensemble Tree Models

3.2.1 Overview

In addition to the lasso penalized logistic regression models, two ensemble-based classification methods were employed: (1) random forest and (2) XGBoost. Ensemble methods combine multiple decision trees to enhance predictive accuracy and robustness, effectively capturing nonlinear interactions among features and addressing overfitting issues common in single-tree models. Eight classification tree models were fit across all SNR bins (*Low* (Model 17 and 21), *Medium* (Model 16 and 20), *High* (Model 15 and 19), and *All* (Model 14 and 18; Figure 3.1).

3.2.2 Random Forest

We implemented random forest [9] as an ensemble-based classification tree method using the `randomForest` package [10] in the R language (Model 14–17; Figure 3.1). Random forest constructs a collection of decision trees and aggregates their individual predictions, thereby achieving lower variance and improved generalization. Random forest provides a robust, flexible, and nonparametric alternative to logistic regression, efficiently capturing complex nonlinear interactions among acoustic features while simultaneously addressing class imbalance (where SRKW >> WCT observations) and providing built-in model evaluation (Out-of-Bag error). An additional benefit of aggregating trees is that the decision distribution (i.e., the fraction of trees favoring each possible outcome) can provide a measure of confidence. The algorithm is summarized by the following steps:

1. **Bootstrapping**: Generate T bootstrap samples (random sampling with replacement) from the original training data and construct a decision tree for each sample.

2. **Randomized Feature Selection**: At each split within a tree, randomly select features to be candidates for splitting, typically $\lfloor\sqrt{p}\rfloor$ for classification problems, where p is the total number of features. This randomness reduces correlation among trees, improving the ensemble's stability.

3. **Node Splitting Criterion (Gini Impurity)**: The optimal split at each node is determined by minimizing Gini impurity ($Gini$), defined as:

$$Gini = \sum_{k=1}^{K} p_k(1 - p_k),$$

   where p_k is the proportion of observations at the node belonging to class k, and K is the total number of classes. For binary classification, $K = 2$ (SRKW or WCT).

4. **Prediction Aggregation**: For classification, predictions from individual trees are aggregated using a "super-majority" voting of 65%. For estimating the probability of belonging to class 1 (SRKW), the random forest prediction is computed as:

$$\hat{P}(Y = 1 \mid X = x) = \frac{1}{T} \sum_{t=1}^{T} I\{h_t(x) = 1\},$$

where $h_t(x)$ denotes the class prediction from the t^{th} tree, and $I\{\cdot\}$ is the indicator function.

Table 3.1: Optimized random forest hyperparameters for Models 14–17.

Hyperparameter	**Value**	**Definition**	**Rationale**
Number of trees (T)	500	Total number of decision trees in the forest	Ensures stable and robust predictions by aggregating over a large ensemble
Minimum terminal node size	5	Minimum number of observations required to split a node	Prevents overfitting by avoiding overly specific splits
Maximum terminal nodes per tree	20	Maximum number of leaf nodes allowed per tree	Limits model complexity and tree depth
Number of features at each split	$\lfloor\sqrt{p}\rfloor = 3$ (where $p = 12$)	Number of features randomly chosen at each split	Reduces correlation between trees and increases ensemble diversity
Class weights	WCT: 0.60, SRKW: 0.40	Relative weight assigned to each class during training	Addresses class imbalance by weighting the minority class more heavily
Classification threshold	0.65	Minimum predicted probability for classifying as SRKW	Reduces false positives by requiring higher certainty to label as SRKW

Random forests also naturally provide an internal error estimate via Out-of-Bag (OOB) observations. The OOB error for a dataset of n observations is defined as:

$$\text{OOB error} = \frac{1}{n} \sum_{i=1}^{n} I\{\hat{y}_i^{(\text{OOB})} \neq y_i\}$$

where $\hat{y}_i^{(\text{OOB})}$ is the prediction for the ith observation derived only from trees that did not include that observation in their training (bootstrap) sample. The train and validation sets were not split by date as in the lasso manual cross-validation methods because no decrease in prediction error was observed in the lasso methods.

The random forest models were implemented with tuned hyperparameters to control complexity and address class imbalance. The four models fit on each SNR-bin (*Low*, *Medium*, *High*, and *All*) were fit using the hyperparameters outlined in Table 3.1.

3.2.3 EXtreme Gradient Boosting

We implemented eXtreme Gradient Boosting (XGBoost) [14] as a more complex non-parametric ensemble-based classification tree method using the `xgb.train` function in the `xgboost` package [15] in the R language (Model 18–21; Figure 3.1). XGBoost is an advanced, highly efficient, gradient boosting method that sequentially builds decision trees to minimize a specific objective function by fitting the negative gradient of the loss function. It incorporates regularization and computational optimizations to prevent overfitting and improve model scalability. XGBoost requires more intensive hyperparameter tuning than random forest methods, however offers increased flexibility and computational efficiency, especially for high-dimensional data. XGBoost is widely used for many machine learning problems and benefits from its ability to build decision trees sequentially in a ‘greedy’ manner to minimize a loss iteratively.

The training procedure for XGBoost involves the following steps:

1. **Define the Objective Function**: XGBoost minimizes a regularized loss function of the form:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t), \tag{3.4}$$

   where n is the number of training observations, y_i is the true label of the i-th training observation, $\hat{y}_i$ is the predicted label for the i-th observation, $l(y_i, \hat{y}_i)$ is a differentiable loss function (we use logistic loss), T is the number of trees in the ensemble, and f_t represents the individual decision tree forming the boosted ensemble model. $\Omega(f_t)$ is a regularization term that penalizes model complexity to prevent overfitting. $\Omega(f_t)$ is defined as:

$$\Omega(f_t) = \gamma L + \frac{1}{2}\lambda\|w\|^2 + \alpha\|w\|_1,$$

where L is the number of leaf nodes in the decision tree f_t, and γL penalizes tree complexity by controlling the number of leaf nodes, where a higher γ discourages deep trees. Leaf weights, w, are the predicted values assigned to each leaf, λ is a regularization parameter controlling the strength of the ℓ_2 regularization, and $\frac{1}{2}\lambda\|w\|^2$ is the ℓ_2-regularization term which penalizes large leaf weights and prevents overfitting. The strength of the ℓ_1 regularization is controlled by a regularization parameter, α, and $\alpha\|w\|_1$ is the ℓ_1-regularization term, which promotes sparsity in the leaf weights by forcing some weights to be exactly zero.

2. **Compute Gradients and Hessians**: The gradient, g_i, and Hessian, h_i, quantify the first- and second-order changes in the loss, guiding the tree growth:

$$g_i = \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i}, \quad h_i = \frac{\partial^2 l(y_i, \hat{y}_i)}{\partial \hat{y}_i^2}.$$

3. **Tree Growth and Node Splitting**: Each new tree is grown by selecting splits that minimize the loss function in equation (3.4). The process continues recursively, constrained by hyperparameters like maximum tree depth and minimum node size.

4. **Ensemble Updates**: New trees are sequentially added, updating the ensemble prediction, $\hat{y}_i^{(t)}$, iteratively at a controlled learning rate (η):

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i).$$

Here, t represents the iteration step or the index of the boosting round, indicating the current stage in the sequential learning process, f_t is the newly added decision tree at iteration t, which corrects the errors from the previous predictions, and x_i is the input feature vector for the i-th training example.

The hyperparameters of the XGBoost models in this implementation were carefully tuned to address class imbalance, improve generalization performance and mitigate overfitting. The four models trained on each SNR-bin (*Low*, *Medium*, *High*, and *All*) were fit using the hyperparameters outlined in Table 3.2.

Table 3.2: Optimized eXtreme gradient boosting (XGBoost) hyperparameters for Models 18–21.

Hyperparameter	**Value**	**Definition**	**Rationale**
Maximum tree depth (d)	3	Maximum depth of each decision tree	Controls model complexity and prevents overfitting
Learning rate (η)	0.001	Step size shrinkage used in update to prevent overfitting	Ensures slow and stable convergence across boosting iterations
Subsample ratio	0.8	Fraction of training data used per boosting round	Reduces variance and prevents overfitting
Column sampling (feature subsample)	0.4	Fraction of features randomly sampled at each tree	Minimizes correlation between trees, improving generalization
Minimum child weight	20	Minimum sum of instance weight needed in a child node	Prevents splits that result in very small leaf nodes
Scale positive weight	$\frac{N_0}{N_1}$	Balancing weight between classes, where N_0 is WCT and N_1 is SRKW	Scales the SRKW class down and the minority class (WCT) up, improving classification accuracy for imbalanced data
ℓ_2-regularization (λ)	1	Penalization term on leaf weights	Reduces overfitting by discouraging large weights
ℓ_1-regularization (α)	1	Sparsity-inducing term on leaf weights	Encourages simpler models with fewer active leaf weights
Early stopping rounds	50	Number of rounds with no improvement before stopping training	Prevents overfitting and reduces computation time
Total number of rounds	1000	Maximum number of boosting iterations	Ensures sufficient training time
‘Best’ number of boosting rounds	49	The best iteration optimized using `xgb.cv`	Uses 5-fold cross-validation to choose the number of rounds in the final model

A classification probability threshold of 0.5 produced the best test performance based on event level confusion matrix values with a priority on high recall, and was therefore used to assign observations to the SRKW class. Thus, the final class prediction for each observation was determined by:

$$\hat{y} = I\left\{\hat{P}(Y = 1 \mid X = x) > 0.5\right\},$$

where $I\{\cdot\}$ is an indicator function that equals 1 if the condition is true and 0 otherwise.

Chapter 4 compares the test performance of the eight tree models and the 13 lasso-penalized logistic regression models on the test data to determine how the modeling approach and SNR-based binning affect predictive performance, and identify the 'best' overall population-specific killer whale echolocation click classifier.

3.3 Model Performance Metrics

All models were trained and tested using consistent dataset splits (train: 74% of events; test: 26% of events) [57]. Table 3.3 summarizes the temporally partitioned echolocation click data by population, dividing echolocation click observations into training and test sets based on total events. To ensure temporal independence, the data were split by unique ordered events, with no overlap between subsets.

Table 3.3: Summary of acoustic observations and recording events used to train and test classifier models for predicting Southern Resident killer whale (SRKW) and West Coast Transient (WCT) echolocation clicks.

Population	**Train**			**Test**		
	Clicks	**Events**	**Events(%)**	**Clicks**	**Events**	**Events(%)**
1 (SRKW)	127,936	19	38%	60,700	5	10%
0 (WCT)	10,539	18	36%	6,562	8	16%
Total	138,475	38	74%	67,262	13	26%

The hold-out test set (26% of events) from the Boundary Pass acoustic echolocation click repository was used to evaluate and compare classification performance across all 21 models developed in this study.

Signal-To-Noise-Ratio (SNR) Bin

To investigate if echolocation click SNR affects prediction, the dataset was stratified by SNR into three bins (*Low*, *Medium*, and *High*) as stated in Section 2.2.2, and eight separate logistic regression models with lasso or UniLasso penalization and eight ensemble-based tree models were fit for each bin, including all bins together using the data outlined in Table 3.4. The pre-trained lasso model was fit using all of the data and then individual models were trained based on SNR bin.

Table 3.4: Dataset statistics by SNR bin and killer whale population.

SNR Bin	Population	Train		Test	
		Clicks	Events	Clicks	Events
Low	1	44,391	19	19,589	5
Low	0	3,003	18	1,760	8
Medium	1	43,882	19	19,496	5
Medium	0	3,080	17	2,039	8
High	1	39,663	19	21,615	5
High	0	4,456	18	2,763	8
Total	1	127,936	19	60,700	5
	0	10,539	18	6,562	8

3.3.1 Performance Metrics

We assessed performance using confusion matrices, ROC-AUC, and a suite of commonly used classification metrics: Precision, Recall, F_2 Score, and Matthews Correlation Coefficient (MCC; Table 3.5). Each performance metric reported in this study provides unique insight into model performance, especially under conditions of class imbalance, where SRKW echolocation click observations outnumber those of WCT, likely due to differing foraging strategies.

In this study, SRKW were defined as the positive class, with WCT as the negative class. Precision quantifies how many of the predicted SRKW detections were correct. This metric is particularly important because misclassifying WCT echolocation clicks as SRKW (i.e., false positives) can cause more slowdowns or rerouting of vessels and be very costly by misinforming population-specific conservation actions. Recall (or True Positive Rate) measures the proportion of actual SRKW instances that were correctly identified. Since SRKW are endangered, missing their presence (i.e., false negatives) is even more costly from a conservation perspective. High recall ensures that the model is sensitive enough to capture most or all SRKW activity. As a result, the F_β Score is calculated with $\beta = 2$ rather than the more commonly used $\beta = 1$, which balances precision and recall. F_2 Score puts higher weight on minimizing false negatives (to avoid missing SRKW presence) than minimizing false positives (to avoid WCT misclassification), however both are still taken into account in the resulting score. It summarizes the model's ability to be both accurate

Table 3.5: Formulas for classification performance metrics calculated in this study.

