frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 04 October 2022
DOI 10.3389/fmars.2022.879145



OPEN ACCESS

EDITED BY
Ana Širovic,
Norwegian University of Science and Technology, Norway

REVIEWED BY
John E. Joseph,
Naval Postgraduate School, United States
Pina Gruden,
University of Hawaii at Mānoa, United States

*CORRESPONDENCE
Ellen L. White
elw1g13@soton.ac.uk

SPECIALTY SECTION
This article was submitted to Ocean Observation, a section of the journal Frontiers in Marine Science

RECEIVED 18 February 2022
ACCEPTED 14 September 2022
PUBLISHED 04 October 2022

CITATION
White EL, White PR, Bull JM, Risch D, Beck S and Edwards EWJ (2022) More than a whistle: Automated detection of marine sound sources with a convolutional neural network. *Front. Mar. Sci.* 9:879145. doi: 10.3389/fmars.2022.879145



# More than a whistle: Automated detection of marine sound sources with a convolutional neural network

Ellen L. White[1]*, Paul R. White[2], Jonathan M. Bull[1], Denise Risch[3], Suzanne Beck[4] and Ewan W. J. Edwards[5]

[1]School of Ocean and Earth Science, University of Southampton, Southampton, United Kingdom, [2]Institute of Sound and Vibration, University of Southampton, Southampton, United Kingdom, [3]Marine Science Department, Scottish Association of Marine Science, Oban, United Kingdom, [4]Agri-Food and Biosciences Institute, Belfast, United Kingdom, [5]Marine Scotland Science, Marine Laboratory, Aberdeen, United Kingdom

The effective analysis of Passive Acoustic Monitoring (PAM) data has the potential to determine spatial and temporal variations in ecosystem health and species presence if automated detection and classification algorithms are capable of discrimination between marine species and the presence of anthropogenic and environmental noise. Extracting more than a single sound source or call type will enrich our understanding of the interaction between biological, anthropogenic and geophonic soundscape components in the marine environment. Advances in extracting ecologically valuable cues from the marine environment, embedded within the soundscape, are limited by the time required for manual analyses and the accuracy of existing algorithms when applied to large PAM datasets. In this work, a deep learning model is trained for multi-class marine sound source detection using cloud computing to explore its utility for extracting sound sources for use in marine mammal conservation and ecosystem monitoring. A training set is developed comprising existing datasets amalgamated across geographic, temporal and spatial scales, collected across a range of acoustic platforms. Transfer learning is used to fine-tune an open-source state-of-the-art 'small-scale' convolutional neural network (CNN) to detect odontocete tonal and broadband call types and vessel noise (from 0 to 48 kHz). The developed CNN architecture uses a custom image input to exploit the differences in temporal and frequency characteristics between each sound source. Each sound source is identified with high accuracy across various test conditions, including variable signal-to-noise-ratio. We evaluate the effect of ambient noise on detector performance, outlining the importance of understanding the variability of the regional soundscape for which it will be deployed. Our work provides a computationally low-cost, efficient framework for mining big marine acoustic data, for information on temporal scales relevant to the

management of marine protected areas and the conservation of vulnerable species.

KEYWORDS

marine soundscapes, CNN - convolutional neural network, passive acoustic monitoring, efficientNet-B0, sound source detection, marine mammal acoustics, Delphinids

## Introduction

The need to manage effectively marine resources and habitats in the Anthropocene era is a current conservation issue that needs addressing as we seek to move to a better balance between exploitation and preservation of marine ecosystems. Maintaining productive coastal seas, and protected areas that conserve the species which reside there, depends on our ability to rapidly detect functional disturbances across multiple spatial scales, and respond in a time-effective manner. Sound, travelling faster and farther in water than in air, is a vital sensory resource for marine animals, used by marine mammals, fish and invertebrates for communication, predation and navigation (Duarte et al., 2021). The production of sound by marine life allows researchers to eavesdrop beneath the surface with Passive Acoustic Monitoring (PAM), increasingly used to record the cacophony of the marine environment (Sousa-Lima et al., 2013; Roch et al., 2017; Howe et al., 2019). Analysis of PAM data collected over long temporal and spatial scales can be used as a reliable indicator of habitat quality by characterizing the biological, anthropogenic and geophonic sound source components within a region's soundscape (Marley et al., 2017; Pittman, 2017; McKenna et al., 2021). Effective analysis of regional soundscape data can provide insights into species composition, long-term changes in species distribution, biodiversity and human activity (Pijanowski et al., 2011; Davis et al., 2020). Knowledge of this is critical for wildlife conservation, particularly for migratory marine species, and in regions designated as marine protected areas (MPAs).

The introduction of anthropogenic sound to the sea alters the acoustic environment which may negatively affect the presence and persistence of populations and species (Kunc et al., 2016; Dunlop, 2016; Stafford et al., 2018; Erbe et al., 2019). Holistic models for studying underwater sound in relation to the detection of marine species allow for the assessment of overall ecosystem health, and of certain elements that cause long-term and chronic adverse effects to marine life such as low-intensity pervasive vessel noise, and site-specific noise contributors (e.g. seismic arrays, echo sounders and Acoustic Deterrent Devices (ADDs)). Geographic shifts in the spatio-temporal distributions of marine mammals have been found to be directly impacted by anthropogenic use of the oceans (Pompa et al., 2011; Cox et al., 2018). The requirement to minimize the environmental impacts of noise on marine organisms has therefore become a part of many international agreements such as the Convention for the Protection of the marine environment of the North-East Atlantic (OSPAR Convention). In addition, underwater noise has become an important aspect of the Marine Strategy Framework Directive (MSFD) adopted by the European Commission in 2008, which considers both the spatial and temporal distribution of loud impulsive noise, as well as trends in low-frequency continuous noise.

Acoustic recorders are a low-cost, non-invasive method for studying a wide range of biological processes, marine organisms and anthropogenic activities within marine habitats over long time scales (Wang et al., 2019) and are increasingly used for regional monitoring of marine species. The enhanced use of PAM has resulted in the growth of existing underwater datasets which can reach scales of terabytes per deployment. As the volume of acoustic data increases, the time required for extracting ecologically important information also increases (Sugai et al., 2019), with derived information potentially delivered to stakeholders long after the monitoring period. The development of algorithms and methodological approaches for exploiting the information embedded within marine soundscapes is essential for effective long-term species monitoring and ecosystem health assessment, on temporal scales relevant to marine management.

Traditional approaches to extracting acoustic signals from PAM data stem from the fields of signal processing and machine learning. Computationally low-cost algorithms such as the band-limited energy sum and the Teager energy operator (Kaiser, 1990; Gillespie, 1997; Kandia and Stylianou, 2006; Kim et al., 2006 and Mae et al., 2010), matched filtering and spectrogram correlation have been used successfully for call extraction in noisy data, static PAM mooring data (Širović et al., 2015) and from autonomous vehicles (Baumgartner and Mussoline, 2011; Baumgartner et al., 2013; Baumgartner et al., 2020). The typical approach is to use some form of generic detector to identify a period of potential interest, extract those periods and then input that signal segment to a classifier. A

classifier places the identified signal into a broad category, species-specific group or specific call type belonging to a single family depending on the application (Bittle & Duncan, 2013). Many signal processing/machine learning approaches have been used for classifying marine mammal calls including discriminant analysis (Steiner, 1981), support vector machines (Jarvis et al., 2008; Roch et al., 2008), generalized linear models, hidden markov models (Roch et al., 2004; Roch et al., 2007; Brown and Smaragdis, 2009; Brown et al., 2010; Roch et al., 2011a; Pace et al., 2012) and classification and regression tree analysis (Oswald et al., 2007).

The aforementioned algorithms rely upon a large amount of human input, often expert, presenting a limitation to their development. They yield systems which are not easily generalizable to broad categories of sound sources, data collected at differing sample rates, geographic locations or on different recording platforms. Reliance on manually selected features to define the signal(s) of interest requires a sophisticated knowledge of signal processing and may not adequately describe the complex and variable time-frequency characteristics of sounds (Jiang et al., 2019). For many marine species, acoustic repertoires are not well understood restricting the ability to define parameters on which these methods depend (Dudzinski et al., 2009; Gruden & White, 2016; Vester et al., 2017).

Acoustically active species reside in all oceanic bodies, but much of the legislation for Marine Protected Areas is focused on protecting coastal regions (water depth<200m) (Jones, 2012). These shallow water areas are characterized by acoustic complexity. The geophonic, anthropogenic and biophonic components of the soundscape share an acoustic space which varies over short spatial and temporal scales. Variation is attributed to changes in regional bathymetry, bottom-substrate type, oceanographic and weather conditions (Kuperman and Lynch, 2004), resulting in ever changing ambient noise conditions. Developments in technology and the ability to mount PAM recorders on a range of static and moving platforms adds to the variation in propagation conditions which can occur in a single PAM dataset (McKenna et al., 2021). In this work, Convolutional Neural Networks (CNNs) are applied to PAM data for the detection of highly varied marine acoustic signals. CNNs are more robust than the previously discussed techniques to fluctuating ambient noise (Xie et al., 2020).

CNNs are end-to-end deep neural networks, which efficiently handle the complexity of 2-Dimensional input data and excel at pattern recognition tasks when input data is noisy (Khan et al., 2020). CNNs have been shown to outperform existing machine learning techniques rivalling human performance at signal detection (LeCun et al., 2015) and are becoming commonplace in the bio-acoustic domain (Stowell, 2022). The bottleneck in acoustic datasets is the labor-intensive task of manually labelling archived PAM data for use in CNN training (Sugai et al., 2019). CNNs learn to discriminate spectro-temporal information directly from a labelled spectrogram used as an image input, removing the dependence on human experts for manual feature extraction, and improving the robustness to variation in signal structure, caller distance and signal-to-noise-ratio (SNR) conditions (Gibb et al., 2019). The success of CNNs has been demonstrated by many studies in the marine domain for binary species detection and multi-class species classification (Belgith et al., 2018; Harvey, 2018; Liu et al., 2018; Bergler et al., 2019; Bermant et al., 2019; Shiu et al., 2020; Yang et al., 2020; Zhong et al., 2020; Allen et al., 2021) advancing the capabilities of mining large PAM datasets for detecting species of interest. Existing work tends to make use of spectrogram representations across a limited bandwidth, which is selected according to the species (or signal) of interest. Herein the full frequency band is used to represent the signal as sources of interest in this application span the complete range of frequencies available. This does render the classification task more challenging, as the proportion of pixels containing information important to the classification task can be quite low compared to that available when the bandwidth is limited (Kahl et al., 2020), suggesting that large training sets may be necessary for the system to learn the task.

In this work we detect sound sources across the frequency spectra, to encompass the soundscape a signal is embedded in, resulting in source signatures occupying a small proportion of the input image. This a difficult problem when access to labelled data is limited. Transfer learning with fine-tuning presents a useful technique for developing a detector where labelled data is scarce (Shin et al., 2016) as a CNN trained for one task or domain is re-purposed for another related task or domain. It is effective as the original model is trained upon a large image dataset to enable the learning of low-level features applicable to many tasks, and is increasingly exploited by work which detects and classifies marine species acoustic signals from PAM data (Ibrahim et al., 2020; Thomas et al., 2020; Shiu et al., 2020; Lu et al., 2021). Fine-tuning is used to tweak the architecture to make it suitable for the new task or domain. In this work we harness the power of transfer learning, exploiting the features learned by pre-trained models, making use of a 'light-weight' architecture, EfficientNet B0 (Tan and Le, 2019). The solution developed is intended to be computationally efficient and suitable for on-platform deployment, while achieving high accuracies across differing sound source categories.

This works aims to demonstrate the application of transfer learning to discriminate major components of the soundscape in shallow waters, which vary in temporal and spectral characteristics, and assess the impact of variations in ambient noise on the detection capabilities of the approach. The resulting CNN model allows for sound source detection across a wide frequency range, to extract marine mammal call types and other soundscape components, as a method toward extracting ecological context within PAM data.

## Study site

Climatic and anthropogenic impacts are the likely cause of northward extensions of warmer species to mid-latitude areas, particularly for the *Delphinidae* family in the British Isles (Pirotta et al., 2015; Marley et al., 2017; Evans and Waggitt, 2020). Recording this shift in habitat use is vital for understanding the ecological importance of UK waters, for an appropriate evaluation of the effectiveness of current MPAs and policy applied to protected species. Twenty-three species of Cetacea have been recorded in Western Scottish waters over the last 25 years, of which eleven are regularly sighted, the majority of which are delphinids. All delphinid species in the region are listed on Annex IV of the Council Directive 92/43/EEC of 21 May 1992, the Habitats Directive, and have led to the designation of several MPAs in the region (Solandt, 2018). The COMPASS project (EU INTERREG) comprises a network of twelve PAM moorings in Western Scotland (Figure 1) operational since 2017, for monitoring protected sites and species, and nested within a suite of marine protected areas.

