Open-source machine learning BANTER acoustic classification of beaked whale echolocation pulses

Shannon Rankin [a,*], Taiki Sakai [b], Frederick I. Archer [a], Jay Barlow [a], Danielle Cholewiak [c], Annamaria I. DeAngelis [c], Jennifer L.K. McCullough [d], Erin M. Oleson [d], Anne E. Simonis [b], Melissa S. Soldevilla [e], Jennifer S. Trickey [d]

[a] Marine Mammal and Turtle Division, Southwest Fisheries Science Center, National Marine Fisheries, National Oceanic and Atmospheric Administration, La Jolla, CA, United States
[b] Ocean Associates, Inc., Arlington, VA, United States
[c] Northeast Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Woods Hole, MA, United States
[d] Pacific Islands Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Honolulu, HI, United States
[e] Southeast Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Miami, FL, United States

ARTICLE INFO

ABSTRACT

Passive acoustic monitoring is increasingly used for assessing populations of marine mammals; however, analysis of large datasets is limited by our ability to easily classify sounds detected. Classification of beaked whale acoustic events, in particular, requires evaluation of multiple lines of evidence by expert analysts. Here we present a highly automated approach to acoustic detection and classification using supervised machine learning and open source software methods. Data from four large scale surveys of beaked whales (northwestern North Atlantic, southwestern North Atlantic, Hawaii, and eastern North Pacific) were analyzed using PAMGuard (acoustic detection), PAMpal (acoustic analysis) and BANTER (hierarchical random forest classifier). Overall classification accuracy ranged from 88% for the southwestern North Atlantic data to 97% for the northwestern North Atlantic. Results for many species could likely be improved with increased sample sizes, consideration of alternative automated detectors, and addition of relevant environmental features. These methods provide a highly automated approach to acoustic detection and classification using open source methods that can be readily adopted for other species and geographic regions.

1. Introduction

Passive acoustic monitoring (PAM) has proven to be a valuable tool for studying populations of marine mammals (Parijs et al., 2009); however, the value of these studies depends on our ability to identify the sources of the sounds we are monitoring. Historically, experienced acousticians have manually identified stereotyped sounds that could reliably be attributed to a given species, based on their spectral or temporal characteristics (Baumann-Pickering et al., 2013; Bittle and Duncan, 2013; Rankin and Barlow, 2005; Soldevilla et al., 2008). However, the substantial increase in the quantity of recordings being made make it impossible for experienced acousticians to manually annotate it all. This has led to an increased need for the development of automated

classification routines that can provide accurate species determinations from acoustic recordings.

Advances in computing have led to a dramatic increase in the opportunities to use machine learning for bioacoustic classification (Bianco et al., 2019; Mutanu et al., 2022; Stowell, 2022). In particular, use of deep neural networks show great promise for large datasets on a wide variety of taxa, including marine mammals (Frainer et al., 2023; Frasier et al., 2017; Li et al., 2020a; Román Ruiz et al., 2023; Zhong et al., 2020). Neural networks typically require significant sample sizes for all classes, although there are increasingly new tools to enhance data or combined disparate datasets (Dufourq et al., 2022; Nanni et al., 2020; Nolasco et al., 2023; Padovese et al., 2023; Shorten and Khoshgoftaar, 2019). There is a significant learning curve to understand how to

properly apply and test neural network classification methods on a novel dataset, although development of more user friendly tools are in development (Kahl et al., 2021; Kirsebom et al., 2022). Despite the great potential for deep learning methods, modification of existing machine learning methods with a lower barrier to entry may be sufficient for many bioacoustic classification tasks.

Random Forest based acoustic classifiers have been found to be relatively easy, efficient, and effective approach to bioacoustic classification (López-Baucells et al., 2019; Ross and Allen, 2014). BANTER (**B**io-**A**coustic eve**NT** classifi**ER**) is a hierarchical Random Forest (Breiman, 2001) event classifier originally developed for species identification of dolphin signals from acoustic recordings (Rankin et al., 2017). BANTER combines information from independent classification models for each call type (e.g., whistles, echolocation pulses, and burst pulses for the dolphin-based model) into a model classifying a full acoustic event, which is defined as a discrete collection of calls during an encounter. The event classifier uses the distribution of species assignment probabilities for each call type, along with any event-level metrics (such as call rate). BANTER is very flexible and can accommodate any number of measures from any number of detectors used to identify calls from any species. BANTER is easy to use and is readily available as a stand-alone R package (Archer, 2022a).

While BANTER was designed to classify events by integrating information from multiple types of call detectors, relatively minor changes to a detector for a specific call type could yield different detector results. In the dolphin-based model (Rankin et al., 2017), settings for a whistle and moan detector were modified to improve performance at detecting burst pulses. This approach of applying output from multiple call detectors (of the same type but with different settings) suggests that perhaps BANTER could be used on species that only produce one call type (or where only one call type is to be analyzed), as long as the settings of the automated call detector were modified such that the results were different. This approach was successfully applied to classified echolocation pulses for narwhals and belugas (Zahn et al., 2021). For species that primarily (or exclusively) produce echolocation pulses, BANTER may serve as an option for automated machine learning classification.

Beaked whales are deep diving marine mammals found in offshore waters; their long dive intervals and cryptic surfacing behavior make them difficult to study using typical shipboard visual observation methods (MacLeod, 2018). However, beaked whales make stereotyped echolocation signals in the form of frequency-modulated (FM) pulses that have temporal and spatial characteristics that can be used to differentiate species (Baumann-Pickering et al., 2013). Detection of beaked whale signals in large datasets is greatly expanding our understanding of their population structure and the potential impact of human activities on these species (Barlow et al., 2022b; Baumann-Pickering et al., 2014; Simonis, 2020). Unfortunately, manual classification of beaked whale echolocation pulses requires analysis by trained analysts using a number of visual aids to examine the call characteristics. Development of automated classification routines, if accurate, serves to improve the efficiency, reduce the subjectivity, and decrease the cost of analyzing large datasets.

Here we applied BANTER acoustic classification to beaked whale detections from large acoustic datasets from the U.S. East Coast, Hawaii, and the U.S. West Coast. Acoustic data were analyzed by experienced acousticians to determine species identity, and these 'ground truth' classifications served as training data for the supervised BANTER models. Our goals were to identify an efficient and accurate automated approach to acoustic species identification of beaked whales, and to provide a framework for analysis that may serve for other PAM studies.

2. Materials and methods

2.1. Field data collection

Passive acoustic data were collected during surveys conducted by the National Oceanographic and Atmospheric Administrations' (NOAA) research operations off the U.S. East Coast, the Hawaiian Islands, and the U.S. West Coast.