Performance Metric	Formula
True Positive (TP)	$y = 1$ and $\hat{y} = 1$
False Positive (FP)	$y = 0$ but $\hat{y} = 1$
True Negative (TN)	$y = 0$ and $\hat{y} = 0$
False Negative (FN)	$y = 1$ but $\hat{y} = 0$
Precision	$\frac{TP}{TP+FP}$
Recall (True Positive Rate)	$\frac{TP}{TP+FN}$
F$_\beta$ Score ($\beta = 2$)	$(1+\beta^2) \times \frac{\text{Precision} \times \text{Recall}}{(\beta^2)\text{Precision} + \text{Recall}}$
Matthews Correlation Coefficient (MCC)	$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
ROC-AUC	Area under the Receiver Operator Characteristic Curve

and sensitive, however it weights sensitivity a bit higher. The MCC incorporates all four elements of the confusion matrix (TP, TN, FP, and FN) offering a more balanced view of overall model performance. MCC is especially useful in imbalanced settings like this one, where metrics such as accuracy can be misleading due to the dominance of SRKW detections. Finally, ROC-AUC measures the model's ability to discriminate between classes across all thresholds. While widely used, ROC-AUC can appear inflated in imbalanced datasets, as it includes the true positive rate, which may be high simply because SRKW are more prevalent. To address this, ROC-AUC was interpreted alongside F$_2$ Score and MCC to ensure the model was not simply capitalizing on class imbalance.

3.3.2 Aggregation of Log Likelihoods for File and Event Level Prediction

To further evaluate classification performance, we aggregated predicted probabilities to compute a log likelihood [55] at two different levels: (1) per five-minute acoustic file and (2) per killer whale acoustic event. For a given file (representing a five-minute segment), the model produces a predicted probability, π_i, for each echolocation click, i. Assuming independence among echolocation clicks, the overall likelihood that the file corresponds to an SRKW is given by the product of the individual probabilities:

$$\mathcal{L}_{\text{file}} = \prod_{i \in \text{file}} \pi_i$$

Taking the natural logarithm transforms this product into a sum:

$$log\, \mathcal{L}_{\text{file}} = \sum_{i \in \text{file}} \log(\pi_i)$$

This aggregated log likelihood serves as a measure of the model's confidence that the entire file includes echolocation clicks from SRKW. A similar aggregation can be applied to predict population at the killer whale acoustic event level. By calculating the optimal threshold from the ROC curve, a final classification can be assigned at the file or event level. This aggregation method was explored to assess whether incorporating the cumulative evidence across events would reduce test error relative to classifying individual echolocation clicks in isolation.

3.3.3 Classifier Generalization

Two classifiers were selected for further evaluation: (1) the top-performing lasso-based regression model and (2) the top-performing ensemble-based tree model, as determined by overall performance across all metrics. These classifiers were subsequently tested on independent datasets from other locations containing SRKW echolocation clicks (Section 2.1.2), with predictions generated at the echolocation click, file, and event levels. These predictions were used to assess the classifiers' ability to generalize to previously unseen data collected from different locations with different equipment, recording quality and sampling rates, thereby providing an estimate of real-world transferability, performance, and robustness.

Chapter 4

Results

Our analysis of killer whale echolocation clicks revealed distinct patterns in the selection of acoustic features across the trained models, as well as differences in echolocation click characteristics between SRKW and WCT populations. Across the 21 models (Table D.1 in Appendix D), classification performance improved from the echolocation click to event level, underscoring the value of temporal aggregation of observations to enhance the robustness and reliability of real-time killer whale population classification. When comparing classifier performance across models trained on *low*, *medium*, *high*, and *all* SNR bin detections, those trained on isolated SNR bins generally performed comparably to models trained on the full dataset, suggesting that high, medium and low SNR echolocation clicks are informative observations and aid in classification of SRKW and WCT. Most models achieved strong performance across metrics at the event level with perfect recall (i.e., all SRKW events correctly classified). As many models produced comparable prediction scores at the event level, the pre-trained model (Model 13) and XGBoost model trained on all SNR bins (Model 18) were further tested on the additional test datasets to evaluate the generalizability of the detector-classifier to other locations in the Salish Sea. In the following sections, we present population-specific acoustic characteristics derived from feature-level analysis and summarize trends in classification performance across the 21 models and prediction aggregation levels.

4.1 Feature Selection

For both lasso and tree-based methods, the 21 models fit in this study consistently prioritized certain acoustic features (Table E.1 in Appendix E). The four most frequently selected acoustic features (by lasso penalty or ranked top four in importance in ensemble methods) were *click rate* (21 models), *high frequency* (21 models), *low frequency* (21 models), and *centroid frequency* (17 models). The least selected features were *3 dB down high frequency* (3 models), *10 dB down high frequency* (2 models), and *10 dB down low frequency* (2 models; Figure 4.1).

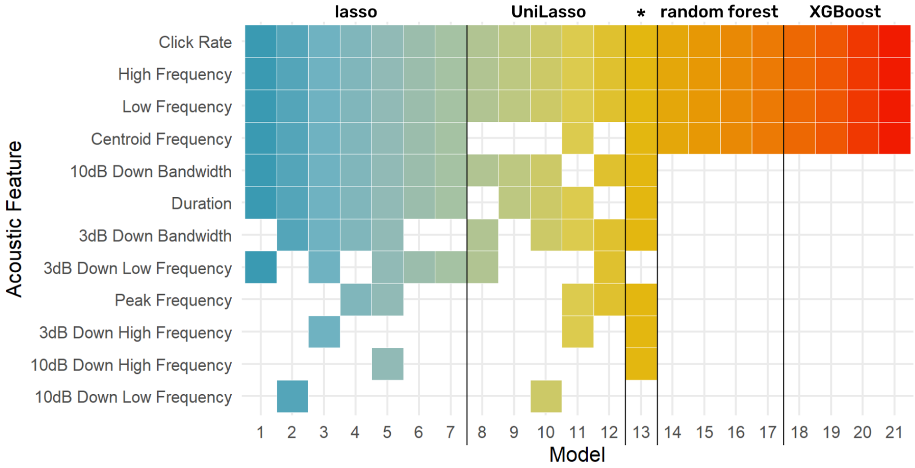

Figure 4.1: Matrix of the number of times each acoustic feature was selected across the 21 models trained (* denotes pre-trained lasso). Boxes are color coded based on model number. For ensemble methods, where all features were retained yet with differing importance, only the top four most important features (those contributing the greatest decrease in Gini impurity) were included in this figure.

Comparative analysis of the 12 acoustic echolocation click features highlighted key distinctions between killer whale populations (Figure 4.2). SRKW had a higher mean *click rate* than WCT, with 414±267 and 201±240 echolocation clicks per five minute file, respectively, where values represent the mean ± standard deviation. Although the standard deviation for WCT exceeds the mean, all values were non-negative, reflecting a highly skewed distribution rather than negative click rates. SRKW had a slightly shorter mean echolocation click *duration* (131 ± 47 µs vs. 134 ± 48 µs). SRKW echolocation clicks consistently exhibited higher values across most frequency-related features, aligning with the shorter mean echolocation click *duration* observed. Compared to WCT, SRKW echolocation clicks had a higher mean *centroid frequency* (15.0 ± 2.5 kHz vs. 14.5 ± 2.5 kHz), *peak frequency* (14.4 ± 3.4 kHz vs. 13.9 ± 3.4 kHz), *low frequency* (9.8 ± 1.8 kHz vs. 9.4 ± 1.7 kHz), and *high frequency* (22.3 ± 3.6 kHz vs. 21.2 ± 3.6 kHz). Similarly, on average 3 dB down frequency features were higher in SRKW echolocation clicks: *3 dB down low frequency* (13.9 ± 3.4 kHz vs. 13.3 ± 3.3 kHz) and *3 dB down high frequency* (15.0 ± 3.5 kHz vs. 14.5 ± 3.5 kHz). However, SRKW had a slightly narrower mean *3 dB down bandwidth* (0.88±0.3 kHz vs. 0.89±0.3 kHz) than WCT. The same trend held for 10 dB down features: SRKW echolocation clicks had a higher mean *10 dB down low frequency* (13.1 ± 3.5 kHz vs. 12.4 ± 3.5 kHz), *10 dB down high frequency* (15.8 ± 3.7 kHz vs. 15.4 ± 3.8 kHz), and a narrower mean *10 dB down bandwidth* (2.4 ± 1.2 kHz vs. 2.6 ± 1.3 kHz). These results demonstrate that on average SRKW echolocation clicks tend to be produced more frequently, with narrower spectral bandwidths, and contain more energy at higher frequencies compared to WCT echolocation clicks.

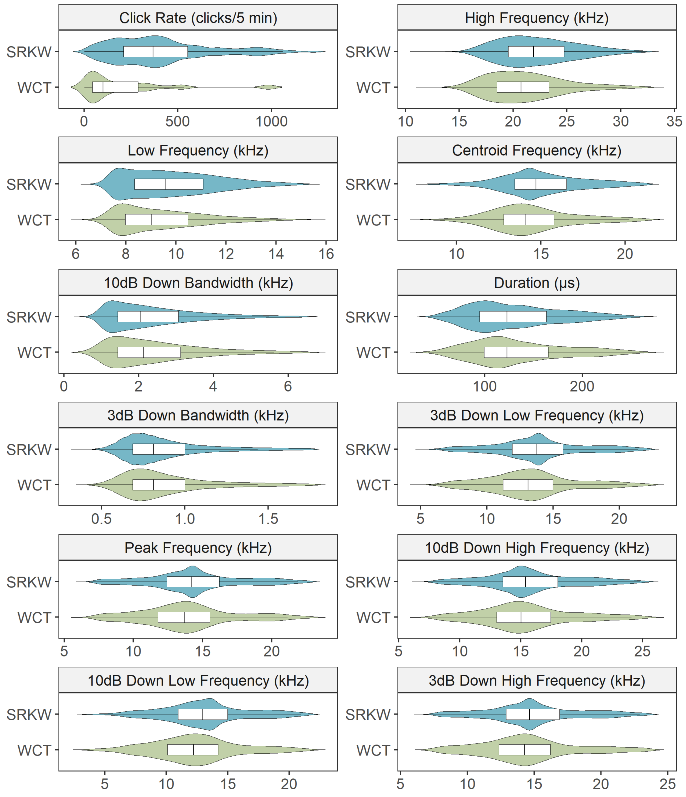

Figure 4.2: Violin plots with inset box plots comparing acoustic features of echolocation clicks between Southern Resident killer whales (SRKW; $n = 183,038$; blue) and West Coast Transients (WCT; $n = 16,559$; green). Boxes represent interquartile ranges (IQR), horizontal lines within boxes indicate medians, whiskers extend to Tukey's fences (1.5×IQR), and the width of violins represent the probability density. Outliers were omitted to focus on the central distribution of the data.

Among the lasso-based modeling approaches, the pre-trained model (Model 13) selected the highest number of features (10), while the UniLasso models trained on *high* SNR detections (Model 9) and on *all* SNR detections (Model 8) produced the sparsest models (Figure 4.3; Table E.1). The UniLasso model trained on *medium* SNR detections, along with the standard lasso models trained on *all* SNR detections (Models 1, 6, and 7), each selected seven features. The UniLasso model incorporating pairwise interactions (Model 12) also demonstrated greater sparsity than its standard lasso counterpart (Model 5), selecting 17 features compared to 71. Notably, all standard lasso models trained on the complete set of SNR detections, regardless of the cross-validation (CV) strategy, converged on the same regularization parameter, λ, and selected an identical feature set. This consistency was observed across Model 1 (using `cv.glmnet`), Model 6 (manual 10-fold CV), and Model 7 (manual 3-fold CV; Table E.1). These results indicate that partitioning CV folds based on event identity or temporal ordering did not influence model training and thus predictive performance on the test set.

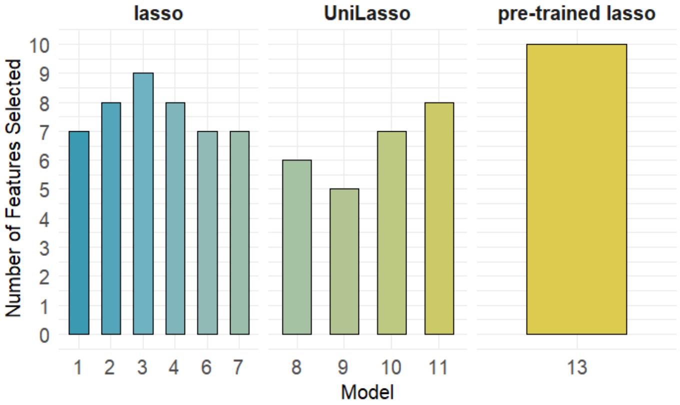

Figure 4.3: Grouped bar plot showing the number of features selected for lasso-based models (not including the pairwise interaction models with >12 features), organized by model type: lasso (Models 1–4, 6–7), UniLasso (Models 8–11), and pre-trained lasso (Model 13). Bars are colour-coded by model number.

4.2 Model Performance

Classification performance calculated based on MCC consistently improved from the echolocation click to file to event level (Figure 4.4; Table E.2, E.3, and E.4 in Appendix E). This demonstrates that individual echolocation clicks may be too variable to make an accurate prediction at the echolocation click level, likely due to propagation effects, source directionality and behavioural state. Nonetheless, once individual echolocation clicks are aggregated to the file level and further to the event level, prediction accuracy increases as more echolocation clicks are analyzed and used to make a prediction [55]. At the event level, MCCs generally exceeded 0.84 across the 21 models, except for Model 5 (lasso with pairwise interactions) and Model 15 (random forest trained on *high* SNR bin detections; Figure 4.4). When comparing lasso to tree-based methods, lasso achieved equivalent performance using fewer model optimization hyperparameters and features, demonstrating that simpler models can achieve similar performance to more flexible complex models in machine learning applications.

