A Comparison of Detection Methods for Identifying the Presence of Active Sonar in Long-Term Passive Acoustic Data: A Case Study Within a Scottish Marine Protected Area

Benedict L. Dell, Ellen L. White, Denise Risch, Jonathan M. Bull, Paul R. White, and Nienke C. F. van Geel

Contents

Abstract

Standardized methods for the detection and classification of active sonar have yet to be developed, hindering research into ecological questions related to sonar use over large spatio-temporal scales using archival Passive Acoustic Monitoring (PAM) data. This chapter compares two pipelines for classification of military sonar presence in 20-min files, designed to be generalizable across soundscapes and types of sonar. Pipelines included adapting a deep learning network designed to classify delphinid vocalizations and vessels at the 3-second level to a 20-min resolution using a decision tree, and a Gradient Boosted Random Forest (GBRF)

OPEN ACCESS with major support from JASCO APPLIED SCIENCES

B. L. Dell (✉) · P. R. White
Institute of Sound and Vibration, University of Southampton, Southampton, UK
e-mail: B.Dell@soton.ac.uk

E. L. White · J. M. Bull
School of Ocean and Earth Science, University of Southampton, Southampton, UK

D. Risch · N. C. F. van Geel
Marine Science Department, Scottish Association of Marine Science, Oban, UK

A. N. Popper et al. (eds.), *The Effects of Noise on Aquatic Life IV*,
https://doi.org/10.1007/978-3-031-94229-7_48-1

using acoustic indices as features. The adapted deep learning and GBRF pipelines achieved F1 scores of 0.57 and 0.74 respectively. The GBRF pipeline was demonstrated to provide usable predictions of sonar presence, reducing a 51-day dataset to a 12-day period with elevated levels of predicted sonar presence, and 54 out of 935 files predicted to contain sonar predictions outside this period which could be subsequently manually verified. This pipeline is a promising approach for identifying active sonar use in large PAM datasets.

Keywords

Passive Acoustic Monitoring (PAM) · Underwater acoustics · Machine learning · Acoustic indices · Sonar detection · Anthropogenic noise

Introduction

Noise impact assessments fundamentally depend on the detection and classification of a noise source in passively collected data, or standardized methods to directly measure it. Direct measures using metrics such as Sound Pressure Levels (SPLs) or Pressure Spectral Densities (PSDs) rely on features of the sound which are unique or dominant within a specified spectral-temporal space (Haver et al. 2018), whilst automated detection and classification methods rely on signal processing, machine learning or deep learning classifiers which require large, diverse training data sets for their development. To date, there are no standardized, generalizable methods for the detection and classification of active sonar, such as Mid Frequency Active Sonar (MFAS), within large Passive Acoustic Monitoring (PAM) datasets.

Sonar is a highly varied sound source, with the types of sonar used changing regionally, temporally and operationally. Types of sonar can be distinguished by differences in the frequency range, frequency modulation, signal duration, inter-signal duration, and operating source levels. Recorded signals on hydrophones deployed independently of planned sonar use can be complex due to propagation effects resulting in signal distortion, propagation loss and overlap with or masking by other sound sources. The resultant diversity in recorded signals, alongside a high range of Signal-to-Noise Ratio (SNR) values and complex soundscapes in which "noise" (such as delphinid whistles) with similar features to sonar are often concurrently present, mean signal processing pipelines for classification often fail to achieve sufficient accuracies. This complexity also means the development of bespoke deep learning methods for the automated detection of sonar is challenging. It is further hindered by limited publicly available datasets containing labelled sonar exemplars and the well documented problem of domain adaptation.