Recordings from the U.S. East Coast (western North Atlantic Ocean) were collected using towed hydrophone arrays during the 2016 AMAPPS II (Atlantic Marine Assessment Program for Protected Species) survey by Northeast Fisheries Science Center and Southeast Fisheries Science Center (United States National Marine Fisheries Service, 2016). AMAPPS II was subdivided into the northern (NAtlantic) and southern (SAtlantic) survey areas. The NAtlantic study area ranged from Massachusetts south to New Jersey (HB1603, Appendix A, (United States National Marine Fisheries Service, 2016)) and recordings used in this analysis were collected with two hydrophones (HTI-96-min, High Tech Inc., Long Beach, MS), which recorded at a 192 kHz sample rate with a 1 kHz high pass filter (National Instruments USB-6356 A/D card, see DeAngelis et al. (2018) for more information). The SAtlantic area ranged from Delaware south to central Florida (GU1605, Appendix B, (United States National Marine Fisheries Service, 2016)) and recordings used in this analysis were collected with two hydrophones (Reson TC4013, Teledyne Marine, Slangerup, Denmark), which recorded at a 500 kHz sample rate (custom 12 channel SailDAQ soundcard) and were decimated to 192 kHz with a 1 kHz high pass filter (see GU1605, Appendix B, (United States National Marine Fisheries Service, 2016) for more information). The NAtlantic and SAtlantic recordings were treated as two separate datasets due to differences in hydrophone frequency responses and sensitivities.

Recordings from Hawaii were collected using drifting acoustic recorders during the 2017 HICEAS (Hawaiian Island Cetacean Ecosystem Assessment Survey, McCullough et al. (2021), Yano et al. (2018)). Drifting acoustic recorders were deployed at randomly selected locations within the Main Hawaiian Islands. Data were collected using SoundTrap ST4300 recorders (Ocean Acoustics, Auckland, New Zealand) sampled at 288 kHz with a duty cycle of 2 min on for every 10 min. Individual drifting recorders were deployed for 11 to 23 days (see McCullough et al. (2021), Yano et al. (2018) for more detailed information).

Recordings from the eastern North Pacific Ocean off the U.S. West Coast (EPacific) were collected using drifting acoustic recorders during the 2018 CCES (California Current Ecosystem Survey) (Simonis, 2020). Drifting recorders were deployed at 23 stations distributed throughout the California Current region; data from 15 of these included high quality acoustic data used in this analysis. Two types of drifting recorders were used: (1) SoundTrap ST4300 recorders (Ocean Acoustics, Auckland, New Zealand) and (2) Wildlife Acoustics SM3M recorders (Wildlife Acoustics, Maynard, MA). Recordings were duty cycled and recorded 2 min out of a variable 'off' time; recorders sampled at a minimum 256 kHz sample rate. Individual drifting recorders were deployed for 5–79 days (see Simonis (2020) for more detailed information).

2.2. Acoustic data analysis

All acoustic recordings were analyzed using PAMGuard software (versions 2.00.15c, 2.00.16e) with a suite of generic click classifiers within the Click Detector module (see Keating and Barlow (2013)). Classifier sets were saved such that any click may be classified as more than one click type (using 'save classifier set' in Pamguard click classification window). Five general click classifiers were treated as spectral band click detectors (i.e., 2–15 kHz, 15–30 kHz, 30–50 kHz, 50–80 kHz, and > 80 kHz). An additional sixth detector within the 30–50 kHz peak frequency range considered the presence of a frequency sweep that is

characteristic of beaked whale pulses (see Keating and Barlow (2013)).

Echolocation signals that were identified as frequency-modulated pulses from beaked whales were then linked to PAMGuard "events" by experienced acoustic analysts. Events included all the pulses that were close to each other in time, had similar spectral and waveform characteristics, and were received at consistent bearing angles. For NAtlantic and SAtlantic datasets, acoustic detections of beaked whales using towed hydrophone arrays were further subdivided into putative individuals by identification of consecutive pulses along the same bearing angle in the bearing time plot (see DeAngelis et al. (2017)).

Experienced acoustic analysts used multiple lines of evidence (PAMGuard click detector bearing time display, click waveform display, click spectrum, and wigner plots) to assess the species identity of the acoustic event. When the available characteristics were inconclusive, the species was considered an unidentified beaked whale. Data were saved within the PAMGuard database and binaries for downstream processing using the *PAMpal package* (Sakai, 2021) in R v2022.07.02 (Team, 2013).

Standard click metrics describing the spectral and temporal characteristics were measured from echolocation pulses using default values in *PAMpal* (Sakai, 2021). Inter-pulse interval (IPI) was calculated using the 'calculateICI' function in *PAMpal* and the mode of the IPI was identified for each beaked whale acoustic event. In addition, for the EPacific data, we used the 'matchEnvData' function in *PAMpal* to include ERDAPP environmental data (Simons and John, 2022), including seafloor depth and seafloor gradient (both from erdSrtm30plusSeafloorGradient), as well as sea surface temperature (from jplMURSST41). A full consideration of environmental data for all regions was beyond the scope of this study.

Multiple models were created to better understand the value (and potential limitations) of the different suite of measurements. Models included standard click metrics (only) for all pulses in a beaked whale acoustic event (EC); the addition of IPI for each event (EC_IPI), and for the EPacific dataset we also included environmental variables (EC_IPI_ENV).

2.3. BANTER acoustic classification

BANTER models were created using the *banter* package (Archer, 2022a) in R v2022.07.02 (Team, 2013). Output from PAMGuard click detector modules were used to create the BANTER detector models, for a total of 6 detector models (one for each click detector used). For each detector model, BANTER creates a mean of the species classification probabilities for each call type, which is applied to the BANTER event model (in conjunction with any event-level variables). A minimum of two events per species and the complete suite of variables (no missing data) were required for training and testing of the classification model. Event level variables included IPI and environmental variables (for the EPacific dataset). The Random Forest models that BANTER is based on are largely driven by two parameters, the number of trees in the forest (*ntree*) and the number of samples randomly drawn to build each tree (*sampsize*). Small values of *sampsize* were selected to improve computational performance, but large values of *ntree* were used to obtain models with stable classification results (Rankin and Archer, 2021). Model results were summarized using the *rfpermute* package (Archer, 2022b).

3. Results

Four datasets from four different regions (NAtlantic, SAtlantic, Hawaii, EPacific) were analyzed for this study, and species included: Baird's beaked whales (*Berardius bairdii*), Blainville's beaked whales (*Mesoplodon densirostris*), Cross Seamount beaked whales (McDonald et al., 2009), Cuvier's beaked whales (*Ziphius cavirostris*), Gervais' beaked whales (*M. europaeus*), Hubbs' beaked whales (*M. carlhubbsi*), Longman's beaked whales (*Indopacetus pacificus*), Sowerby's beaked

whales (*M. bidens*), Stejneger's beaked whales (*M. stejnegeri*), True's beaked whales (*M. mirus*), and unidentified beaked whale 'BW43' (Baumann-Pickering et al., 2013). Analyses were conducted separately for the different study areas due to differences in the data collection methods that precluded combining of datasets. Species and sample sizes varied by study area.