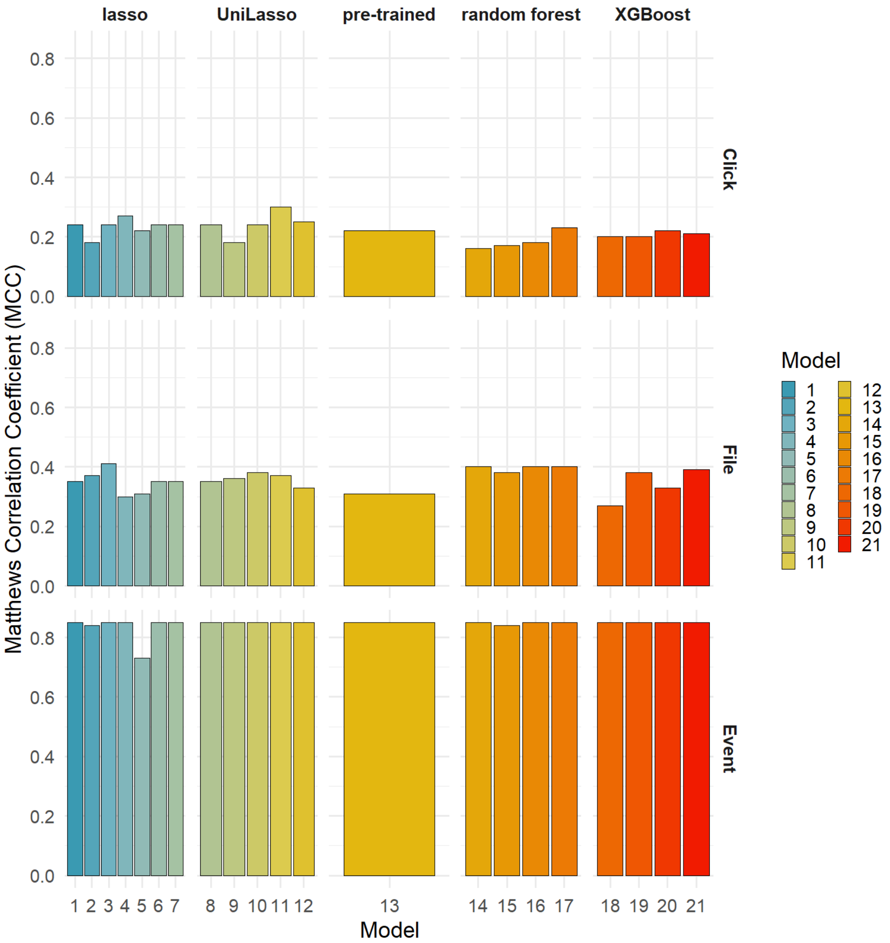

Figure 4.4: Grouped bar plots comparing Matthews Correlation Coefficients (MCCs) for the 21 trained models at the echolocation click (top), file (middle), and event level (bottom).

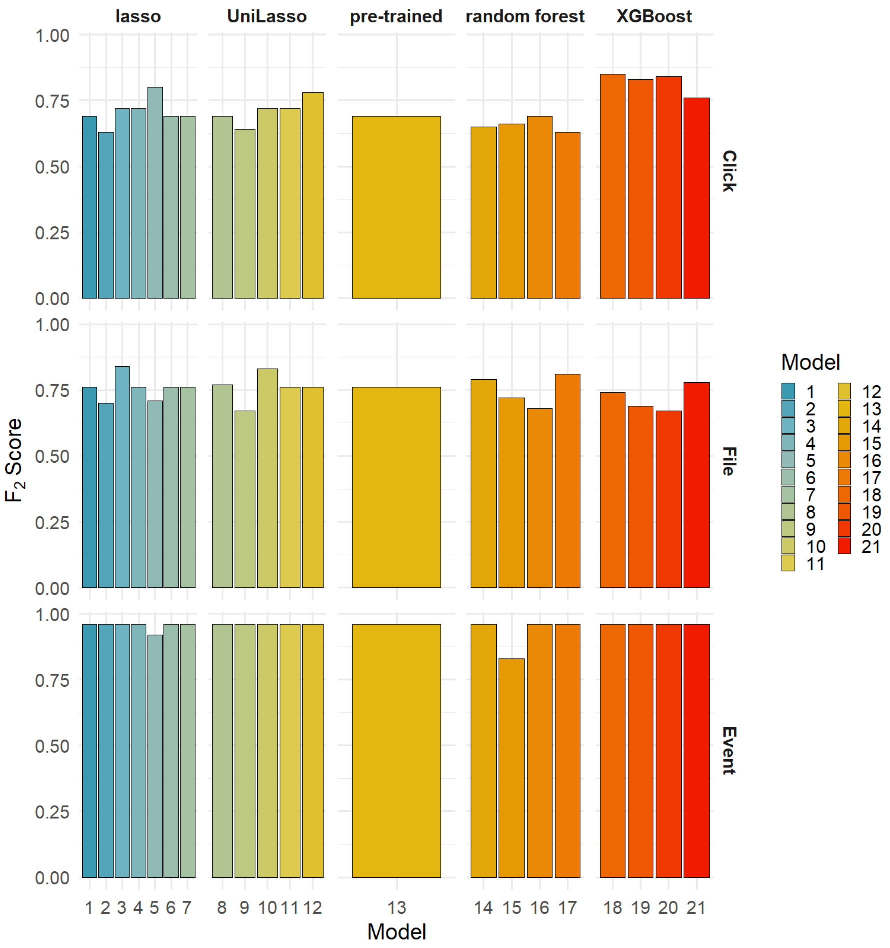

Figure 4.5: Grouped bar plots comparing F_2 Scores for the 21 trained models at the echolocation click (top), file (middle), and event level (bottom).

When comparing F_2 Scores, performance generally increased from the echolocation click to file to event level, similar to MCC. Instances where the echolocation click level F_2 Score exceeded the file level likely reflect the metric being more reliant on precision and recall than MCC and leveraging class imbalance (Figure 4.5). At the echolocation click level, precision remained high (>0.90), yet metrics such as recall, F_2 Score, MCC, and ROC-AUC exhibited lower scores across models (Table E.2). This demonstrates the importance of comparing multiple performance metrics, especially when evaluating the classification performance of models trained and validated on imbalanced datasets. A summary of the 21 models trained is included in Table D.1 in Appendix D and complete test performance metrics including confusion matrix values (TP, FP, TN, and FN), precision, recall, F_2 Score, MCC, and ROC-AUC are reported for the 21 fitted models for each of three hierarchical prediction levels (echolocation click, file, and event) in Table E.2, E.3, and E.4, respectively.

Misclassifications occurred consistently on one specific event across all models. In particular, the WCT event on November 5, 2024 was repeatedly misclassified by most models. This misclassification is not surprising as the WCT event on November 5 contained the highest number of echolocation clicks (4,018) of any WCT event in the dataset, which may have resulted in a *click rate* that more closely resembled SRKW (Table 3.3). An additional WCT misclassification occurred once on September 7, 2024 (Model 5) and a SRKW misclassification occurred once on July 24, 2024 (Model 15; Table E.1). Most of the models performed well when classifying SRKW. Models 1–14 and 16–21 (95% of all models) correctly classified all SRKW events. Meanwhile, Model 15 correctly classified all WCT events. However, no single model achieved perfect classification for both SRKW and WCT events.

Interestingly, the overall performance of most classifiers was not strongly influenced by SNR, indicating the detector-classifiers' robustness to high, medium and low SNR echolocation clicks. A potential explanation is the consistent selection of *click rate* as a key feature across all models, where it ranked as the most important feature across all ensemble methods, with substantially higher relative importance than any other feature (Figure 4.6). The stability of *click rate* across SNR bins suggests that the consistent and comprehensive acoustic environmental scanning behavior of killer whales (i.e., producing echolocation clicks in numerous directions) makes *click rate* a robust feature for predicting population across all SNR bins, including high, medium and low SNR echolocation click observations.

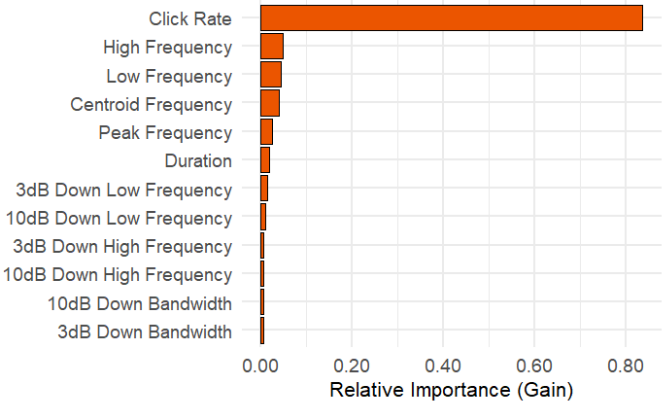

Figure 4.6: EXtreme gradient boosting (XGBoost) Model 18 relative feature importance plot.

Most models consistently classified all SRKW events correctly (i.e., perfect recall) across all SNR bins, highlighting their robustness and suitability for real-time SRKW detection. Given that $n >> p$, this strong performance is not unexpected for ensemble-based tree models and generally suggests that the dataset can support a complex model without overfitting. In contrast, this performance is more surprising for lasso-based models as they typically succeed when $p >> n$. Given the high performance across both lasso and tree-based methods, the pre-trained lasso model (Model 13) and XGBoost model trained on all SNR detections (Model 18) were selected as candidate classifiers for evaluation on the additional test datasets. These models represent rigorous yet interpretable approaches (pre-training with embedded feature selection vs. gradient boosting with feature importance) and were chosen to maximize the number of usable detections across a broad range of acoustic conditions, specifically all SNR levels. In the case of the pre-trained lasso, the model fine-tunes to a specific SNR bin, in contrast the XGBoost model trains all SNR bins together. Incorporating as many SNR level detections as possible into the models is important for reliably detecting endangered SRKW regardless of their distance from the hydrophone, directionality, or the presence of high background noise (e.g., from passing vessels). Model 13 and 18 achieved perfect classification of SRKW events while misclassifying only 12.5% (1/8 events) of WCT events (Figure 4.7) and achieved a precision of 0.83, perfect recall (1.00; 5/5 events), an F_2 Score of 0.96, an MCC of 0.85 and a ROC-AUC of 0.98 using aggregation thresholds of -451.3067 (pre-trained lasso)/-230.2585 (XGBoost) and -14626.56 (pre-trained lasso)/-10776.1 (XGBoost) for the file and event level, respectively. These results indicate that Model 13 and 18 predicted classes demonstrate high overall agreement with observed classes.

The choice between model simplicity and flexibility (complexity) did not impact predictive performance. The more flexible ensemble methods (random forests and XGBoost) performed similarly to the simpler lasso-penalized logistic regression models. This demonstrates the ensemble methods were fit with sufficient training data as no signs of overfitting were present once weights and thresholds were optimized. While lasso-based approaches offer greater interpretability by performing feature selection, tree-based ensemble methods may obscure individual feature effects due to their aggregated structure. Nonetheless, there was agreement in the top three acoustic features (*click rate*, *high frequency*, and *low frequency*) selected across all 21 models. XGBoost provided advantages in computational efficiency and trained models quicker (<1 minute) than random forest and lasso-based models (2.5–30 minutes). However, for our application of predicting killer whale population from echolocation clicks, computational efficiency was not an issue in the prediction step. Yet, if retraining the classifier is necessary for acoustic data from new locations, where echolocation click usage and acoustic behaviour may differ, XGBoost may offer an advantage due to its computational efficiency and hyperparameter tuning capabilities.

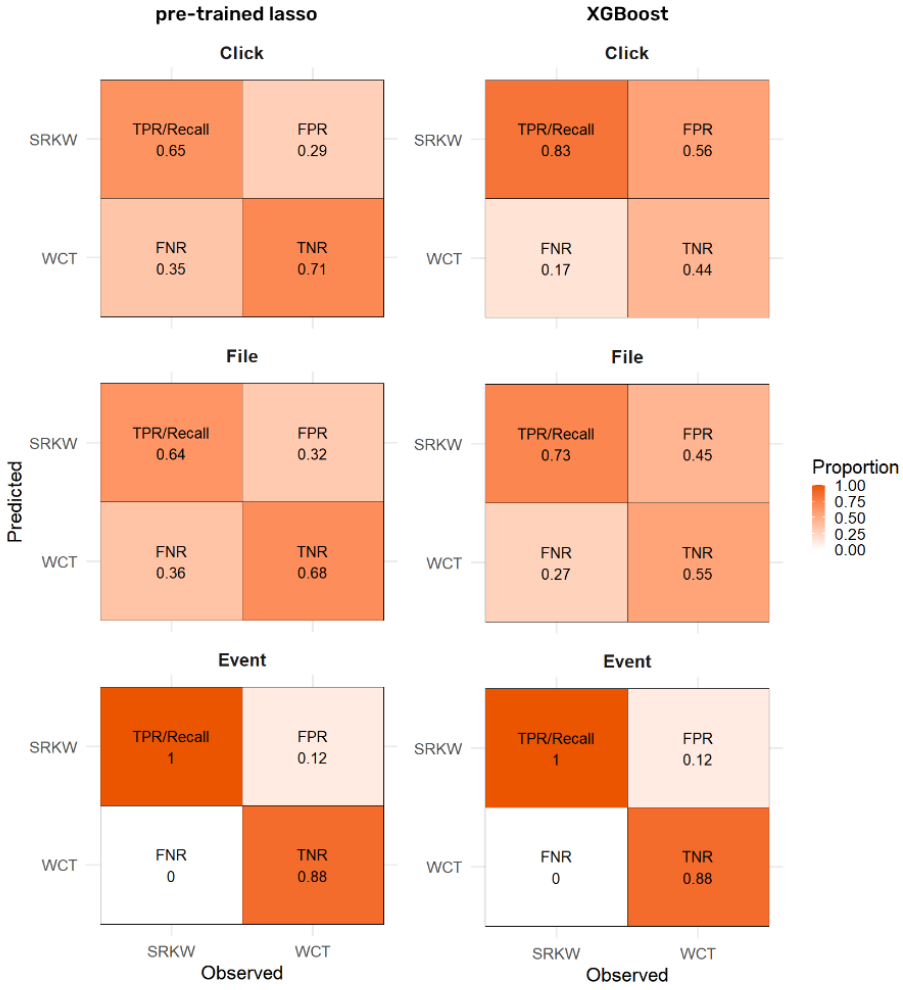

Figure 4.7: Normalized confusion matrices at the echolocation click (top), file (middle), and event level (bottom) for Model 13 (left; pre-trained lasso classifier) and 18 (right; eXtreme gradient boosting (XGBoost) classifier trained on all SNR bin detections). The color intensity corresponds to the proportion of events, with darker shades indicating higher classification confidence.

