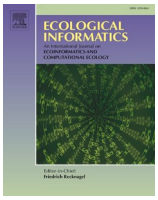
Ecological Informatics 78 (2023) 102363

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



One size fits all? Adaptation of trained CNNs to new marine acoustic environments

Ellen L. White [a,*], Holger Klinck [b], Jonathan M. Bull [a], Paul R. White [c], Denise Risch [d]

[a] School of Ocean and Earth Science, University of Southampton, UK
[b] K. Lisa Yang Center for Conservation Bioacoustics, Cornell Lab of Ornithology, Cornell University, NY, USA
[c] Institute of Sound and Vibration, University of Southampton, UK
[d] Marine Science Department, Scottish Association of Marine Science, Oban, UK

ARTICLE INFO

Keywords:
Bioacoustics
Deep learning
Domain adaptation
Marine acoustics
Marine mammal detection
Soundscapes

ABSTRACT

Convolutional neural networks (CNNs) have the potential to enable a revolution in bioacoustics, allowing robust detection and classification of marine sound sources. As global Passive Acoustic Monitoring (PAM) datasets continue to expand it is critical we improve our confidence in the performance of models across different marine environments, if we are to exploit the full ecological value of information within the data. This work demonstrates the transferability of developed CNN models to new acoustic environments by using a pre-trained model developed for one location (West of Scotland, UK) and deploying it in a distinctly different soundscape (Gulf of Mexico, USA). In this work transfer learning is used to fine-tune an existing open-source 'small-scale' CNN, which detects odontocete tonal and broadband call types and vessel noise (operating between 0 and 48 kHz). The CNN is fine-tuned on training sets of differing sizes, from the unseen site, to understand the adaptability of a network to new marine acoustic environments. Fine-tuning with a small sample of site-specific data significantly improves the performance of the CNN in the new environment, across all classes. We demonstrate an improved performance in area-under-curve (AUC) score of 0.30, across four classes by fine-training with only 50 spectrograms per class, with a 5% improvement in accuracy between 50 frames and 500 frames. This work shows that only a small amount of site-specific data is needed to retrain a CNN, enabling researchers to harness the power of existing pre-trained models for their own datasets. The marine bioacoustic domain will benefit from a larger pool of global data for training large deep learning models, but we illustrate in this work that domain adaptation can be improved with limited site-specific exemplars.

1. Introduction

Climatic and human pressures are leading to shifts in the size, structure, spatial range and seasonal abundance of marine populations, knowledge of which is essential for effective wildlife conservation. Sound provides a mechanism for communication and a source of information which is exploited by a wide range of marine taxa (Haver et al., 2017). Passive acoustic monitoring (PAM), a technique which allows us to eavesdrop on the marine environment, is being increasingly used to continuously monitor temporal and spatial variation in the characteristics of regional soundscapes. Ocean soundscapes are the characterisation of ambient noise in terms of spatial, temporal and frequency attributes, and the types of sources contributing to the sound field, with the aggregation of geophysical, biological and anthropogenic sounds present (Pijanowski et al., 2011; IOS, 2014).

Advances in data acquisition, storage, battery power, and processing techniques (Howe et al., 2019); has resulted in a widespread adoption of PAM globally, recording the oceans soundscape over larger temporal and spatial scales (Wall et al., 2021; Wang et al., 2019). As the volume of acoustic data recorded globally increases, the time required to extract ecologically important information from the soundscape grows (Sugai et al., 2018). The information derived from PAM data often gets delivered to the research community and stakeholders long after the sensor has been recovered.

Machine learning (ML) solutions for analysing acoustic signals are effective tools for long-term ecological monitoring over timescales appropriate for marine management. ML can be used to classify sounds within a recording according to the source that generate them. In this

* Corresponding author.
E-mail address: elw1g13@soton.ac.uk (E.L. White).

https://doi.org/10.1016/j.ecoinf.2023.102363
Received 23 August 2023; Received in revised form 30 October 2023; Accepted 1 November 2023
Available online 7 November 2023

way ML allows the automation of tasks previously considered as requiring manual processing (Stowell, 2022). For instance, marine mammal calls have been classified with a wealth of ML algorithms including support vector machines (Jarvis et al., 2008; Roch et al., 2008), generalised linear models, hidden Markov models (Brown et al., 2010; Pace et al., 2012; Roch et al., 2011) and classification and regression tree analysis (Oswald et al., 2003). These advances have led to a rise in the number of published trained CNN models available for researchers to download and use as tools. The performance of these CNNs has been found to rival human performance at signal recognition (LeCun et al., 2015). CNNs can learn to discriminate spectro-temporal information directly from a labelled spectrogram, used as an image input. The success of CNNs within the marine bioacoustic field has been demonstrated by studies for binary 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; White et al., 2022).

Although CNNs can provide good performance in adverse conditions, the underlying assumption is that the training and testing datasets used to develop the model are extracted from the same distributions. Whilst a model may perform well on the test data, it may not hold up in real-world applications where data has a different underlying distribution (Farahani et al., 2021). Acoustically active species reside in all ocean bodies, ranging from shallow coastal habitats to offshore deep waters, with acoustic repertoires varying geographically (Tyack and Miller, 2002). Recorded soundscapes vary significantly between sites, they are dependent upon site-specific bathymetry, sediment type, mooring depth, hydrophone type, and local anthropogenic activity, as well as natural physical processes such as currents, tides, surge, storms and winds.

There is a lack of labelled data within the marine bioacoustics' domain, available training sets for CNN models tend to be biased towards specific geographic locations. Data can span multiple years, moorings, depths and method of collection and the characteristics of this recording system becomes learned features within the model. Post model development, the question of 'domain transferability' remains – how well does a model perform in a novel environment with differing ambient conditions and new acoustic sources present? The ability to transfer knowledge learned by a CNN to a new marine domain is critical for researchers to harness the power of existing models.

Within the field of marine bioacoustics few studies have considered the performance of broadband CNNs in environments outside those in which they were trained (Shiu et al., 2020; Padovese et al., 2023). Best et al. (2020) demonstrate the difficulty of improving model performance on new data collected within the same geographic region as the original training data for orca vocalisation detection, determining the recording system itself considerably disturbs the previously successful model. The terrestrial domain currently has several 'global' acoustic species classification systems e.g. BirdNet (Kahl et al., 2021) where the training data is extracted from a global corpus of acoustic recordings, from which training samples vary in terms of hardware, species present, geography and seasonality. The diversity within these large-scale training sets allows for the feature representations to be more generalisable to new acoustic environments. Lauha et al. (2022) describe how large global models can benefit from fine-tuning (re-training) on local data, collected in the same acoustic environment the model is to be used in. Fine-tuning allows the model to generalise to the specific habitat it is to be used in. Learning the feature representation of the local soundscape can increase the confidence we have in model output.

This work explores the transferability of CNNs to unseen ocean soundscapes. Domain adaptation of CNN models refers to the process of training a deep learning model on a particular environment (source domain) and then transferring the knowledge learned to another environment (target domain) to improve the generalization and accuracy of the model in the target domain (Csurka, 2017). We take a base model developed using data from the West coast of Scotland (White et al., 2022) and deploy it on PAM data collected in the Gulf of Mexico, USA (Fig. 1), considering its performance in the new environment with limited re-training. Starting with a pre-trained model (EfficientNet B0, Tan and Le, 2019), we freeze the networks original feature extractor. We then train fully connected layers, which operate on the output of the EfficientNet B0 network, to classify the PAM data frames into the specified sound source classes. Audio data is input to the network in a novel spectrogram representation (White et al., 2022). Networks pre-trained on image data have been proven to perform well on classification tasks using spectrograms as image input, transformed from raw audio data (Allen et al., 2021; Shiu et al., 2020; Stowell, 2022). In this work we fine-tune the parameters of the fully connected layers iteratively on small batches of training data acquired from a new region. Through an experimental approach to retraining we demonstrate the performance capabilities of a domain-specific CNN when trained and applied to new marine environments. We demonstrate the ability to utilise existing bioacoustic models in new and unseen marine environments.

2. Methods

This section provides a detailed summary of (i) Experimental design, (ii) Data acquisition and (iii) Model Evaluation.

2.1. Base model

The base model used in this work was originally developed to aid processing of a large PAM dataset (COMPASS project) from the west coast of Scotland (White et al., 2022). Input frames to the model comprise spectrogram representations of 3 s clips of acoustic data, sampled at 96 kHz and are classified them into one of four broad classes: Ambient Noise (AN), Biological Clicks (BC), Delphinid Tonal (DT) and Vessel Noise (VN). The spectrogram amplitudes are displayed on a dB scale and employ a linear frequency axis.

The base model makes use of a 'stacked' spectrogram input, developed to take advantage of the amount of information available to the model for a single input, per channel (White et al., 2022). Each channel of the input RGB image corresponds to a single spectrogram computed at one of 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) standardized for the sampling rate 96 kHz. The spectrogram values are standardized to correspond to the range − 80 to 0 dBfs.

A diverse training set, utilising data collected by a variety of organisations, under differing survey protocols and across a range of geographic locations and temporal scales was used; see White et al. (2022) for details. The training and test sets for this work made use of acoustic data collected at three moorings: Stanton Banks (56.097°N, −8.022°W), Tolsta (58.394°N -6.012°W) and Garvellachs (56.235°N -5.756°W), collected between 2017 and 2019 (Fig. 1). The physical conditions at each site are quite different, contributing to distinct soundscapes (Fig. 2).

To enhance the robustness of the base model to seasonal variation within the local soundscape, evaluation data used during model development has been used to re-train the model on a larger pool of local data. The final training set is made up of 46,749 training frames and 4673 validation frames (Table 1). Classes are imbalanced, reflecting their presence within the local soundscape during manual annotation.