Data from the COMPASS project, in common with many other PAM datasets, has few manual annotations, and analysis of this data requires automation. The COMPASS data provides a case study for investigating the use of transfer learning in developing an automated detection model for multi-sound source classification of signals, which vary in temporal and spectral characteristics, in shallow water environments. We demonstrate the use of multi-class detection with application to the acoustic repertoire of delphinids and vessel sounds, combining labelled data from multiple data sources which span geographic, seasonal and temporal ranges, to develop a bespoke training set. Further, the effect of variable ambient noise conditions due to inter-site variability on detector performance is assessed.

## Methods

A computationally compact CNN model is developed for the detection and classification of marine sound sources spanning a wide frequency bandwidth, using real data representative of variable soundscapes. This section provides a detailed summary of the (i) Data acquisition and annotation, (ii) Data pre-processing, (iii) Model architecture and (iv) Training and testing.

### (i) Data acquisition

To obtain a robust network, the training data should represent the full diversity of each class. To increase the diversity with our training set we utilize data collected by a variety of organizations, under differing survey protocols and across a range of geographic locations and temporal scales, the technical details of these data sources are summarized in Table 1. The COMPASS data is supported by 4 additional sources. The

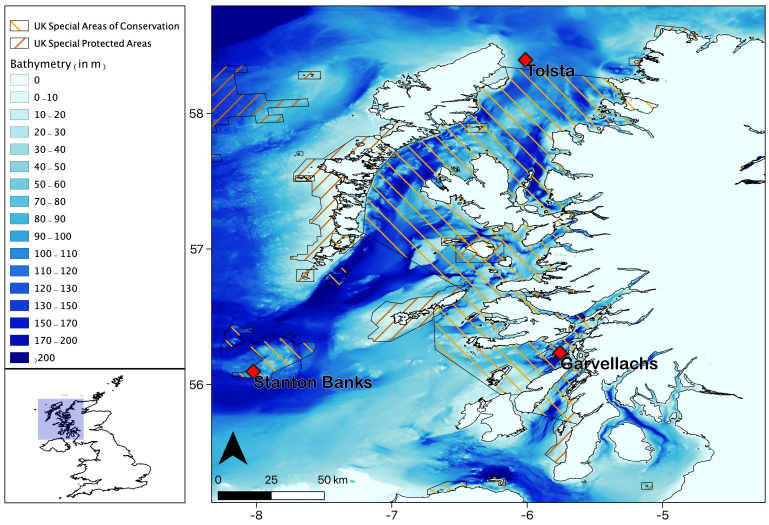


FIGURE 1
The location of the three hydrophone moorings used in this work, located off the west coast of Scotland, UK. Stanton Banks, Tolsta and Garvellachs recorder water depths were 72m, 102m and 95m respectively and are part of the EU INTERREG COMPASS project which aims to provide effective monitoring and management of Marine Protected Areas, including both Special Areas of Conservation and Special Protected Areas. Bathymetry data was sourced from GEBCO and Digimap.

TABLE 1 Description of the recording apparatus, and acoustic settings used to obtain the PAM data for DCLDE, HWDT, Solent and COMPASS data collections used in this study.

| Dataset | Temporal Period | Recording Platform | Sensitivity (dB) | Recorder Depth | Sample Rate (kHz) | Gain/Pre-amp | Bandwidth for analysis (kHz) |
|---|---|---|---|---|---|---|---|
| DCLDE Hawaii | 2017 | HTI-96-min | 14-85kHz ±5 dB at 158 dB re 1V/µPa | 0-30m | 500 | 37dB gain 1500Hz High Pass Filter | 0 - 250 |
| DCLDE Oregon | 2006 2007 | ITC 1042 HS150 | Flat frequency response (±3 dB) 1-100kHz | 10-30m | 192 | NA | 0 - 96 |
| HWDT | 2019 | HS150 | -204 dB re 1V/µPa | 4-10m | 96 | 29-35 DB gain | 0 - 48 |
| Sandown Bay | Aug-Sept 2020 | Wildlife Acoustic Song meter SM4M | -164.5 dB re 1V/µPa | 12m | 48<br>96 | 10dB<br>0-15dB | 0 -24<br>0 - 48 |
| COMPASS | 2017-present | SoundTrap 300 HF | 121 dB re 1V/µPa | 72-102m | 96 | NA | 0 - 48 |

additional datasets are available with different levels of annotation to mitigate the lack of annotations available for the COMPASS dataset. Two additional sources are from the open-access DCLDE conference datasets, DCLDE 2011 (Oregon) and DCLDE 2022 (Honolulu, Hawaii), both collected in the Pacific Ocean with annotations of interest being those for delphinid vocalizations. The Hebridean Whale and Dolphin Trust (HWDT) provided delphinid PAM recordings from West Scotland and archived acoustic data collected in Sandown Bay, Isle of Wight was provided by the University of Southampton (Table 1) which provide a rich source of ship noise data. The final source of data comes from the COMPASS project itself: it provides the greatest number of examples in the training set, ensuring the model learns to differentiate signals of interest from the ambient noise present in the deployment region (Figure 2, Table 2).

## Compass

This work makes use of acoustic data collected at three of the twelve COMPASS moorings: Stanton Banks, Tolsta and Garvellachs, collected between 2017 and 2019 (Figure 1). Each site possesses specific geographic and bathymetric conditions resulting in a distinct soundscape. The sites are located on the outer boundaries of the COMPASS array (Table 3). Stanton

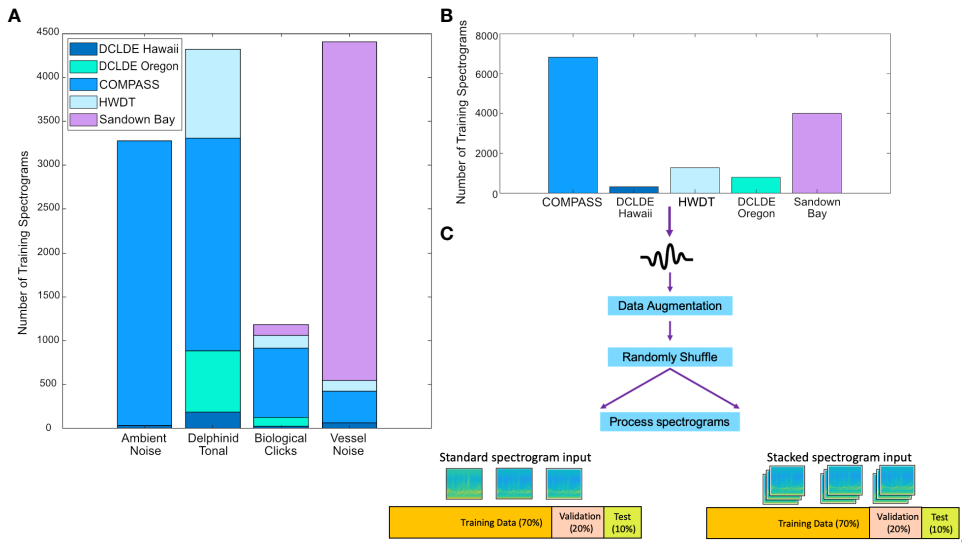


FIGURE 2
Creation of the training, validation and test sets. Sound files came from a range of data sources with differing recording hardware, survey platforms and location (see main text for more information). **(A, B)** Raw audio was manually collated into three-second frames each containing an example of ambient noise, delphinid call types or vessel noise. **(C)** To train and evaluate the model the sound sources, post augmentation, were randomly shuffled and divided into three folders; Training – 70%, Validation – 20% and Testing – 10%. Two datasets were computed for the standard spectrogram input and the stacked spectrogram input. The validation set was used to evaluate model performance during each iteration of training. The test set was shown to the model after training to test its performance on unseen data.

TABLE 2 Composition of training set source signatures, breakdown by dataset of origin, before and after augmentation. .

| | Ambient | Delphinid Tonal | Biological Clicks | Vessel Noise | Totals |
|---|---|---|---|---|---|
| **COMPASS** | 3288 | 2422 | 792 | 361 | **6863** |
| **DCLDE Hawaii** | 36 | 190 | 29 | 62 | **317** |
| **HWDT** | 0 | 1014 | 145 | 102 | **1261** |
| **DCLDE Oregon** | 0 | 693 | 86 | 0 | **779** |
| **Sandown Bay** | 0 | 0 | 123 | 3855 | **3978** |
| **Totals** | **3324** | **4319** | **1175** | **4380** | **13,198** |
| **Augmented** | **9972** | **12,957** | **3525** | **13,140** | **39,594** |
| **Training Set** | **6981** | **9070** | **2468** | **9198** | **27,717** |
| **Validation Set** | **1994** | **2591** | **705** | **2628** | **7918** |
| **Test Set** | **997** | **1295** | **353** | **1314** | **3959** |

Bold: emphasise the total values from the raw values.

TABLE 3 Contribution of PAM files from COMPASS to model development.

| COMPASS site | Model use | No. of wav files | Hours of recording | Season/year obtained |
|---|---|---|---|---|
| Tolsta | Training | 152 | 50.6 | Nov – March 2018 |
| Stanton Banks | Training | 59 | 19.6 | June – Nov 2017 |
| Tolsta | Testing | 322 | 107 | Dec, March, April, June - 2019 |
| Stanton Banks | Testing | 322 | 107 | Dec, March, April, June - 2019 |
| Garvellachs | Testing | 322 | 107 | Dec, March, April, June - 2019 |

Banks (56.097 N, -8.022 W), is an exposed site which sits at a water depth of 104m, the most western component of the network, with a bottom substrate composed of mud to muddy sand. Tolsta is a sheltered site to the North of the Inner Hebrides (58.394 N, -6.012 W) at a water depth of 94m, on a bed of sand. Garvellachs sits within a harbor at a depth of 76m (56.235 N, -5.756 W), the bottom type for this location is undescribed. For each mooring a single omnidirectional acoustic broadband recorder is moored 3-5m above the seafloor, recording on a 20/40 minutes on/off duty cycle, saved in 20-minute wav files captured at a sample rate of 96 kHz.

## DCLDE datasets

The DCLDE 2011 Oregon dataset contains calls from short-beaked and long-beaked common dolphins (*Delphinus delphis* and *D. capensis*), bottlenose dolphins (*Tursiop truncatus*) and spinner dolphins (*Stenella longirostris*), which were used to develop the training set with ground-truthed delphinid signals. Recorded in the Southern California Bight, the dataset encompasses both echolocations click and tonal calls (Roch et al., 2011b).

The DCLDE 2022 dataset includes annotated PAM recordings from Hawaiian waters in 2017, featuring delphinid calls, both identified and unidentified to species level, (Yano et al., 2018). Acoustic data is collected on a six-channel towed hydrophone from a large survey vessel, only channels 5 and 6 are included within the training set; these channels being those furthest from the towing vessel, resulting in the lowest noise levels from the vessel.

## HWDT

Acoustic data collected during dedicated visual and acoustic line surveys, within the waters surrounding the Inner and Outer Hebrides, Scotland, was analyzed. PAM files recorded during both the summer and winter seasons, 2019, were provided by HWDT together with associated timestamps for delphinid tonal and broadband calls, classified to species level. A sailing vessel (*Silurian*) collected the data and use of this platform resulted in a low level of vessel noise present within the field recordings of odontocetes.

## Sandown bay

Archive acoustic data collected in August 2020 within Sandown Bay, UK (50.690 N, -1.258 W) was provided as a basis for vessel detection, recorded on a static seabed mooring deployed at 12m water depth. This data source presents vessel signatures to the model recorded in a shallow water acoustic environment. Across the recording period, a range of vessel types were present in large numbers due to the proximity of shipping lanes.

## Data annotation

Four broad sound source classes are defined to test the feasibility of using a small-scale transfer learning developed

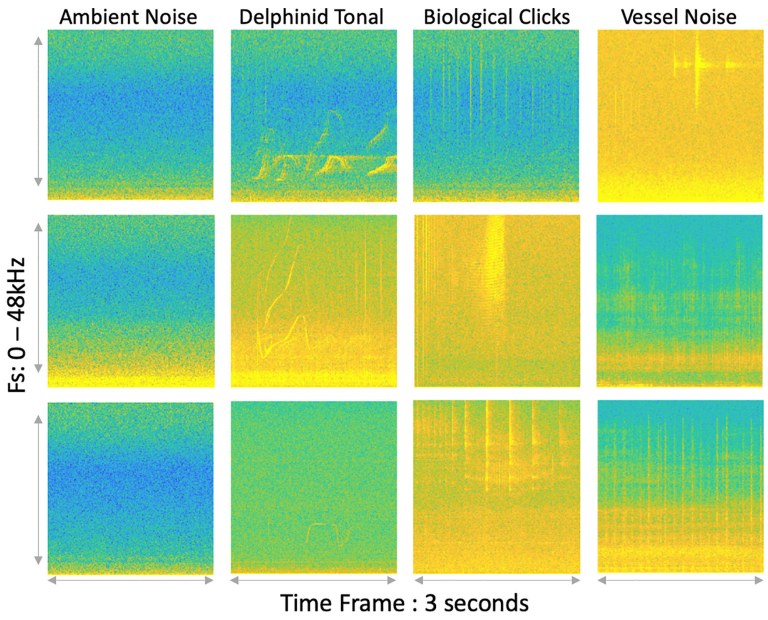


FIGURE 3
Representative spectrograms illustrating the four classes used as input for training; Ambient noise, Delphinid tonal, Biological clicks (echolocation and burst pulses) and Vessel noise. Each spectrogram had an image resolution of 224x224 pixels and is computed using a Hanning window, 75% overlap, and an FFT size of 2048, standardised across the various sampling rates. Time and Frequency are on the horizontal and vertical axes, respectively. Three different examples are shown for each sound class as spectrogram characteristics vary across temporal, spatial and geographic scales introducing variance during training, essential for generalisation to new unseen datasets.