The results indicate that the acoustic distinctions between individual SRKW and WCT echolocation clicks are relatively subtle, and classification performance improves when moving from evaluating individual echolocation clicks to aggregated event level predictions. The superior classification results achieved through implementing an aggregated log-likelihood framework effectively reduced uncertainty from highly-variable single echolocation clicks, showing that combining multiple detections is critical for capturing the signal differentiating the two killer whale populations using PAM. As a result, the pre-trained lasso (Model 13) and XGBoost classifier trained on all SNR bin detections (Model 18) will be used to make predictions on additional test datasets to evaluate the robustness of the classifier in the next section.

4.3 Classifier Generalization

4.3.1 Saturna Island Marine Research and Education Society (SIMRES)

Saturna Island Marine Research and Education Society (SIMRES) operates an icListen hydrophone with a sampling rate of 128,000 samples per second at a water depth of 20 m (Table 2.1). The JASCO odontocete click detector was run on 28 files recorded from this hydrophone, and detected 686 killer whale echolocation clicks. PAMlab successfully computed 12 acoustic echolocation click features for 681 of the echolocation clicks in 24 files (Table 4.1). Seven of the 28 files did not contain killer whale echolocation clicks, only sea urchin (*Strongylocentrotus* spp.) broadband crunching sounds, and most of the files contained vessel noise. Thus, three files were classified in which no killer whale echolocation clicks were present. The Model 13 classifier (pre-trained lasso) performed poorly and predicted half of the files to be WCT and half to be SRKW, while it predicted the whole event to be WCT. However, the Model 18 (XGBoost) classifier predicted seven (29.2%) of the files as WCT and 17 (70.8%) as SRKW (Figure 4.8). At the event level, the classifier predicted SRKW. Thus, the Model 18 (XGBoost) classifier performed much better on the additional test set than the pre-trained lasso model. This may demonstrate that the pre-trained lasso classifier was fine-tuned too heavily to the ULS site and captured information specific to that location. Thus, we recommend the Model 18 classifier over the Model 13 classifer for the SIMRES hydrophone. If this detector-classifier is used on SIMRES data, it is recommended to either reduce the sensitivity of the JASCO detector, thereby minimizing false positives, or to implement a threshold whereby classification only occurs if a five-minute file contains more than 20 detected echolocation clicks.

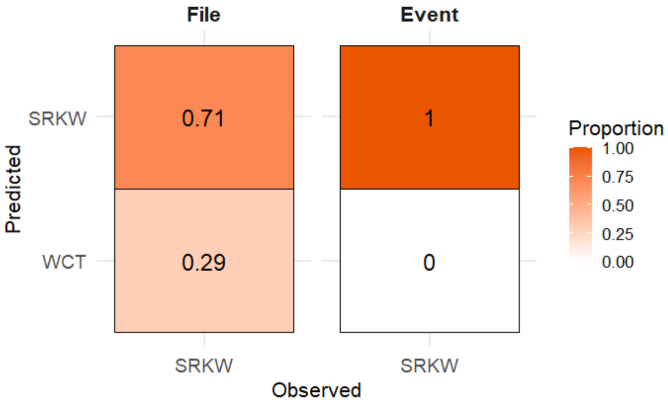

Figure 4.8: Normalized confusion matrices at the event level for the Model 18 classifier tested on the Saturna Island Marine Research and Education Society (SIMRES) data. The colour intensity corresponds to the proportion of files or events, with darker shades indicating higher classification confidence.

Table 4.1: Summary of additional test observations from other locations collected using the JASCO odontocete click detector. Details include acoustic recorder (EP01 = Saturna Island, OS-J-NB = San Juan Island), true positive, false positive, and total number of detected killer whale (KW) echolocation clicks.

Acoustic Recorder	Detected KW Clicks	True Positives	False Positives
EP01	686	299	387
OS-J-NB	38	16	22

4.3.2 Orcasound

Orcasound operates an Aquarian hydrophone with a sampling rate of 48,000 samples per second in 8 m water depth (Table 2.1). The JASCO odontocete click detector was run on 27 files, detected 38 killer whale echolocation clicks (although many more were present) in 19 of the files, and calculated all 12 acoustic features for 37 of the echolocation clicks (Table 4.1). All of the files contained SRKW echolocation clicks, and many of the files contained vessel noise. The low number of killer whale echolocation click detections is attributed to the 10.7 times lower sampling rate (48,000 samples per second) of the acoustic equipment. As a result, the JASCO odontocete click detector misclassified many killer whale echolocation clicks as those of lower-frequency clicking odontocetes, specifically sperm whales (*Physeter macrocephalus*). Of the 710 click detections, 345 were labeled as 'sperm whale', only 38 as 'killer whale', and the remaining 327 as 'unknown click'. The Model 13 and 18 classifier predicted 19 (100%) of the files as WCT and 0 (0%) as SRKW. At the event level, the classifiers predicted the file incorrectly as WCT. Consequently, we do not recommend applying either of the population-specific killer whale echolocation click detector-classifiers to Orcasound data. The significantly lower sampling rate (48,000 samples per second) of the Orcasound recordings, compared to the 512,000 samples per second used to train the classifiers, results in insufficient detection of echolocation clicks. This discrepancy leads to inaccurate or incomplete representation of high-frequency acoustic features, which are essential for reliable classification using killer whale echolocation clicks. Instead, Orcasound data are better suited for detecting lower frequency vocalizations such as killer whale pulsed calls and tonal whistles.

4.3.3 Summary

The XGBoost classifier trained on all SNR detections (Model 18) demonstrated good generalization to new acoustic files from Saturna Island, correctly identifying killer whale population presence even in the presence of impulsive broadband sea urchin sounds due to a shallower depth. Despite the lower sampling rate (128,000 samples per second) compared to the training data (512,000 samples per second), classification performance remained promising, particularly at the event level. In contrast, performance on Orcasound recordings, which were sampled at 48,000 samples per second, highlighted a critical limitation of sampling rate. This testing allowed us to identify a practical minimum sampling rate threshold of approximately 128,000 samples per second (64 kHz) for reliable echolocation click-based classification. Although this meant we could not fully validate performance at an entirely new location, the results are encouraging in demonstrating that less expensive recording systems operating above this threshold may still support accurate classification and monitoring efforts. As a result, the XGBoost classifier trained on all SNR detections (Model 18) will be incorporated into the final proposed population-specific killer whale echolocation click detector-classifier algorithm described in the next section.

4.4 Summary of the Workflow for the Population-Specific Killer Whale Echolocation Click Detector-Classifier Algorithm

The automated near real-time population-specific killer whale echolocation click detector-classifier algorithm follows the following steps (Figure 4.9):

1. A five-minute WAV file is recorded, ensuring proper deployment (calibration) information is linked for accurate acoustic measurements.

2. JASCO's automated odontocete click detector is applied to the five-minute WAV file. This click detector selects individual killer whale echolocation clicks in a box using the following steps: (1) a high pass filter of 5 kHz is applied to remove any low frequency vessel noise, (2) a Teager-Kaiser energy detector identifies possible click events, (3) zero-crossing characteristics of the detection are extracted, (4) the detection characteristics are compared to a killer whale-specific zero-crossing echolocation click characteristic template, and (5) the detection is classified as a species-level killer whale echolocation click if under a Mahalanobis distance threshold [45, 46].

3. If >20 species-level killer whale echolocation clicks are detected in the five-minute file, 12 acoustic features are calculated and extracted for each echolocation click using JASCO's PAMlab (v11.4.2) and R (v4.4.1).

4. The XGBoost classifier-model trained on all SNR killer whale detections (Model 18) is used to predict killer whale population at the echolocation click level using the 12 calculated acoustic features.

5. The predicted probabilities are aggregated to compute a log likelihood per five-minute acoustic file. Predictions are classified as a SRKW if the aggregation threshold is >-230.2585. Otherwise, the file is classified as a WCT.

6. If >30 minutes pass following a classified file without another classification, the acoustic event is considered to have ended. At that point, all predicted probabilities associated with the event are aggregated to calculate a log-likelihood score. The event

is classified as an SRKW if the aggregated log likelihood >-10776.1; otherwise, it is classified as a WCT.

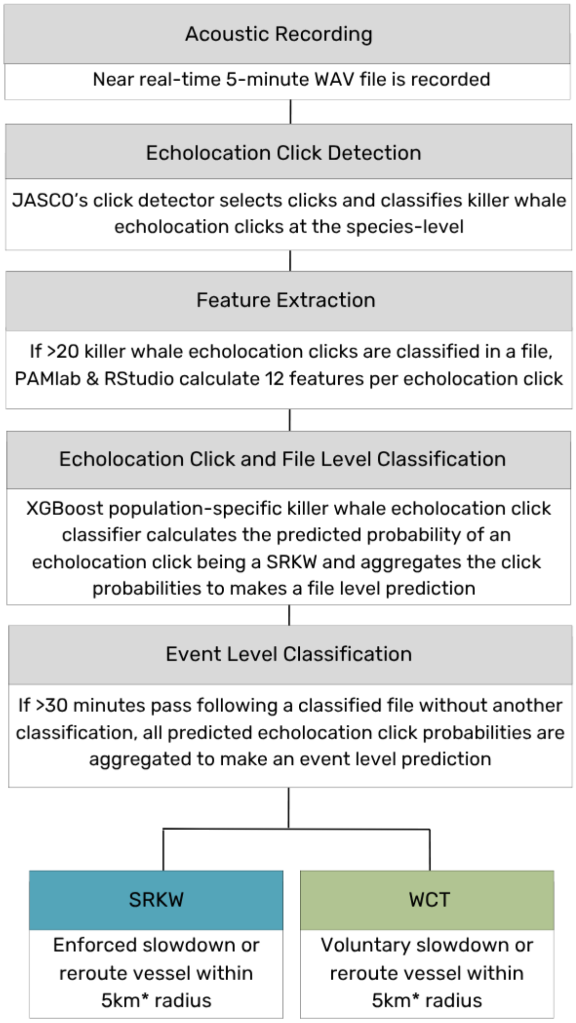
Acoustic Recording
Near real-time 5-minute WAV file is recorded
Echolocation Click Detection
JASCO's click detector selects clicks and classifies killer whale echolocation clicks at the species-level
Feature Extraction
If >20 killer whale echolocation clicks are classified in a file, PAMlab & RStudio calculate 12 features per echolocation click
Echolocation Click and File Level Classification
XGBoost population-specific killer whale echolocation click classifier calculates the predicted probability of an echolocation click being a SRKW and aggregates the click probabilities to makes a file level prediction
Event Level Classification
If >30 minutes pass following a classified file without another classification, all predicted echolocation click probabilities are aggregated to make an event level prediction
SRKW
Enforced slowdown or reroute vessel within 5km* radius
WCT
Voluntary slowdown or reroute vessel within 5km* radius

Figure 4.9: Flowchart outlining the population-specific killer whale echolocation click detector-classifier algorithm used to predict Southern Resident killer whale (SRKW) or West Coast Transient (WCT) presence. *The 5 km mitigation radius was selected based on SRKW median swimming speeds of 1.6–1.7 m/s [71], allowing for approximately 3 km of travel within 30 minutes post-detection, combined with a 2 km detection range of the Underwater Listening Station (ULS).

Chapter 5

Discussion

This study demonstrates that killer whale echolocation clicks can train a structured, interpretable machine learning pipeline for reliable event level classification of sympatric killer whale populations. Specifically, we developed and evaluated a shallow classification framework that distinguishes SRKW from WCT within Boundary Pass, using 12 features extracted from automatically detected killer whale echolocation clicks. Our findings show that distinct inter-population differences are preserved in echolocation click characteristics. These differences can be successfully leveraged for classification, offering a promising pathway for applying acoustic monitoring tools to conservation efforts in ecologically sensitive and increasingly trafficked marine environments like the Salish Sea.

The automated classification framework developed in this study has strong potential for real-time killer whale monitoring, particularly in high-traffic areas such as Boundary Pass. By training models on a large dataset of echolocation clicks identified through an established odontocete click detector, we achieved high event level classification performance (F_2 Scores: $\geq$0.83; MCCs: $\geq$0.73) using a compact set of acoustic features. Although precision remained high for echolocation click level predictions, F_2 Score and MCC did not, thus underscoring the importance of evaluating multiple performance metrics when assessing model performance. Previous study on performance metrics supports our findings and suggests that accuracy, precision, recall, and ROC-AUC can be misleading when individually reported for machine learning models trained and validated on imbalanced datasets [34]. Notably, we found that oversampling techniques were unnecessary; instead, adjusting class weights and decision thresholds yielded event level predictions that closely aligned with observed labels. The ability to automatically distinguish between killer whale populations in near real-time has direct implications for conservation policy. Both SRKW and WCT populations were found to be acoustically active and regularly present in Boundary Pass throughout the three years analyzed (Table A.1). This is especially significant given that SRKW are a critically endangered population, with designated critical habitat in both Canadian and U.S. waters, and face growing threats from increasing maritime activity [64].