The soundscape contribution of sonar and its potential effects on marine life is a question of global significance. In response to observed mass strandings of cetaceans, over the past two decades there have been a number of studies examining both direct and indirect impacts of sonar signals, with many of these relying on playback experiments (Halvorsen et al. 2013; Harris et al. 2016; Southall et al. 2016) or using data from extensive hydrophone arrays such as the Atlantic Undersea Testing and Evaluation Center (AUTEC) (Moretti et al. 2014; Joyce et al. 2020). These studies have demonstrated that the effects of sonar on cetaceans can be diverse and depend on physical characteristics of the environment, behavioral context and sound exposure level (Harris et al. 2018). While some studies have reported minimal observed impacts (Casey et al. 2024), many others have demonstrated changes in behavioral activity, including changing diving patterns (Moretti et al. 2014; Joyce et al. 2020), area occupancy (Tyack et al. 2011; Wensveen et al. 2019; Joyce et al. 2020) and vocalization behavior (Tyack et al. 2011; Stanistreet et al. 2022), as well as tissue damage (Southall et al. 2016) and death leading to mass strandings (Filadelfo et al. 2009; Southall et al. 2016).

Outside of navy ranges, there are few studies examining the impact of military sonar over large spatio-temporal scales. This means baseline questions ranging from the contribution of sonar to local soundscapes to potential behavioral changes of key species when exposed to varying densities, types and SNRs of sonar over long time periods can only be extrapolated from controlled experiments (Harris et al. 2018; Stanistreet et al. 2022). There is an opportunity to exploit decades of archival broadband acoustic recordings from many regions worldwide to address some of these knowledge gaps. However, a generalizable, easily adaptable and user-friendly detection and classification system needs to be developed first.

The authors present initial results from a wider study which aims to develop generalizable detection and classification pipelines of sonar presence at a 20-min resolution. Classification pipelines were developed to allow an initial analysis of large PAM data sets for the presence of incidental sonar within wav files, rather than the classification of individual sonar signals. The lower temporal resolution reduces the manual cost of labelling and is sufficient for many ecological questions. These pipelines are designed to be used to filter data for periods of sonar presence, reducing the manual burden for analyses at a finer temporal resolution such as manual labelling at a higher temporal resolution for the quantification of sonar signals and development of deep learning models bespoke to the region and type of sonar present.

Methods

Data Description

Data for initial model development and testing were taken from a single mooring deployed as part of the COMPASS project within a Marine Protected Area in the Northwest of Scotland (https://compass-data-portal-marineinstitute.hub.arcgis.com/pages/story-maps). A single omnidirectional acoustic broadband recorder (SoundTrap 300 HF; Ocean Instruments, New Zealand) with an end-to-end sensitivity of −172.6 dB re 1 V/μPa was moored 3–5 m above the seafloor at a water depth of 100 m. The hydrophone was set to a sample rate of 96 kHz and a duty cycle of 20/40 min on/off. Calibration tones present in the first 3 s of the recording were excluded from the analysis. Acoustic data were calibrated using values provided by the manufacturer.

Data from the 10th of March to 30th of April 2019 were manually labelled using Audacity (version 3.2.1, 2022) for the presence or absence of sonar in 20-min wav files by two expert acousticians. Spectrograms were viewed using a Hann window of size 2048 samples on a mel-scale axis. Labelling was carried out with 30 s of data visible on the screen to ensure temporal resolution was standardized between the labelers. A confidence score was assigned to each label, with scores lower than the most confident subsequently re-labelled by the other labeler. Re-labelled files were only labelled as "sonar present" if both labelers assigned that label to reduce errors in the training set.

Other noise sources present within files were noted during labelling. These included vessel noise which varied from Lloyd's mirror effects (the interference between the direct path and the sea surface reflection of the signal leading to a frequency-dependent interference pattern) to low frequency tonal noise, echosounders, delphinid whistles, clicks as well as burst pulses and abiotic noise such as wind, rain and mooring related noise. For the purposes of this study, echosounders were distinguished from sonar by the peak frequency being above 30 kHz and were excluded from the sonar class.