3.1. NAtlantic

Species encountered in the NAtlantic dataset include Cuvier's beaked whales ($n = 120$), Gervais' beaked whales ($n = 4$), Sowerby's beaked whales ($n = 6$), and True's beaked whales ($n = 76$). BANTER classification models included (1) echolocation pulses (EC), and (2) echolocation pulses and IPI (EC_IPI).

For the NAtlantic dataset, the most accurate model included echolocation pulse metrics with IPI; this model had a modest increase in accuracy over the EC (only) model (EC_IPI model, Detector Model: *sampsize* = 3, *ntree* = 10,000; Event Model: *sampsize* = 3, *ntree* = 100,000). This model provided an overall correct classification rate of 97.5% for all four species (pct.correct in Confusion Matrix, Fig. 1 a). Classification scores ranged from 75% for Gervais' beaked whales to 100% for Cuvier's and Sowerby's beaked whales. All classification results were greater than expected (Priors in Fig. 1 a). Results from the EC_IPI model are presented here; results from the EC model were similar and can be found in supplementary materials (Supplement Fig. S1).

The proximity plot (Fig. 1 b) provides a view of the distribution of events within the classification model space. For each event in the plot, the color of the central dot represents the true species identity, and the circle represents the BANTER classification result. This plot shows the degree of overlap between Gervais', Sowerby's, and True's beaked whales. Species classifications of events are resolved using additional features (mean assignment probabilities for each of the detectors in the detector model, as well as any event level variables), as seen in the ten most important features for predicting each species (importance heat map, Fig. 1 c). The strength of classifications for each event can be seen in the distribution of votes for each species across the forest (Fig. 1 d). In each frame of this figure, the events are represented as vertical slices along the x-axis, and the percentage of votes for each species is represented by their color along that vertical slice in the y-axis. This distribution shows strong assignment probabilities for most Cuvier's and Sowerby's beaked whale events, with greater variability in assignment for Gervais' and True's beaked whale events.
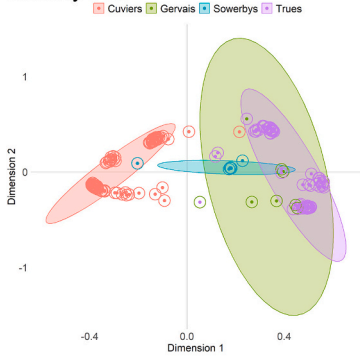
3.2. SAtlantic

Species encountered in the SAtlantic study dataset include Blainville's beaked whales ($n = 46$), Cuvier's beaked whales ($n = 111$), Gervais' beaked whales ($n = 77$), Sowerby's beaked whales ($n = 2$), and True's beaked whales ($n = 13$). BANTER classification models included (1) echolocation pulses (EC), and (2) echolocation pulses and IPI (EC_IPI).

For the SAtlantic dataset, the most accurate model considered the echolocation pulse metrics and the IPI for an event (EC_IPI Model, Detector Model: *sampsize* = 4, *ntree* = 10,000; Event Model: *sampsize* = 1, *ntree* = 100,000) and provided an overall correct classification rate of 88.7% for all five species. Classification scores ranged from a low of 50% (Sowerby's beaked whale) to a high of 100% (True's beaked whale, Fig. 1); all classification results were greater than expected (Priors in Fig. 1 a). The sample sizes for the BANTER model were kept low (event sampsize = 1) to retain Sowerby's beaked whale, which had a sample size of 2. Analysis with a larger sample size resulted in improved classification results for Cuvier's and Gervais' beaked whale (see Supplement Fig. S3 for EC_IPI_ALT), but resulted in the loss of Sowerby's in the final model. Results from the EC_IPI model are presented here; results from the EC and EC_IPI_ALT model can be found in supplementary materials (Supplement Fig. S2, S3).
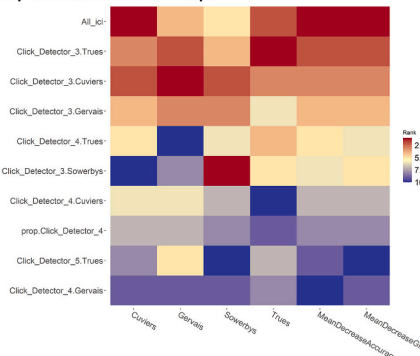
a) Confusion Matrix

	Cuviers	Gervais	Sowerbys	Trues	pct.correct	LCI_0.95	UCI_0.95	priors
Cuviers	120	0	0	0	100.00	96.97	100.00	58
Gervais	0	3	0	1	75.00	19.41	99.37	2
Sowerbys	0	0	6	0	100.00	54.07	100.00	3
Trues	0	4	0	72	94.74	87.07	98.55	37
Overall	**NA**	**NA**	**NA**	**NA**	**97.57**	**94.43**	**99.21**	**48**

b) Proximity Plot c) Importance Heat Map d) Vote Plot
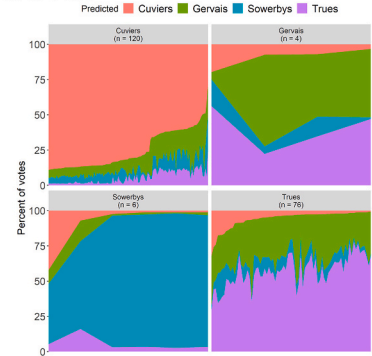
Fig. 1. BANTER classification results from the NAtlantic dataset including echolocation pulses and inter-pulse interval (EC_IPI). Confusion matrix (a) provides the percent correct classification for each species (pct.correct), lower confidence intervals (LCI_0.95), upper confidence intervals (UCI_0.95), and priors (expected error rate). Proximity plot (b) for species events from BANTER model (central dot color represents true species identity; color of circle surrounding dot represents BANTER species classification). Heat map (c) for ranks of ten most important variables; colors scale from most important predictors (dark red) to least important predictors (dark blue). Vote Plot (d) shows the vote distribution for each event (vertical slice) for each species; distribution of votes by species is shown by their representative color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

For this dataset, the two most important features are insufficient to differentiate species, as seen by the overlap in species distribution with these feature spaces in the proximity plot (Fig. 2 b). Importance variables varied by species (importance heat map Fig. 2 c), with IPI measurements having relatively low importance (Fig. 2 c). The distribution of votes (Fig. 2 d) shows relatively weak classification results; most

events consisted of <50% of the trees voting for the correct species.