The model is constructed using the EfficientNet B0 network (Tan and Le, 2019) which had been trained for generic image classification. The EfficientNet feature extraction layers are frozen (transfer learning) with only the weights of the final dense classification layers updated during training. Training was conducted within the Google Collaboratory 'Colab' platform (Bisong, 2019), using the Tesla K80 GPU, accessed through cloud computing. An Adam optimizer was used to control gradient descent during training (Kingma and Ba, 2014), with

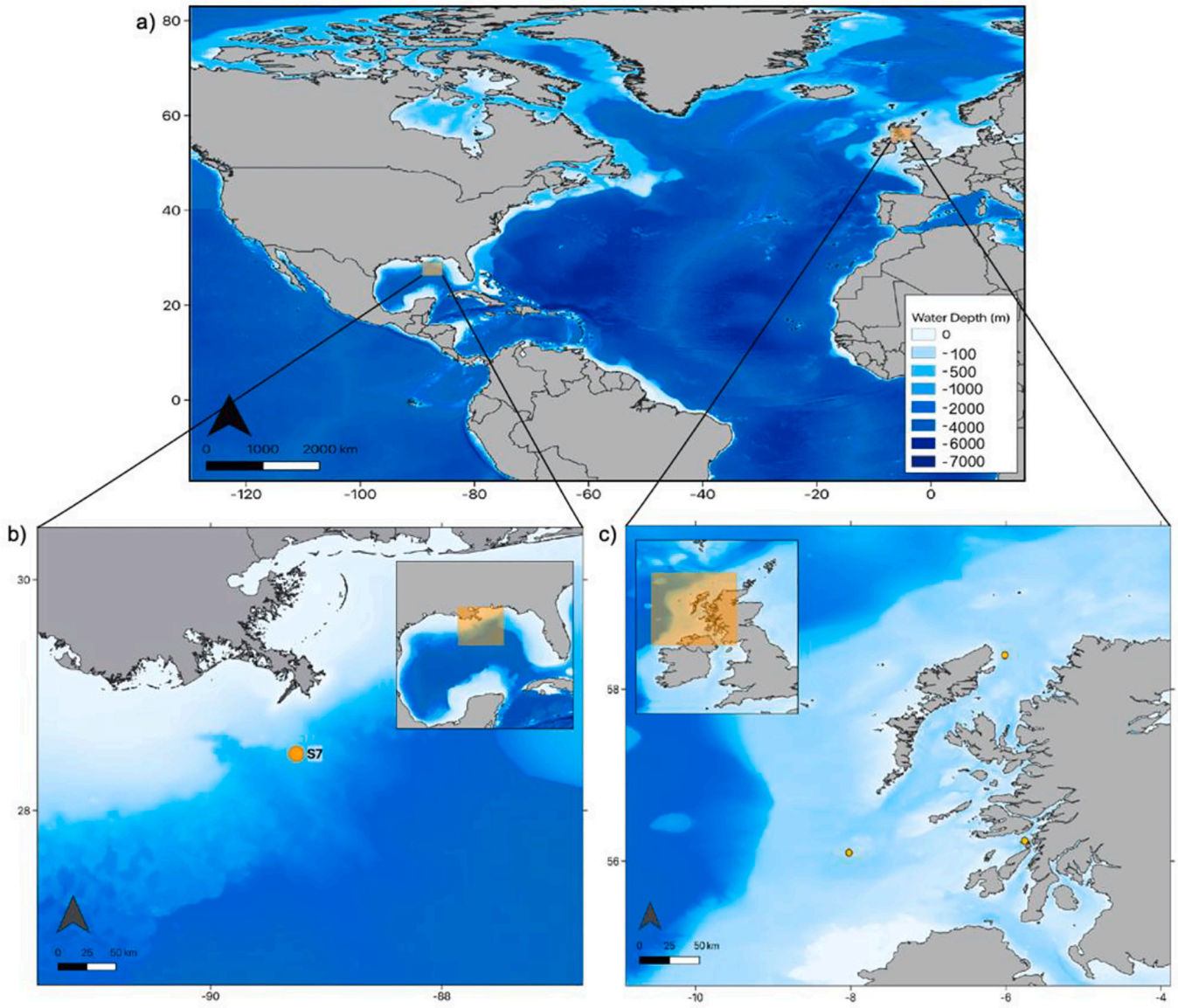


Fig. 1. a) Map of the locations used to collect PAM data used in this study. b) The location of the hydrophone mooring, within the Gulf of Mexico, USA, depth of 440 m, used in this work. The inset map depicts the location of the mooring within the larger setting of the Gulf Mexico, adjacent to the Mississippi Canyon. c) The location of the PAM moorings in the West of Scotland used to collect training data for the base model, depths ranging 25 m – 150 m. The inset map displays the moorings location within British and Irish waters. Bathymetry data was sourced from GEBCO.

parameters set to: learning rate of 0.001, decay factor of 0.75 and a step size of 8. A dropout rate of 0.5 is used to regularise this network. Categorical cross-entropy was employed as the loss function (Koidl, 2013). The base model trained for 37 epochs, with a training loss value of 0.32 (validation loss 0.22) and categorical accuracy 0.84 (validation categorical accuracy 0.95).

2.1.1. Experimental design

To evaluate the effect of local training data on model performance we fine-tune a base model with randomly pooled training sets of increasing size (sets of 50, 100, 200, 300 and 500- frames) from the Gulf of Mexico acoustic data collected December 3rd 2019 (Fig. 3). An extra training set is considered consisting of 500 frames of randomly sampled ambient noise from December 3rd 2019.

The models are validated during training by combining the original base model validation set (Table 1) with 100 frames per class from the labelled December 3rd 2019 training set, randomly pooled per experiment. In this way the model's performance is assessed across both environments simultaneously(Fig. 4).

2.2. Model fine-tuning

To fine-tune the base model with Gulf of Mexico data the feature extractor remains frozen. Models are trained with the same parameters as the base model but we use a cyclical learning rate of 0.0004 and is set to run for 50 epochs, with early stopping set to deploy if the validation loss does not improve within 10 epochs. A dropout rate of 0.2 is used during fine-tuning, and DropConnect is employed. Drop out layers randomly discard the output of the hidden nodes during training, DropConnect randomly discards the input of the hidden layer (Sun et al., 2022).

2.2.1. Data acquisition

The data used to carry out fine-tuning experiments in this work is collected from the northern Gulf of Mexico, at the mooring S7 (28.92 N, −89.26 W, Fig. 1). These data were collected between November 2019

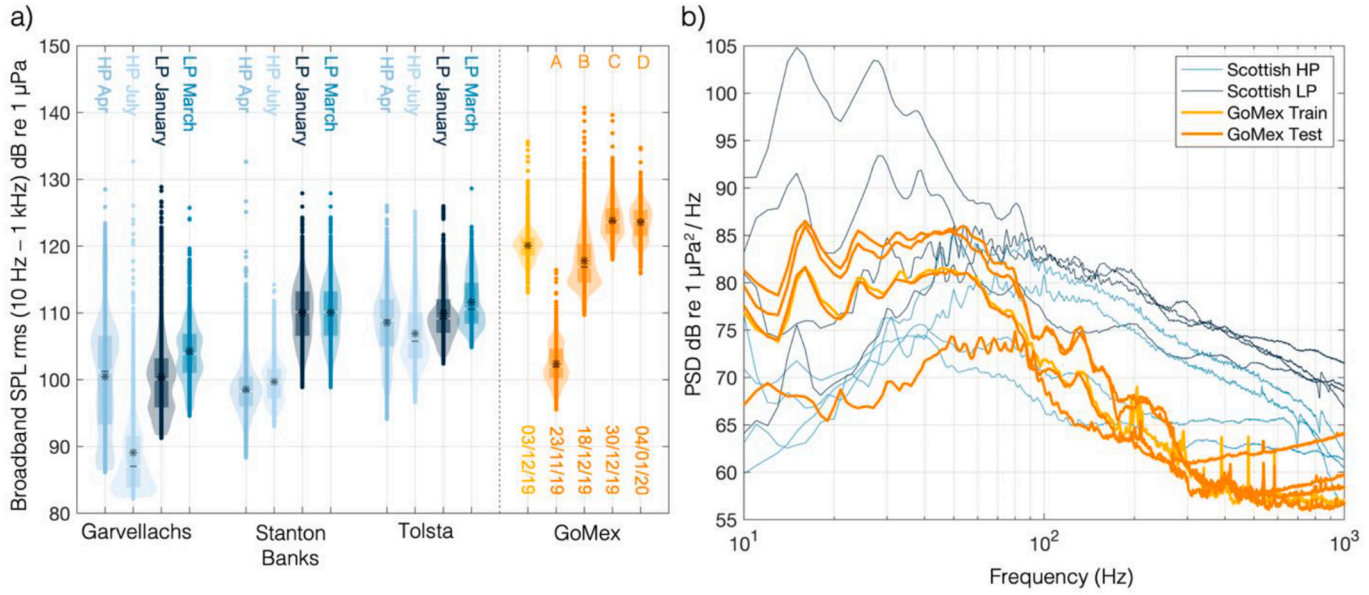


Fig. 2. Comparison of soundscapes used in training the base model and the Gulf of Mexico. a) Median Broadband sound pressure level (SPL) (10 Hz – 1 kHz) for the seasonal periods used in White et al. (2022), extracted from PAM data spanning December 2018 – November 2019. The median broadband SPL is compared to that of the Gulf of Mexico training data (yellow) and test sets (orange) to demonstrate variation in the ambient low frequency spectra between the two environments. b) Median power spectral density (PSD) (10 Hz – 1 kHz) for each seasonal period and Gulf of Mexico training and test sets. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Contribution of sound sources to the training and validation sets for developing the base model.