Significant variation between sonar types and frequency of occurrence within a certain period was observed during labelling. Features which varied between individual signals included the frequency modulation, frequency range, signal duration, the presence of harmonics, and the SNR due to both low received levels and significant background noise. The number of sonar types in a single file and occurrence frequency in both the time and frequency domain varied significantly. Some files included near-continuous sonar within a bandwidth of 1 kHz, with other files including multiple types of sonar simultaneously recorded between 500 Hz and 30 kHz. Signals were recorded at different inter-signal durations, varying frequency of occurrence across single files and with varied SNRs.

Sonar Classification Pipelines

A signal processing approach to the classification of sonar was initially developed using metrics developed based on observed signal features during labelling. This was not successful due to the complexity of sonar signals, overlap in metric values between sonar and other sources such as delphinid whistles, and the high variance in the recorded SNRs of sonar within this dataset. As a result, two classification pipelines were developed, one which adapted an existing Convolutional Neural Network (CNN) trained on Ambient, Whistle, Vessel and Click classes (White et al. 2022), and one which used acoustic indices as features input to a Gradient Boosted Random Forest (GBRF).

A flowchart of the vital components of each pipeline is shown in Fig. 1, with full details beyond the scope of this paper. The adapted CNN pipeline uses an existing EfficientNet B0 based CNN (full training details can be found in (White et al. 2022))

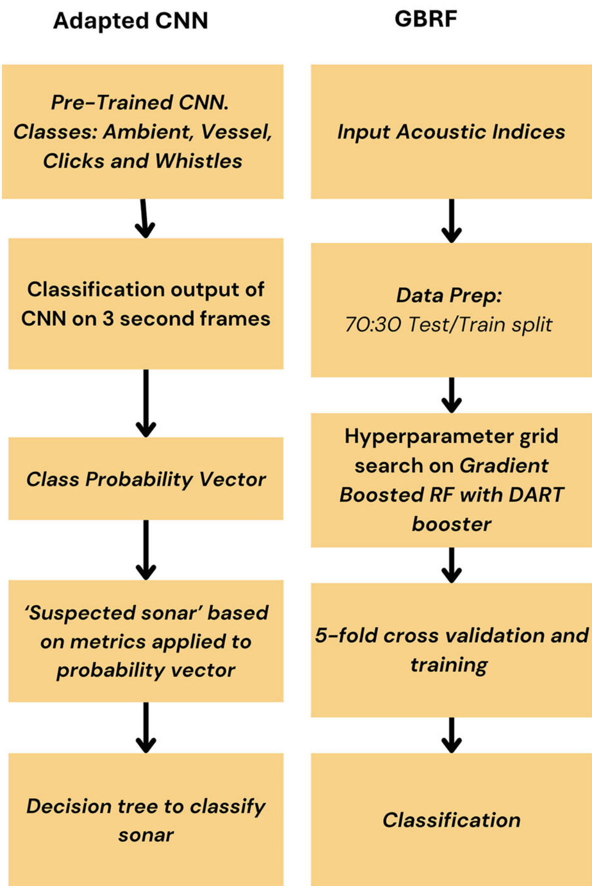

Fig. 1 Illustrative workflows of each pipeline. The Adapted CNN applied multiple metrics to the class probability vector from an existing model which was observed to misclassify sonar as whistles, followed by a decision tree to adapt it to file level

which was observed to misclassify sonar as whistles. The increasing use of CNNs in underwater acoustics means a pipeline adapted from an existing CNN would be broadly applicable. Training a new CNN for classification at a 20-min resolution was infeasible due to the prohibitive input data sizes and corresponding model size. Re-training the CNN with a new class for sonar was not feasible as training data did not match the 3-s resolution which the model was originally trained with. To leverage the deep learning-based classifier, a decision tree was therefore built based on the 3 s output of the CNN. This "Adapted CNN" pipeline identified model outputs as "suspected sonar" based on measures applied to the output probability vector of the model, with the distribution of "suspected sonar" in frequency and time domains used to classify the file.