Although spatial travel patterns were not formally incorporated into the classification models because detector-classifier generalizability across the Salish Sea was of interest, observational data revealed ecologically meaningful behavioural differences between populations. SRKW echolocation clicks exhibited higher received amplitudes than those of WCT, suggesting that SRKW typically travel closer to and more directly in line with the ULS (i.e., on-axis) compared to WCT. Specifically, SRKW were observed to travel through the centre of the shipping lanes, whereas WCT tended to navigate closer to the cliffs of Saturna Island, away from high-traffic routes. These patterns were corroborated by land-based observations from the Southern Gulf Island Whale Sightings Network (SGIWSN) on Saturna Island. Future research could investigate the potential of using received amplitude as a proxy for estimating travel paths, adding a spatial layer to acoustic classification models. However, such approaches would likely depend on consistent movement patterns and may be location-specific in application.

Recent and ongoing developments, including the completion of the Trans Mountain Expansion Pipeline terminal (2023), construction of (and continuous vessel traffic from) the Woodfibre Liquid Natural Gas terminal, and the approved Roberts Bank Terminal 2 expansion ($\sim$6 year construction) to berth post-Panamax vessels $>$294 m in length, have already or are expected to elevate ambient underwater noise levels in the Salish Sea. In this context, deploying real-time acoustic detection-classification tools such as the one developed in this study could offer valuable support for mitigation initiatives that enforce slowing down or rerouting vessels when SRKW are detected. Local initiatives using classification algorithms on acoustic data recorded from hydrophones near shipping lanes could be sent to Ocean Wise's Whale Report Alert System (WRAS) and used as an input to real-time forecasts of SRKW movement in the region. Together, commercial ship captains could be alerted in advance to mitigate noise disturbance or potential collision risk with the endangered killer whales.

The XGBoost detector-classifier framework proposed is adaptable to other regions of the Salish Sea, offering a scalable and location-specific solution for population-level killer whale monitoring. However, deploying in new locations and on different equipment types with ≥128,000 samples per second may require recalibration of the underlying echolocation click detector or retraining of the classifier to reflect local acoustic conditions and location-specific killer whale behavioural variation. Due to the limited timeline of this thesis, more testing is required in new locations as only one event was tested from Saturna Island and San Juan Island. Additionally, as the classifier includes the *click rate* from Boundary Pass as a feature, caution is warranted when applying it to other ecologically distinct areas, particularly those with higher whale densities (i.e., higher echolocation click rates), without additional model retraining, as this may lead to misclassification.

This work contributes to ongoing discussions in applied machine learning regarding the trade-offs between model complexity, interpretability, and predictive performance. While it is not surprising that ensemble-based XGBoost models consistently delivered high classification performance at the event level on this 'tall' dataset, logistic regression models with lasso penalization achieved comparable performance on the ULS test data using fewer hyperparameters and without the need for tuning, even though they are better suited for 'wide' data due to their feature selection capabilities [55]. Most UniLasso models selected fewer features than standard lasso-based methods, aligning with previous findings in the literature [13]. Although the pre-trained lasso method was not chosen as the 'best' performing model with generalizability to other locations, we recommend the pre-trained lasso model (Model 13) as a sophisticated ULS site-specific classifier that implements the powerful paradigm of pre-training from neural nets, yet still produces an interpretable model which is crucial to validating conservation-orientated machine learning. The enhanced interpretability of lasso models is particularly valuable in ecological applications, where understanding which acoustic features drive classification can yield biologically meaningful insights.

Nonetheless, feature importance metrics from tree-based ensemble models provided a level of interpretability that was closely aligned with lasso-based feature selection and demonstrated that *click rate* was driving the ensemble-based models. Notably, manual exclusion of amplitude features was key to producing a generalizable classifier, which highlights the continued importance of expert knowledge in classifier development. From a conservation and regulatory standpoint, the transparency of interpretable models enhances stakeholder trust and supports integration into existing acoustic monitoring frameworks.

This study further supports the presence of subtle but consistent acoustic differences in echolocation clicks between SRKW and WCT populations. Prior research by Leu et al. [48] reported SRKW echolocate more and at higher frequencies than WCT, consistent with our results. However, we observed slightly higher mean peak frequencies in SRKW (14.4 ± 3.4 kHz vs. Leu et al.: 13.7 ± 2.6 kHz) and WCT echolocation clicks (13.9 ± 3.4 kHz vs. Leu et al.: 12.8±2.7 kHz) compared to Leu et al. [48], yet within their standard deviation measured. Additionally, Leu et al. [48] reported a second higher-frequency peak in SRKW echolocation clicks around 18.8 ± 2.2 kHz, also noted in earlier work on NRKW by Au et al. [3]. Our dataset however, did not consistently capture this bimodal pattern. This can be attributed to PAMlab (v11.4.2) extracting only a single peak frequency with the highest energy per echolocation click. Thus, while the upper-frequency peak may be present, the lower-frequency sub-peak may consistently have a higher spectral level.

The frequency differences observed between the echolocation clicks of sympatric killer whale populations may reflect adaptations to their respective prey types. SRKW primarily hunt fish, while WCT target much larger marine mammals. Higher-frequency echolocation clicks may be advantageous for detecting smaller prey with gas-filled swim bladders, such as fish, whereas lower-frequency echolocation clicks could be more effective for locating larger targets (e.g., the larger air volume in the lungs of marine mammals). The broader auditory range of marine mammals (<10 Hz to >100 kHz), compared to the narrower range of most fish species (∼80 Hz to 1 kHz), as summarized by Southall [62], may further explain this divergence, suggesting that predators optimize their echolocation to remain undetected by their prey. However, differences in travel patterns between SRKW and WCT may also affect the received echolocation click characteristics. Au [2] demonstrated echolocation clicks from bottlenose dolphins (*Tursiops truncatus*) recorded at off-axis angles (e.g., >45°) exhibited reduced received source levels (∼ 25 dB re 1 µPa lower) and frequency shifts toward lower values. Furthermore, Zimmer et al. [75] demonstrated high-frequency components of echolocation clicks attenuate more rapidly over distance, especially when the vocalizing animal is far from the hydrophone. These factors highlight the importance of accounting for variability in echolocation click characteristics caused by orientation and distance when designing detection and classification algorithms. This supports our approach of incorporating recordings from multiple hydrophones and frame orientations to capture a broader range of signal variation associated with different killer whale positions and behaviors. It also underscores the need to retrain classifiers when deploying them in new locations, where travel paths and behavioral patterns differ.

Future work should continue investigating these inter-population differences in acoustic echolocation click characteristics, ideally in conjunction with behavioral observations to contextualize the function of echolocation clicks (e.g., foraging vs. navigating). A possible next step involves applying the developed detector-classifier to the full archive of ULS data collected in Boundary Pass since 2020. This would enable a retrospective analysis to determine how many echolocation-only events may have been missed by the currently deployed JASCO contour detector used to select killer whale acoustic events across the three years included in this study. This would not only enrich our understanding of killer whale presence patterns, but also underscore the added value of echolocation click-based detection methods in complementing tonal detection frameworks or movement models [70], especially for the WCT population that seems to vocalize less frequently.

Although echolocation clicks may provide a reliable acoustic basis for classifying killer whale populations, several classifier advances could be investigated. As Roch et al. [59] aptly caution, acoustic machine learning applications must critically assess algorithmic outputs and ensure that model predictions correspond to real-world patterns, rather than artifacts of data bias or model overfitting. Although the classifier model was trained on a diverse dataset in Boundary Pass (e.g., numerous events over multiple years, seasons, and environmental conditions), future iterations could benefit from incorporating additional metadata when available to enhance robustness. For instance, estimated distance to receiver and bearing angle could improve model robustness and prediction at a finer scale (i.e., five minute file or echolocation click level) and allow for the incorporation of additional amplitude features. Integrating complementary vocalization types, such as pulsed calls, into the classification pipeline may also enhance confidence and reduce ambiguity in cases where multiple call types co-occur. Retraining the detector classifier on a broader dataset comprising many locations [56] would also be advantageous for population-specific killer whale detection across broader regions of the Salish Sea. More generally, this work contributes to the growing body of literature exploring the application of machine learning to conservation and ecological monitoring.

Future research can build on this work through several promising directions. First, testing the detector-classifier on the ULS data in near real-time is crucial to quantify false positive rates more rigorously, refine decision thresholds, and assess real-world performance. Second, dividing the dataset into distinct echolocation click categories, such as click train versus buzz train clicks, could provide deeper insight into population-specific acoustic characteristics. Third, retraining and validating the classifier using data collected from different acoustic environments, such as deeper channels or more complex underwater topographies, would test its robustness and enable scalable location-specific killer whale monitoring across the broader Salish Sea and North Pacific.

In conclusion, this thesis demonstrates that interpretable machine learning models trained on echolocation clicks can effectively distinguish killer whale populations in complex, real-world underwater acoustic environments. By leveraging automated detections from a long-term dataset and selecting acoustically meaningful features, we developed a robust and scalable classification system with potential applications in conservation monitoring, acoustic ecology, and real-time mitigation. These findings contribute to a growing interdisciplinary field at the intersection of biology, acoustics, and artificial intelligence, and highlight how well-designed, transparent models can help address urgent conservation needs in increasingly noisy oceans. As the acoustic landscape of the Salish Sea continues to change, tools such as the one proposed here will be essential to inform data-driven marine policy, improve protection for endangered species, and support long-term ecological resilience.

Bibliography

[1] W. W. Au. Echolocation. In Bernd Würsig, J.G.M. Thewissen, and Kit M. Kovacs, editors, *Encyclopedia of Marine Mammals (Third Edition)*, pages 289–299. Academic Press, 2018.

[2] W. W. Au, B. Branstetter, P. W. Moore, and J. J. Finneran. The biosonar field around an atlantic bottlenose dolphin (*Tursiops truncatus*). *The Journal of the Acoustical Society of America*, 131(1):569–576, 2012.

[3] W. W. Au, J. K. Ford, J. K. Horne, and K. A. N. Allman. Echolocation signals of free-ranging killer whales (orcinus orca) and modeling of foraging for chinook salmon (oncorhynchus tshawytscha). *The Journal of the Acoustical Society of America*, 115(2):901–909, 2004.

[4] W. W. Au and M. C. Hastings. *Principles of Marine Bioacoustics*, volume 510. Springer, New York, 2008.

[5] L. G. Barrett-Lennard, J. K. Ford, and K. A. Heise. The mixed blessing of echolocation: differences in sonar use by fish-eating and mammal-eating killer whales. *Animal Behaviour*, 51(3):553–565, 1996.

[6] S. Baumann-Pickering, M. A. McDonald, A. E. Simonis, A. Solsona, K. P. Merkens, E. M. Oleson, M. A. Roch, S. M. Wiggins, S. Rankin, T. M. Yack, and J. A. Hildebrand. Species-specific beaked whale echolocation signals. *Journal of the Acoustical Society of America*, 134(3):2293–2301, 2013.

[7] C. Bergler, H. Schröter, R.X. Cheng, H. Ritter, R. Tiedemann, T. Hothorn, and H. Klinck. Orca-spot: An automatic killer whale sound detection toolkit using deep learning. *Scientific Reports*, 9:10997, 2019.

[8] M. A. Bigg, P. F. Olesiuk, G. M. Ellis, J. K. B. Ford, and K. C. Balcomb. Social organization and genealogy of resident killer whales (*Orcinus orca*) in the coastal waters of british columbia and washington state. *Report of the International Whaling Commission*, 12:383–405, 1990.

[9] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[10] L. Breiman, A. Cutler, A. Liaw, and A. Wiener. *randomForest: Breiman and Cutlers Random Forests for Classification and Regression*, 2022. R package version 4.7-1.2.

[11] M. Brunoldi, G. Bozzini, A. Casale, P. Corvisiero, D. Grosso, N. Magnoli, J. Alessi, C.N. Bianchi, A. Mandich, C. Morri, and P. Povero. A permanent automated real-time passive acoustic monitoring system for bottlenose dolphin conservation in the mediterranean sea. *PLOS ONE*, 11(1):e0145362, 2016.

[12] Center for Whale Research. Orca Survey: Southern Resident Killer Whales ID Guide. `https://www.whaleresearch.com`, 2023. [Identification Guide, Friday Harbor, WA].

[13] S. Chatterjee, T. Hastie, and R. Tibshirani. Univariate-guided sparse regression. *arXiv preprint*, 2025.

[14] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[15] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, Y. Li, J. Yuan, and XGBoost contributors. *xgboost: Extreme Gradient Boosting*, 2024. R package version 1.7.8.1.

[16] E. Craig and R. Tibshirani. *ptLasso: Pretrained Lasso*, 2025. R package version 1.0.

[17] E. Craig, C. Zhou, Q. Berthet, and S. Wager. Pretraining and the lasso. *arXiv preprint*, 2024.

[18] V. B. Deecke. *The vocal behaviour of mammal-eating killer whales: communicating with costly calls*. Phd thesis, University of St. Andrews, 2003.

[19] V. B. Deecke, J. K. B. Ford, and P. J. B. Slater. The vocal behaviour of mammal-eating killer whales: communicating with costly calls. *Animal Behaviour*, 69(2):395–405, 2005.

[20] R. Dewey, T. Dakin, X. Mouy, and I. R. Urazghildiiev. A regional hydrophone network: monitor, detect and track. In *Underwater Acoustics Conference and Exhibition*, Crete, Greece, June 2015.