	Training data	Validation data
Ambient	14,826	1482
Biological Clicks	4675	467
Delphinid Tonal	11,841	1184
Vessel	15,407	1540
Totals	**46,749**	**4673**

and January 2020 using a Rockhopper unit, deployed on the seabed, the hydrophone located approximately 10 m above the seafloor in a water depth of 440 m (Klinck et al., 2020). PAM data was recorded continuously, at a sampling rate of 197 kHz and 24 bit resolution. The dynamic range of the system is approximately 107 dB, the system noise floor is below 35 dB re 1µPa^2/Hz at frequencies above 1 kHz (Klinck et al., 2020).

2.3. Training data

A training set was developed from the Gulf of Mexico data consisting of data from a single day: December 3rd 2019 (selected arbitrarily). The 24 h of PAM data was divided into frames of 3 s, reviewed visually (spectrograms) and aurally using Audacity software (version 3.0.02, 2021) and classified into one of the four sound source categories. Following the principles detailed in White et al., 2022, frames were assigned a single label based on a hierarchy 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; (ii) biological clicks are only labelled as such in the absence of whistles; (iii) a sound source is labelled if any detection is made by an analyst regardless of signal strength in the frame in respect to the ambient noise. After annotation a training set of 28,800 frames is available across the four classes.

Within the soundscape of the Gulf of Mexico there are signals present which are not present in the base model training data, or the Scottish soundscape, specifically Sperm Whales, airguns and delphinid harmonic burst pulses (Fig. 5). Sperm whale acoustic signals present during Dec 3rd were included within the biological click class, and make up 42% of the frames labelled biological clicks (Fig. 5). Airguns are present within the training data across all classes, no separate class was created to represent them. Harmonic burst pulses are labelled as 'delphinid tonal' and make up 11% of training frames in that class.

The existence of class noise in the training set, due to mislabelling, is a common issue and results in a marginal decrease in the accuracy of the classifier when the mislabelling error rate is low (Nazari et al., 2018). To ensure consistent labelling, a subset of the training data was blindly reviewed by an analyst. An error rate of 2% is estimated within the training set based on the manual verification.

To fine-tune the base model, we required 600 frames per class (500 training, 100 validation), (Fig. 2). Data augmentation is applied to ensure sufficient frames per class from December 3rd 2019 are available. Specifically, augmentation is applied only to the delphinid tonal and biological click classes for which only 497 and 163 frames were available respectively. Augmentation was performed by applying three randomly selected signal transformations to the audio prior to forming the spectrogram: pitch shift, time-shift and added noise, not limited to one per category. The parameters for each transformation were randomly selected from a pre-defined range: pitch-shift 0.5–1.5, time-shift −300 ms and 300 ms and Gaussian white noise added with powers between 1 and 2. After augmentation all frames were manually reviewed to ensure the class label remained appropriate.

Stacked spectrograms processed for the Gulf of Mexico were computed as per the base model. RGB images were created by combining three spectrograms created using different window lengths and stacking each spectrogram into a three-channel matrix. The constituent spectrograms were computed using FFT sizes of 1024, 2048 and 4096, for RGB channels respectively, with a sample rate of 197 kHz and employing a Hanning window, with 50% overlap. Each spectrogram was transformed onto a decibel energy scale and normalised, whereupon they were resized to 224 × 224 pixels, the required input size of the network, and inserted into the appropriate colour channel of the image.

2.4. Test data

To evaluate the performance of each fine-tuned base model a test set

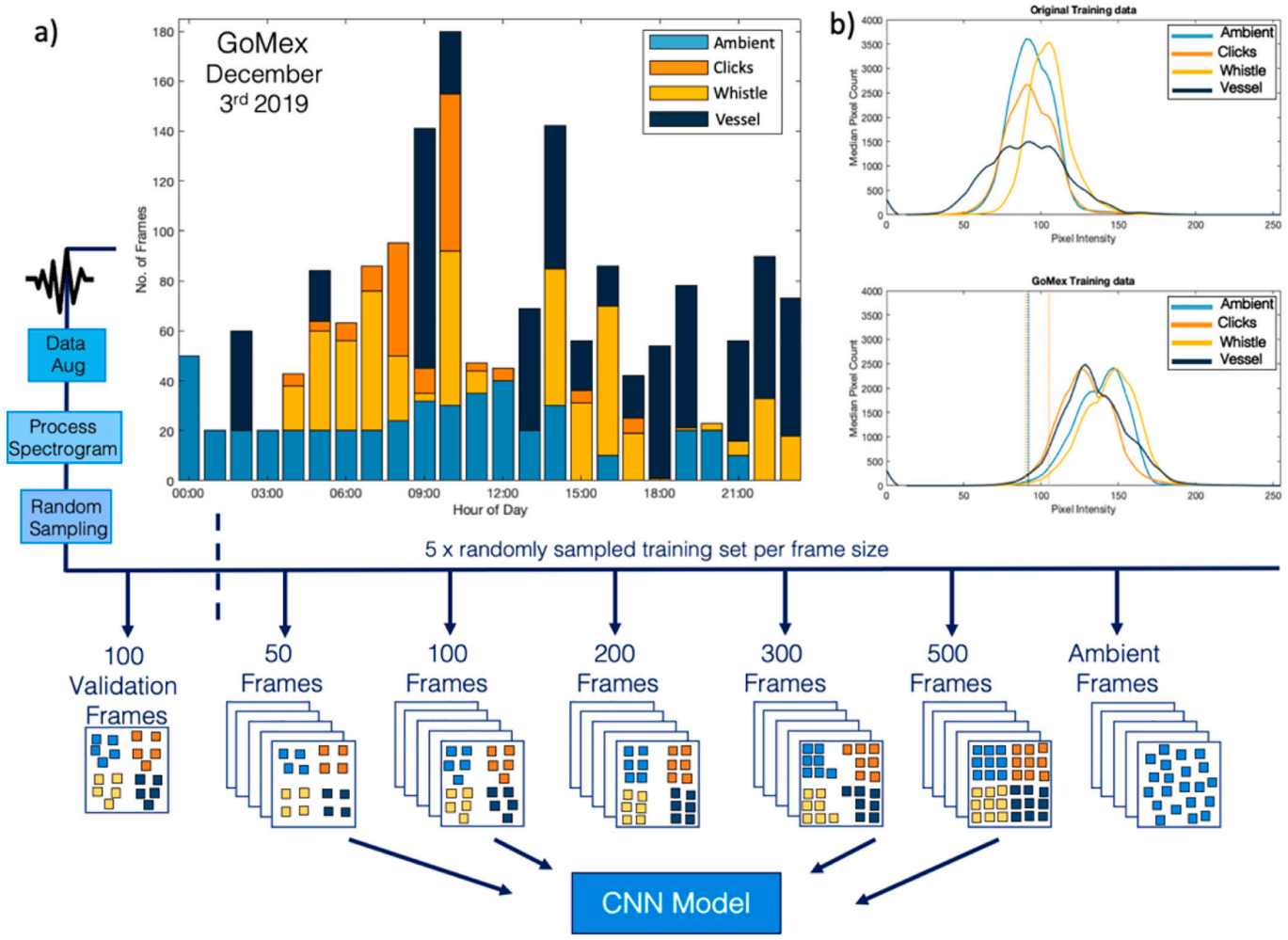


Fig. 3. Experimental design. a) A training set is developed by manually labelling 3 s frames of PAM data collected on December 3rd 2019 in the Gulf of Mexico. The training set is amplified using data augmentation to create 600 frames per class. Three-channel RGB stacked spectrograms are produced per frame. A randomly sampled training set is computed per experimental batch size; 50, 100, 200, 300 & 500 frames, repeated five times per batch size. Similarly, a validation set of 100 frames is extracted from the randomly sampled pool. An ambient only experimental batch is randomly sampled, taking 500 ambient frames from the larger pool. The base model is fine-tuned on each individual training set. b) A comparison of the mean pixel intensity of 100 randomly sampled spectrograms per class, compared between the base model and the Gulf of Mexico training sets. The base model mean per class is marked with a dotted line, demonstrating the intensity shift present across all classes between the two domains.

is created, comprising four full days of annotated data from the Gulf of Mexico; Day A: 23rd November 2019, Day B: 18th December 2019, Day C: 30th December 2019 and Day D: 4th January 2020, Table 2. The minimum time between test datasets is 4 days. Each 24-h period of PAM data equates to 28,800 frames, a total of 115,200 frames for the four days (Table 2). In this dataset the delphinid tonal class is least represented, accounting for only 0.6% of the frames.

2.4.1. Model evaluation

Each fine-tuned model's performance was evaluated using *Precision (P)*, *Recall (R)* and *Accuracy (A)*, (Mesaros et al., 2016). These are calculated based on the number of true and false positive detections, N_{TP} and N_{FP}, along with the number of true and false negative detections, N_{TN} and N_{TP}:

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{1}$$

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}}, \tag{2}$$

$$A = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{FP}}. \tag{3}$$

The F1 score, the harmonic mean of the precision and recall, is also considered. Further we consider the one-vs-all Receiver Operating Characteristic (ROC) curves to summarise performance (Hildebrand et al., 2022). ROC curves are especially useful for domains with skewed class distribution, as found in our test sets (Table 2). The area-under-curve (AUC) is used as a summary statistic for these curves, per class (Stowell, 2022). Models are randomly initialised and re-trained 5 times per batch size. The standard error reported is the variance of the mean performance metrics for models within each batch size.

To understand the effect local fine-tuning has on the model's ability to classify a sound source we used the final convolution feature map of each model to identify which parts of the input spectrogram impact the classification score. In particular we are interested in the evolution of the feature map with more site-specific training data added. Gradient-weighted Class Activation Mapping (Grad-CAM, Selvaraju et al., 2017), is used to create importance maps to highlight the critical regions of the spectrogram, critical to forming classification outputs.