The other pipeline incorporated the signal processing approach with a learnt component by using acoustic indices as generalizable features input to a Gradient Boosted Random Forest (GBRF) implemented using XGBoost in Python (Chen et al. 2025). Acoustic Indices are signal processing metrics which measure varied time-frequency features of input data, such as measures of entropy or the zero-crossing rate. A GBRF was used for classification as it outperforms many classifiers on tabular data and is robust to redundant features. Therefore, the input acoustic indices do not need to be highly bespoke for a specific soundscape or sonar type, provided a subset of them respond to sonar presence in each domain and for each sonar type. A set of eight commonly used acoustic indices were augmented by using multiple window lengths and frequency ranges. A standard approach was taken to training the model by using five-fold stratified cross-validation with an automated grid search implemented using scikit-learn (Pedregosa et al. 2011) for key parameters in the booster and random forest. Data were split 70% for training and 30% for testing.

Model performance was evaluated using the F1 score which is a harmonic mean of precision and recall (Mesaros et al. 2016) defined as:

$$\mathrm{F1} = \frac{2 * \mathrm{TP}}{2 * \mathrm{TP} + \mathrm{FP} + \mathrm{FN}},$$

where TP is the number of true positives, FP is the number of false positives and FN is the number of false negatives. Precision-recall curves alongside Receiver Operating Characteristic (ROC) curves were used to show model performance of the GBRF classifier with the Area Under the Curve (AUC) reported as a measure of overall model performance.

Results

The GBRF outperformed the Adapted CNN, with PR-AUC of 0.78 and ROC-AUC of 0.86 (Fig. 2a, b); note that the Adapted CNN returned single values in these plots as classifications of the CNN were subsequently passed through a decision tree. When applied to unseen test data, the F1 scores were 0.74 and 0.57 for the GBRF and Adapted CNN respectively.

To highlight a potential use of a classifier predicting sonar presence at a 20-minute resolution, the output from the GBRF classifier was plotted across the entire data period, including data used for training (Fig. 3). The number of hours of predicted and labelled sonar per day was plotted to aid in the interpretation of results. The moving average showed a peak in predicted sonar presence between 29th March and 12th April, corresponding to a peak in labelled sonar between 1st and 12th April. As expected from the confusion matrix (Fig. 2c), FPs were present throughout the data.

To illustrate how the predicted sonar presence could be interpreted, Fig. 3 was split into three regions. In the green region, daily sonar exceeded 5 files per day for a period of 12 days, totaling 124 out of 312 files with predicted sonar. This period could therefore be designated as a period of active sonar use despite a FN rate of 0.35 over the test data set (Fig. 2c). The red region was used to indicate differences in the