CNN to detect soundscape components across a wide bandwidth; 'Ambient noise', 'Vessel noise' and two classes relating to the presence of delphinids, 'Delphinid tonal' and 'Biological clicks' (Figure 3), representing delphinid vocal repertoires, anthropogenic sounds, and a negative class – ambient noise. Ambient noise spans the entire analysis bandwidth and is described as the soundscape of the water column in the absence of one of the distinct sound sources. Delphinid tonal includes tonal frequency-modulated whistles with a typical frequency range of 1 kHz – 40 kHz. The species most frequently encountered in the region of the COMPASS array are: bottlenose dolphins (*Tursiops truncatus*), killer whales (*Orcinus orca*), Risso's dolphin (*Grampus griseus*), white-beaked dolphin (*Lagenorhynchus albirostris*), Atlantic white-sided dolphin (*L. acutus*), and striped dolphin (*S. coeruleoalba*) (Hague et al., 2020). Biological clicks for this work are defined as echolocation click trains and burst pulse delphinid signals, spanning frequency ranges from 8 kHz to 100 kHz. Due to the limited analysis bandwidth of 48 kHz the echolocation clicks recorded on the system are not a faithful representation of the pulses in the water. This class does not include the low-frequency clicking sounds which are labelled as part of the noise class. The precise source of this low-frequency clicking remains uncertain off the UK coastline but is qualitatively similar to sounds associated with snapping shrimp: such species are usually associated with more temperate waters (Au and Banks, 1998). Vessel noise encompasses low-frequency vessel noise and high-frequency signals produced from echo-sounders.

PAM files are divided into blocks of 3 s, a length chosen as a compromise, with the sound sources selected in this work having durations over different time-scales: milliseconds (individual echolocation clicks), seconds (whistles and echo-sounders) and several minutes (vessel passage). The 3 s blocks were reviewed visually (spectrograms) and aurally using Audacity software (Audacity version 3.0.02, 2021) to classify each block into one of the four categories, developing a training set of 13,198 spectrograms (Table 3). Spectrograms are not time-centered on the detected signal, so the signal can occur in part or in full within the 3 s time window.

DCLDE and HWDT data were provided with associated weak labels, labels which identified PAM files containing signals of interest but not timestamps matching sounds pertaining to that label. Consequently, these datasets were included in the annotation process with the selection of PAM files based on the weak labels. Data blocks within the COMPASS, (Table 3), and Sandown Bay datasets were selected randomly from the entire datasets, providing a sampling of the soundscape across temporal and seasonal scales.

Annotations were labelled by a team, under the supervision of the lead author according to a set of rules: (i) if a whistle is present in the 3 s frame the label is 'Delphinid Tonal', regardless of the presence of another sound source, e.g. clicks; (ii)

echolocation clicks in the same frame as vessel noise is labelled as clicks to reduce delphinid false negatives; (iii) a sound source is labelled if any detection is made by an analyst regardless of signal strength in the frame in comparison to the ambient noise. All labels were reviewed by the lead author to ensure annotation rules were followed.

The existence of class noise in the training set, due to mislabeling, is a common issue and results in a marginal decrease in the accuracy of the classifier when the error rate is low (Nazari et al., 2018). Retrospectively, a strategy was employed for annotation verification with 20% (n = 2639) of the spectrograms within the training set sampled randomly. Two analysts independently (blind) annotated the spectrograms providing their own associated labels (Supplementary Table 1). Across the verification data an error metric of 3.3% is reported, reflecting the mean per class of the two analysts compared to the original training labels. Most discrepancies occurred between the 'Ambient Noise' (6.3%) and 'Vessel Noise' (3.7%) classes (Supplementary Table 1) where distant vessels were difficult to identify within the ambient soundscape. We expect there remains an approximate error of 3% within the overall training set.

## (ii) Data pre-processing:

When using data from disparate sources ideally one would standardize collections protocols, for instance using a common sample rate and common/calibrated sensor systems. We wish to exploit datasets whose collection was not collected according to a common protocol and so have to deal with inconsistencies between recording configurations. A set of pre-processing steps were applied to ensure a consistent overall dataset. First the mean data was centered (the mean value of a recording subtracted from all the samples). Not all of the systems were calibrated, with the result that different gains applied to the acoustic data from different datasets. In particular, the dynamic range of the COMPASS data was significantly smaller than that of the other data. This was mitigated by applying 30 dB gain to the COMPASS data, which yielded spectrograms which were subjectively judged to be comparable with those from the other data sources. The effect of these steps is visualized in Figure 4.

In this work the 3 s wav files were processed as spectrograms (using a linear frequency axis with energy represented in dB and computed based on a Hanning window with 50% overlap) and are used as input to the CNN. EfficientNet is originally trained on RGB images and requires a three-channel input. Here we present a 'stacked spectrogram' that takes advantage of the three channels to increase the information available to the network for a single input. The input channels correspond to three spectrograms computed at three different time-frequency resolutions (frequency bins of widths 93.75 Hz, 46.88 Hz and 23.44 Hz, corresponding to FFT sizes of 1024, 2048 and 4096 at 96 kHz sample rate). The FFT window sizes were adjusted to ensure a consistent time and frequency spacing independent of

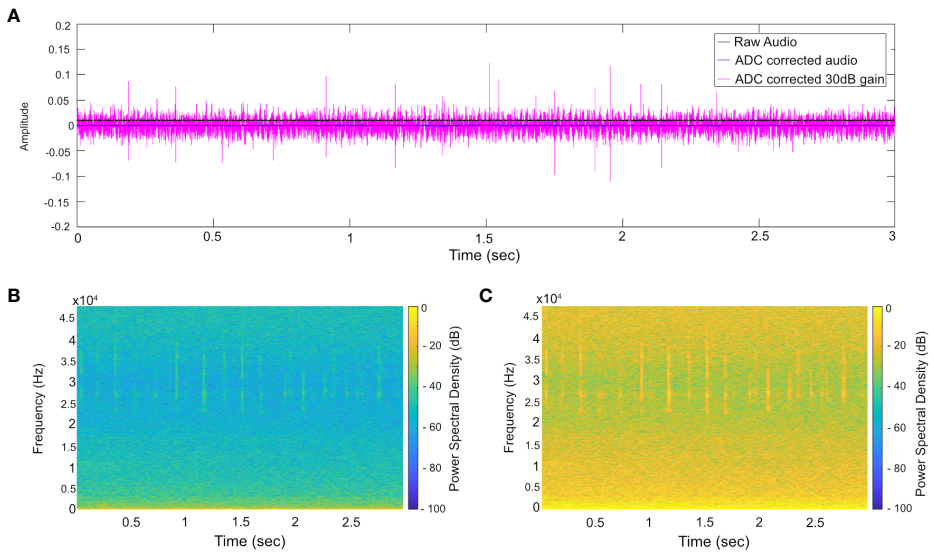


FIGURE 4
Conditioning applied to COMPASS acoustic files, using exemplar data containing echolocation clicks. **(A)** time domain showing raw data (black line) with positive bias (+0.01), corrected data (blue line), and data after application of 30 dB gain (magenta). **(B)** Raw spectrogram without bias correction, or gain. Clicks are only faintly visible. **(C)** Spectrogram after bias and application of a gain, emphasizing spectral content of which the model will detect ensuring consistency with the other data sources being used.

the sample rates used. For convenience we refer to FFT sizes of 1024, 2048 and 4096 which pertain to the 96 kHz.

In datasets with sample rates greater than 96 kHz, the spectrogram points corresponding to frequencies above 48 kHz were discarded, so that the CNN is only presented with data corresponding to 0-48 kHz. A small portion of Sandown Bay data was sampled at 48 kHz resulting in a spectrogram with a bandwidth 0-24 kHz (n = 300). Vessel source signatures were the only sound source category extracted from this dataset, and this is the only portion of data within the training set sampled below 96 kHz. For the 300 frames sampled at 48 kHz spectrograms were created with a frequency bandwidth 0 – 24 kHz with zero padding used to maintain the consistent pixel resolution required as the models input.

Spectrograms used to train the model each possess time × frequency dimensions of 1122 × 512 (FFT: 1024), 558 × 1024 (FFT: 2048) and 277 × 2048 (4096). Each color channel of the image contains a greyscale spectrogram processed with one of the FFT values described above, corresponding to R-1024, G-2048, B-4096, with the FFT value corrected for the data source sampling rate for consistency (Figure 5). All spectrograms have a frequency bandwidth spanning 0 – 48 kHz. The portion of data recorded in Sandown Bay sampled at 48 kHz, with an analysis bandwidth of 24 kHz, were included in this work, with the FFT window sizes adjusted to ensure a time and frequency spacing consistent with the rest of the data.

Each spectrogram channel (R-1024, G-2048, B-4096) is resized to 224 × 224 pixels and combined for the stacked spectrogram. In this manner we ensure that the RGB value for each pixel corresponds to the same time – frequency point in spectrograms computed with different window lengths and different sampling rates. The spectrogram values are standardized so that they correspond to the range -100 to 0 dB.

A second training set was computed to evaluate the performance of the stacked spectrogram. Pilot studies were conducted using spectrograms computed using a single frequency resolution. Three CNN models were trained, each receiving input data computed at a single frequency resolution. A comparison of inter-class performance between the three models trained on spectrograms computed at specific FFT window sizes (1024, 2048 or 4096), determined variation in model performance for each sound source category. A combined approach was developed, a single model is trained on input images processed at one of the three chosen FFT window sizes (1024, 2048 or 4096) for all acoustic files. The following sections refer to this training set as the standard spectrogram approach. Using the same correction process described above the full set of 3 s wav files are processed at each of the three FFT sizes 1024, 2048 and 4096, outputting three stand-alone spectrograms to be stored as images for input. The trained model received each 3 s frame input at three different frequency resolutions, independent from one another. These spectrograms used the colormap 'parula' to present a three-channel RGB image as input, corresponding to the range -100 to 0 dB. This method enhanced the training set three-fold, and was originally explored as a form of audio image augmentation, to introduce variance with respect to the temporal and frequency resolution of each image produced.

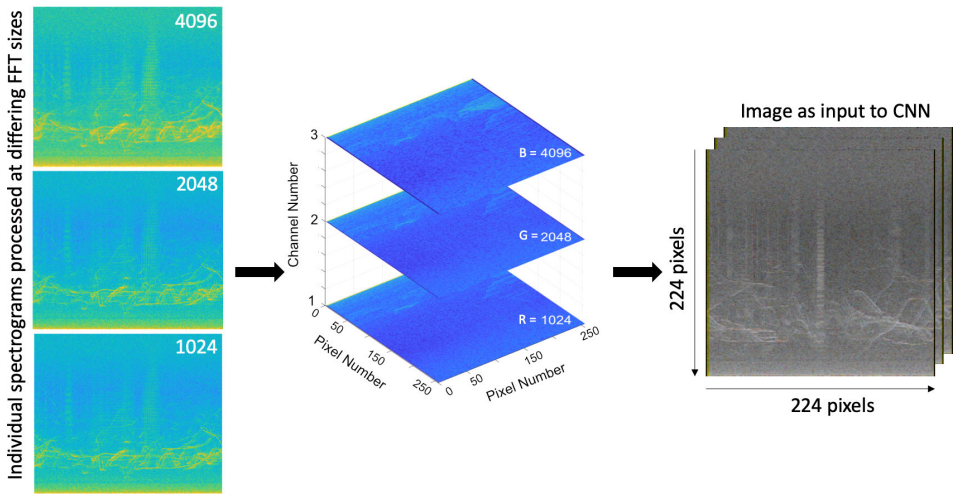


FIGURE 5
Graphic demonstrating the structure of the input to the model, a 3-channel greyscale spectrogram and is resized after processing to 224x224. Three spectrograms were produced for each wav file with different FFT size parameters as input: NFFT values were standardised for the equivalent of 1024, 2048 and 4096 points per each sample rate with a constant frequency increment of 93.75 Hz, 46.88 Hz and 23.44 Hz respectively. Spectrograms were computed using a Hanning window and 75% overlap. The three images were stacked within one of three color channels of the image in order, allowing the CNN to 'see' each sound source represented at various resolutions.

## (iii) Model architecture

This work harnessed the power of transfer learning, using pre-trained layers and blocks from an existing architecture to develop a CNN pipeline, for automated sound source detection. The network used in this work was the EfficientNet model (Tan and Le, 2019). EfficientNet is a family of networks which use model scaling, balancing network depth, width and resolution to output state-of-the-art accuracies with relatively few parameters (Tan and Le, 2019). Our work utilized EfficientNet B0 the smallest of these networks. Using 5.3 million parameters it is 8.4× smaller and 6.1× faster than other commonly adopted architectures (Tan and Le, 2019). EfficientNet B0 is pretrained on the ImageNet database (Deng et al., 2009) with 1000 classes. Figure 6 outlines the EfficientNet B0 architecture as used in this study, the original feature extractor is left frozen, the weights and biases determined through training on the ImageNet database are not updated during training for this work. The final layers of the architecture, the classifier, have been replaced through fine-tuning and trained on the dataset developed for this work.