3.3. Hawaii

The Hawaii dataset consisted of 13 drifting recorders deployed within the Main Hawaiian Islands; species included Blainville's beaked
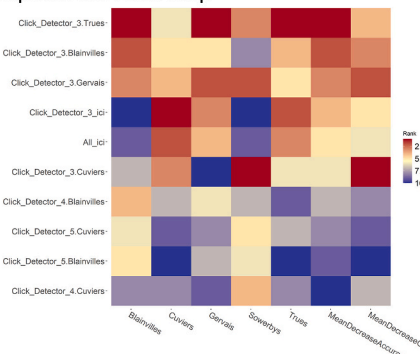
a) Confusion Matrix

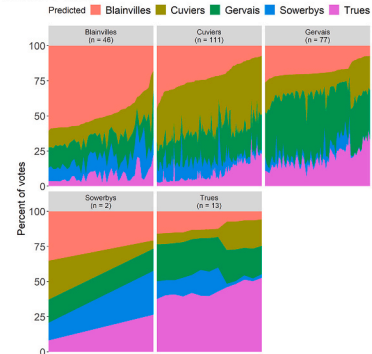|	Blainvilles	Cuviers	Gervais	Sowerbys	Trues	pct.correct	LCI_0.95	UCI_0.95	priors
Blainvilles	43	1	1	1	0	93.48	82.10	99	18
Cuviers	9	92	7	2	1	82.88	74.57	89	45
Gervais	0	1	72	0	4	93.51	85.49	98	31
Sowerbys	1	0	0	1	0	50.00	1.26	99	1
Trues	0	0	0	0	13	100.00	75.29	100	5
Overall	**NA**	**NA**	**NA**	**NA**	**NA**	**88.76**	**84.16**	**92**	**33**
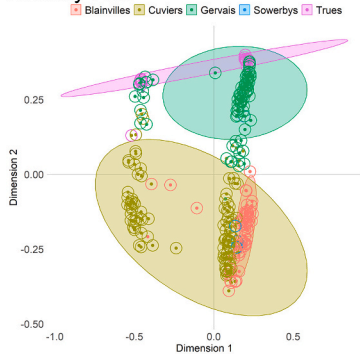
b) Proximity Plot c) Importance Heat Map d) Vote Plot

Fig. 2. BANTER classification results from the SAtlantic dataset (EC_IPI). Confusion matrix (a) provides the percent correct classification for each species (pct. correct), lower confidence intervals (LCI_0.95), upper confidence intervals (UCI_0.95), and priors (expected error rate). Proximity plot (b) for species events from BANTER model (central dot color represents true species identity; color of circle surrounding dot represents BANTER species classification). Heat map (c) for ranks of ten most important variables; colors scale from most important predictors (dark red) to least important predictors (dark blue). Vote Plot (d) shows the vote distribution for each event (vertical slice) for each species; distribution of votes by species is shown by their representative color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

whale (*n* = 521), Cross Seamount beaked whale (*n* = 76), Cuvier's beaked whale (*n* = 201), and Longman's beaked whale (*n* = 122). BANTER classification models included (1) echolocation pulses (EC), and (2) echolocation pulses and IPI (EC_IPI).

The most accurate model considered the echolocation pulses only (EC Model, Detector Model: *sampsize* = 10, *ntree* = 5000; Event model: *sampsize* = 4, *ntree* = 10,000) and provided an overall correct classification rate of 92.3% for all four species. Classification scores ranged from a low of 86% (Cuvier's beaked whale) to 97.5% (Longman's beaked whale, Fig. 3). All classification results were greater than expected (Prior in Fig. 3 a). Results from the EC model are presented here; results from EC_IPI model can be found in supplementary materials (Supplement Fig. S4).

Distinctive clusters can be seen for Longman's and Blainville's beaked whales, although there is overlap on these first two dimensions of the clusters for Cuvier's and Cross Seamount beaked whales (Fig. 3 b). This overlap is resolved with different importance features for Cuvier's and Cross Seamount beaked whales as seen in the importance heat map (Fig. 3 c), resulting in the low misclassification rates for these species (Fig. 3 a confusion matrix). The distribution of votes (Fig. 3 d) show strong classification results for most Blainville's, Cross Seamount, and Longman's beaked whales, with lower classification strength for many Cuvier's beaked whale events.

3.4. EPacific

The EPacific dataset consisted of 15 drifting recorders deployed off the U.S. West Coast, and species included Baird's beaked whale (*n* = 29), unidentified beaked whale 'BW43' (*n* = 125), Cross Seamount beaked whale (*n* = 6), Cuvier's beaked whale (*n* = 926), Hubbs' beaked whale (*n* = 66), and Stejneger's beaked whale (*n* = 42). BANTER classification models included (1) echolocation pulses (EC), (2) echolocation pulses and IPI (EC_IPI), and (3) echolocation pulses, IPI, and environmental features (EC_IPI_ENV).

The best model considered the echolocation pulses, IPI, and environmental features (EC_IPI_ENV Model, Detector Model: *sampsize* = 1,

ntree = 10,000; Event model: *sampsize* = 5, *ntree* = 10,000) and provided an overall correct classification rate of 91.9% for all six species. Classification scores ranged from a low of 91.2% (BW43) to 100% for Cross Seamount beaked whale (Fig. 4 a). All classification results were greater than expected (Prior in Fig. 4 a). Results from the EC_IPI_ENV model are presented here; results from the EC and EC_IPI model can be found in supplementary materials (Supplement Figs. S5, S6).

There is considerable overlap in the primary feature space for all but BW43, as shown in the proximity plot (Fig. 4 b). IPI was among the most important classification variable for all species (Fig. 4 c importance heat map). The distribution of votes (Fig. 4 d) show relatively strong classification results for Baird's, Hubbs', and Stejneger's beaked whales.

4. Discussion

Development of automated acoustic classifiers is critical for efficient analysis of large passive acoustic datasets; these classification routines require representative training data created by experienced acoustic analysts. While beaked whales species can be identified based on their stereotyped frequency-modulated echolocation clicks (Baumann-Pickering et al., 2013), not all acoustic events include sounds that can be readily attributed to a specific species. Moreover, there are a number of beaked whales where little or nothing is known about their acoustic repertoire and new species have been detected in recent years (Barlow et al., 2022a; DeAngelis et al., 2018). Therefore, development of acoustic classifiers for beaked whales should accommodate small and/or unbalanced sample sizes and have sufficient flexibility to allow for variable call characteristics.