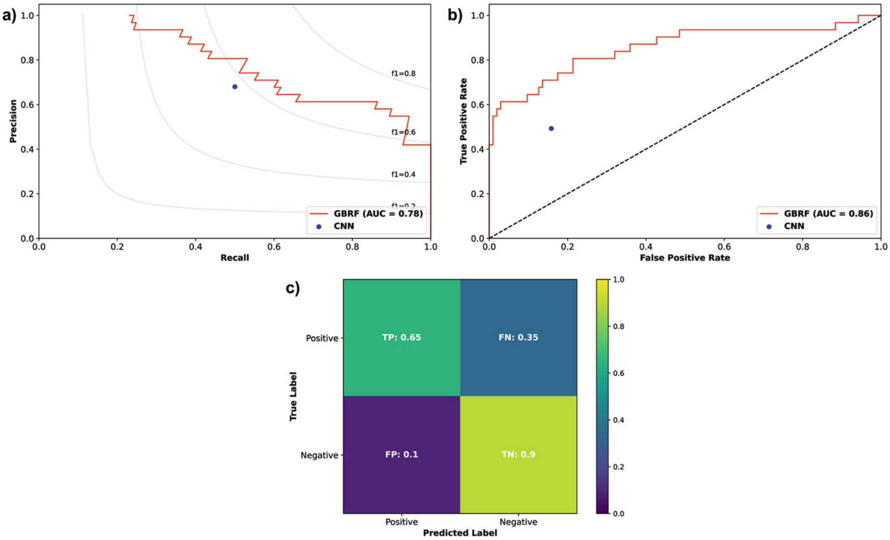

Fig. 2 Plots of (**a**) Precision-Recall curves with contours for different F1 scores, (**b**) Receiver Operating Characteristic (ROC) curves and (**c**) the confusion matrix for the Gradient Boosted Random Forest (GBRF) pipeline, normalized row-wise to show classification proportion relative to the total number of true labels for that category. Curves in (**a**) and (**b**) indicate that the GBRF outperforms the Adapted CNN. Note that the Adapted CNN is only a single point as a manual decision tree is applied to the output of a trained CNN. The confusion matrix corresponds to an F1 score of 0.74

exact timing of the onset and offset of this period of elevated sonar use between the predicted and labelled data.

The amber region was used to show sporadic predicted sonar presence which were treated as potential FPs. These covered a total 935 files with 54 positive predictions, which could be manually verified in approximately 4 hours by an experienced labeler using open-source software such as Audacity. This “human-in-the-loop” aspect allowed for an arbitrarily high precision outside of the period of more frequent sonar use. The illustrative interpretation of the classifier output presented here therefore reduced 1247 files over 52 days to a predicted period of 12 days with daily sonar use, and 54 files which would have required further manual verification.

Discussion

This study compared two pipelines for the classification of sonar at a 20-min resolution, corresponding to a single wav file. The first approach adapted a CNN trained for the classification of delphinid vocalizations and vessel presence at a 3-s resolution to a 20-min resolution by applying a decision tree to the output probability vector of the CNN. The second approach trained a GBRF with acoustic indices as