### Fine-tuning

The final classification layers are a set of fully connected layers, attached *via* a Global Average Pooling layer (GAP) which reduces the number of features to 1280 (Figure 6). The GAP layer takes the average of each feature map in the last convolutional layer of the EfficientNet B0 architecture and flattens the output of the feature extractor into a vector, which can be used as a feature descriptor and fed into the fully connected layers of the classifier. This process is described as bottlenecking as we are summarizing the learned features in the EfficientNet B0 architecture into a single vector to be used as input to our classifier. The bottlenecked information is passed through three fully connected layers with 512, 256 and 4 neurons in sequence. Between each layer are dropout layers of 50% and 20% respectively. Dropout works by randomly setting a set percentage of neurons in the fully connected layer to zero during each training update, helping to generalize model performance and avoid overfitting, important where small training sets are being used. The 512 and 256 dense layers use a ReLU activation function (Krizhevsky et al., 2012), a simple function that returns the input value if it is positive or sets the value to 0 if the input is negative. The final four neuron fully connected layer classifies the input spectrogram into one of the four classes, using a softmax activation function. The softmax function outputs the pseudo-probability of an image belonging to each of the four classes and the network assigns a label based on the highest pseudo-probability value. The final model architecture had a total of 4.8 million parameters, of which 788,228 are trainable and 4,049,564 remain frozen during training updates.

## (iv) Training

Data augmentation is used to enhance variance within the training set and increase the number of inputs, without requiring further manual annotation effort (Figure 7). From each 3 s audio

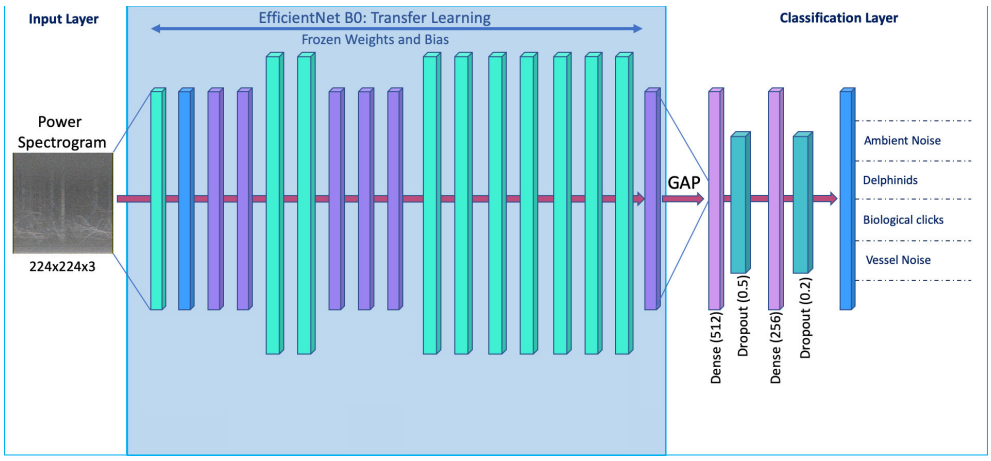


FIGURE 6
Schematic diagram of the CNN model. Transfer learning was used to build the model, freezing weights and biases from the original EfficientNet B0 architecture, represented by the blue shaded region. White regions are fine-tuned layers, trained on our dataset. The model learns representations of the input at each stage of the CNN, the learned features are bottlenecked in the Global Average Pooling (GAP) layer, before being fed into the custom layers. Layer composition is described within the main text. Drop out layers were introduced to enhance the model's ability to generalize to unseen data. The final layer has four neurons per sound source category, with the ability to adjust this layer to new categories of interest.

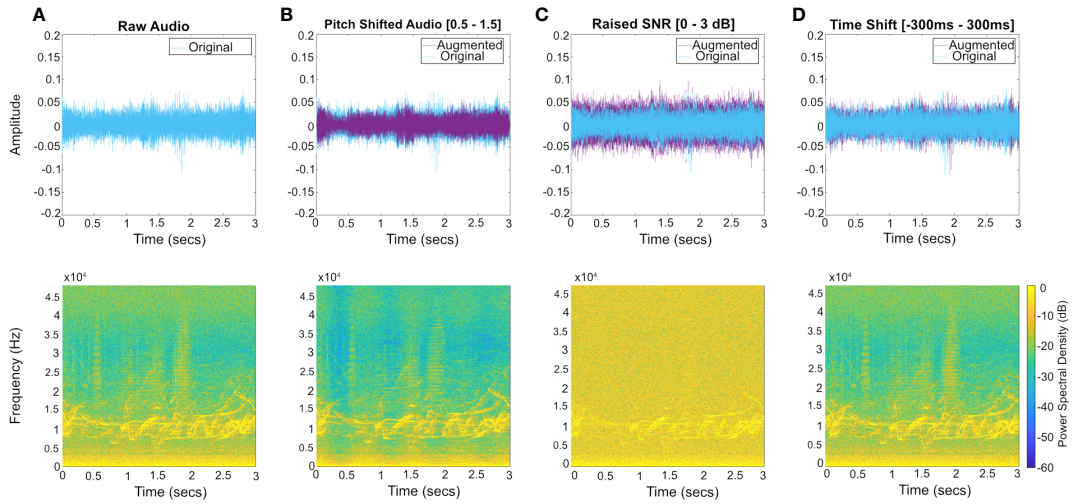


FIGURE 7
Temporal and frequency domain examples of three methods of audio augmentation used to enhance variability within the training dataset. Audio augmentation was performed on raw audio, in the time domain. **(A)** raw acoustic data; **(B)** pitch shift applied by a random factor between 0.5 to 1.5; **(C)** reduction in SNR (up to 3dB); **(D)** time shift (between -300ms and 300ms). Augmentation maximum and minimum values introduce variance to the training set while maintaining the spectral content within the frame. Each input audio file was subjected to two augmentations, the range of augmentation methods and the value parameters for their implementation are randomised. Spectrograms were produced using a Hanning window, FFT size 2048 and a 75% overlap. Augmentation increases the training set three-fold, from 13,198 to 39,594 images.

clip in the training set, 2 additional spectrograms were created through augmentation. The spectrograms were created by applying one of 3 randomly selected signal transformations to the audio before forming the spectrogram: these being pitch shifts, time shifts, and adding noise, not limited to one per category. Each transformation was parameterized and the parameter value was randomly selected from a defined range. Specifically, the factors for pitch shifting were selected between 0.5 and 1.5, time shifts are selected between -300 ms and 300 ms and Guassian white noise is added with powers between 0 and 3 dB. Note that a limited range of time shifts were used to reduce the possibility of the sound source moving out of the window and in doing so changing the correct class for the clip. After augmentation the training set increased three-fold. Table 2 details the contribution of each data source to each sound source category.

Labelled spectrograms from each data source are combined to form a single data set. The full set of spectrograms is divided into three subsets, training (70%), validation (20%) and testing (10%), Figure 2, for use in model training and evaluation. Spectrograms are randomly isolated per class to remove the likelihood of one sound source being underrepresented in each of the data subsets, with respect to the imbalance in quantity across the classes, and to prevent leakage between training and test sets.

The model was developed and trained within the Google Collaboratory 'Colab' platform using the Tesla K80 GPU, allowing training to occur on a personal computer rather than relying on expensive hardware. Custom written python scripts made use of Tensorflow and Keras (Chollet, 2018) libraries for model architecture, finetuning of layers, model training and testing. Training data were shown to the model in small batches (n = 32), until the whole dataset has been revealed, one full epoch, for mini-batch gradient descent. During training a horizontal flip, an image augmentation method was applied randomly to images during training (one per batch of 32). This method does not increase the size of the training set. The image was either flipped, or not flipped, along the horizontal axis as it is fed into the model to increase variance.

Successful training is reliant on the successful choice of a learning rate (Murphy, 2012); an Adam optimizer was used, controlling gradient descent during training (Kingma and Ba, 2014). A cyclical learning rate was applied, using a learning rate of $10^{-3}$, decaying by 0.75 every 90 steps (Smith, 2017). The loss function used is categorical cross-entropy (Koidl, 2013). During training, model performance was evaluated in real-time from reported training and validation loss and accuracy values, once per epoch, to observe the potential for overfitting. The model was set to train for 50 epochs, but training was stopped early if the validation accuracy and loss did not improve over 8 epochs (early-stopping). The delay of 8 epochs was due to the observation that improvements in model performance are stochastic during training and it can take several epochs to realize the potential benefit of a specific learning rate (Ruff et al., 2021). For both training sets, the standard spectrogram and the stacked spectrogram inputs, the model architecture and training

procedure were the same. The final model using the stacked input took 64 minutes to train within Google Colab, and ran for 33 epochs before early stopping. The final model using the standard spectrogram input trained for 41 epochs before early stopping, taking 112 minutes to train.

Precision (*P*), Recall (*R*) and Accuracy (*A*) were used to compare overall performance, as well as cross-class performance. *P* and *R* are calculated from the number of true positives (TP), or correct classifications, $N_{TP}$, false positives (FP), or incorrect classifications, $N_{FP}$, and false negatives (FN), or missed detections, $N_{FN}$:

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}}, \qquad R = \frac{N_{TP}}{N_{TP} + N_{FN}}.$$

A perfectly performing detector across all classes results in 1, *P*=*R*=1. Accuracy was based on F1 scores which are used to combine precision and recall, F1 being the harmonic mean of *P* and *R*, with good performance indicated by values close to 1. There are two methods for combining metrics across multiple classes. These are: the macro-average, calculated by computing the F1 metric for each class and then finding the unweighted mean of those values (Mesaros et al., 2016) and the micro-averaged scores for which the values of $N_{TP}$, $N_{FP}$ and $N_{FN}$ are accumulated across the classes and the metric evaluated using these combined values. We plotted Receiver Operating Characteristic (ROC) curves to summarize performance within a class across a range of threshold levels, with the areas-under-curve (AUC) used as a summary statistic for these curves (Stowell, 2022). As a baseline, a random classifier is expected to output a diagonal line on the ROC curve, that is the false positive rate (*FPR*) is equal to true positive rate (*TPR*). For calculating the statistical metrics Scikit-learn was used (Pedregosa et al., 2011).

# Evaluation

In this section, we evaluate the performance of the model by investigating, (i) the effect of the CNN input on model accuracy using in-sample test data, (ii) the effect of variable signal-to-noise-ratio, and (iii) the effect of extreme weather conditions.

## (i) In-sample test set

The first evaluation of the model was performed on the in-sample test set, i.e. the 10% of data (n = 3959, Table 2) not seen during the training and validation processes. The test set consisted of 997 ambient noise images, 1295 delphinid tonal images, 353 biological click images and 1314 vessel noise images, totaling 3959 input images. The aim of this test was to assess the developed model's performance on unseen data and to validate the use of the stacked spectrogram input developed for this work. The model using the stacked spectrogram input trained for 33 epochs before stopping, reporting a loss of 0.14, whereas the model trained on standard spectrogram inputs stopped after 41 epochs, with a loss of 0.42.

Figure 8 shows the performance of both models on the test set with respect to the confusion matrix and the ROC curves. For the in-sample test data set the macro-averaged accuracy for the standard approach was 80% with a lower micro-average accuracy of 66%, the relatively large difference between these accuracies stems from the poor performance on the smallest class, the delphinid clicks (Figure 8). In contrast, the results for the network using the stacked spectrogram input had similar accuracies of 94% (macro-averaged) and 96% (micro-averaged), reflecting the success of this approach across all sound sources. In particular, the stacked spectrogram outperformed the standard approach in both delphinid classes (clicks and tonal calls) and the ambient noise class. Both approaches make incorrect decisions regarding the vessel noise class, the stacked approach suffers from FNs, failing to detect the presence of a vessel and incorrectly classifying clips as ambient noise, whereas the standard approach has a tendency to generate FPs, mistaking ambient noise for a vessel.

The ROC curve (Figure 8) illustrates the performance of the network across a range of threshold values to evaluate model performance with respect to positive detection rates. For the stacked spectrogram approach the AUC scores are above 0.93 for all classes, maximized in the delphinid class (AUC = 0.97). The standard spectrogram input performs poorly as a delphinid classifier, achieving an *F1* score of 0 for clicks and 0.05 for tonal calls. The poor performance derives from it misclassifying 93% of the images as ambient noise and labelling 87% of the clicks as tonal calls. Supplementary Table 5 details the calculated macro-*P* macro-*R* and macro-*F1* scores for both input types, highlighting the success of the stacked spectrogram approach for source signals that vary dramatically in nature.