The results for the four study areas varied from a low of 88.7% correct classification in the SAtlantic dataset to a high of 97.5% correct classification in the NAtlantic dataset. All species in all datasets had classification rates well above those expected by chance; however Cuvier's (SAtlantic, Hawaii), Sowerby's (SAtlantic), and Gervais' beaked whales (NAtlantic) had classification rates below 90%. Recent beaked whale classification studies using deep learning methods exhibit more robust results (Cohen et al., 2022; LeBien and Ioup, 2018; Li et al.,
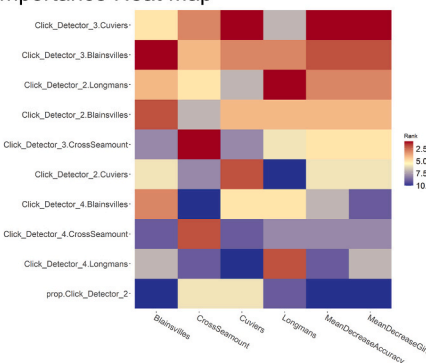
a) Confusion Matrix

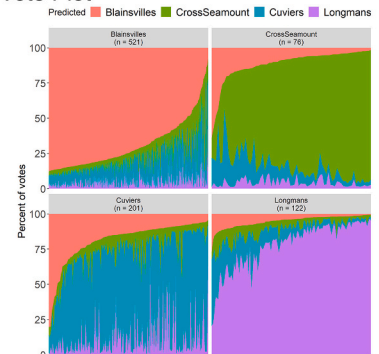|	Blainsvilles	CrossSeamount	Cuviers	Longmans	pct.correct	LCI_0.95	UCI_0.95	priors
Blainsvilles	487	10	23	1	93.47	91.00	95.44	57
CrossSeamount	3	71	2	0	93.42	85.31	97.83	8
Cuviers	16	0	173	12	86.07	80.50	90.54	22
Longmans	0	0	3	119	97.54	92.98	99.49	13
Overall	**NA**	**NA**	**NA**	**NA**	**92.39**	**90.48**	**94.02**	**39**
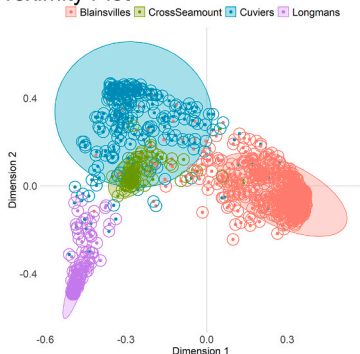
b) Proximity Plot
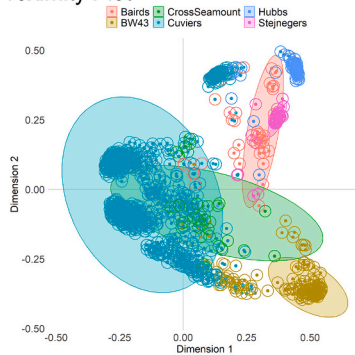c) Importance Heat Map
d) Vote Plot

Fig. 3. BANTER classification results from the Hawaii dataset (EC). Confusion matrix (a) provides the percent correct classification for each species (pct.correct), lower confidence intervals (LCI_0.95), upper confidence intervals (UCI_0.95), and priors (expected error rate). Proximity plot (b) for species events from BANTER model (central dot color represents true species identity; color of circle surrounding dot represents BANTER species classification). Heat map (c) for ranks of ten most important variables; colors scale from most important predictors (dark red) to least important predictors (dark blue). Vote Plot (d) shows the vote distribution for each event (vertical slice) for each species; distribution of votes by species is shown by their representative color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
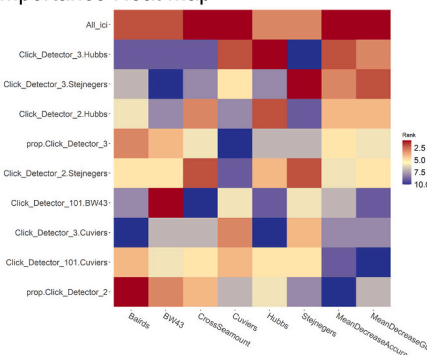
a) Confusion Matrix

	Bairds	BW43	CrossSeamount	Cuviers	Hubbs	Stejnegers	pct.correct	LCI_0.95	UCI_0.95	priors
Bairds	28	0	0	0	1	0	96.55	82	100	2
BW43	0	114	9	2	0	0	91.20	85	96	10
CrossSeamount	0	0	6	0	0	0	100.00	54	100	1
Cuviers	12	31	23	846	11	3	91.36	89	93	78
Hubbs	0	0	1	0	65	0	98.48	92	100	6
Stejnegers	1	0	0	0	2	39	92.86	81	99	4
Overall	**NA**	**NA**	**NA**	**NA**	**NA**	**NA**	**91.96**	**90**	**93**	**62**
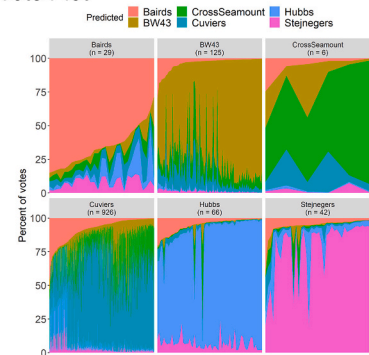
Fig. 4. BANTER classification results from the EPacific dataset with environmental data (EC_IPI_ENV). Confusion matrix (a) provides the percent correct classification for each species (pct.correct), lower confidence intervals (LCI_0.95), upper confidence intervals (UCI_0.95), and priors (expected error rate). Proximity plot (b) for species events from BANTER model (central dot color represents true species identity; color of circle surrounding dot represents BANTER species classification). Heat map (c) for ranks of ten most important variables; colors scale from most important predictors (dark red) to least important predictors (dark blue). Vote Plot (d) shows the vote distribution for each event (vertical slice) for each species; distribution of votes by species is shown by their representative color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2020b; Ziegenhorn et al., 2022); however, these models may not be appropriate for our dataset. Several studies included incomplete beaked whale species for the study area (Li et al. (2020b), Ziegenhorn et al. (2022)) and all studies has sample large sample sizes (>200) for each species class. The deep learning model developed by Cohen et al. (2022) included strong classification results with all beaked whale species found in our Atlantic models; however, it is unclear how this dataset trained on seafloor recorders would perform with towed array datasets. The BANTER models presented here only requires a minimum of two samples per class and can therefore include all species within each region, making them useful to assess the presence of all beaked whales in the region. While deep learning methods may provide improved results once there are sufficient sample sizes for all relevant species within a study area, our results suggest that reasonable classification results can be obtained for all species using BANTER classification methods with as few as two samples per class.

Several of our datasets included very low sample sizes for some species; however, low sample sizes did not always result in low classification scores. In the NAtlantic dataset, Sowerby's beaked whale ($n = 6$) was correctly classified in 100% of the events, with strong assignment probabilities. The Sowerby's beaked whale events in the SAtlantic study and the Gervais' beaked whale events in the NAtlantic study consisted of weak classifications. This is likely due to both a small number of events ($n = 2$ for Sowerby's in SAtlantic and $n = 4$ for Gervais' in NAtlantic) combined with the small *sampsize* used in the BANTER model. In general, larger sample sizes should lead to improved overall classification results, and improved strength of these classification results (as indicated in the vote distributions).