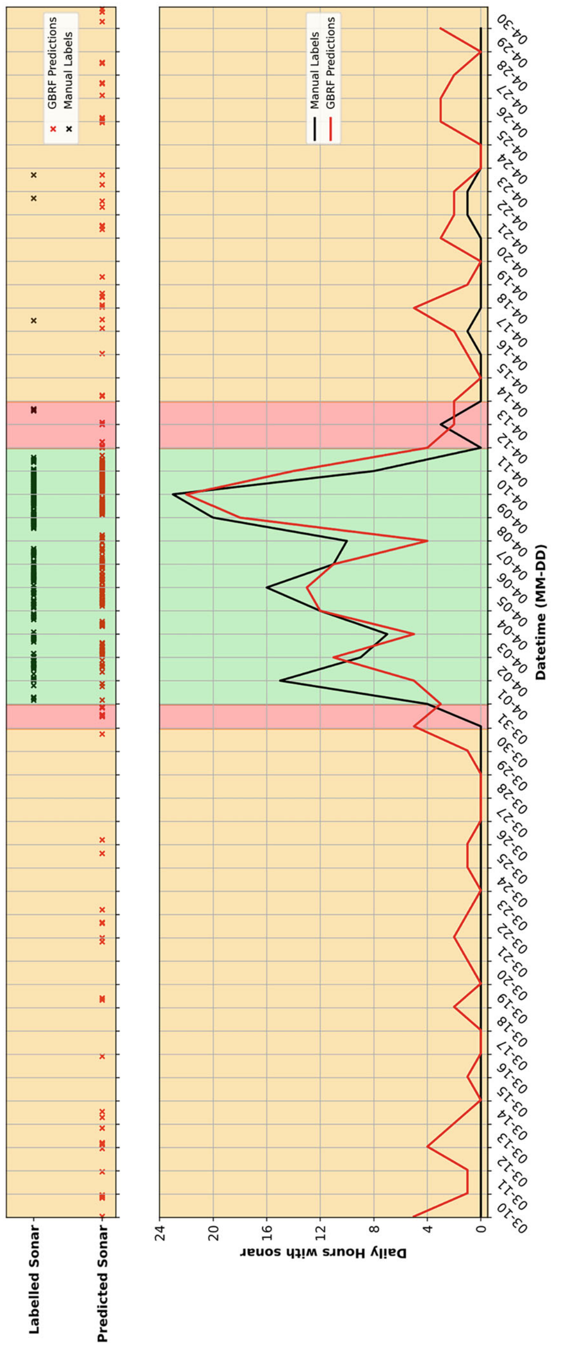

Fig. 3 Positive sonar predictions (top panel) for the entire data period between 10th March and 30th April from the GBRF (red crosses) and manual positive labels (black crosses). Corresponding trendlines show the sum of labelled (black) and predicted (red) sonar within each day (bottom panel). Colored boxes are used to aid interpretation of the plot, and to indicate how the classifier could be used in practice. The green box shows a period of elevated predicted and labelled sonar presence between 1st and 12th April. The red boxes indicate differences between the exact predicted and labelled onset and offset of this period. Amber boxes span periods with sporadic sonar prediction, totaling 54 out of 935 files with predicted sonar, which would be treated as potential false positives

input features. The pipelines were explicitly designed to be generalizable to new soundscapes and robust when previously unseen sonar is present by adapting widely used deep learning models or by using features theoretically applicable to any soundscape. Generalizability in machine learning is well documented to come at a cost in classifier performance, meaning F1 scores of 0.74 and 0.57, which would typically be considered low for a state-of-the-art classifier, were anticipated.

The design of the Adapted CNN pipeline was based on a CNN trained on Ambient, Vessel, Click and Whistle classes (White et al. 2022), which was observed to misclassify sonar as whistles. It was therefore hypothesized that the output class probability vector could be used to identify "potential sonar" signals, with a decision tree based on the coherent use of sonar over 20-min periods used to subsequently classify entire files. An F1score of 0.57 was deemed too low for further use of this pipeline, however.

The assessment of any classifier performance is dependent on the context of its intended use, with classification at a file level allowing for a human-in-the-loop aspect to the pipeline without prohibitive manual labelling costs. Output classifications from the GBRF pipeline predicted a period of sonar use between the 29th March and 12th April, with 124 positive predictions out of 312 files in this period. Predicted sonar before and after this period did not exceed five files per day and totaled 54 out of 934 files, or 5.8% of files. These were treated as potential false positives and could be manually verified in approximately 4 hours by an experienced acoustician. Fifty-one days of data were therefore reduced to a period of 12 days where sonar was predicted to be in use, with a precision of 1.0 outside of this period after further manual labelling.

This demonstrates how questions relating to the presence of sonar within large PAM datasets could be investigated using predictions from the GBRF. Questions related to the relative contribution of sonar to the soundscape and impacts on marine species could be investigated by considering periods before, during and after the predicted onset of sonar use for example, with further labelling effort used to find the exact start and end dates of the predicted period of sonar use if required. This classifier could also be used to inform further work at a higher temporal resolution, with the description of sonar types present in the data or questions around ecological responses to sonar presence at the resolution of individual sonar signals able to discard redundant data from within the 52-day PAM deployment based on output predictions from the GBRF classifier.

At a 20-min input scale, a deep learning method such as a bespoke CNN would not be feasible due to the size of the input data and corresponding size of the model. A bespoke deep learning method trained on higher temporal resolution data would be expected to outperform the GBRF approach, but would require greater manual labelling effort, in part due to the requirement of sufficient examples of each type of sonar in the training data set. For the initial detection of sonar presence in large archival PAM data sets therefore, a deep learning approach was not deemed necessary. An anticipated use of a lower accuracy classifier such as the GBRF would be to identify periods of sonar to allow targeted manual labelling at a higher temporal resolution, from which deep learning classifiers could be developed for ecological questions on the scale of individual sonar signals.