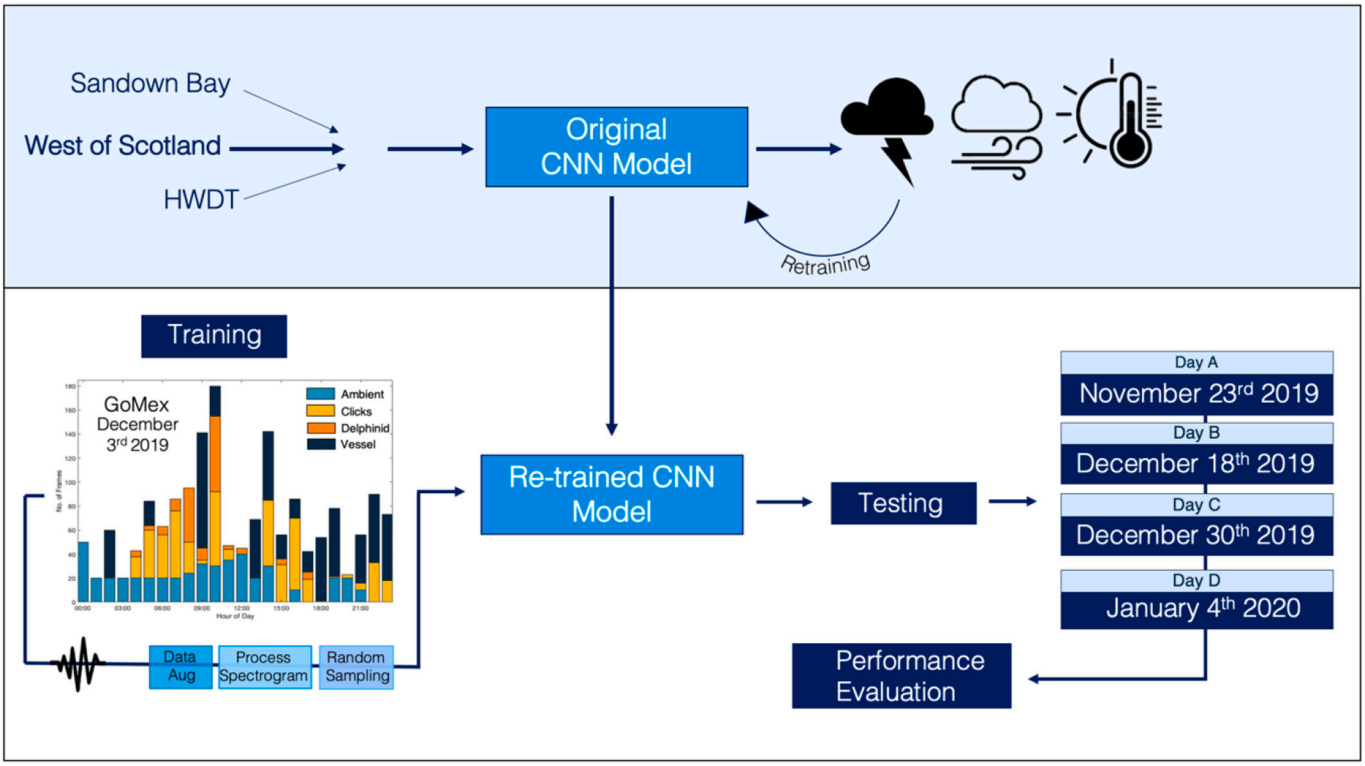


Fig. 4. Workflow of model development and experimental retraining. The original base model is trained on multiple global datasets. The base model is used as the framework for experimentation on the Gulf of Mexico dataset. PAM files from December 3rd 2019 are used as a training set. Each experimental run is tested on four days of data: November 23rd, December 18th, December 30th 2019, and January 4th 2020.

3. Results

Here we evaluate the base model performance as a result of fine-tuning on local data by (i) assessing variability between soundscape characteristics, (ii) analysing performance metrics across the whole test set and (iii) inspecting per day variation within the test set.

3.1. Variation in soundscape characteristics

Ambient sound levels differ between the base model training sites and the Gulf of Mexico (Fig. 2a). Overall median broadband SPLs (10 Hz – 1 kHz) are higher in the Gulf of Mexico than the Scottish sites: median broadband SPLs are between 118 and 125 dB re 1 μPa for the training and testing periods (Fig. 2a). Scottish broadband SPLs were computed for two low pressure and two high pressure weather periods, per site, representing the extremities of the soundscape characteristics (Fig. 2). High pressure conditions generate the lowest ambient sound conditions, with a median broadband SPL between 89 and 110 dB re 1 μPa. Low pressure conditions report the highest SPLs, ranging between 100 and 112 dB re 1 μPa. The acoustic data used for training and testing demonstrate diurnal variation in ambient conditions, with a 17 dB difference between November 23rd 2019 (testing) and December 3rd 2019 (training), due to variation in seismic activity. Fig. 3b illustrates the variation in median spectrum levels between 10 Hz and 1 kHz within each local soundscape. The median spectrum comparison highlights the variability between the Gulf of Mexico and the west of Scotland.

3.2. Entire test set

Site-specific fine-tuning significantly improves the performance of the model in a new environment (Fig. 6). The original base model reports AUC scores of 0.51–0.57 across the four classes (Table 3). Adding 50 frames per class of site-specific data to the training set during fine-tuning improves mean AUC scores by >0.30 across the classes (Fig. 6). The AUC score is above 0.75 for all classes with ≥50 training frames added. Batch sizes of 500 frames result in the highest mean AUC scores across all classes, with the lowest variation between model runs, Fig. 6. Increasing the number of frames per class used for fine-tuning from 50 to 500, improved F1 scores by 0.01, 0.08, 0.27 and 0.30 for AN, BC, DT and VN, respectively (Table 3). However, the F1 score standard error for 500 frames is greater than observed when 50 frames are employed.

The mean AUC score per class for 100 frames is lower than scores for 50 frames, with a greater standard error for all classes, (Fig. 6). As batch sizes increase to 200 and 300 frames per class there is no improvement in mean AUC score for ambient noise and biological click frames. AUC scores for the vessel noise class fluctuate between 100 and 300 frames, reporting a mean AUC 0.06 lower than models trained on 50 frames and 0.11 lower than models trained on 500 frames. A severe decrease in recall is reported for the vessel class (Table 3). There is subtle difference between the AUC and F1 scores for the models trained on ambient or 50 frames of each class (Table 3), for AN, BC and VN. For the delphinid class the ambient only model has a higher AUC score, 0.86, than the 50 frames model. The 500 frames mean AUC score for the delphinid class also reports 0.86.

3.3. Spectrogram regions of importance

Using GradCam to inspect model predictions shows that our model is capable of focusing on important regions of the spectrogram for each class (Fig. 7). Fine-tuning with local data allows the model to adapt, learning to extract specific regions of the image to improve classification of data from a new domain. Fine-tuning with 50 frames improves the models class prediction (Fig. 7a), with increasing classification scores (confidence) after increasing site-specific data. Low SNR and temporal overlap with other signals of interest can result in poorer model performance. As batch sizes increase to 500 frames the regions of