## (ii) The effect of signal-to-noise-ratio

Marine datasets contain ambient noise levels which vary depending on the local activity and prevailing weather conditions. Classification becomes more challenging as the SNR reduces, we explored the influence of SNR on the behavior of the proposed method to evaluate the performance of our model outside of the test set and in conditions reflective of real-world acoustic data. The model was deployed, without further training, on a set of manually selected files containing delphinid tonal, biological clicks and vessel noise in high and low SNR conditions, recorded at Tolsta, Stanton Banks and Garvellachs. Note that no data from Garvellachs was included in the training, validation or test data, representing a previously unseen data source. The time periods for selected files from Tolsta and Stanton Banks did not occur within, or overlap with

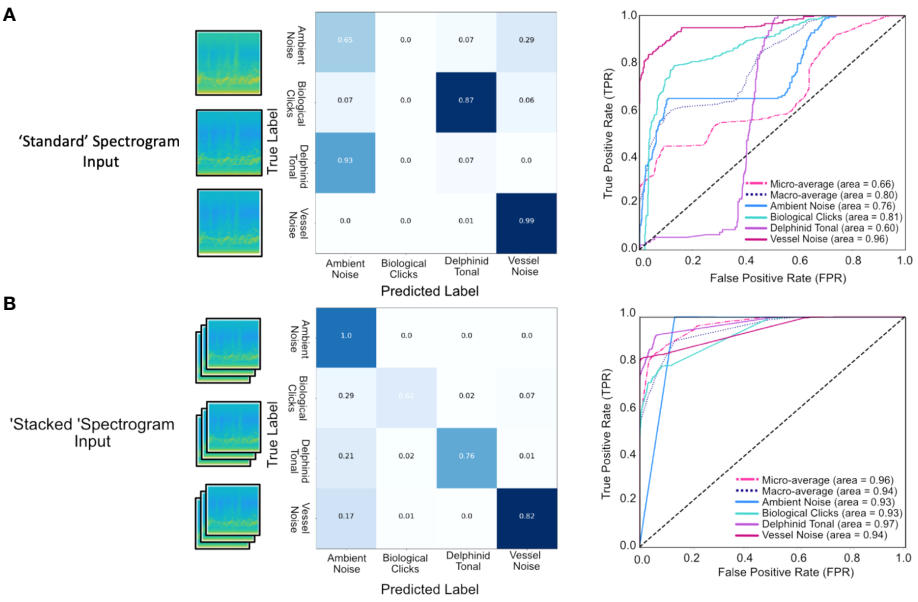


FIGURE 8
Performance of the CNN on the test set (10% of the initial training data) for each of the four sound classes as illustrated using Confusion Matrices and ROC curves for **(A)** Standard spectrogram input and **(B)** Stacked spectrogram input. The stacked input outperformed the standard spectrogram input reporting a macro average of 0.94, compared to 0.80. Performance was high for all sound classes with the highest accuracy for ambient noise and vessel noise which are the classes comprising the greatest proportion of the training set based on true vs predicted labels for the stacked input. Although making up the smallest component of the training set, biological click detection remained high, with an AUC of 0.93, with false negatives in this class attributed to single click occurrences.

the time periods used for training, testing or validation and statistical work was carried out to ensure the distribution of each data set are independent of one another (see Supplementary Material). High and low SNR conditions were identified by calculating a per frame SNR to account for variation in signal types, using per frame ambient noise as the reference signal. High SNR frames were defined as encounters with loud clear signals SNR>10dB above the ambient noise level of the PAM file from which they were extracted. Low SNR frames were defined as signals having SNR<5dB, visually appearing to be of poor quality in the spectrogram. See Supplementary Material for exemplar spectrograms belonging to both high and low SNR conditions.

For high SNR data, a macro-average of 0.95 (ROC) was reported and a macro-average accuracy score computed at 0.91, see Figure 9. Classification of biological clicks was high, reporting an AUC score of 0.96 and an F1 score of 0.95 (Table 4, Figure 9A). As SNR varies in the click class the model reported few false detections as ambient noise or vessel noise (Figure 9A), a common issue for other detection algorithms. Delphinid tonal calls had a lower recall compared to clicks (0.78, Table 4), because of the method of manual annotation for a true label; if a whistle was present the frame is labelled delphinid tonal. During periods of high SNR the delphinid clicks in many frames presented visually as a stronger signal than the whistles, the model was correctly identifying the presence of a click signal but was penalized in the evaluation metrics.

An accuracy of 0.87 was achieved on the low SNR data (Table 4). Ambient noise and delphinid whistle classifications were high, (Ambient noise: AUC = 0.93, F1 score=0.90, Delphinid: AUC=0.96, F1 score=0.88, Table 4). No false positives occurred between clicks and tonal calls, 18% of whistles were missed and recorded as ambient noise as a result of call masking, Figure 9B. Performance of biological clicks was poor, $F1$ = 0.34, AUC=0.8, as a result of very low recall (Table 4, Figure 9B). Detections missed (68%) were manually inspected and determined to be instances of singular strong clicks or few clicks, not present in a click train. The classification of true click frames reported an AUC score of 0.79 (Table 4, Figure 9B). Whistle classification thresholding was successful at low *FPR*, AUC=0.96 in low SNR conditions, Table 4, Figure 9B. No vessel noise frames were used from the selected data for this test and are not included in Figure 9 or Table 4. Discrimination of vessel noise and ambient noise was not straightforward, and full analysis would need to include ship automatic identification system (AIS) data for the testing region to identify true low SNR vessel signatures.

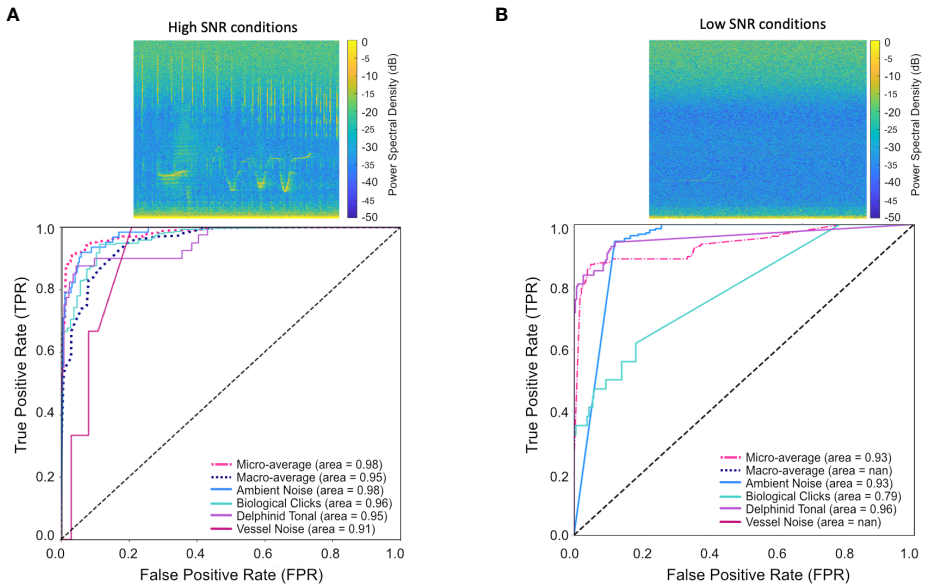


FIGURE 9
Comparison of CNN performance for periods of **(A)** high, and low **(B)** signal to noise ratio, represented as receiver operating characteristic (ROC) curves. Our model was able to classify delphinid calls with 80% accuracy in low-quality data, and 95% accuracy in high-quality data. Reduced accuracy occured in the biological click class in low-quality data, AUC = 0.79. Manual analysis determined many missed detections are due to singular higher strength clicks amongst low signal strength clicks. Spectrograms **(A, B)** were computed using a Hanning window, FFT size 2048 and a 75% overlap.

TABLE 4 Performance metrics computed for stacked spectrogram input, the standard spectrogram input, high SNR and low SNR evaluations.

| Test | Sound source | Precision | Recall | F1 Score | Average | Macro-Average (ROC curve) | AUC |
|---|---|---|---|---|---|---|---|
| Stacked Spectrogram input | Ambient | 0.64 | 1.00 | 0.78 | 0.84 | 0.94 | 0.93 |
| | Clicks | 0.86 | 0.62 | 0.72 | | | 0.93 |
| | Delphinid Tonal | 0.99 | 0.76 | 0.86 | | | 0.97 |
| | Vessel | 0.98 | 0.82 | 0.89 | | | 0.94 |
| Standard Spectrogram input | Ambient | 0.45 | 0.65 | 0.54 | | 0.80 | 0.76 |
| | Clicks | 0 | 0 | 0 | | | 0.60 |
| | Delphinid Tonal | 0.03 | 0.10 | 0.05 | | | 0.60 |
| | Vessel | 0.76 | 0.99 | 0.86 | | | 0.96 |
| High SNR>10dB | Ambient | 0.83 | 0.81 | 0.82 | 0.91 | 0.95 | 0.98 |
| | Clicks | 0.94 | 0.96 | 0.95 | | | 0.96 |
| | Delphinid Tonal | 0.82 | 0.78 | 0.79 | | | 0.95 |
| | Vessel | 0 | 0 | 0 | | | 0.91 |
| Low SNR<5dB | Ambient | 0.82 | 1.00 | 0.90 | 0.87 | *na* | 0.93 |
| | Clicks | 1.00 | 0.21 | 0.34 | | | 0.79 |
| | Delphinid Tonal | 0.97 | 0.82 | 0.88 | | | 0.96 |
| | Vessel | *na* | *na* | *na* | | | *na* |

## (iii) Soundscape variation

The final evaluation of model performance was tested across variable soundscape conditions at Stanton Banks, Tolsta and Garvellachs. Temporal periods correlating to above average high-pressure (HP) and low-pressure (LP) weather systems were identified by reviewing archive weather reports for the South Uist weather station on the Outer Hebrides (Figure 10A).

Two LP periods are identified, December – Storm Diedre (15th – 18th, 2018) and March – Storm Gareth (12th – 13th, 2019) with maximum wind speeds of 71 mph and 61 mph, respectively. The storm in December resulted in 67 mm of rainfall in contrast to 14 mm in March. The HP systems identified occurred in April (19th – 22nd, 2019) and June (26th – 28th, 2019) with maximum wind speeds of 32 mph and 26 mph and rainfall of 3 mm and 0 mm, respectively. The site-specific characteristics surrounding each PAM mooring vary in water depth, bottom type and bathymetry affecting the acoustic complexity of the soundscape under variable weather conditions. Analysis of the sound pressure level per site within third-octave level bands between 10 Hz – 10 kHz demonstrated the regional variation in ambient noise under differing weather conditions (Figure 10A). Low frequency ambient noise was present in the wide-bandwidth spectral input, we assessed the effect of this on sound source detection.

Each weather period comprised differing temporal lengths in hours, April: 32, June: 24, December: 32, March: 16 for a total number of 38,304 (April), 28,728 (June), 38,304 (December) and 19,152 (March) input spectrograms. Human analysts manually labelled each frame prior to model input, for comparison to model output post detection.

Overall macro-averages were consistent across the four seasons, ranging between 0.73 and 0.87 (Figure 10, Table 5), with December reporting the highest average score of 0.87 (Figure 10B). Averages calculated from confusion matrices, without macro-averaging, determined that performance in LP conditions (Dec: 0.87, Mar: 0.85) were higher than the summer (HP) seasons (April: 0.55, June: 0.65), Table 5. This variation indicates per-frame variation in classification ability as a result of soundscape diversity and reflects the unbalanced nature of the datasets with respect to images per class skewing overall averages. For this reason, macro-averages reported from the ROC-curves offer a better assessment of model performance.

Within each seasonal period AUC scores fluctuate per class, with inter-class performance consistent across each period (Table 5, Figure 10B). Seasonality was expected to affect model performance, with storm periods affecting the model output. Ambient noise, our negative class, reported the highest *F1* score during the March storm period, 0.92, and the lowest in April, 0.32, with an average *F1* score of 0.36 in HP systems and 0.9 for LP systems (Table 5). Precision scores were high for ambient noise, with an overall average of 0.83, varying between 0.78 for HP systems and 0.88 for LP systems (Table 5). The overall average was impacted by a low recall score of 0.24 across April and June (Table 5), compared with 0.93 for December and March (Table 5). Ambient noise classification accuracy is affected by seasonal variation in soundscape characteristics per site.

Soundscape variability due to weather conditions affects the model's capability to detect delphinid tonal calls in particularly bad weather (March, $P = 0.19$; Table 5, Figure 10B); the severity of weather conditions is important to consider when interpreting results. In the three other temporal periods, including Storm Diedre, the model detected delphinid tonal calls with high precision April ($P = 0.98$), June ($P = 0.95$), December ($P = 0.97$), indicating high confidence as a whistle detection model, Table 5. Tonal calls score an average *R* of 0.92 in HP systems and 0.72 LP systems with an average *F1* score of 0.94 for April, June and December. AUC scores were high, reporting 0.99 for both April and June, and 0.97 for December, Table 5. The detector is effective but is skewed by seasonal and diurnal variation in soundscape characteristics.