[21] P. L. Edds-Walton. Acoustic communication signals of mysticetes whales. *Bioacoustics*, 8(1–2):47–60, 1997.

[22] C. Erbe, A. MacGillivray, and R. Williams. Mapping cumulative noise from shipping to inform marine spatial planning. *The Journal of the Acoustical Society of America*, 132(5):EL423–EL428, 2012.

[23] J. K. B. Ford. A catalogue of underwater calls produced by killer whales (orcinus orca) in british columbia. Canadian data report of fisheries and aquatic sciences no. 633, Department of Fisheries and Oceans, Fisheries Research Branch, Pacific Biological Station, 1987.

[24] J. K. B. Ford. Acoustic behaviour of resident killer whales (orcinus orca) off vancouver island, british columbia. *Canadian Journal of Zoology*, 67(3):727–745, 1989.

[25] J. K. B. Ford, G. M. Ellis, L. G. Barrett-Lennard, A. B. Morton, R. S. Palm, and K. C. Balcomb. Dietary specialization in two sympatric populations of killer whales (*Orcinus orca*) in coastal british columbia and adjacent waters. *Canadian Journal of Zoology*, 76(8):1456–1471, 1998.

[26] J. K. B. Ford, E. H. Stredulinsky, G. M. Ellis, J. W. Durban, and J. F. Pilkington. Offshore killer whales in canadian pacific waters: distribution, seasonality, foraging ecology, population status and potential for recovery. Technical report, Canadian Science Advisory Secretariat (CSAS), 2014.

[27] F. Frazao, O. Kirsebom, M. Dowd, and R. Joy. Open-source deep learning models for detection and classification of orcas. In *Workshop on Machine Learning and Movement Models*, Burnaby, BC, August 2022.

[28] F. Frazao, O. S. Kirsebom, A. Houweling, J. Wladichuk, J. Kanes, R. Joy, and M. Dowd. Using a sequence deep learning model to increase the acoustic context of a killer whale detector. *The Journal of the Acoustical Society of America*, 155(3_Supplement):A87, 2024.

[29] J. Friedman, T. Hastie, R. Tibshirani, B. Narasimhan, K. Tay, N. Simon, J. Qian, and J. Yang. *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*, 2023. R package version 4.1-8.

[30] G. V. Frisk. Noiseonomics: the relationship between ambient noise levels in the sea and global economic trends. Technical Report 1, 2012.

[31] E. C. Garland, A. W. Goldizen, M. L. Rekdahl, R. Constantine, C. Garrigue, N. D. Hauser, M. M. Poole, J. Robbins, and M. J. Noad. Dynamic horizontal cultural transmission of humpback whale song at the ocean basin scale. *Current Biology*, 21(8):687–691, 2011.

[32] E. C. Garland, L. Rendell, L. Lamoni, M. M. Poole, and M. J. Noad. Song hybridization events during revolutionary song change provide insights into cultural transmission in humpback whales. *Proceedings of the National Academy of Sciences*, 114(30):7822–7829, 2017.

[33] G. Giovannini, P. J. O. Miller, P. J. Wensveen, and F. I. P. Samarra. Sound production during feeding in icelandic herring-eating killer whales (*Orcinus orca*). *Ethology Ecology & Evolution*, 2025.

[34] Q. Gu, L. Zhu, and Z. Cai. Evaluation measures of the classification performance of imbalanced data sets. In *Computational Intelligence and Intelligent Systems: 4th International Symposium, ISICA 2009, Huangshi, China, October 23–25, 2009. Proceedings 4*, pages 461–471, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[35] G. Gubnitky and R. Diamant. Detecting the presence of sperm whales' echolocation clicks in noisy environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[36] G. Gubnitsky and R. Diamant. Inter-pulse estimation for sperm whale click detection. In *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, June 2023.

[37] B. Hendricks, M. K. Pine, G. Baer, M. Welton, H.K. Symonds, D. T. Dakin, H. M. Alidina, C. R. Picard, and J. Wray. Quantifying vessel noise and acoustic habitat loss in marine soundscapes. *Marine Pollution Bulletin*, 219:118150, 2025.

[38] M. M. Holt, M. B. Hanson, C. K. Emmons, D. K. Haas, D. A. Giles, and J. T. Hogan. Sounds associated with foraging and prey capture in individual fish-eating killer whales, orcinus orca. *The Journal of the Acoustical Society of America*, 146(5):3475–3486, 2019.

[39] M. M. Holt, D. P. Noren, V. Veirs, C. K. Emmons, and S. Veirs. Speaking up: Killer whales (*Orcinus orca*) increase their call amplitude in response to vessel noise. *The Journal of the Acoustical Society of America*, 125(1):EL27–EL32, 2009.

[40] M. M. Holt, J. B. Tennessen, B. M. Hanson, C. K. Emmons, D. A. Giles, J. T. Hogan, and M. J. Ford. Vessels and their sounds reduce prey capture effort by endangered killer whales (orcinus orca). *Marine Environmental Research*, 170:105429, 2021.

[41] International Organization for Standardization. Iso/cd 23990: Underwater acoustics — bioacoustical terminology. ISO/TC 43/SC 3, 2024. Under development; a draft is being reviewed by the committee.

[42] R. Joy, D. Tollit, J. Wood, A. MacGillivray, Z. Li, K. Trounce, and O. Robinson. Potential benefits of vessel slowdowns on endangered southern resident killer whales. *Frontiers in Marine Science*, 6:344, 2019.

[43] M. B. Kaplan and S. Solomon. A coming boom in commercial shipping? the potential for rapid growth of noise from commercial ships by 2030. *Marine Policy*, 73:119–121, 2016.

[44] O. S. Kirsebom, F. Frazao, B. Padovese, S. Sakib, Y. Su, and S. Matwin. Meridian open-source software for deep learning-based acoustic data analysis. *The Journal of the Acoustical Society of America*, 151(4_Supplement):A27–A27, 2022.

[45] K. A. Kowarski, J. Delarue, B. Martin, J. O'Brien, R. Meade, O. Ó Cadhla, and S. Berrow. Signals from the deep: spatial and temporal acoustic occurrence of beaked whales off western ireland. *PLOS ONE*, 13(6):e0199431, 2018.

[46] K. A. Kowarski, S. B. Martin, E. E. Maxner, C. B. Lawrence, J. J. Y. Delarue, and J. L. Miksis-Olds. Cetacean acoustic occurrence on the us atlantic outer continental shelf from 2017 to 2020. *Marine Mammal Science*, 39(1):175–199, 2023.

[47] M. M. Krahn, P. R. Wade, S. T. Kalinowski, M. E. Dahlheim, B. L. Taylor, M. B. Hanson, G. M. Ylitalo, R. P. Angliss, J. E. Stein, and R. S. Waples. Status review of southern resident killer whales (*Orcinus orca*) under the endangered species act. NOAA Technical Memorandum NMFS-NWFSC-54, U.S. Department of Commerce, NOAA, 2002.

[48] A. A. Leu, J. A. Hildebrand, A. Rice, S. Baumann-Pickering, and K. E. Frasier. Echolocation click discrimination for three killer whale ecotypes in the northeastern pacific. *The Journal of the Acoustical Society of America*, 151(5):3197–3206, 2022.

[49] P. T. Madsen, U. Siebert, and C. P. H. Elemans. Toothed whales use distinct vocal registers for echolocation and communication. *Science*, 379(6635):928–933, 2023.

[50] M. Matei. Killerclick: Automatic detection and classification of the echolocation clicks of the southern resident killer whales. Master's thesis, University of St Andrews, 2022.

[51] M. F. McKenna, S. M. Wiggins, D. Ross, and J. A. Hildebrand. Underwater radiated noise from modern commercial ships. *The Journal of the Acoustical Society of America*, 131(1):92–103, 2012.

[52] J. E. Moloney, C. A. Hillis, X. Mouy, I. R. Urazghildiiev, and T. Dakin. Autonomous multichannel acoustic recorders on the venus ocean observatory. In *OCEANS 2014*, pages 1–6, St. John's, NL, Canada, September 2014. IEEE.

[53] P. A. Morin, M. L. McCarthy, C. W. Fung, J. W. Durban, K. M. Parsons, W. F. Perrin, B. L. Taylor, T. A. Jefferson, and F. I. Archer. Revised taxonomy of eastern north pacific killer whales (*Orcinus orca*): Bigg's and resident ecotypes deserve species status. *Royal Society Open Science*, 11(3):231368, 2024.

[54] National Oceanic and Atmospheric Administration. Endangered and threatened wildlife and plants: endangered status for southern resident killer whales. Rules and Regulations, 50 CFR Part 224, 2005.

[55] K. J. Palmer, K. Brookes, and L. Rendell. Categorizing click trains to increase taxonomic precision in echolocation click loggers. *The Journal of the Acoustical Society of America*, 142(2):863–877, 2017.

[56] K. J. Palmer, E. Cummings, M. Dowd, K. Frasier, F. Frazao, A. Harris, A. E. Houweling, J. Kanes, O.S. Kirsebom, H. Klinck, H.T. LeBlond, L. Laturnus, C. Matkin, O. Murphy, H. Myers, D. Olsen, C. Oneill, B. Padovese, J. Pilkington, L. Quale, A. Riera, K. Trounce, S. Vagle, S. Viers, V. Viers, J. Wladichuk, J. Wood, T. Yack, H. Yurk, and R. Joy. Annotated *Orcinus orca* acoustic dataset for detection and ecotype classification. *Scientific Data*, 2025.

[57] R. R. Picard and K. N. Berk. Data splitting. *The American Statistician*, 44(2):140–147, 1990.

[58] A. Rice, V. B. Deecke, J. K. Ford, J. F. Pilkington, E. M. Oleson, and J. A. Hildebrand. Spatial and temporal occurrence of killer whale ecotypes off the outer coast of washington state, usa. *Marine Ecology Progress Series*, 572:255–268, 2017.

[59] M. A. Roch, S. M. Kerosky, S. Baumann-Pickering, M. S. Soldevilla, and J. A. Hildebrand. Compensating for the effects of site and equipment variation on delphinid species identification from their echolocation clicks. *The Journal of the Acoustical Society of America*, 137(1):22–29, 2015.

[60] Y. Shiu, K. J. Palmer, M. A. Roch, E. Fleishman, X. Liu, E. M. Nosal, T. A. Helble, D. Cholewiak, D. Gillespie, and H. Klinck. Deep neural networks for automated detection of marine mammal species. *Scientific Reports*, 10(1):607, 2020.

[61] M. S. Soldevilla, E. E. Henderson, G. S. Campbell, S. M. Wiggins, J. A. Hildebrand, and M. A. Roch. Classification of risso's and pacific white-sided dolphins using spectral properties of echolocation clicks. *The Journal of the Acoustical Society of America*, 124(1):609–624, 2008.

[62] B. L. Southall, A. E. Bowles, W. T. Ellison, J. J. Finneran, R. L. Gentry, C. R. Jr Greene, D. Kastak, D. R. Ketten, J. H. Miller, P. E. Nachtigall, W. J. Richardson, J. A.

Thomas, and P. L. Tyack. Marine mammal noise exposure criteria: initial scientific recommendations. *Aquatic Mammals*, 33:411–521, 2007.

[63] J. E. Stanistreet, D. P. Nowacek, S. Baumann-Pickering, J. T. Bell, D. M. Cholewiak, J. A. Hildebrand, L. E. Hodge, H. B. Moors-Murphy, S. M. Van Parijs, and A. J. Read. Using passive acoustic monitoring to document the distribution of beaked whale species in the western north atlantic ocean. *Canadian Journal of Fisheries and Aquatic Sciences*, 74(12):2098–2109, 2017.

[64] J. B. Tennessen, M. M. Holt, M. B. Hanson, D. A. Giles, C. K. Emmons, J. T. Hogan, and S. E. Parks. Males miss and females forgo: auditory masking from vessel noise impairs foraging efficiency and success in killer whales. *Global Change Biology*, 30(9):e17490, 2024.

[65] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[66] R. Tibshirani and T. Hastie. *uniLasso: Univariate-guided sparse regression*, 2025. R package version 2.4.

[67] J. R. Towers, G. J. Sutton, T. J. H. Shaw, M. Malleson, D. Matkin, B. Gisborne, J. Forde, D. Ellifrit, G. M. Ellis, J. K. B. Ford, and T. Doniol-Valcroze. Photo-identification catalogue, population status, and distribution of bigg's killer whales known from coastal waters of british columbia, canada. Technical Report 3311, Canadian Technical Report of Fisheries and Aquatic Sciences, 2019.

[68] K. Trounce, O. Robinson, A. MacGillivray, D. Hannay, J. Wood, D. Tollit, and R. Joy. The effects of vessel slowdowns on foraging habitat of the southern resident killer whales. In *Proceedings of Meetings on Acoustics*, volume 37. AIP Publishing, 2019.

[69] P. L. Tyack and C. W. Clark. *Communication and acoustic behavior of dolphins and whales*, pages 156–224. Springer, New York, 2000.

[70] T. Wei. Detection and classification of whale calls using machine learning. Master's thesis, Simon Fraser University, December 2023. `https://www.sfu.ca/~rjoy/TengWeiMScThesisDec2023.pdf`.

[71] R. Williams, E. Ashe, K. A. Nielsen, H. H. Nollens, S. Reiss, K. Wold, and J. K. Gaydos. Respiratory intervals and swimming speed as remotely sensed health metrics in free-ranging killer whales (*Orcinus orca*). *Journal of Wildlife Diseases*, 61(1):17–29, 2025.