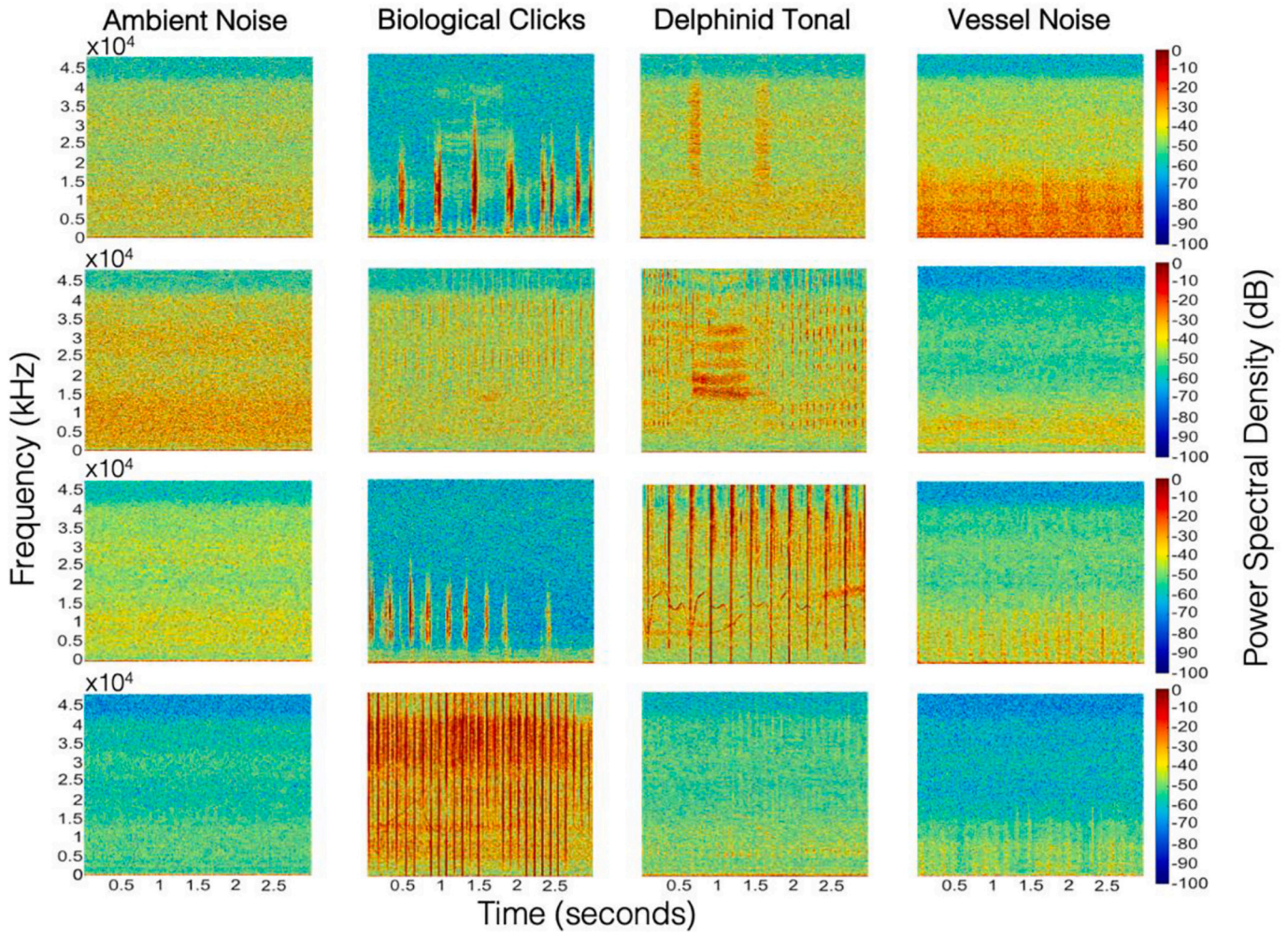


Fig. 5. Representative spectrograms of the four classes used as input for training; Ambient noise, Biological clicks (echolocation, burst pulses and sperm whales), Delphinid tonal (whistles and harmonic pulses) and Vessel noise. Each spectrogram computed with a Hanning window, 75% overlap and an FFT size of 4211. Time and frequency are on the horizontal and vertical axis, respectively.

Table 2
Test set class distribution. Number of frames per class labelled for each test day.

Day	Date	Ambient Noise	Biological Clicks	Delphinid Tonal	Vessel Noise
Day A	23rd Nov 2019	22,934 (79.6%)	2077 (7.2%)	346 (1.2%)	3443 (12%)
Day B	18th Dec 2019	27,582 (95.7%)	363 (1.3%)	0 (0%)	828 (3%)
Day C	30th Dec 2019	20,492 (71.3%)	715 (2.6%)	360 (1.4%)	7101 (24.7%)
Day D	4th Jan 2020	24,052 (83.6%)	1511(5.2%)	12 (0.04%)	3199 (11.1%)
	Totals	**95,060** (82.4%)	**4666** (4%)	**718** (0.6%)	**14,571** (13%)

importance of the input spectrogram become apparent and class activation scores improve (Fig. 7). Detection of new and unfamiliar sound sources is complex, the delphinid harmonic pulses output high classification scores (>0.90) for the incorrect class (Fig. 7d).

3.3.1. *Inspecting per day performance*

In this section we inspect model performances for each individual test day to evaluate model performance when fine-tuned on data from one day, 3rd December 2019, and tested on data from other days.

Day A represents the richest soundscape, Table 2, with the ambient sound spectra significantly lower between 10 Hz and 200 Hz than the training data (Fig. 2). Performance metrics for are high across all sound sources for all batch sizes, outperforming the base model (Fig. 8a). All classes report AUC scores above 0.70 after training with 50 frames per class. Training using ambient only data improves the class prediction by >0.11 from the base model (Fig. 8a). The biological click class mean AUC score improves by 0.25 with ambient only training data, equivalent to the mean AUC score 0.89 after training with 50 frames per class. Fine-tuning on site-specific data adds the biggest performance boost to the delphinid class, from an AUC score of 0.59 by the base model, to 0.78 ± 0.04 with 50 frames and 0.84 ± 0.01 with 500 frames per class. Precision and recall become more balanced for each class by using fine-tuning, with the greatest performance improvements within the delphinid class (Fig. 9).

The frequency spectra of Day C subjectively mirrors that of the training data, with an increase of 4–6 dB between 10 Hz – 100 Hz, and a median broadband SPL 5 dB higher than the training data (Fig. 3). The base model is not able to detect biological clicks or delphinid tonal calls within the soundscape (Fig. 9). Fine-tuning using site specific data improves mean AUC scores by 0.24, 0.31, 0.35 and 0.21 for AN, BC, DT and VN, respectively, with only 50 frames per class. The detection of vessel noise signatures is poor resulting in a high number of false positives and negatives (Fig. 9c). Fine-tuning on increasing batch sizes reduces performance scores for each class when 200 and 300 frames per class are used (Fig. 8c). Training sets consisting of ambient only frames

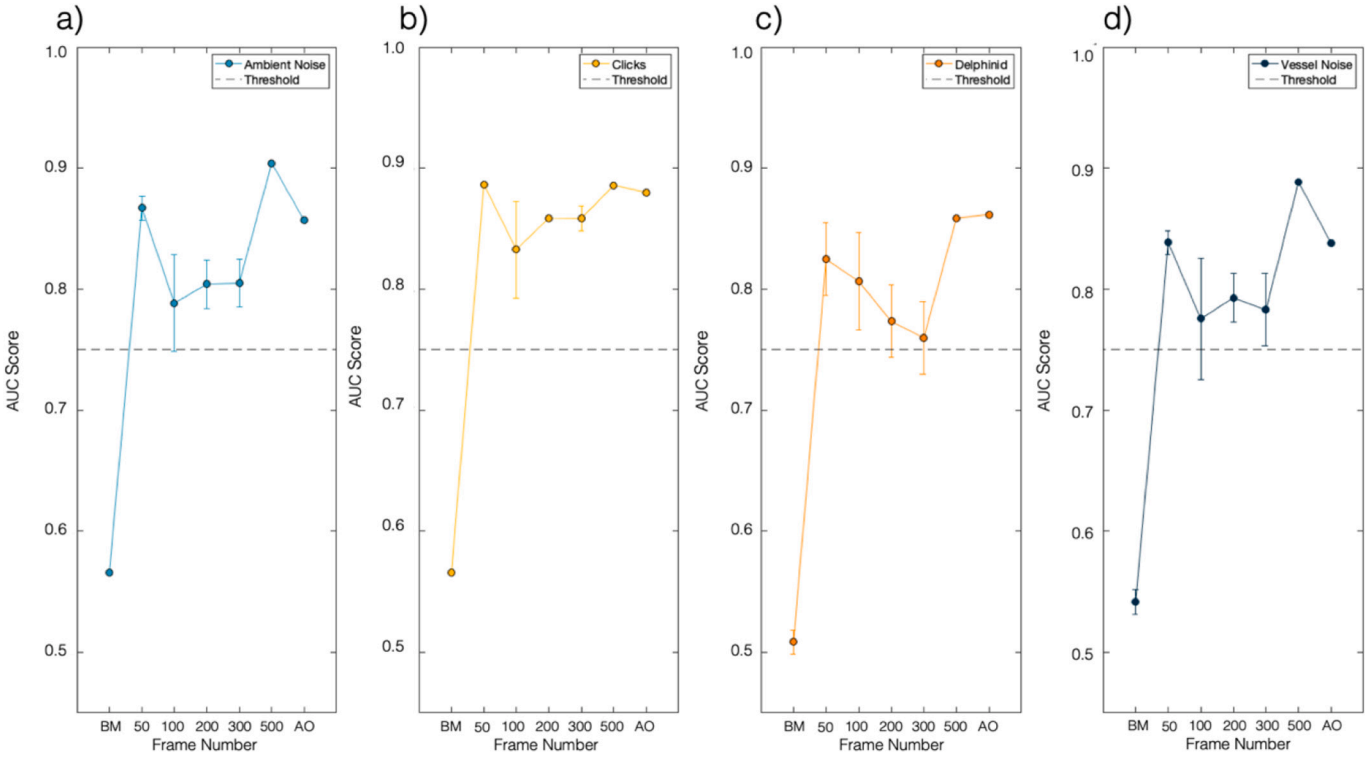


Fig. 6. The importance of site-specific data demonstrated across each of the four sound source classes. Mean AUC scores per training set are plotted per class: a) Ambient Noise, b) Biological Clicks, c) Delphinid Tonal, d) Vessel Noise. Confidence intervals plotted represent the standard error of the mean. Model performance is evaluated for acoustic data collected on November 23rd, December 18th, December 30th 2019 and January 4th 2020. Fine-tuning with 50 frames provides an improvement across all classes from the base model, fine-tuning with 500 frames per class results in lower error rates.

Table 3
Mean performance metrics for each experimental batch size, with standard error, for the entire test set.

Batch size	Sound source	Precision		Recall		F1 Score		Mean AUC (Roc curve)	
Base Model (BM)	Ambient	0.49	–	0.96	–	0.66	–	0.57	±0.00
	Clicks	0.65	–	0.45	–	0.55	–	0.57	±0.00
	Delphinid	0.55	–	0.25	–	0.35	–	0.51	±0.01
	Vessel	0.73	–	0.76	–	0.75	–	0.54	±0.01
50 Frames	Ambient	0.89	±0.00	0.98	±0.00	0.93	±0.00	0.87	±0.01
	Clicks	0.58	±0.01	0.58	±0.00	0.58	±0.00	0.87	±0.00
	Delphinid	0.59	±0.02	0.34	±0.00	0.43	±0.00	0.83	±0.03
	Vessel	0.78	±0.05	0.26	±0.02	0.38	±0.02	0.84	±0.01
100 Frames	Ambient	0.88	±0.01	0.98	±0.00	0.93	±0.00	0.79	±0.04
	Clicks	0.48	±0.04	0.53	±0.07	0.50	±0.05	0.83	±0.04
	Delphinid	0.35	±0.06	0.32	±0.06	0.33	±0.06	0.81	±0.04
	Vessel	0.85	±0.02	0.19	±0.01	0.30	±0.02	0.78	±0.05
200 Frames	Ambient	0.89	±0.00	0.97	±0.00	0.93	±0.00	0.80	±0.02
	Clicks	0.54	±0.02	0.57	±0.01	0.55	±0.01	0.86	±0.00
	Delphinid	0.23	±0.07	0.36	±0.07	0.24	±0.06	0.77	±0.03
	Vessel	0.83	±0.03	0.25	±0.01	0.38	±0.02	0.79	±0.02
300 Frames	Ambient	0.89	±0.00	0.95	±0.03	0.92	±0.01	0.81	±0.02
	Clicks	0.53	±0.04	0.55	±0.03	0.54	±0.01	0.86	±0.01
	Delphinid	0.34	±0.10	0.47	±0.08	0.32	±0.07	0.76	±0.03
	Vessel	0.86	±0.00	0.25	±0.00	0.39	±0.00	0.78	±0.03
500 Frames	Ambient	0.91	±0.01	0.67	±0.01	0.94	±0.01	0.90	±0.00
	Clicks	0.66	±0.05	0.67	±0.06	0.66	±0.04	0.89	±0.00
	Delphinid	0.77	±0.03	0.67	±0.09	0.70	±0.07	0.86	±0.00
	Vessel	0.79	±0.03	0.63	±0.11	0.68	±0.07	0.89	±0.00
Ambient Only	Ambient	0.90	–	1.00	–	0.90	–	0.86	–
	Clicks	0.50	–	0.60	–	0.60	–	0.88	–
	Delphinid	0.60	–	0.30	–	0.40	–	0.86	–
	Vessel	0.90	–	0.20	–	0.30	–	0.84	–