The use of a classifier robust to redundant features in the GBRF pipeline means acoustic indices do not need to be manually tailored for each soundscape and type of sonar. This pipeline is therefore theoretically highly applicable across soundscapes and sonar types provided a subset of acoustic indices are influenced by sonar in a sufficiently unique way relative to other noise sources in the soundscape. The data used in these initial results was a "challenging" domain as it included significant vessel noise and delphinid vocalization overlapping in frequency and time with sonar, along with echosounders with similar features but a different frequency range. As such the GBRF pipeline is considered promising approach.

Results and pipelines presented here are a preliminary study designed to test the feasibility of three approaches to the problem of generalizable sonar classification at low temporal resolutions. Further pipeline development and testing is required. The theoretical robustness of the GBRF pipeline across soundscapes and sonar types has yet to be explicitly tested, nor has the performance of the model when specific types of sonar are excluded from the data sets. As such the results reported here should be seen as a baseline from which further development will be carried out.

Conclusion

This preliminary study was designed as part of a wider project to see if a generalizable sonar classification pipeline could be developed that would be easily adaptable to new soundscapes and robust to new sonar types. We presented initial results comparing different pipelines which are theoretically generalizable, with a GBRF pipeline performing the best. Adapting an existing CNN from 3-s classification to 20-min classifications returned poor results with an F1 score of 0.57. Despite an F1 score of 0.74, which would be considered low relative to state-of-the-art deep learning techniques, the potential use of the GBRF was demonstrated using a human-in-the-loop approach to validate periods of sporadic positive predictions. Model outputs over the 52-day deployment revealed a 12-day period of dense sonar use, before and after which predicted sonar could be efficiently manually validated meaning this period had an effective precision of 1.0. These predictions would allow a user to discard irrelevant data during further analysis efforts at finer temporal resolutions or answer broader ecological questions such as the soundscape contribution of sonar in this period. Further testing around pipeline generalizability and model refinement is required, but these results suggest a GBRF with acoustic indices as input features provides a promising approach to identify the presence of sonar signals in large PAM datasets in which incidental sonar occurs.

Acknowledgments This work was supported by the Natural Environmental Research Council [grant number NE/S007210/1]. The COMPASS project has been supported by the EU's INTERREG VA Programme, managed by the Special EU Programmes Body. The views and opinions expressed in this document do not necessarily reflect those of the European Commission or the Special EU Programmes Body (SEUPB).

Competing Interest Declaration The author(s) has no competing interests to declare that are relevant to the content of this manuscript.

References

Casey C, Fregosi S, Oswald JN, Janik VM, Visser F, Southall B (2024) Common dolphin whistle responses to experimental mid-frequency sonar. PLoS One 19(4):e0302035. https://doi.org/10.1371/JOURNAL.PONE.0302035

Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T, Li M, Xie J, Lin M, Geng Y, Li Y, Yuan J, Cortes D (2025) xgboost: extreme gradient boosting. R package version 3.0.5.1. https://github.com/dmlc/xgboost

Filadelfo R, Mintz J, Michlovich E, D'Amico A, Tyack PL, Ketten DR (2009) Correlating military sonar use with beaked whale mass strandings: what do the historical data show? Aquat Mamm 35(4). https://doi.org/10.1578/AM.35.4.2009.435

Halvorsen MB, Zeddies DG, Chicoine D, Popper AN (2013) Effects of low-frequency naval sonar exposure on three species of fish. J Acoust Soc Am 134(2):EL205–EL210. https://doi.org/10.1121/1.4812818

Harris CM, Thomas L, Sadykova D, DeRuiter SL, Tyack PL, Southall BL, Read AJ, Miller PJO (2016) The challenges of analyzing behavioral response study data: an overview of the MOCHA (Multi-Study Ocean acoustics human effects analysis) project. Adv Exp Med Biol 875:399–407. https://doi.org/10.1007/978-1-4939-2981-8