The 'click' class reported variable results across the seasonalities, with higher F1 scores for December (0.92) and March (0.73), than the HP systems April (F1 = 0.33) and June (F1 = 0.38), Table 5. For April and June both *P* and *R* scores were below 0.5, with *P* values highest in June (0.44), Table 5. For this sound source class AUC scores remained high at 0.90 and 0.79 for April and June (Figure 10B). Contributing the lowest number of training images, the click class reported an AUC score of 0.98 for December and 0.97 in March during intense storm periods (Table 5, Figure 10B) demonstrating the capability of the model for high-frequency sound sources in variable soundscape conditions.

Classification of vessels varied seasonally with low AUC scores for each period (Table 5). As a result of increased geophonic noise during winter and storm periods, vessel *P* and *R* scores were low in December (0, 0, respectively) and March (0.48, 0.10, respectively), the poorest preforming class, Table 5. High *R* scores in April and June were reported, 0.99 and 0.87 respectively, *P* scores were lower at 0.36 and 0.44 for April and June (Table 5), determining missed detections. Vessel classification trends in the ROC curves strongly mimicked the ambient noise class under HP weather systems (Figure 10B), with both sound sources affected by lower rates of low-frequency noise in respect to spectrogram visualization. This trend was reversed during storm periods as vessel noise detections reported AUC scores of 0.64 and 0.45 for December and March (Table 5, Figure 10B), while ambient noise classifications had an AUC score of 0.91 and 0.71 (Table 5, Figure 10B).

## Discussion

We provide proof of concept that transfer learning can produce a powerful model for multi-sound source detection in the marine domain across a wide frequency bandwidth and is capable of mining large data sets for information of value to ecosystem assessment and management. As the field of soundscape ecology (Pijanowski et al., 2011) matures, and PAM datasets grow ever larger, we present an adaptable framework for processing acoustic data across the frequency spectra, for the characterization of soundscape components. The work presented here addresses the challenge of detecting

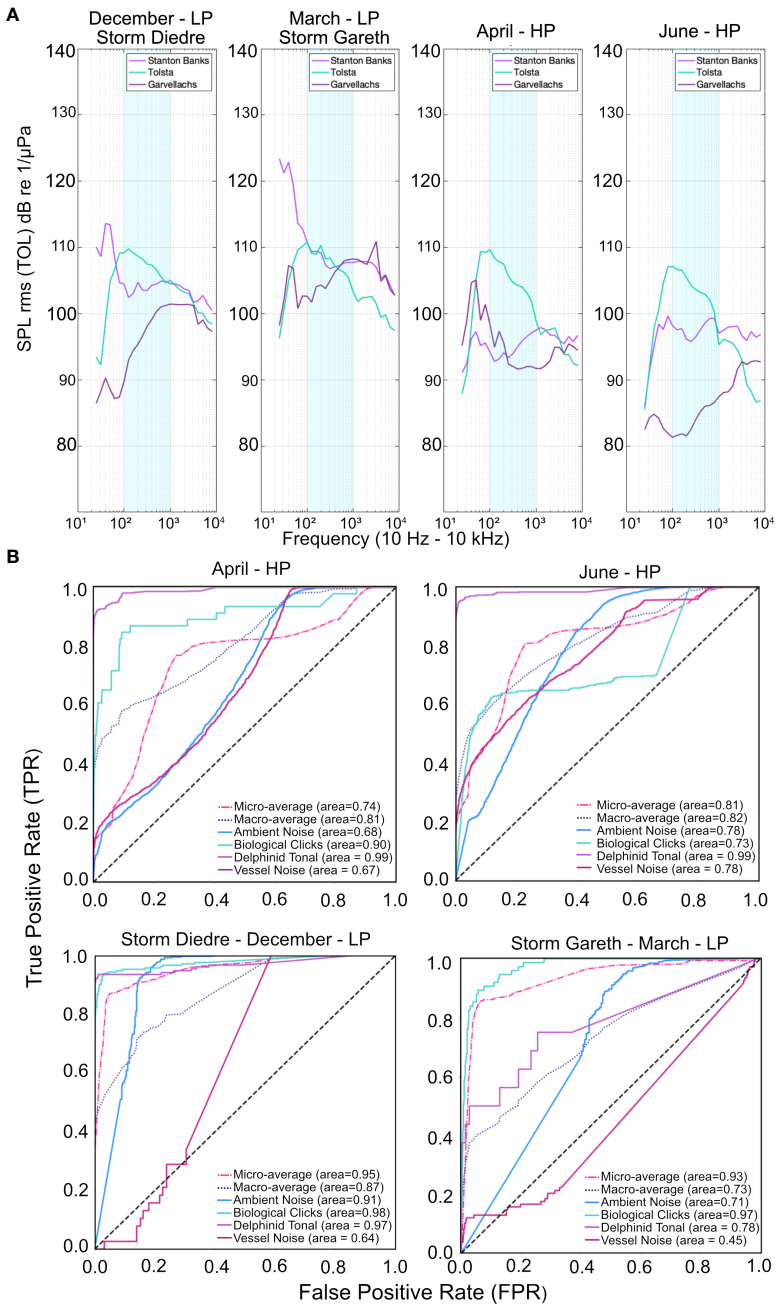


FIGURE 10
Comparison of fidelity of results of the CNN for different weather conditions which produce differing ambient noise conditions per site. **(A)** Median sound pressure levels for third octave bands ranging from 10 Hz to 10 kHz for Stanton Banks, Tolsta and Garvellachs within the four weather periods selected; Low Pressure (LP) systems - December (Storm Diedre) and March (Storm Gareth), and high pressure (HP) systems - April and June, demonstrating variation in soundscape characteristics per season and per site. **(B)** Performance metrics of the CNN are represented as ROC curves. Sound source class detection accuracy fluctuates across seasonalities, with low-frequency geophony sounds affecting the detectability of vessel noise during storm periods (AUC = 0.64, AUC = 0.45).

multiple sound sources in soundscape recordings, and demonstrates high model precision across a range of ambient sound levels present across the varied seasons. The input spans a large frequency bandwidth from 0 to 48 kHz, allowing the model to account for many acoustic signatures within the water column. Although only four categories of sound source are considered here, the large acoustic bandwidth allows for additional noise sources to be accounted for in future

TABLE 5 Performance metrics computed for model evaluation in high pressure (HP) and low pressure (LP) weather conditions, demonstrating seasonal variation.

| Temporal period | Sound source | Precision | Recall | F1 score | Average confusion matrix | Macro-Average ROC curve | AUC |
|---|---|---|---|---|---|---|---|
| 2019 April (HP) | Ambient | 0.87 | 0.20 | 0.32 | 0.55 | 0.81 | 0.68 |
| | Clicks | 0.34 | 0.33 | 0.33 | | | 0.90 |
| | Delphinid Tonal | 0.98 | 0.89 | 0.94 | | | 0.99 |
| | Vessel | 0.36 | 0.99 | 0.53 | | | 0.67 |
| 2019 June (HP) | Ambient | 0.69 | 0.28 | 0.40 | 0.65 | 0.82 | 0.78 |
| | Clicks | 0.44 | 0.33 | 0.38 | | | 0.79 |
| | Delphinid Tonal | 0.95 | 0.94 | 0.94 | | | 0.99 |
| | Vessel | 0.44 | 0.87 | 0.58 | | | 0.78 |
| 2019 December (LP) | Ambient | 0.86 | 0.91 | 0.88 | 0.87 | 0.87 | 0.91 |
| | Clicks | 0.94 | 0.90 | 0.92 | | | 0.98 |
| | Delphinid Tonal | 0.97 | 0.93 | 0.95 | | | 0.97 |
| | Vessel | 0 | 0 | 0 | | | 0.64 |
| 2019 March (LP) | Ambient | 0.90 | 0.94 | 0.92 | 0.85 | 0.73 | 0.71 |
| | Clicks | 0.67 | 0.81 | 0.73 | | | 0.97 |
| | Delphinid Tonal | 0.19 | 0.50 | 0.27 | | | 0.78 |
| | Vessel | 0.48 | 0.10 | 0.17 | | | 0.45 |

iterations of model development such as seismic array guns and Acoustic Deterrent Devices. The limiting factor in the inclusion of these sound sources in this work is the lack of available labelled data for anthropogenic sound sources. Due to their consistent spectral characteristics, it is easier to develop automated labelling systems for these sound sources compared to biological components and further work will incorporate regionally specific anthropogenic signals.

Evaluating the model performance on the widest range of soundscape conditions provides an indication of likely success when deploying the model on annual data archives. The detector outputs high-frequency sound sources with high accuracy (including Odontocete broadband and tonal calls). Site-specific temporal soundscape variation affects per-frame performance for low frequency (e.g. anthropogenic) sound sources. Performance of the detector was expected to be high when the signal strength is high and activity levels output multiple calls per frame, but confidence in our method for multi-source sound detection is high as a result of its performance per-class in poor signal conditions. This work demonstrates the impressive capabilities of open-source 'small' CNN architectures and the opportunities for application to soundscape component classification in marine datasets. We highlight the importance of understanding the soundscape characteristics on which a model will be deployed in order to trust the output, particularly for shallow water environments.

This work presents a proof-of-concept approach to detecting signals which occupy small areas of the image input across the frequency spectra and continuous sound sources (e.g. vessel noise) which vary in their temporal and frequency characteristics over time of their occurrence. The highly successful capabilities of CNNs, in the presence of a small annotated training set are invaluable for extracting valuable ecological information from within large datasets, relevant to marine mammal conservation and ecosystem management. We present a new method for inputting audio to the CNN in the form of a stacked input, using three individually processed spectrograms within the RGB channels of the input image. The sound sources used in this work possess differing temporal and spectral characteristics that make defining one set of spectrogram parameters applicable to all a difficult task. Increasing the volume of input data by using a stacked spectrogram approach enhanced the model performance, as indicated by the lower loss value reported in the results. Another study has used a similar approach for low-frequency baleen whales, interpolating three STFT spectrograms at varying window sizes (Thomas et al., 2019), also reporting an improvement in performance over a one set parameter input, emphasizing the advantage of presenting the model with more information to learn from in the form of a stacked input.

Few previous studies have focused on multiple call types with the incorporation of anthropogenic sound sources. A study from 2020 classifying grouper species (*Epinephelinae)* by their vocalizations presents a similar task and workflow to ours, incorporating the detection of vessel signatures and bioacoustic signals (Ibrahim et al., 2020). In contrast to our work which encompasses a large spectral input for the soundscape as a whole, Ibrahim et al. (2020) focuses on the frequency range 10 – 400 Hz, narrowing the focus of the spectral input to the signals of interest. As the bandwidth of interest

becomes larger the spatial coverage of a signal of interest reduces within the input image presenting a challenge for narrow-band signals. Our work demonstrates the success of transfer learning in the pipeline of detecting multi-class sound sources, extending the spectral domain and detecting signals which are present in many frequency ranges relevant to overall ecosystem assessment.

Existing work using CNNs for the detection and classification of delphinid signals, which classify to species level have reported accuracies of 99.75% (Liu et al., 2018) for echolocation clicks, 93% for both whistles and clicks of a single species (Bergler et al., 2019) and outperform existing general mixed model efforts (Roch et al., 2011a) achieving high accuracies at multi-species click classification (Yang et al., 2020), making use of larger architectures and labelled training sets. Using an open-source 'light-weight' architecture and a small annotated training set we demonstrate similar overall accuracies of our model and in-depth exploration of seasonal variation on model performance to present researchers with an insight into the reliability of CNNs across annual cycles. Existing work on CNNs typically addresses the detection of signals at species level, whereas in this work we looked at family level. Species level detection is invaluable to regional marine management and presents a complex task for a CNN to address. Our results demonstrate the ability of a CNN to extract higher level components of the soundscape for an evaluation of regional systems, beneficial to marine management, policy and stakeholders. It is not possible to directly compare the efficiency of our model to existing species level classifiers as our approach is to analyze the broad components of the soundscapes, arguably an easier task to obtain higher accuracies. We note that broad level taxonomic groups are useful in areas which lack the knowledge required for a species-level classifier. A single CNN model will never be suitable for all bioacoustic research needs, the need for models which incorporate soundscape elements should be used in tandem with complex species-level classifiers to meet the desired research needs. The advantage of using both tonal and impulsive call types of the Delphinid family allows for more confident determination of their temporal presence over single call type approaches and the methodology described here could be adapted to more complex repertoires of other marine species.

Work that incorporates multi-sound sources is scarce at the time of writing; Belgith et al. (2018) demonstrates the capability of CNNs at discriminating between baleen whale calls, odontocete echolocation clicks and anthropogenic noise sources, through the use of custom CNN networks, achieving overall accuracies of 66.4% with a site-specific training set. Our work has a higher accuracy with an overall macro-average of 94% on the test set, exemplifying that with small training sets for the detection of multi-sound sources, a multi-channel spectrogram input combined with transfer learning of a high performing architecture enhances model performance. We acknowledge accuracy score comparison is not straightforward due to differing test metrics and training sets which cannot be compared (Hildebrand et al., 2022). This work reports ROC curves alongside confusion matrices to address per-class performance for an understanding of the effect of regional soundscape variation on performance metrics for specific signal types.