outperform both of these batch sizes scoring >0.75 AUC.

for each class (Fig. 8c). Fine-tuning with 500 frames per class produces mean AUC scores >0.80 for all classes with small standard error ranges. The precision and recall scores are most balanced for 500 training frames per class, across all of the classes (Fig. 9c).

Test days B and D have less diverse soundscapes, both dominated by sperm whale activity. Sperm whale clicks make up the vast majority of the biological click class rather than delphinid click types. No delphinid

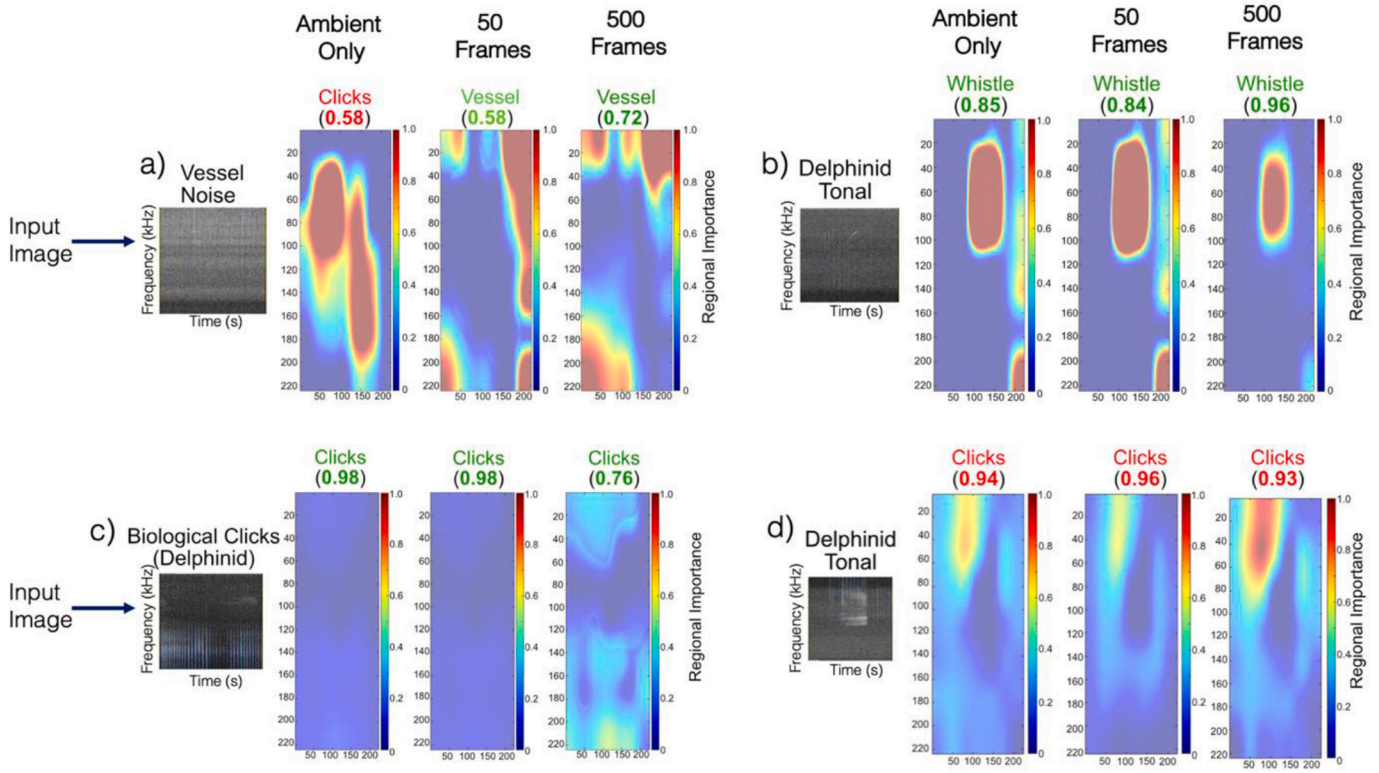


Fig. 7. Exemplar model classifications, illustrated with GradCAM. GradCAM extracts the activation map of the final convolutional layer of each model, highlighting regions of importance within the spectrogram to the model's classification prediction for a) vessel noise, b) low SNR delphinid whistle, c) delphinid clicks within vessel noise and d) delphinid harmonic burst pulse. Each GradCAM plot displays the predicted class label and score by the model, coloured in red if the label is incorrect and green if the label is correct. Darker green indicates a more confident prediction. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

tonal signals are found on Day B and only 12 low SNR whistles (0.04%) are found throughout Day D (see Supplementary Fig. 1). Both days B and C report a median broadband SPL of 125 dB (Fig. 3). The lower diversity soundscapes benefit from site specific fine-tuning, improving mean AUC scores across all classes after fine-tuning on 50 frames, (Fig. 8). Results for Day B show a large performance increase for ambient noise (0.43) and vessel noise (0.42), with a smaller improvement in the biological click class (0.11). Fine-tuning on 500 frames per class, or ambient only frames reports high performance metrics, with ambient only frames outperforming 500 frames by 0.11 for the vessel class (Fig. 8b). Fine-tuning on increasing batch sizes has a lesser effect on overall performance for the Day D soundscape (Fig. 8d). Fine-tuning with ambient frames only does not offer a benefit over using 50 frames per class (Fig. 8d), with 500 frames per class scoring the highest for all classes (Fig. 8d). Fine-tuning offers minor improvements over the base model for the delphinid class but AUC scores are low across all batch sizes (Fig. 8d). The precision and recall scores provide insight to the poor performance with and without fine-tuning on local data of any quantity, with no correct delphinid detections in nearly all experiments (Fig. 9).

Overall, our results demonstrate that local site-specific data is necessary for successful deployment of the detection model, with performance improving significantly after only 50 frames.

4. Discussion

Data labelling in sufficient quantities to train deep neural networks demands significant manual effort, and so the ability to apply a trained model on a specific ocean soundscape to other recording configurations and locations is greatly beneficial to the bioacoustic community. Here we consider the challenge of deploying a CNN developed for one marine environment in a new environment, for acoustic detection where class distribution shifts are significant. Our results illustrate that addition of site-specific training data improves the performance of a CNN model across all defined classes. With as little as 50 frames per class we demonstrate a significant improvement in model performance, outperforming the base model AUC score by 0.30. The addition of greater site-specific frames to the training set reduces the variation in model performances between runs. Fine-tuning with randomly selected data from the local soundscape improves performance in the new marine habitat without cherry-picking data.

There have been relatively few prior studies on fine-tuning for domain adaptation with which we can compare our work. Lauha et al. (2022) found that additional site-specific training frames were necessary for improving global bird detection CNN models on local environments with niche species present. We demonstrate similar degrees of success to Lauha et al. (2022) with a small-scale model (fewer parameters) and a restricted training set. By fine-tuning regional models on subsets of data from the local soundscape we aim to move the field towards utilising CNN models in regions for which they are not trained, to extract signals of interest which could be advantageous to the development of global models. Matching quantitative success of small-scale models with global models should encourage researchers to tailor existing marine CNN networks to their own regions through fine-tuning as a form of domain adaptation.