[72] R. Williams, R. C. Lacy, E. Ashe, L. Barrett-Lennard, T. M. Brown, J. K. Gaydos, F. Gulland, M. MacDuffee, B. W. Nelson, K. A. Nielsen, and H. Nollens. Warning sign of an accelerating decline in critically endangered killer whales (*Orcinus orca*). *Communications Earth & Environment*, 5:173, 2024.

[73] B. M. Wright, V. B. Deecke, G. M. Ellis, A. W. Trites, and J. K. Ford. Behavioral context of echolocation and prey-handling sounds produced by killer whales (*Orcinus orca*) during pursuit and capture of pacific salmon (*Oncorhynchus* spp.). *Marine Mammal Science*, 37(4):1428–1453, 2021.

[74] M. J. Zahn, M. Ladegaard, M. Simon, K. M. Stafford, T. Sakai, and K. L. Laidre. Accurate species classification of arctic toothed whale echolocation clicks using one-third octave ratios. *The Journal of the Acoustical Society of America*, 155(4):2359–2370, 2024.

[75] W. M. X. Zimmer. *Passive Acoustic Monitoring of Cetaceans*. Cambridge University Press, Cambridge, UK, 2011.

Appendix A

Acoustic Dataset Summary

Appendix A summarizes the selected acoustic recordings from the Canadian waters of Boundary Pass in the Salish Sea. Table A.1 includes the number of WAV files analyzed, their recording date, from which ULS frame (A or B), the number of detected killer whale echolocation clicks, and any visual confirmations from Saturna Island provided by Spyhopper.

Table A.1: Summary of selected acoustic recordings from Boundary Pass, provided by JASCO. Visual sighting data from Saturna Island (population and estimated number of killer whales), provided by Spyhopper. Frame A/B indicates WAV files were taken from both frames.

Date	**Location**	**WAV Files**	**SRKW Clicks**	**WCT Clicks**	**Visual Sighting**
2022-Feb-09	Frame A/B	17	0	373	None
2022-Feb-12	Frame A/B	26	0	517	None
2022-Mar-18	Frame A/B	4	0	23	WCT (5)
2022-Mar-20	Frame A/B	17	0	167	None
2022-Jul-17	Frame A/B	19	0	114	WCT (7)
2022-Jul-23	Frame A/B	24	8,060	0	SRKW (12-40)
2022-Aug-4	Frame A/B	76	16,559	0	SRKW (10-25)
2022-Oct-3	Frame A	5	2,809	0	SRKW (20-70)
2022-Nov-02	Frame A/B	49	0	2,674	WCT (5)
2022-Nov-11	Frame A/B	36	4,075	0	SRKW (12-40)
2022-Nov-21	Frame A/B	11	5,042	0	SRKW (7-30)
2022-Dec-27/28	Frame A/B	59	8,092	0	SRKW (6-10)
2023-Jan-01	Frame A/B	8	0	212	None
2023-Mar-31	Frame A/B	52	0	3,628	WCT (6-25)
2023-Jun-01	Frame A	4	1,917	0	SRKW (3-20)
2023-Jun-30	Frame A	8	3,361	0	SRKW (8-20)
2023-Oct-20	Frame A/B	11	0	32	None
2023-Nov-20	Frame A	3	45	0	SRKW (40)
2023-Dec-15	Frame A/B	79	0	226	WCT (5)
2023-Dec-21	Frame A/B	31	957	0	None
2024-Jan-20	Frame A	2	0	5	None
2024-Feb-02	Frame A	13	311	0	None
2024-Feb-23	Frame A/B	10	0	300	WCT (2-3)
2024-Mar-12	Frame A/B	3	0	62	None
2024-Mar-26	Frame A	5	0	34	WCT (3)
2024-Apr-03	Frame A/B	9	0	62	WCT (2-10)
2024-Apr-19	Frame A	16	1,791	0	SRKW (3-25)
2024-May-10	Frame A/B	76	16,413	0	None
2024-May-20	Frame B	3	0	9	WCT (5-6)
2024-May-21	Frame A	11	2,076	0	SRKW (20)
2024-May-29	Frame A/B	64	13,161	0	SRKW (25)
2024-Jun-01	Frame A	16	178	0	SRKW (40)
2024-Jun-13	Frame A/B	100	16,732	0	SRKW (10-25)
2024-Jun-18/19	Frame B	5	576	0	SRKW (15)
2024-Jun-21*	Frame A/B	7	0	52	None
2024-Jun-21*	Frame A/B	64	18,472	0	SRKW (15-20)
2024-Jun-24	Frame A	4	0	10	WCT (5)
2024-Jun-26	Frame A/B	46	23,967	0	SRKW (25)
2024-Jul-06	Frame A/B	14	0	179	WCT (3)
2024-Jul-24	Frame A/B	25	2,313	0	SRKW (20-25)
2024-Jul-28/29	Frame A/B	120	9,818	0	SRKW (10)
2024-Aug-29	Frame A/B	8	0	140	WCT (5)
2024-Sep-07	Frame A/B	26	0	1,123	WCT (4-8)
2024-Sep-15	Frame A/B	76	19,696	0	SRKW (15-20)
2024-Sep-29	Frame A/B	14	0	51	WCT (2-10)
2024-Oct-16	Frame B	2	0	14	None
2024-Nov-05	Frame A/B	33	0	4,018	None
2024-Nov-25	Frame A/B	43	4,906	0	None
2025-Jan-11	Frame A/B	42	0	487	WCT (5)
2025-Jan-23	Frame A/B	16	0	550	WCT (6-8)
Total 50	48/39	1,401	188,636	17,101	37

*SRKW and WCT killer whale events were recorded approximately 14 hours apart.

Appendix B

Logistic Regression Model Assumptions

Linearity of the Logit

Figure B.1, a plot of deviation residuals versus fitted values, revealed a mild curvature, indicating a slight nonlinearity. However, the residuals appeared symmetrically spread around zero, suggesting no major violations.

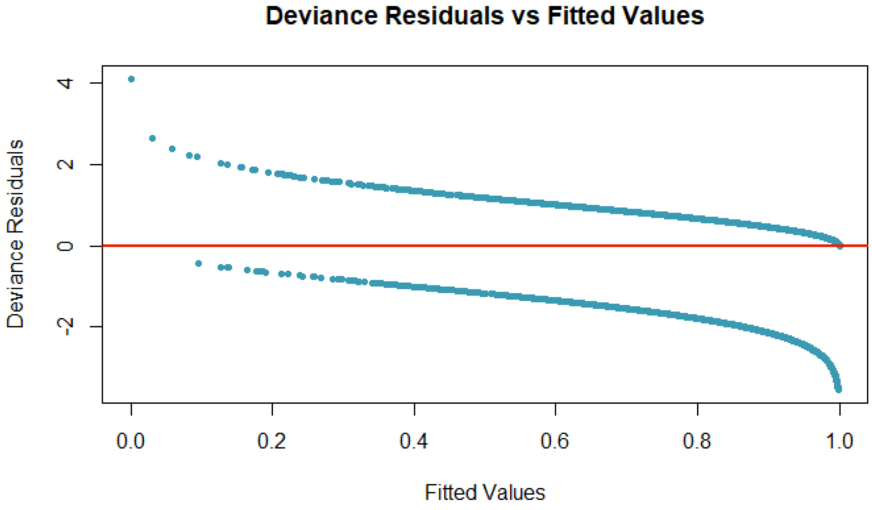

Figure B.1: Deviance residuals versus fitted values for the logistic regression model.

Independence of Errors

Figure B.2, an autocorrelation function (ACF) plot of the deviation residuals, did not show a significant autocorrelation between the lags, supporting the assumption that the observations are independent.

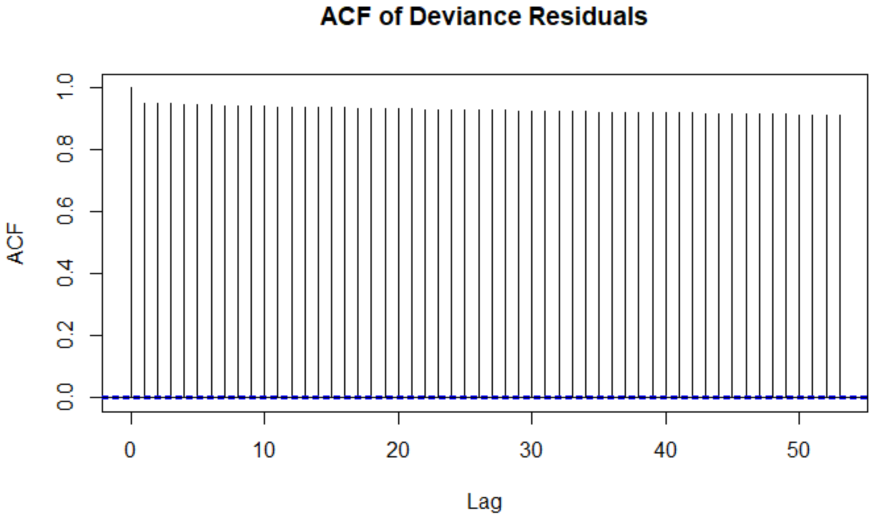

Figure B.2: Autocorrelation function (ACF) plot of deviance residuals for the logistic regression model.

Influence and Leverage

Figure B.3 and B.4 provide a Cook's distance and leverage (hat value) plot, which identified only a few slightly influential or high-leverage points, all of which were real data points.

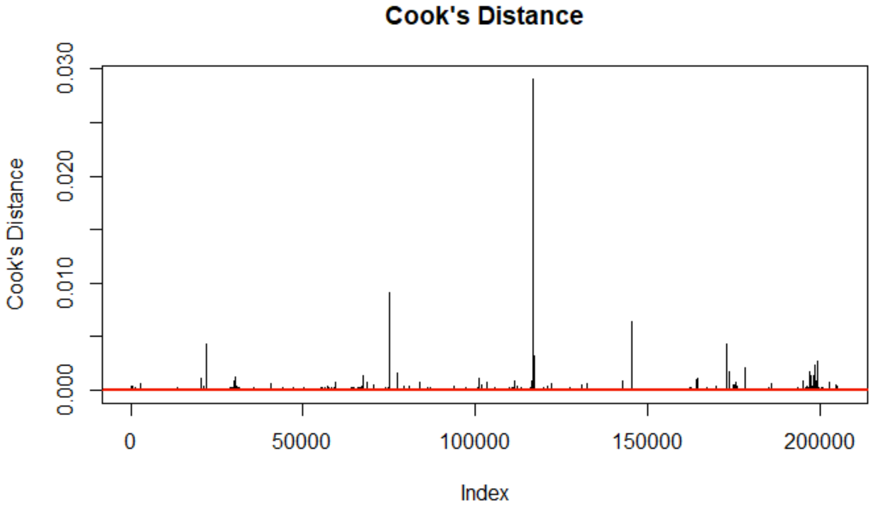

Figure B.3: Cook's distance plot for the logistic regression model. Only a few observations show slightly elevated influence.

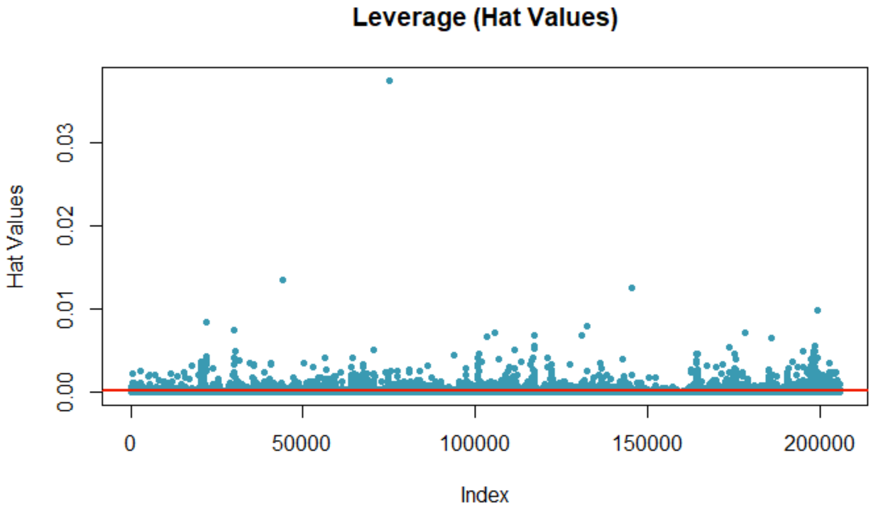

Figure B.4: Leverage (hat values) plot for the logistic regression model.

Calibration

A calibration plot (Figure B.5) indicated good agreement between predicted probabilities and observed outcomes, especially in higher probability ranges.

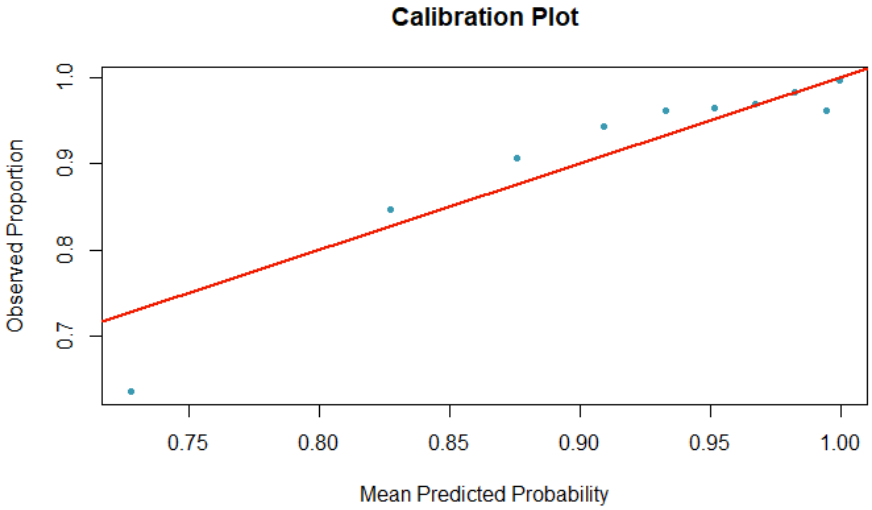

Figure B.5: Calibration plot for the logistic regression model.