Our work has developed a computationally low-cost approach to mining PAM recordings for data of interest to marine management and species monitoring. There is a demand in the bioacoustics domain for real-time detection, and algorithms which perform detection of acoustic signals of interest on-board marine robotics e.g. gliders and autonomous surface vehicles. We use EfficientNet B0, currently the smallest architecture available for off-the-shelf transfer learning (Tan and Le, 2019), to develop a computationally low-cost CNN model for multi-sound source detection,. Other architectures such as ResNet (He et al., 2016), VGG (Simonyan and Zisserman, 2014) and AlexNet (Krizhevsky et al., 2012) are common choices for transfer learning within marine mammal species detection and classification studies (Bergler et al., 2019; Rasmussen & Širović., 2021; Allen et al., 2021; Lu et al., 2021). These architectures possess more trainable parameters making them computationally more expensive, and studies have shown that larger networks do not always obtain higher accuracies (Bergler et al., 2019). Our evaluation of EfficientNet is not quantifiably comparable to existing work, due to the detection of multiple sound sources rather than a binary classifier, and the variation in spectral input. Our approach overall is computationally efficient, despite the requirement to use three individual spectrograms during input to achieve the best accuracy. Performance metrics per class are high illustrating that low-cost CNN architectures are suitable for bioacoustic tasks, and are appropriate for embedding on board autonomous platforms, enabling a progression in soundscape characterization and species conservation with marine robotics.

The testing protocol implemented is designed to assess the model's ability to generalize to unseen data across temporal and spatial scales within the COMPASS region to provide an understanding of the reliability of model output per deployment. Model performance varied across seasonal extremities, with specific classes outperforming others under differing weather conditions, likely due to the variation in soundscapes which change over time and the introduction of low-frequency geophony above normal levels (Figure 10A). Limited success in detecting impulsive signals (echolocation clicks) within high pressure systems can be attributed to the presence of other sound sources which are visually similar to biological clicks. Through manual review of the test data we attribute the results to the contribution of 'snapping shrimp' to the acoustic environment, which the model was not trained to identify as a separate class. The effect of sound sources not defined during training emphasizes the importance of incorporating year-round data exemplars within the training

data to develop a robust detector, particularly for the ambient noise class. This outcome demonstrates the necessity for training data to include seasonal variations for each site in order for any model to learn to adapt to changing soundscapes. Garvellachs is a geographic site subjected to high rates of snapping shrimp, which is not found at Stanton banks or Tolsta, skewing our model performance. Seasonal variation in 'snapping shrimp' presence, and fluctuating ambient noise, reduces the effect of this in the low-pressure weather periods. Generally, for the click class missed detections are a result of a single high strength click per frame, followed by lower SNR clicks, but not part of a train, or click trains that have been split across frames, a difficult task for any algorithm to identify. Vessel detections are affected by the relative short duration (3-seconds) of the analysis window relative to the length of the source signature, resulting in a lower precision than would be afforded by a longer analysis window. Future work will aim to incorporate temporal context to improve the detection capabilities of vessel noise in fluctuating ambient noise conditions.

Overall, the model performed well across the seasonal soundscapes for each class, in a range of variable SNR conditions, reporting high precision, recall and area-under-curve scores which outperform existing signal-processing methods for multi-class work. This work has demonstrated the potential of multi-channel input to CNNs particularly when looking to detect a range of signal types and could be applicable to the wider bioacoustics' community. Future work will be investigating the advantages of pooling frames as a method of improving evaluation metrics and model sensitivity. Currently the model is evaluated on a per-frame basis which can skew the results when few sound source exemplars are present in a clip. A single 3 s missed whistle frame in a 20-minute file of ambient noise can present a poor confusion matrix as 100% of the whistles are missed. This is an important factor to note when assessing performance between confusion matrix statistics and ROC, which are more robust to unbalanced datasets. The diversity in signal temporal characteristics as noted above for the vessel class can result in poor performance metrics as not each 3-second frame of the vessel passing is detected, however as long as >1 frame is detected the model can accurately detect vessel presence.

## Conclusion

This work demonstrates the capability of developing multi-sound source deep learning models, harnessing the power of open-source CNN architectures, and novel input methods, to develop a low-cost framework for analyzing large marine acoustic datasets for more than single call types with high accuracy. We demonstrate the importance of assessing performance across different SNR conditions and inter-seasonal soundscape variations, with emphasis on the need to understand the region of deployment. Inter-annual soundscape variation plays a dominant role in the success of automated detection systems, trained models must account for site-specific deviation. The outcomes of this study should encourage the development of automated detection and classification models which account for 'more than a whistle'. By incorporating multiple sound sources and moving away from a reliance on running multiple detectors and algorithms in a multi-stage process we can begin to account for anthropogenic and geophonic components of the soundscape. Expansion on the type of information we seek to extract from acoustic data is imperative to robust ecosystem-level marine management and will aid our efforts to monitor regional health in the age of the Ocean Decade.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors upon request.

## Author contributions

EW, PW and JB conceptualized the study. EW organized the dataset for model training, preformed data processing and trained the model. At each stage of work PW and JB contributed to method development and iteration. EW carried out evaluation of the model. EW wrote the first manuscript draft, with PW and JB contributing to all sections of the manuscript early drafts. DR, SB and EE conceptualized the COMPASS project and carried out all data collection, survey design and data organization relating to this work. All authors contributed to the article and approved the submitted version.

## Funding

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).

## Acknowledgments

We thank H.Gower, M.Poile, S.Ponnet, V.Skripek, A.Szwarczynska, I.Thompson, L.Watton and J.Wells for their

annotation assistance, the Hebridean Whale and Dolphin Trust for providing acoustic data, and those responsible for providing open-source annotated marine mammal datasets in the form of DCLDE Oregon and Honolulu. This manuscript has been greatly improved by the constructive comments from the editor Ana Širović, and three anonymous reviewers.

## Conflict of interest

The reviewer PG declared a past collaboration with the author PW to the handling editor.

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note



## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmars.2022.879145/full#supplementary-material

## References

Allen, A. N., Harvey, M., Harrell, L., Jansen, A., Merkens, K. P., Wall, C. C., et al. (2021). A convolutional neural network for automated detection of humpback whale song in a diverse, long-term passive acoustic dataset. *Front. Mar. Sci.* 8, 165. doi: 10.3389/fmars.2021.607321

Au, W. W., and Banks, K. (1998). The acoustics of the snapping shrimp *Synalpheus parneomeris* in kaneohe bay. *J. Acoust. Soc. Am.* 103 (1), 41–47. doi; 10.1121/1.423234

Baumgartner, M. F., Bonnell, J., Corkeron, P. J., Van Parijs, S. M., Hotchkin, C., Hodges, B. A., et al. (2020). Slocum Gliders provide accurate near real-time estimates of baleen whale presence from human-reviewed passive acoustic detection information. *Front. Mar. Sci.* 7, 100. doi: 10.3389/fmars.2020.00100

Baumgartner, M. F., Fratantoni, D. M., Hurst, T. P., Brown, M. W., Cole, T. V., Van Parijs, S. M., et al. (2013). Real-time reporting of baleen whale passive acoustic detections from ocean gliders. *J. Acoust. Soc. Am.* 134 (3), 1814–1823. doi: 10.1121/1.4816406

Baumgartner, M. F., and Mussoline, S. E. (2011). A generalized baleen whale call detection and classification system. *J. Acoust. Soc. Am.* 129 (5), 2889–2902. doi: 10.1121/1.3562166

Belgith, E. H., Rioult, F., and Bouzidi, M. (2018). "Acoustic diversity classifier for automated marine big data analysis," in *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*. (Volos, Greece: IEEE), 130–136.

Bergler, C., Schröter, H., Cheng, R. X., Barth, V., Weber, M., Nöth, E., et al. (2019). ORCA-SPOT: An automatic killer whale sound detection toolkit using deep learning. *Sci. Rep.* 9 (1), 1–17. doi: 10.1038/s41598-019-47335-w

Bermant, P. C., Bronstein, M. M., Wood, R. J., Gero, S., and Gruber, D. F. (2019). Deep machine learning techniques for the detection and classification of sperm whale bioacoustics. *Sci. Rep.* 9 (1), 1–10. doi: 10.1038/s41598-019-48909-4

Bittle, M., and Duncan, A. (2013). "A review of current marine mammal detection and classification algorithms for use in automated passive acoustic monitoring" in *Proceedings of Acoustics, Victor Harbor, Australia* Proc. Acoust, 1–8.

Brown, J. C., and Smaragdis, P. (2009). Hidden Markov and Gaussian mixture models for automatic call classification. *J. Acoust. Soc. Am.* 125 (6), EL221–EL224. doi: 10.1121/1.3124659

Brown, J. C., Smaragdis, P., and Nousek-McGregor, A. (2010). Automatic identification of individual killer whales. *J. Acoust. Soc. Am.* 128 (3), EL93–EL98. doi: 10.1121/1.3462232

Chollet, F. (2018). Keras: The python deep learning library. *Astrophys. Source Code Libr.*, 1806.

Cox, S., Embling, C. B., Hosegood, P. J., Votier, S. C., and Ingram, S. N. (2018). Oceanographic drivers of marine mammal and seabird habitat-use across shelf-seas: a guide to key features and recommendations for future research and conservation management. *Estuar. Coast. Shelf Sci.* 212, 294–310. doi: 10.1016/j.ecss.2018.06.022

Davis, G. E., Baumgartner, M. F., Corkeron, P. J., Bell, J., Berchok, C., Bonnell, J. M., et al. (2020). Exploring movement patterns and changing distributions of baleen whales in the western north Atlantic using a decade of passive acoustic data. *Global Change Biol.* 26 (9), 4812–4840. doi: 10.1111/gcb.15191

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., and Fei-Fei, L. (2009). "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. (Miami FL: IEEE)248–255.

Duarte, C. M., Chapuis, L., Collin, S. P., Costa, D. P., Devassy, R. P., Eguiluz, V. M., et al. (2021). The soundscape of the anthropocene ocean. *Science* 371 (6529), eaba4658. doi: 10.1126/science.aba4658

Dudzinski, K. M., Thomas, J. A., and Gregg, J. D. (2009). "Communication in marine mammals," in *Encyclopedia of marine mammals* (Academic Press), 260–269.

Dunlop, R. A. (2016). The effect of vessel noise on humpback whale, megaptera novaeangliae, communication behaviour. *Anim. Behav.* 111, 13–21. doi: 10.1016/j.anbehav.2015.10.002

Erbe, C., Marley, S. A., Schoeman, R. P., Smith, J. N., Trigg, L. E., and Embling, C. B. (2019). The effects of ship noise on marine mammals–a review. *Front. Mar. Sci.* 606 (6), 606. doi: 10.3389/fmars.2019.00606

Evans, P. G. H., and Waggitt, J. J. (2020). Impacts of climate change on marine mammals, relevant to the coastal and marine environment around the UK. *MCCIP Sci. Rev.* 2020, 421–455. doi: 10.14465/2020.arc19.mmm

Gibb, R., Browning, E., Glover-Kapfer, P., and Jones, K. E. (2019). Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods Ecol. Evol.* 10 (2), 169–185. doi: 10.1111/2041-210X.13101

Gillespie, D. (1997). An acoustic survey for sperm whales in the southern ocean sanctuary conducted from the RSV aurora Australis. *Rep. Int. Whaling Comm.* 47, 897–907.

Gruden, P., and White, P. R. (2016). Automated tracking of dolphin whistles using Gaussian mixture probability hypothesis density filters. *J. Acoust. Soc. Am.* 140 (3), 1981–1991. doi: 10.1121/1.4962980

Hague, E. L., Sinclair, R. R., and Sparling, C. E. (2020). Regional baselines for marine mammal knowledge across the north Sea and Atlantic areas of Scottish waters. *Scottish Mar. Freshw. Sci.* 11 (12), 305. doi: 10.7489/12330-1

Harvey, M. (2018). Acoustic detection of humpback whales using a convolutional neural network. *Google AI Blog*. Available at: https://ai.googleblog.com/2018/10/acoustic-detection-of-humpback-whales.html.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (Las Vegas: IEEE), 770–778.

Hildebrand, J. A., Frasier, K. E., Helble, T. A., and Roch, M. A. (2022). Performance metrics for marine mammal signal detection and classification. *J. Acoust. Soc. Am.* 151 (1), 414–427. doi: 10.1121/10.0009270

Howe, B. M., Miksis-Olds, J., Rehm, E., Sagen, H., Worcester, P. F., and Haralabus, G. (2019). Observing the oceans acoustically. *Front. Mar. Sci.* 6, 426. doi: 10.3389/fmars.2019.00426

Ibrahim, A. K., Zhuang, H., Cherubin, L. M., Schärer-Umpierre, M. T., Nemeth, R. S., Erdol, N., et al. (2020). Transfer learning for efficient classification of grouper sound. *J. Acoust. Soc. Am.* 148 (3), 260–266. doi: 10.1121/10.0001943

Jarvis, S., DiMarzio, N., Morrissey, R., and Moretti, D. (2008). A novel multi-class support vector machine classifier for automated classification of beaked whales and other small odontocetes. *Can. Acoust.* 36 (1), 34–40.