Fine-tuning on local data is challenging: There is a balance between maintaining the learned features from the original base model, while presenting new information for the model to learn from which is inclusive of the numerical ranges present within spectrograms from the new domain. In the common case that labelled data is scarce for the site of interest or the signal of interest is rare, training with only ambient noise frames proves beneficial where signal types between marine environments are comparable. The Gulf of Mexico is subject to high levels

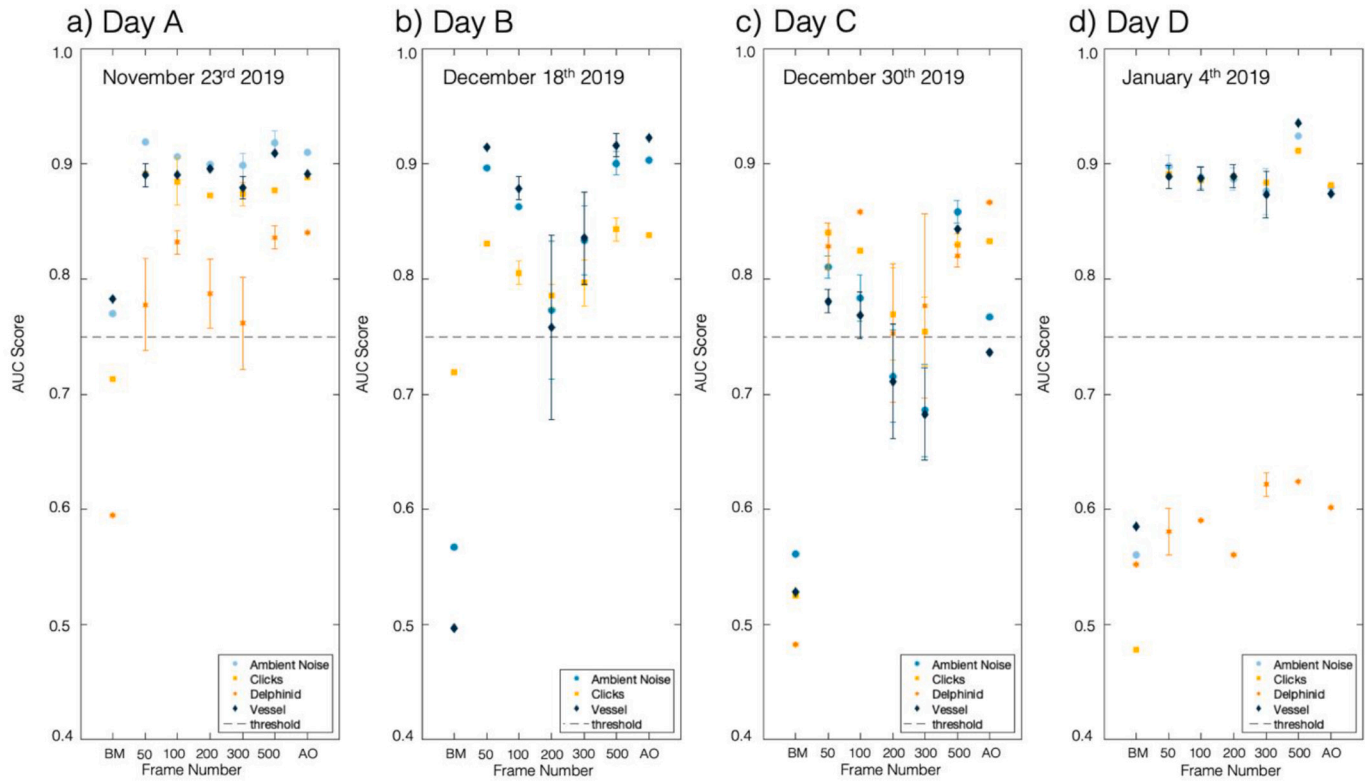


Fig. 8. The importance of fine-tuning on site-specific training data, demonstrated across each of the four sound source classes per individual test day. Mean AUC scores are plotted per class for each test day, a) Day A, b) Day B, c) Day C & d) Day D when the training data is fed in in batches of 50/100/200/300/500 frames, plus ambient only frames. Confidence intervals plotted represent the standard error of the mean of the AUC scores across model runs. In a diverse soundscape performance metrics are high, with mean AUC scores 0.10–0.30 higher than those reported by the base model (a & b). See supplementary Table 2 for raw data.

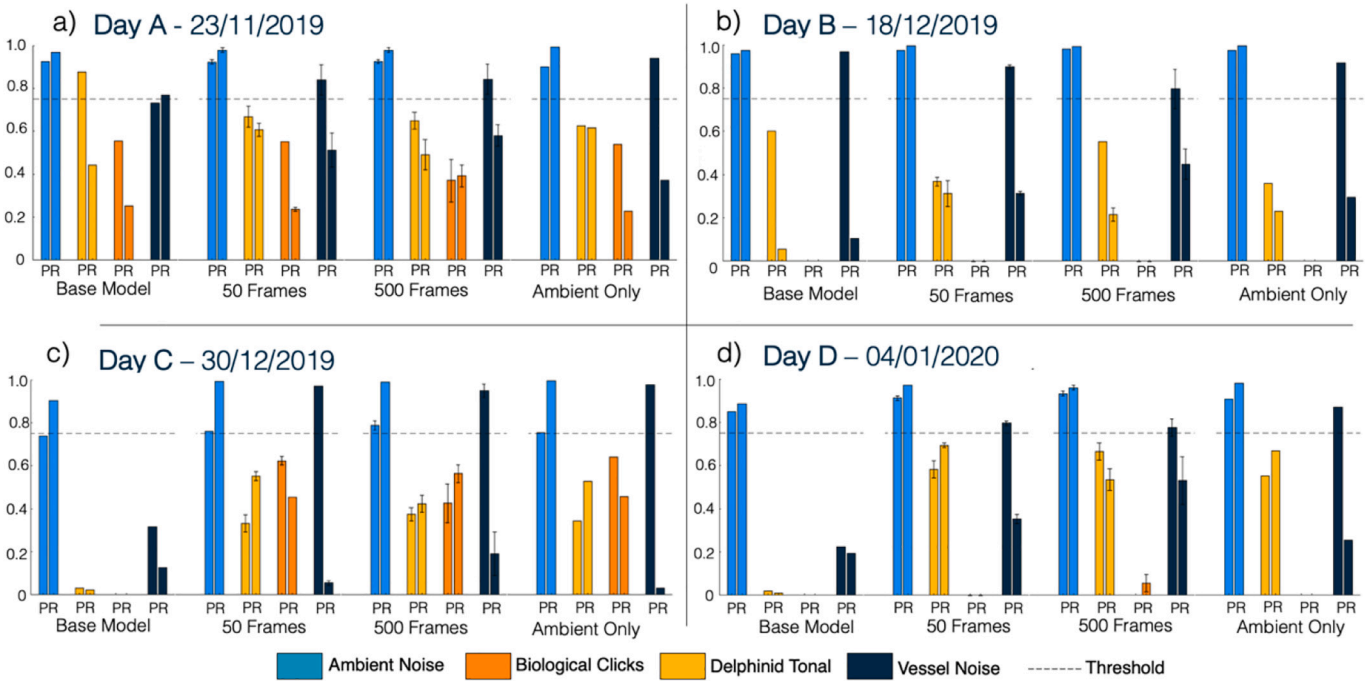


Fig. 9. The success of site-specific fine-tuning demonstrated by the precision (P) and recall (R) scores for each of the four sound source classes for each individual test period: a) November 23rd, b) December 18th, c) December 30th and d) January 4th 2020. Mean precision and recall scores are plotted per class for the base model, 50, 500 and Ambient only frames. Error bars are plotted for each batch size, representing the standard deviation of the mean for the precision and recall scores across each model run. Some standard deviation values are so small error bars are not visible, due to the axis range 0–1, see supplementary table 2. Site specific training data improves precision and recall scores above the base model, with scores most balanced for 500 training frames across all classes.

of seismic activity, with airguns found throughout the training and test periods, a loud contributor to the soundscape not present within the base model training set (Klinck et al., 2020). We found that through fine-tuning with 500 frames of ambient noise we can achieve similar results across all classes as when the model is trained on batches of each class. Performance metrics indicate that by using ambient noise data to fine-tune the base model we optimise the learned feature representation to the local environment, improving performance across all classes after only 50 frames per class (2.5 min of data).

Our results show that selecting a diverse set of training frames for local fine-tuning is essential to capture the variability in local soundscape characteristics. Where a signal is not present within the selected 24-h training period, we do not show the signal to the model. Sperm whale activity is high in the Gulf of Mexico (Collum and Fritts, 1985; Farmer et al., 2018; Miller et al., 2009), a species not found in the data from the West of Scotland. Successful model deployment here is dependent on the training batches including both delphinid and Sperm whale clicks to allow the model to learn the salient features across a wide frequency spectrum. Within 50 frames the model is capable of detecting sperm whales, with increasing batch sizes the precision and recall for click detection improve. Within the training data, a novel type of whistle was present, not found in the original training data, specifically a harmonic buzzing tonal call (see Supplementary fig. 1). This buzzing call occurs at high frequencies, occupying a similar spatial region within the spectrogram as the biological click class. Model predictions across the test set are mixed with respect to this signal type. After 500 frames of fine-tuning the model still found difficulty detecting high SNR signals, illustrated by the results of Day D (Supplementary fig. 1). The random pooling used to curate each training set allows for the possibility that scarce or novel signals are not included, resulting in a large degree of error between model runs with small batch sizes. CNNs are robust to signal fluctuations but for unique and complex signal types found within the new environment it is critical that efforts are taken to label a pool of frames representative of the specific signal under various local conditions, for optimum model performance. Knowledge of the long-term temporal and spatial characteristics of a soundscape could be used to select training frames, and maximise the information contained in the training spectrograms. Future efforts will explore the performance benefits of selecting training periods based upon soundscape analyses.

The variability between ambient spectra is attributed to be the reason for poor precision and recall scores within specific classes. The ambient noise and vessel classes are closely coupled, such that we observe similar trends in performance metrics across test sets. Within the Gulf of Mexico there is a high level of anthropogenic activity which varies day-to-day, resulting in the ambient soundscape fluctuating on short temporal periods. The point at which a vessel becomes a vessel, and is no longer ambient noise remains difficult to pinpoint, and can cause irreducible error as a result of annotation bias. The low precision and recall scores for the vessel class are a result of the model depicting the start and end of a vessel passing as different to that of a manual annotator, as a result the performance metrics are penalised. Future efforts on automated solutions for marine signal analysis will aim to define more concrete boundaries for the distinction between vessels and ambient noise, to improve overall performance metrics.