Multicollinearity

A correlation matrix showed moderate to high correlations among several frequency-related features (e.g., centroid frequency, 3dB down, and 10dB down frequencies, with some correlations exceeding 0.9). This demonstrates the need for applying lasso penalization and further feature selection and weighting implemented in all models.

Appendix C

UniLasso Algorithm

1. For $j = 1, 2, \ldots, p$, fit the p separate univariate least-squares models $\hat{\eta}_j(x_j) = \hat{\beta}_{0j} + \hat{\beta}_j x_j$ to the response y. For $j = 1, 2, \ldots, p$ and each $i = 1, 2, \ldots, n$, compute the leave-one-out (LOO) fitted value $\hat{\eta}_j^{-i} = \hat{\beta}_{0j}^{-i} + \hat{\beta}_j^{-i} x_{ij}$, resulting in an $n \times p$ feature matrix $F = \{\hat{\eta}_j^{-i}\}$.

2. Fit the Lasso, with an intercept, no standardization, and non-negativity constraints, to target y and these LOO fits as features, giving coefficients $\hat{\theta} = (\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_p)^T$:

$$\min_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \theta_0 - \sum_{j=1}^{p} \theta_j \hat{\eta}_j^{-i} \right)^2 + \lambda \sum_{j=1}^{p} |\theta_j| \right\}, \quad \text{subject to } \theta_j \geq 0,\, j = 1, \ldots, p. \tag{1}$$

   and return the composite model:

$$\hat{\eta}(x) = \hat{\theta}_0 + \sum_{j=1}^{p} \hat{\theta}_j \hat{\eta}_j(x_j). \tag{2}$$

   Since each of the constituent models $\hat{\eta}_j(x_j)$ are linear, this final model can be written:

$$\hat{\eta}(x) = \hat{\gamma}_0 + \sum_{j=1}^{p} \hat{\gamma}_j x_j, \tag{3}$$

   with $\hat{\gamma}_j = \hat{\theta}_j \hat{\beta}_j$, and $\hat{\gamma}_0 = \hat{\theta}_0 + \sum_{\ell=1}^{p} \hat{\theta}_\ell \hat{\beta}_{0\ell}$.

Appendix D

Summary of Models

Appendix D summarizes the 21 models trained in this study in Table D.1.

Table D.1: Model number and description for the 21 models trained using lasso, UniLasso, pre-trained lasso, random forest, and eXtreme gradient boosting (XGBoost) methods.

Model Number	Model Type	SNR	Cross-validation
1	lasso	All	10-fold
2	lasso	High	10-fold
3	lasso	Medium	10-fold
4	lasso	Low	10-fold
5	lasso (pairwise interactions)	All	10-fold
6	lasso	All	manual 10-fold
7	lasso	All	manual 3-fold
8	UniLasso	All	10-fold
9	UniLasso	High	10-fold
10	UniLasso	Medium	10-fold
11	UniLasso	Low	10-fold
12	UniLasso (pairwise interactions)	All	10-fold
13	pre-trained lasso	All, High, Medium, Low	10-fold
14	random forest	All	Out-of-bag error
15	random forest	High	Out-of-bag error
16	random forest	Medium	Out-of-bag error
17	random forest	Low	Out-of-bag error
18	XGBoost	All	5-fold
19	XGBoost	High	5-fold
20	XGBoost	Medium	5-fold
21	XGBoost	Low	5-fold

Appendix E

Model Performance Metrics

Appendix E summarizes the classification performance of the 21 models. Table E.1 details the number of misclassifications and selected features across models. Table E.2, E.3, and E.4 summarize performance metrics for echolocation click, file, and event level predictions (True Positive, False Positive, True Negative, and False Negative counts, and Precision, Recall, F_2 Score, Matthew's Correlation Coefficient, and Area under the Receiver Operator Characteristic Curve.

Table E.1: The number of misclassified killer whale events and the corresponding dates and features selected for all 21 models fit in this study along with their corresponding $\hat{\lambda}$ values for the lasso models. For ensemble methods, where all features were retained yet with differing importance, only the top four most important features (those contributing the greatest decrease in Gini impurity) were included in this figure Acoustic features are labeled as follows: 1: Click Rate, 2: High Frequency, 3: Low Frequency, 4: Centroid Frequency, 5: 10 dB Down Bandwidth, 6: Duration, 7: 3 dB Down Bandwidth, 8: 3 dB Down Low Frequency, 9: Peak Frequency, 10: 10 dB Down High Frequency, 11: 10 dB Down Low Frequency, 12: 3 dB Down High Frequency.

Model	**Misclassified SRKW Events (Day)/Total**	**Misclassified WCT Events (Day)/Total**	**Features Selected (Total)**	**'Best'** λ
1	0/5	1 (2024-11-05)/8	1, 2, 3, 4, 5, 6, 8 (7)	1.92e-04
2	0/5	1 (2024-11-05)/8	1, 2, 3, 4, 5, 6, 7, 11 (8)	1.58e-03
3	0/5	1 (2024-11-05)/8	1, 2, 3, 4, 5, 6, 7, 8, 12 (9)	1.96e-04
4	0/5	1 (2024-11-05)/8	1, 2, 3, 4, 5, 6, 7, 9 (8)	4.21e-04
5	0/5	2 (2024-09-07, 2024-11-05)/8	1,2,3,4,5,6,7,8,9,10 (10+61 interactions)	1.42e-05
6	0/5	1 (2024-11-05)/8	1, 2, 3, 4, 5, 6, 8 (7)	1.92e-04
7	0/5	1 (2024-11-05)/8	1, 2, 3, 4, 5, 6, 8 (7)	1.92e-04
8	0/5	1 (2024-11-05)/8	1, 2, 3, 5, 7, 8 (6)	1.29e-05
9	0/5	1 (2024-11-05)/8	1, 2, 3, 5, 6 (5)	3.91e-04
10	0/5	1 (2024-11-05)/8	1, 2, 3, 5, 6, 7, 11 (7)	1.74e-05
11	0/5	1 (2024-11-05)/8	1, 2, 3, 4, 6, 7, 9, 12 (8)	2.35e-05
12	0/5	1 (2024-11-05)/8	1,2,3,5,7,8,9 (7+10 interactions)	1.29e-05
13	0/5	1 (2024-11-05)/8	1, 2, 3, 4, 5, 6, 7, 9, 10, 12 (10)	NA
14	0/5	1 (2024-11-05)/8	1, 2, 3, 4 (top 4)	NA
15	1 (2024-07-24)/5	0/8	1, 2, 3, 4 (top 4)	NA
16	0/5	1 (2024-11-05)/8	1, 2, 3, 4 (top 4)	NA
17	0/5	1 (2024-11-05)/8	1, 2, 3, 4 (top 4)	NA
18	0/5	1 (2024-11-05)/8	1, 2, 3, 4 (top 4)	NA
19	0/5	1 (2024-11-05)/8	1, 2, 3, 4 (top 4)	NA
20	0/5	1 (2024-11-05)/8	1, 2, 3, 4 (top 4)	NA
21	0/5	1 (2024-11-05)/8	1, 2, 3, 4 (top 4)	NA

Table E.2: Test performance metrics at the echolocation click level for the 21 models. TP: True Positive; FP: False Positive; TN: True Negative; FN: False Negative; F_2 Score; MCC: Matthew's Correlation Coefficient; ROC-AUC: Area under the Receiver Operator Characteristic Curve.

Model	**TP**	**FP**	**TN**	**FN**	**Precision**	**Recall**	**F_2 Score**	**MCC**	**ROC-AUC**
1	39535	1723	4839	21165	0.96	0.65	0.69	0.24	0.74
2	12469	809	1954	9146	0.94	0.58	0.63	0.18	0.68
3	13257	603	1436	6239	0.96	0.68	0.72	0.24	0.74
4	12966	335	1425	6623	0.97	0.66	0.72	0.27	0.81
5	46582	2842	3720	14118	0.94	0.77	0.80	0.22	0.72
6	39535	1723	4839	21165	0.96	0.65	0.69	0.24	0.74
7	39535	1723	4839	21165	0.96	0.65	0.69	0.24	0.74
8	39615	1713	4849	21085	0.96	0.65	0.69	0.24	0.74
9	12712	852	1911	8903	0.94	0.59	0.64	0.18	0.68
10	13241	598	1441	6255	0.96	0.68	0.72	0.24	0.74
11	12995	336	1424	6594	0.97	0.66	0.72	0.30	0.81
12	45330	2436	4126	15370	0.95	0.75	0.78	0.25	0.73
13	39244	1930	4632	21456	0.95	0.65	0.69	0.22	0.72
14	36431	2219	4343	24269	0.94	0.60	0.65	0.16	0.67
15	13373	1001	1762	8242	0.93	0.62	0.66	0.17	0.67
16	12615	706	1333	6881	0.95	0.65	0.69	0.18	0.68
17	11438	306	1454	8151	0.97	0.58	0.63	0.23	0.74
18	50454	3675	2887	10246	0.93	0.83	0.85	0.20	0.69
19	17398	1502	1261	4217	0.92	0.81	0.83	0.20	0.67
20	15998	1040	999	3498	0.94	0.82	0.84	0.22	0.69
21	14261	672	1088	5328	0.96	0.73	0.76	0.21	0.75

Table E.3: Test performance metrics at the file level for all 21 models. TP: True Positive; FP: False Positive; TN: True Negative; FN: False Negative; F_2 Score; MCC: Matthew's Correlation Coefficient; ROC-AUC: Area under the Receiver Operator Characteristic Curve.

Model	**TP**	**FP**	**TN**	**FN**	**Precision**	**Recall**	**F_2 Score**	**MCC**	**ROC-AUC**
1	233	61	94	77	0.79	0.75	0.76	0.35	0.73
2	200	35	97	101	0.85	0.66	0.70	0.37	0.77
3	256	59	70	45	0.81	0.85	0.84	0.41	0.76
4	228	49	86	79	0.82	0.74	0.76	0.30	0.75
5	215	58	97	95	0.79	0.69	0.71	0.31	0.67
6	233	61	94	77	0.79	0.75	0.76	0.35	0.72
7	233	61	94	77	0.79	0.75	0.76	0.35	0.72
8	235	63	92	75	0.79	0.76	0.77	0.35	0.72
9	200	35	97	101	0.85	0.66	0.67	0.36	0.77
10	250	59	70	51	0.81	0.83	0.83	0.38	0.75
11	228	49	86	79	0.82	0.74	0.76	0.37	0.75
12	232	63	92	78	0.79	0.75	0.76	0.33	0.68
13	199	49	106	111	0.64	0.80	0.76	0.31	0.68
14	244	59	96	66	0.81	0.79	0.79	0.40	0.73
15	208	38	94	93	0.85	0.69	0.72	0.38	0.75
16	195	28	101	106	0.87	0.65	0.68	0.40	0.78
17	247	54	81	60	0.82	0.80	0.81	0.40	0.76
18	225	69	86	85	0.77	0.73	0.74	0.27	0.65
19	200	34	98	101	0.85	0.66	0.69	0.38	0.67
20	190	29	100	111	0.87	0.63	0.67	0.33	0.78
21	235	49	86	72	0.83	0.77	0.78	0.39	0.78

Table E.4: Test performance metrics at the event level for all 21 models. TP: True Positive; FP: False Positive; TN: True Negative; FN: False Negative; F_2 Score; MCC: Matthew's Correlation Coefficient; ROC-AUC: Area under the Receiver Operator Characteristic Curve.

Model	**TP**	**FP**	**TN**	**FN**	**Precision**	**Recall**	**F_2 Score**	**MCC**	**ROC-AUC**
1	5	1	7	0	0.83	1.00	0.96	0.85	0.98
2	5	1	7	0	0.83	1.00	0.96	0.85	0.98
3	5	1	7	0	0.83	1.00	0.96	0.85	0.98
4	5	1	7	0	0.83	1.00	0.96	0.85	0.98
5	5	2	6	0	0.71	1.00	0.92	0.73	0.93
6	5	1	7	0	0.83	1.00	0.96	0.85	0.98
7	5	1	7	0	0.83	1.00	0.96	0.85	0.98
8	5	1	7	0	0.83	1.00	0.96	0.85	0.98
9	5	1	7	0	0.83	1.00	0.96	0.85	0.98
10	5	1	7	0	0.83	1.00	0.96	0.85	0.98
11	5	1	7	0	0.83	1.00	0.96	0.85	0.98
12	5	1	7	0	0.83	1.00	0.96	0.85	0.98
13	5	1	7	0	0.83	1.00	0.96	0.85	0.98
14	5	1	7	0	0.83	1.00	0.96	0.85	0.98
15	4	0	8	1	1.00	0.80	0.83	0.84	0.95
16	5	1	7	0	0.83	1.00	0.96	0.85	0.98
17	5	1	7	0	0.83	1.00	0.96	0.85	0.98
18	5	1	7	0	0.83	1.00	0.96	0.85	0.98
19	5	1	7	0	0.83	1.00	0.96	0.85	0.93
20	5	1	7	0	0.83	1.00	0.96	0.85	0.95
21	5	1	7	0	0.83	1.00	0.96	0.85	0.98