Jiang, J. J., Bu, L. R., Duan, F. J., Wang, X. Q., Liu, W., Sun, Z. B., et al. (2019). Whistle detection and classification for whales based on convolutional neural networks. *Appl. Acoust.* 150, 169–178. doi: 10.1016/j.apacoust.2019.02.007

Jones, P. J. (2012). Marine protected areas in the UK: challenges in combining top-down and bottom-up approaches to governance. *Environ. Conserv.* 39 (3), 248–258. doi: 10.1017/S0376892912000136

Kahl, S., Clapp, M., Hopping, W., Goëau, H., Glotin, H., Planqué, R., et al. (2020). "Overview of birdclef 2020: Bird sound recognition in complex acoustic environments," in *CLEF 2020-11th International Conference of the Cross-Language Evaluation Forum for European Languages* (Thessaloniki, Greece).

Kaiser, J. F. (1990). "On a simple algorithm to calculate the 'energy' of a signal," in *International conference on acoustics, speech, and signal processing IEEE* (Albuquerque NM: IEEE), 381–384.

Kandia, V., and Stylianou, Y. (2006). Detection of sperm whale clicks based on the teager–kaiser energy operator. *Appl. Acoust.* 67 (11-12), 1144–1163. doi: 10.1016/j.apacoust.2006.05.007

Khan, A., Sohail, A., Zahoora, U., and Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artif. Intell. Rev.* 53, 1–62. doi: 10.1007/s10462-020-09825-6

Kim, K. H., Hursky, P., Porter, M. B., Hildebrand, J. A., and Elizabeth, E. (2006). "Automated passive acoustic tracking of dolphins in free-ranging pods," in *Proceedings of the Eighth European Conference on Underwater Acoustics, 8th ECUA* (Carvoeiro, Portugal).

Kingma, D. P., and Ba, J. (2014). *Adam: A method for stochastic optimization* (arXiv) preprint. arxiv:1412.6980

Koidl, K. (2013). *Loss functions in classification tasks* (Dublin: School of Computer Science and Statistic Trinity College).

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* 25, 1097–105. doi: 10.1145/3065386

Kunc, H. P., McLaughlin, K. E., and Schmidt, R. (2016). Aquatic noise pollution: implications for individuals, populations, and ecosystems. *Proc. R. Soc. B: Biol. Sci.* 283 (1836), 08–39. doi: 10.1098/rspb.2016.0839

Kuperman, W. A., and Lynch, J. F. (2004). Shallow-water acoustics. *Phys. Today* 57 (10), 55–61. doi: 10.1063/1.1825269

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521 (7553), 436–444. doi: 10.1038/nature14539

Liu, J., Yang, X., Wang, C., and Tao, Y. (2018). "A convolution neural network for dolphin species identification using echolocation clicks signal," in *2018 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*. (Qingdao, China: IEEE)1–4.

Lu, T., Han, B., and Yu, F. (2021). Detection and classification of marine mammal sounds using AlexNet with transfer learning. *Ecol. Inf.* 62, 101277. doi: 10.1016/j.ecoinf.2021.101277

Mae, T., Collison, N., Theriault, J., Hood, J., Pecknold, S., and Bougher, B. (2010). Detection of precise time events for marine mammal clicks. *Can. Acoust.* 38 (3), 30–31.

Marley, S. A., Erbe, C., Salgado Kent, C. P., Parsons, M. J., and Parnum, I. M. (2017). Spatial and temporal variation in the acoustic habitat of bottlenose dolphins (*Tursiops truncatus*) within a highly urbanized estuary. *Front. Mar. Sci.* 4, 197. doi: 10.3389/fmars.2017.00197

McKenna, M. F., Baumann-Pickering, S., Kok, A., Oestreich, W. K., Adams, J. D., Barkowski, J., et al. (2021). Advancing the interpretation of shallow water marine soundscapes. *Front. Mar. Sci.* 9. doi: 10.3389/fmars.2021.719258

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

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective* (Cambridge, Massachusetts: MIT press).

Nazari, Z., Nazari, M., Sayed, M., and Danish, S. (2018). Evaluation of class noise impact on performance of machine learning algorithms. *IJCSNS Int. J. Comput. Sci. Netw. Secur.* 18, 149. doi: 201808/20180823

Oswald, J. N., Rankin, S., Barlow, J., Oswald, M., and Lammers, M. O. (2003). Realtime call classification algorithm (ROCCA): software for species identification of 26 delphinid whistles. *Detect. classif. localization Mar. mamm. using passive acoust.* 120 (1), 587–95. doi: 10.1121/1.2743157

Pace, F., White, P., and Adam, O. (2012). "Hidden Markov modeling for humpback whale (*Megaptera novaeanglie*) call classification," in *Proceedings of Meetings on Acoustics ECUA2012*. Edinburgh, Scotland, 2 - 6 July. The Journal of the Acoustical Society of America, 1-8. doi: 10.1121/1.4772751

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. doi: 10.48550/arXiv.1201.0490

Pijanowski, B. C., Villanueva-Rivera, L. J., Dumyahn, S. L., Farina, A., Krause, B. L., Napoletano, B. M., et al. (2011). Soundscape ecology: the science of sound in the landscape. *BioScience* 61 (3), 203–216. doi: 10.1525/bio.2011.61.3.6

Pirotta, E., Merchant, N., Thompson, P., Barton, T., and Lusseau, D. (2015). Quantifying the effect of boat disturbance on bottlenose dolphin foraging activity. *Biol. Conserv.* 181, 82–89. doi: 10.1016/j.biocon.2014.11.003

Pittman, S. J. (2017). *Seascape ecology* (Oxford, England: John Wiley & Sons).

Pompa, S., Ehrlich, P. R., and Ceballos, G. (2011). Global distribution and conservation of marine mammals. *Proc. Natl. Acad. Sci.* 108 (33), pp.13600–13605. doi: 10.1073/pnas.1101525108

Rasmussen, J. H., and Širović, A. (2021). Automatic detection and classification of baleen whale social calls using convolutional neural networks. *J. Acoust. Soc. Am.* 149 (5), 3635–3644. doi: 10.1121/10.0005047

Roch, M. A., Klinck, H., Baumann-Pickering, S., Mellinger, D. K., Qui, S., Soldevilla, M. S., et al. (2011a). Classification of echolocation clicks from odontocetes in the southern California bight. *J. Acoust. Soc. Am.* 129 (1), 467–475. doi: 10.1121/1.3514383

Roch, M. A., Miller, P., Helble, T. A., Baumann-Pickering, S., and Širović, A. (2017). Organizing metadata from passive acoustic localizations of marine animals. *J. Acoust. Soc. Am.* 141 (5), 3605–3605. doi: 10.1121/1.4987714

Roch, M. A., Scott Brandes, T., Patel, B., Barkley, Y., Baumann-Pickering, S., and Soldevilla, M. S. (2011b). Automated extraction of odontocete whistle contours. *J. Acoust. Soc. Am.* 130 (4), 2212–2223. doi: 10.1121/1.3624821

Roch, M. A., Soldevilla, M. S., Burtenshaw, J. C., Henderson, E. E., and Hildebrand, J. A. (2007). Gaussian Mixture model classification of odontocetes in the southern California bight and the gulf of California. *J. Acoust. Soc. Am.* 121 (3), 1737–1748. doi: 10.1121/1.2400663

Roch, M., Soldevilla, M., and Hildebrand, J. (2004). Automatic species identification of odontocete calls in the southern California bight. *J. Acoust. Soc. Am.* 116 (4), 2614–2614. doi: 10.1121/1.4785425

Roch, M. A., Soldevilla, M. S., Hoenigman, R., Wiggins, S. M., and Hildebrand, J. A. (2008). Comparison of machine learning techniques for the classification of echolocation clicks from three species of odontocetes. *Can. Acoust.* 36 (1), 41–47.

Ruff, Z. J., Lesmeister, D. B., Appel, C. L., and Sullivan, C. M. (2021). Workflow and convolutional neural network for automated identification of animal sounds. *Ecol. Indic.* 124, 107–419. doi: 10.1016/j.ecolind.2021.107419

Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., et al. 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35 (5), 1285–98. doi: 10.1109/TMI.2016.2528162

Shiu, Y., Palmer, K. J., Roch, M. A., Fleishman, E., Liu, X., Nosal, E. M., et al. (2020). Deep neural networks for automated detection of marine mammal species. *Sci. Rep.* 10 (1), 1–12. doi: 10.1038/s41598-020-57549-y

Simonyan, K., and Zisserman, A. (2014). "Very deep convolutional networks for large-scale image recognition," (arXiv preprint). doi: 10.48550 /arXiv:1409.1556

Širović, A., Johnson, S. C., Roche, L. K., Varga, L. M., Wiggins, S. M., and Hildebrand, J. A. (2015). North pacific right whales (Eubalaena japonica) recorded in the north-Eastern pacific ocean in 2013. *Mar. Mamm. Sci.* 31 (2), 800–807. doi: 10.1111/mms.12189

Smith, L. N. (2017). "Cyclical learning rates for training neural networks," in *2017 IEEE winter conference on applications of computer vision (WACV)*. (Santa Rosa, CA: IEEE)464–472.

Solandt,J.L. (2018). A stocktake of England's MPA network–taking a global perspective approach. *Biodiversity* 19 (1-2), 34–41. doi: 10.1089/14888386.2018.1464950

Sousa-Lima, R. S., Norris, T. F., Oswald, J. N., and Fernandes, D. P. (2013). A review and inventory of fixed autonomous recorders for passive acoustic monitoring of marine mammals. *Aquat. Mamm.* 39 (1), 23–53. doi: 10.1578/AM.39.1.2013.23

Stafford, K.M., Castellote, M., Guerra, M., and Berchok, C.L. (2018). Seasonal acoustic environments of beluga and bowhead whale core-use regions in the Pacific Arctic. *Deep Sea Research Part II: Topical Studies in Oceanography*, 152, 108–120. doi: 10.1016/j.dsr2.2017.08.003

Steiner, W. W. (1981). Species-specific differences in pure tonal whistle vocalizations of five western north Atlantic dolphin species. *Behav. Ecol. Sociobiol.* 9 (4), 241–246. doi: 10.1007/BF00299878

Stowell, D. (2022). Computational bioacoustics with deep learning: a review and roadmap *PeerJ* 10, e13152. doi: 10.7717/peerj.1315.

Sugai, L., Silva, T., Ribeiro, J., and Llusia, D. (2019). Terrestrial passive acoustic monitoring: Review and perspectives. *BioScience* 69 (1), 15–25. doi: 10.1093/biosci/biy147

Tan, M., and Lee, Q. (2019). "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. (Long Beach, CA: PMLR)6105–6114.

Thomas, M., Martin, B., Kowarski, K., Gaudet, B., and Matwin, S. (2020). Marine mammal species classification using convolutional neural networks and a novel acoustic representation. In: U. Brefeld, E. Fromont, A. Hotho, A. Knobbe, M. Maathuis and C. Robardet(eds) Machine learning and knowledge discovery in databases. ECML PKDD 2019. Lecture notes in computer science, vol 11908. (Cham: Springer). doi: 10.1007/978-3-030-46133-1_18

Thomas, M., Martin, B., Kowarski, K., Gaudet, B., and Matwin, S. (2019). Marine mammal species classification using convolutional neural networks and a novel acoustic representation. *Joint Eur. Conf. Mach. Learn. knowl. Discovery Database*, 290–305. doi: 10.1007/978-3-030-46133-1_18

Vester, H., Hallerberg, S., Timme, M., and Hammerschmidt, K. (2017). Vocal repertoire of long-finned pilot whales (*Globicephala melas*) in northern Norway. *J. Acoust. Soc. Am.* 141 (6), 4289–4299. doi: 10.1121/1.4983685

Wang, Z. A., Moustahfid, H. A., Mueller, A., Mowlem, M. C., Michel, A. P. M., Glazer, B. T., et al. (2019). Advancing observation of ocean biogeochemistry, biology, and ecosystems with cost-effective *in situ* sensing technologies. *Front. Mar. Sci.* 6, 519. doi: 10.3389/fmars.2019.00519

Xie, J., Hu, K., Zhu, M., and Guo, Y. (2020). Bio-acoustic signal classification in continuous recordings: Syllable-segmentation vs sliding-window. *Expert Syst. Appl.* 152, 113390. doi: 10.1016/j.eswa.2020.113390

Yang, W., Luo, W., and Zhang, Y. (2020). Classification of odontocete echolocation clicks using convolutional neural network. *J. Acoust. Soc. Am.* 147 (1), 49–55. doi: 10.1121/10.0000514

Yano, K. M., Oleson, E. M., Keating, J. L., Ballance, L. T., Hill, M. C., Bradford, A. L., et al. (2018). Cetacean and seabird data collected during the Hawaiian islands cetacean and ecosystem assessment survey (HICEAS) 72. doi: 10.25923/7-avn-gw82

Zhong, M., Castellote, M., Dodhia, R., Lavista Ferres, J., Keogh, M., and Brewer, A. (2020). Beluga whale acoustic signal classification using deep learning neural network models. *J. Acoust. Soc. Am.* 147 (3), 1834–1841. doi: 10.1121/10.0000921