Harris CM, Thomas L, Falcone EA, Hildebrand J, Houser D, Kvadsheim PH, Lam FPA, Miller PJO, Moretti DJ, Read AJ, Slabbekoorn H, Southall BL, Tyack PL, Wartzok D, Janik VM (2018) Marine mammals and sonar: dose-response studies, the risk-disturbance hypothesis and the role of exposure context. J Appl Ecol 55(1):396–404. https://doi.org/10.1111/1365-2664.12955

Haver SM, Gedamke J, Hatch LT, Dziak RP, van Parijs S, McKenna MF, Barlow J, Berchok C, DiDonato E, Hanson B, Haxel J, Holt M, Lipski D, Matsumoto H, Meinig C, Mellinger DK, Moore SE, Oleson EM, Soldevilla MS, Klinck H (2018) Monitoring long-term soundscape trends in U.S. waters: the NOAA/NPS Ocean Noise Reference Station Network. Mar Policy 90: 6–13. https://doi.org/10.1016/j.marpol.2018.01.023

Joyce TW, Durban JW, Claridge DE, Dunn CA, Hickmott LS, Fearnbach H, Dolan K, Moretti D (2020) Behavioral responses of satellite tracked Blainville's beaked whales (*Mesoplodon densirostris*) to mid-frequency active sonar. Mar Mamm Sci 36(1):29–46. https://doi.org/10.1111/MMS.12624

Mesaros A, Heittola T, Virtanen T (2016) Metrics for polyphonic sound event detection. Appl Sci 6(6):162. https://doi.org/10.3390/app6060162

Moretti D, Thomas L, Marques T, Harwood J, Dilley A, Neales B, Shaffer J, McCarthy E, New L, Jarvis S, Morrissey R (2014) A risk function for behavioral disruption of Blainville's beaked whales (*Mesoplodon densirostris*) from mid-frequency active sonar. PLoS One 9(1):e85064. https://doi.org/10.1371/JOURNAL.PONE.0085064

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830. https://doi.org/10.48550/arXiv.1201.0490

Southall BL, Nowacek DP, Miller PJO, Tyack PL (2016) Experimental field studies to measure behavioral responses of cetaceans to sonar. Endanger Species Res 31(1):293–315. https://doi.org/10.3354/ESR00764

Stanistreet JE, Beslin WAM, Kowarski K, Martin SB, Westell A, Moors-Murphy HB (2022) Changes in the acoustic activity of beaked whales and sperm whales recorded during a naval training exercise off eastern Canada. Sci Rep 12(1):1–13. https://doi.org/10.1038/S41598-022-05930-4

Tyack PL, Zimmer WMX, Moretti D, Southall BL, Claridge DE, Durban JW, Clark CW, D'Amico A, DiMarzio N, Jarvis S, McCarthy E, Morrissey R, Ward J, Boyd IL (2011) Beaked whales respond to simulated and actual navy sonar. PLoS One 6(3):e17009. https://doi.org/10.1371/JOURNAL.PONE.0017009

Wensveen PJ, Isojunno S, Hansen RR, von Benda-Beckmann AM, Kleivane L, van IJsselmuide S, Lam FPA, Kvadsheim PH, DeRuiter SL, Curé C, Narazaki T, Tyack PL, Miller PJO (2019) Northern bottlenose whales in a pristine environment respond strongly to close and distant navy sonar signals. Proc R Soc B-Biol Sci 286(1899):20182592. https://doi.org/10.1098/RSPB.2018.2592

White EL, White PR, Bull JM, Risch D, Beck S, Edwards EWJ (2022) More than a whistle: Automated detection of marine sound sources with a convolutional neural network. Frontiers in Marine Science, 9. https://doi.org/10.3389/fmars.2022.879145