While some species have distinctive differences in their pulse characteristics that can lead to strong classification despite small sample sizes (e.g. Sowerby's beaked whales in NAtlantic), other species, such as Cuvier's beaked whales, have significant variation in their pulse measurements. For species with high variability in the predictor variables and an overlap in the range of these variables with other species, a large

sample size is required to describe the true variability of these call measurements. Even with reasonably large sample sizes, the classification may suffer (e.g., SAtlantic Cuvier's beaked whale = 82.8% correct classification). We found that by increasing the *sampsize* in BANTER, we could increase the classification rate for this species in this area to 91.8% (see Supplement Fig.S3). However, this resulted in the inability to include Sowerby's beaked whale in the final model. So, an increased sample size in Sowerby's would likely lead to increased classification results for other species in the model.

Previous application of BANTER to dolphin species in the California Current found that large sample sizes could result in strong classification of species where experienced analysts are unable to differentiate species (e.g., long-beaked and short-beaked common dolphins, Rankin et al. (2017)). This suggests that large increases in sample sizes may improve classification results for Cuvier's beaked whales.

Gervais' and True's beaked whales have similar pulse characteristics that make them difficult to differentiate and they require considerable expertise to classify manually (DeAngelis et al., 2018). Our results suggest that BANTER may serve as an efficient and effective means of classifying and differentiating these two species. Despite modest sample sizes, Gervais' and True's beaked whales showed high classification scores in the SAtlantic (93.5% and 100%, respectively), and True's performed well in the NAtlantic. Unfortunately, we were unable to combine the Atlantic datasets (which would increase sample sizes of True's beaked whales) due to differences in hydrophone characteristics that resulted in differences in call metrics. Calibration of signals (to make them comparable) may allow for combining datasets to improve sample sizes.

In addition to poor classification of Gervais' in the NAtlantic study area, the plot votes show that for one of the four events, only a few of the 100,000 trees 'voted' for Gervais', and the majority of the votes were for True's beaked whales. Analysts' notes indicated uncertainty in manual species classification of this event in this dataset. While small sample size can at times provide good classification scores, as we found for

Sowerby's beaked whale in this SAtlantic dataset, they can be heavily influenced by inaccuracies in the training data.

These two examples of small sample sizes highlight a conflict in preferred protocol. Ideally, (1) all <u>species</u> would be included in a classification model, so as to better represent the local species diversity, (2) all <u>events</u> would ideally be included in the classification model, so as to better represent the variability found in the area, and (3) only confident 'ground truth' classifications would be considered in the training model. Unfortunately, in species where identity must be determined based on call characteristics (rather than visual confirmation of species identity in the field), it can sometimes be difficult to confidently determine species identity when calls are highly variable and do not include at least one call that provides high confidence of species identity. In the case of beaked whales, we recommend that training data include a high level of confidence for inclusion. This is especially critical for species with small sample sizes. An alternative is to require agreement from multiple analysts. This concern about accurate labeling of training data is further complicated by the potential existence of species that have not yet been identified. For example, Barlow et al. (2022a) recently discovered what appears to be a new species of beaked whale off Baja California, Mexico; this putative new species may have been detected but misclassified in other datasets (expert analysts did not identify these in the EPacific data presented here).

Inter-pulse interval (IPI) was found to be the most important variable for NAtlantic and EPacific, and IPI was the fourth and fifth ranked variables for SAtlantic (see importance heat maps for each survey area). While the addition of IPI did not improve results for data from Hawaii, results were similar and IPI ranked #9 in importance for the Hawaii study area. Baumann-Pickering et al. (2013) showed an overlap in the IPI between Cuvier's and Blainville's beaked whales, but a strong difference in IPI between these species and Cross Seamount beaked whales. It is possible that classification of these species could be improved by refining an event to annotation at the individual level, as was done with the NAtlantic and SAtlantic datasets. Subdividing pulse trains into individuals in PAMGuard is time consuming, but alternative options include consideration of the PAMGuard click train detector module as an additional first stage detector in BANTER, or by developing a similar function in R to apply to events.

For this study, only a few simple environmental features were considered for one of the study areas (seafloor depth, seafloor gradient, and sea surface temperature). In the EPacific study area, inclusion of these variables increased the overall classification rate from 91.1% to 91.9%. These variables had the most impact on classification results for the Cross Seamount beaked whale ($n = 6$), raising the classification scores from 83.3% to 100%. Alternative environmental features may improve results for other species.

While BANTER provides an efficient and consistent approach to classification, there are significant limitations that must be considered. BANTER is a supervised machine learning tool and requires reliable training data for success. Training data should consist of labels with strong confidence in species identity (ideally determined by agreement from more than one analyst) and sample sizes should be large enough to explain the natural variability in the data.

Until deep learning methods become more accessible and practical for applications involving rare species and low sample sizes, or their applicability across recording platforms can be guaranteed, machine learning approaches such as BANTER can serve as an accessible means of developing acoustic classifiers for complicated datasets. The workflow presented here provides a highly automated approach to detection of acoustic events (PAMGuard), integration of environmental data (PAMpal), and acoustic event classification (BANTER). These methods significantly reduce manual analysis, provide more consistent classification results with fewer biases, and provide an estimate of classification error. The greatest improvement to classification results for beaked whales would likely result from improved sample sizes, and examination of individuals to accurately measure IPI. Consideration of additional detectors (e.g., matched filter detector, click train detector) or additional environmental variables may further improve classification results. Improved alignment of detection methods across studies (i.e., definition of events, hydrophone calibration) may allow for combining data within geographic regions to improve assessments within regions for all species, and across regions for species with global distributions, such as Cuvier's beaked whales. These highly automated methods may allow for analysis of data across large spatial and temporal scales to address large ecological and population level questions.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2024.102511.

CRediT authorship contribution statement

Shannon Rankin: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Writing – original draft, Writing – review & editing. **Taiki Sakai:** Software, Writing – review & editing. **Frederick I. Archer:** Software, Writing – review & editing. **Jay Barlow:** Validation, Writing – review & editing. **Danielle Cholewiak:** Validation, Writing – review & editing. **Annamaria I. DeAngelis:** Validation, Writing – review & editing, Data curation. **Jennifer L.K. McCullough:** Validation, Writing – review & editing, Data curation. **Erin M. Oleson:** Resources, Writing – review & editing. **Anne E. Simonis:** Data curation, Validation, Writing – review & editing. **Melissa S. Soldevilla:** Data curation, Validation, Writing – review & editing. **Jennifer S. Trickey:** Validation, Writing – review & editing.

Declaration of competing interest

None.

Data availability

Partial data available on Fighshare, Github

References

Archer, Frederick, 2022a. Banter: BioAcoustic eveNT classifiER. https://cran.r-project.org/web/packages/banter/index.html.

Archer, Frederick, 2022b. rfPermute: Estimate Permutation p-Values for Random Forest Importance Metrics. https://cran.r-project.org/web/packages/rfPermute/index.html.

Barlow, Jay, Gustavo Cárdenas-Hinojosa, E., Henderson, Elizabeth, Breese, Dawn, López-Arzate, Diana, Pla, Eva Hidalgo, Taylor, Barbara L., 2022a. Unique morphological and acoustic characteristics of beaked whales (Mesoplodon Sp.) off the west coast of Baja California, Mexico. Mar. Mamm. Sci. 38 (1), 383–390. https://doi.org/10.1111/mms.12853.

Barlow, Jay, Moore, Jeffrey E., McCullough, Jennifer L.K., Griffiths, Emily T., 2022b. Acoustic-based estimates of Cuvier's beaked whale (Ziphius Cavirostris) density and abundance along the u.s. west coast from drifting hydrophone recorders. Mar. Mamm. Sci. 38 (2), 517–538. https://doi.org/10.1111/mms.12872.

Baumann-Pickering, Simone, McDonald, Mark A., Simonis, Anne E., Berga, Alba Solsona, Merkens, Karlina P.B., Oleson, Erin M., Roch, Marie A., et al., 2013. Species-specific beaked whale echolocation signals. J. Acoust. Soc. Am. 134 (3), 2293–2301. https://doi.org/10.1121/1.4817832.

Baumann-Pickering, Simone, Roch, Marie A., Brownell Jr, Robert L., Simonis, Anne E., McDonald, Mark A., Solsona-Berga, Alba, Oleson, Erin M., Wiggins, Sean M., Hildebrand, John A., 2014. Spatio-temporal patterns of beaked whale echolocation signals in the North Pacific. PLoS One 9 (1), e86072. https://doi.org/10.1371/journal.pone.0086072.

Bianco, Michael J., Gerstoft, Peter, Traer, James, Ozanich, Emma, Roch, Marie A., Gannot, Sharon, Deledalle, Charles-Alban, 2019. Machine learning in acoustics: theory and applications. J. Acoust. Soc. Am. 146 (5), 3590–3628. https://doi.org/10.1121/1.5133944.

Bittle, Michael, Duncan, Alec, 2013. A review of current marine mammal detection and classification algorithms for use in automated passive acoustic monitoring. Proc. Acoust. 2013, 8.

Breiman, Leo, 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Cohen, Rebecca E., Frasier, Kaitlin E., Baumann-Pickering, Simone, Wiggins, Sean M., Rafter, Macey A., Baggett, Lauren M., Hildebrand, John A., 2022. Identification of Western North Atlantic odontocete echolocation click types using machine learning and spatiotemporal correlates. PLoS One 17 (3), e0264988. https://doi.org/10.1371/journal.pone.0264988.

DeAngelis, Annamaria Izzi, Valtierra, Robert, Van Parijs, Sofie M., Cholewiak, Danielle, 2017. Using multipath reflections to obtain dive depths of beaked whales from a towed hydrophone array. J. Acoust. Soc. Am. 142 (2), 1078–1087. https://doi.org/10.1121/1.4998709.

DeAngelis, Annamaria Izzi, Stanistreet, Joy E., Baumann-Pickering, Simone, Cholewiak, Danielle M., 2018. A description of echolocation clicks recorded in the presence of True's beaked whale (Mesoplodon Mirus). J. Acoust. Soc. Am. 144 (5), 2691–2700. https://doi.org/10.1121/1.5067379.

Dufourq, Emmanuel, Batist, Carly, Foquet, Ruben, Durbach, Ian, 2022. Passive acoustic monitoring of animal populations with transfer learning. Eco. Inform. 70 (September), 101688 https://doi.org/10.1016/j.ecoinf.2022.101688.

Frainer, Guilherme, Dufourq, Emmanuel, Fearey, Jack, Dines, Sasha, Probert, Rachel, Elwen, Simon, Gridley, Tess, 2023. Automatic detection and taxonomic identification of dolphin vocalisations using convolutional neural networks for passive acoustic monitoring. Eco. Inform. 78 (December), 102291 https://doi.org/10.1016/j.ecoinf.2023.102291.

Frasier, Kaitlin E., Roch, Marie A., Soldevilla, Melissa S., Wiggins, Sean M., Garrison, Lance P., Hildebrand, John A., 2017. Automated classification of dolphin echolocation click types from the Gulf of Mexico. PLoS Comput. Biol. 13 (12), e1005823 https://doi.org/10.1371/journal.pcbi.1005823.

Kahl, Stefan, Wood, Connor M., Eibl, Maximilian, Klinck, Holger, 2021. BirdNET: a deep learning solution for avian diversity monitoring. Eco. Inform. 61 (March), 101236 https://doi.org/10.1016/j.ecoinf.2021.101236.

Keating, Jennifer L., Barlow, Jay, 2013. Summary of PAMGuard Beaked Whale Click Detectors and Classifiers Used During the 2012 Southern California Behavioral Response Study {NOAA} Tech Memo NOAA-TM-NMFS-SWFSC-517.

Kirsebom, Oliver S., Frazao, Fabio, Padovese, Bruno, Sakib, Sadman, Yue, Su, Matwin, Stan, 2022. MERIDIAN open-source software for deep learning-based acoustic data analysis. J. Acoust. Soc. Am. 151 (4), A27. https://doi.org/10.1121/10.0010545.

LeBien, Jack G., Ioup, Juliette W., 2018. Species-level classification of beaked whale echolocation signals detected in the northern Gulf of Mexico. J. Acoust. Soc. Am. 144 (1), 387–396. https://doi.org/10.1121/1.5047435.

Li, Kun, Sidorovskaia, Natalia A., Tiemann, Christopher O., 2020a. Model-based unsupervised clustering for distinguishing Cuvier's and Gervais' beaked whales in acoustic data. Eco. Inform. 58 (July), 101094 https://doi.org/10.1016/j.ecoinf.2020.101094.

Li, Kun, Sidorovskaia, Natalia A., Tiemann, Christopher O., 2020b. Model-based unsupervised clustering for distinguishing Cuvier's and Gervais' beaked whales in acoustic data. Eco. Inform. 58 (July), 101094 https://doi.org/10.1016/j.ecoinf.2020.101094.

López-Baucells, Adrià, Torrent, Laura, Rocha, Ricardo, Bobrowiec, Paulo E.D., Palmeirim, Jorge M., Meyer, Christoph F.J., 2019. Stronger together: combining automated classifiers with manual post-validation optimizes the workload vs reliability trade-off of species identification in bat acoustic surveys. Eco. Inform. 49 (January), 45–53. https://doi.org/10.1016/j.ecoinf.2018.11.004.

MacLeod, Colin D., 2018. Beaked whales, overview. In: Würsig, Bernd, Thewissen, J.G. M., Kovacs, Kit M. (Eds.), Encyclopedia of Marine Mammals, Third edition. Academic Press, pp. 80–83. https://doi.org/10.1016/B978-0-12-804327-1.00062-5.

McCullough, Jennifer L.K., Oleson, Erin M., Barlow, Jay, Allen, Ann N., Merkens, Karlina P.B., 2021. An acoustic survey in the Main Hawaiian islands using drifting recorders. In: PIFSC Administrative Report H-21-04: 26. https://doi.org/10.25923/RZZZ-0V38.

McDonald, Mark A., Hildebrand, John A., Wiggins, Sean M., Johnston, David W., Polovina, Jeffrey J., 2009. An acoustic survey of beaked whales at cross seamount near Hawaii. J. Acoust. Soc. Am. 125 (2), 624–627. https://doi.org/10.1121/1.3050317.

Mutanu, Leah, Gohil, Jeet, Gupta, Khushi, Wagio, Perpetua, Kotonya, Gerald, 2022. A review of automated bioacoustics and general acoustics classification research. Sensors 22 (21), 8361. https://doi.org/10.3390/s22218361.

Nanni, Loris, Maguolo, Gianluca, Paci, Michelangelo, 2020. Data augmentation approaches for improving animal audio classification. Eco. Inform. 57 (May), 101084 https://doi.org/10.1016/j.ecoinf.2020.101084.

Nolasco, Ines, Singh, Shubhr, Morfi, Veronica, Lostanlen, Vincent, Strandburg-Peshkin, Ariana, Vidaña-Vila, Ester, Gill, Lisa, et al., 2023. Learning to detect an animal sound from five examples. Eco. Inform. 77 (November), 102258 https://doi.org/10.1016/j.ecoinf.2023.102258.

Padovese, Bruno, Kirsebom, Oliver S., Frazao, Fabio, Evers, Clair H.M., Beslin, Wilfried A.M., Theriault, Jim, Matwin, Stan, 2023. Adapting deep learning models to new acoustic environments - a case study on the North Atlantic right whale Upcall. Eco. Inform. 77 (November), 102169 https://doi.org/10.1016/j.ecoinf.2023.102169.

Parijs, Sofie M., Van, Chris W., Clark, Renata S., Sousa-Lima, Susan E., Parks, Shannon Rankin, Risch, Denise, Van Opzeeland, Ilse C., 2009. Management and research applications of real-time and archival passive acoustic sensors over varying temporal and spatial scales. Mar. Ecol. Prog. Ser. 395 (December), 21–36. https://doi.org/10.3354/meps08123.

Rankin, Shannon, Archer, Frederick, 2021. BANTER: A User's Guide to Acoustic Classification. https://taikisan21.github.io/PAMpal/banterGuide.html.

Rankin, Shannon, Barlow, Jay, 2005. Source of the North Pacific 'boing' sound attributed to Minke whales. J. Acoust. Soc. Am. 118 (5), 3346–3351. https://doi.org/10.1121/1.2046747.

Rankin, Shannon, Archer, Frederick, Keating, Jennifer L., Oswald, Julie N., Oswald, Michael, Curtis, Alex, Barlow, Jay, 2017. Acoustic classification of dolphins in the California current using whistles, echolocation clicks, and burst pulses. Mar. Mamm. Sci. 2 (33), 520–540. https://doi.org/10.1111/mms.12381.

Román Ruiz, M., Rossi, C., Esteban, J.A., 2023. Fin whale pulse detection with deep neural networks. Eco. Inform. 77 (November), 102243 https://doi.org/10.1016/j.ecoinf.2023.102243.

Ross, Jesse C., Allen, Paul E., 2014. Random forest for improved analysis efficiency in passive acoustic monitoring. Ecol. Inform. Ecol. Acoust. 21 (May), 34–39. https://doi.org/10.1016/j.ecoinf.2013.12.002.

Sakai, Taiki, 2021. PAMpal: Load and Process Passive Acoustic Data. https://cran.r-project.org/web/packages/PAMpal/index.html.

Shorten, Connor, Khoshgoftaar, Taghi M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6 (1), 60. https://doi.org/10.1186/s40537-019-0197-0.

Simonis, Anne E., 2020. Passive Acoustic Survey of Deep-Diving Odontocetes in the California Current Ecosystem 2018: Final Report. NOAA-TM-NMFS-SWFSC-630. https://doi.org/10.25923/W5XX-JZ73.

Simons, R.A., John, Chris, 2022. ERDDAP. NOAA/NMFS/SWFSC/ERD, Monterey, CA. https://coastwatch.pfeg.noaa.gov/erddap.

Soldevilla, Melissa S., Elizabeth Henderson, E., Campbell, Gregory S., Wiggins, Sean M., Hildebrand, John A., Roch, Marie A., 2008. Classification of Risso's and Pacific white-sided dolphins using spectral properties of echolocation clicks. J. Acoust. Soc. Am. 124 (1), 609–624. https://doi.org/10.1121/1.2932059.

Stowell, Dan, 2022. Computational bioacoustics with deep learning: a review and roadmap. PeerJ 10 (March), e13152. https://doi.org/10.7717/peerj.13152.

Team, R. Core, 2013. R: A Language and Environment for Statistical Computing.

United States National Marine Fisheries Service, 2016. 2016 Annual Report of a Comprehensive Assessment of Marine Mammal, Marine Turtle, and Seabird Abundance and Spatial Distribution in US Waters of the Western North Atlantic Ocean – AMAPPS II. https://doi.org/10.25923/GBAP-G480.

Yano, Kymberly M., Oleson, Erin M., Keating, Jennifer L., Ballance, Lisa, Hill, Marie C., Bradford, Amanda, Allen, Ann N., Joyce, Trevor W., Moore, Jeffrey E., Henry, Annette E., 2018. Cetacean and Seabird Data Collected During the Hawaiian Islands Cetacean and Ecosystem Assessment Survey (HICEAS), July–December 2017 {OCS} Study {BOEM} OCS Study BOEM 2018–044. Bureau of Ocean Energy Management. https://repository.library.noaa.gov/view/noaa/18660.

Zahn, Marie J., Rankin, Shannon, McCullough, Jennifer L.K., Koblitz, Jens C., Archer, Frederick, Rasmussen, Marianne H., Laidre, Kristin L., 2021. Acoustic differentiation and classification of wild belugas and narwhals using echolocation clicks. Sci. Rep. 11 (1), 22141. https://doi.org/10.1038/s41598-021-01441-w.

Zhong, Ming, Castellote, Manuel, Dodhia, Rahul, Ferres, Juan Lavista, Keogh, Mandy, Brewer, Arial, 2020. Beluga whale acoustic signal classification using deep learning neural network models. J. Acoust. Soc. Am. 147 (3), 1834–1841. https://doi.org/10.1121/10.0000921.

Ziegenhorn, Morgan A., Frasier, Kaitlin E., Hildebrand, John A., Oleson, Erin M., Baird, Robin W., Wiggins, Sean M., Baumann-Pickering, Simone, 2022. Discriminating and classifying Odontocete echolocation clicks in the Hawaiian islands using machine learning methods. PLoS One 17 (4), e0266424. https://doi.org/10.1371/journal.pone.0266424.