In this work we use two environments that are significantly different with respect to their soundscape. The training and testing protocol used is restricted to five 24-h periods of acoustic data, within which we do not select periods of high quality, or high activity data. Although training batches are equal across classes, the contributing frames to each training batch cannot be equal due to overrepresentation of ambient frames. We attribute the variable results between batch sizes 100 and 300 across all performance metrics to the bias-variance trade-off (Neal et al., 2018). Specifically, at small sample sizes, noise can create fluctuations in the data that look like genuine patterns. Here we are using a small-scale model pre-trained on 40,000 frames. During fine-tuning, as the number of frames per class reaches 500, the model is able to learn to separate the noise from the patterns within each signal type across the two marine environments, improving performance.

Our results show that as fine-tuning batch sizes increase, the AUC scores and confidence intervals converge across the signal types. We demonstrate that fine-tuning on even small portions of local PAM data enables the classifier to reach good performance with a considerably smaller training volume, helping to overcome the data bottleneck within marine bioacoustics.

5. Conclusion

Automated analysis of passive acoustic data is essential if we are to take advantage of the large volumes of acoustic data available to exploit the ecological information they contain. We demonstrate that with small subsets of site-specific data researchers can harness the power of existing bioacoustic detection models and tailor them to other recording environments, without cherry-picking high-quality data. We report a 0.30 improvement to AUC scores after fine-tuning with only 50 frames per class. Our approach highlights the ability to detect new signal types not found within the original marine environment, aiding the detection of new species and signal types with minimal annotation effort. Fine-tuning using only ambient noise frames can produce results which rival training on larger pools of training data representing each class. This is beneficial in the absence of labelled data, or when working with sparsely occurring signal types. Our analysis shows that models trained on one geographic region can be applied to another specific region with a small amount of additional labelled data, and that an initial globally trained model is not required. We hope to encourage the adaptation of existing bioacoustic CNN architectures to new marine regions and soundscapes in order to exploit the wealth of information which is held in PAM datasets globally.

Author contributions

EW and HK conceptualised the study. HK conceptualised the Gulf of Mexico dataset and carried out data storage and transfer. EW re-trained the base model, preformed data processing, organised the datasets and carried out model fine-tuning. EW carried out model evaluation and wrote the first manuscript draft. HK, PW and JB contributed to all sections of the manuscript early drafts. Denise Risch was part of the consortium conceptualising the COMPASS project (including all data collection, survey design and data organisation), which forms the bulk of the training data for the base model, and helped with initial analysis concept discussion. All authors contributed to the article and approved the submitted version.

Data sources (where appropriate)

The data used within this manuscript was provided for previous work (White et al., 2022). The COMPASS project acoustic array, developed and managed by SAMS, the Marine Directorate Scotland and the Agri-Food and Biosciences Institute, provided the bulk of the training data for the base model, this data can be made available upon request. We thank the Hebridean Whale and Dolphin Trust for providing acoustic data which aided the development of the base model in this work. The data from the Gulf of Mexico is provided by the K. Lisa Yang Centre for Conservation Bioacoustics, and can be made available upon request to the authors.

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).

Declaration of Competing Interest

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.

Data availability

The raw data supporting the conclusions of this article will be made available by the authors upon request. Requests for accessing the labelled data used to develop the models in this work will be reviewed individually.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2023.102363.

References

Allen, A.N., Harvey, M., Harrell, L., Jansen, A., Merkens, K.P., Wall, C.C., Cattiau, J., Oleson, E.M., 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.

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

Bergler, C., Schröter, H., Cheng, R.X., Barth, V., Weber, M., Nöth, E., Hofer, H., Maier, A., 2019. ORCA-SPOT: an automatic killer whale sound detection toolkit using deep learning. Sci. Rep. 9 (1), 1–17.

Bermant, P.C., Bronstein, M.M., Wood, R.J., Gero, S., Gruber, D.F., 2019. Deep machine learning techniques for the detection and classification of sperm whale bioacoustics. Sci. Rep. 9 (1), 1–10.

Best, P., Ferrari, M., Poupard, M., Paris, S., Marxer, R., Symonds, H., Spong, P., Glotin, H., 2020, July. Deep learning and domain transfer for orca vocalization detection. In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–7.

Bisong, E., 2019. Google colaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley, CA.

Brown, J.C., Smaragdis, P., Nousek-McGregor, A., 2010. Automatic identification of individual killer whales. J. Acoust. Soc. Am. 128 (3), 93–98.

Collum, L.A., Fritts, T.H., 1985. Sperm whales (*Physeter catodon*) in the Gulf of Mexico. Southwest. Nat. 101–104.

Csurka, G. (Ed.), 2017. Domain Adaptation in Computer Vision Applications, Vol. 2. Springer International Publishing, Cham.

Farahani, A., Voghoei, S., Rasheed, K., Arabnia, H.R., 2021. A brief review of domain adaptation. In: Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020, pp. 877–894.

Farmer, N.A., Noren, D.P., Fougères, E.M., Machernis, A., Baker, K., 2018. Resilience of the endangered sperm whale *Physeter macrocephalus* to foraging disturbance in the Gulf of Mexico, USA: a bioenergetic approach. Mar. Ecol. Prog. Ser. 589, 241–261.

Harvey, M., 2018. Acoustic Detection of Humpback Whales Using a Convolutional Neural Network. Google AI Blog.

Haver, S.M., Klinck, H., Nieukirk, S.L., Matsumoto, H., Dziak, R.P., Miksis-Olds, J.L., 2017. The not-so-silent world: measuring Arctic, equatorial, and Antarctic soundscapes in the Atlantic Ocean. Deep-Sea Res. I Oceanogr. Res. Pap. 122, 95–104.

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

Howe, B.M., Miksis-Olds, J., Rehm, E., Sagen, H., Worcester, P.F., Haralabus, G., 2019. Observing the oceans acoustically. Front. Mar. Sci. 6, 426.

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

Kahl, S., Wood, C.M., Eibl, M., Klinck, H., 2021. BirdNET: a deep learning solution for avian diversity monitoring. Eco. Inform. 61, 101–236.

Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Klinck, H., Winiarski, D., Mack, R.C., Tessaglia-Hymes, C.T., Ponirakis, D.W., Dugan, P. J., Jones, C., Matsumoto, H., 2020, October. The ROCKHOPPER: A compact and extensible marine autonomous passive acoustic recording system. In: Global Oceans 2020: Singapore–US Gulf Coast. IEEE, pp. 1–7.

Koidl, K., 2013. Loss Functions in Classification Tasks. School of Computer Science and Statistic Trinity College, Dublin.

Lauha, P., Somervuo, P., Lehikoinen, P., Geres, L., Richter, T., Seibold, S., Ovaskainen, O., 2022. Domain-specific neural networks improve automated bird sound recognition already with small amount of local data. Methods Ecol. Evol. 13 (12), 2799–2810.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Liu, J., Yang, X., Wang, C., 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). IEEE, Qingdao, China, pp. 1–4.

Mesaros, A., Heittola, T., Virtanen, T., 2016. Metrics for polyphonic sound event detection. Appl. Sci. 6 (6), 162.

Miller, P.J., Johnson, M.P., Madsen, P.T., Biassoni, N., Quero, M., Tyack, P.L., 2009. Using at-sea experiments to study the effects of airguns on the foraging behavior of sperm whales in the Gulf of Mexico. Deep-Sea Res. I Oceanogr. Res. Pap. 56 (7), 1168–1181.

Nazari, Z., Nazari, M., Sayed, M., Danish, S., 2018. Evaluation of class noise impact on performance of machine learning algorithms. IJCSNS Int. J. Comput. Sci Netw. Secur. 18, 149.

Neal, B., Mittal, S., Baratin, A., Tantia, V., Scicluna, M., Lacoste-Julien, S., Mitliagkas, I., 2018. A modern take on the bias-variance tradeoff in neural networks. arXiv preprint:1810.08591.

Oswald, J.N., Rankin, S., Barlow, J., Oswald, M., Lammers, M.O., 2003. Realtime call classification algorithm (ROCCA): software for species identification of 26 delphinid whistles. Detection, classification and localization of marine mammals using passive acoustics, 2013 (10).

Pace, F., White, P., Adam, O., 2012. Hidden Markov modeling for humpback whale (*Megaptera novaeanglie*) call classification. In: Proceedings of Meetings on Acoustics ECUA2012. The Journal of the Acoustical Society of America, p. 17.

Padovese, B., Kirsebom, O.S., Frazao, F., Evers, C.H., Beslin, W.A., Theriault, J., Matwin, S., 2023. Adapting deep learning models to new acoustic environments-A case study on the North Atlantic right whale upcall. Ecol. Informa. 77, 102169.

Pijanowski, B.C., Villanueva-Rivera, L.J., Dumyahn, S.L., Farina, A., Krause, B.L., Napoletano, B.M., Gage, S.H., Pieretti, N., 2011. Soundscape ecology: the science of sound in the landscape. BioScience 61 (3), 203–216.

Roch, M.A., Soldevilla, M.S., Hoenigman, R., Wiggins, S.M., 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.

Roch, M.A., Klinck, H., Baumann-Pickering, S., Mellinger, D.K., Qui, S., Soldevilla, M.S., Hildebrand, J.A., 2011. Classification of echolocation clicks from odontocetes in the Southern California bight. J. Acoust. Soc. Am. 129 (1), 467–475.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626.

Shiu, Y., Palmer, K.J., Roch, M.A., Fleishman, E., Liu, X., Nosal, E.M., Helble, T., Cholewiak, D., Gillespie, D., Klinck, H., 2020. Deep neural networks for automated detection of marine mammal species. Sci. Rep. 10 (1), 1–12.

Stowell, D., 2022. Computational bioacoustics with deep learning: a review and roadmap. PeerJ 10, e13152.

Sugai, L., Silva, T., Ribeiro, J., Llusia, D., 2018. Terrestrial passive acoustic monitoring: review and perspectives. BioScience 69 (1), 15–25. Support Vector Machine Classifier for Automated Classification of Beaked Whales and.

Sun, X., Li, G., Qu, P., Xie, X., Pan, X., Zhang, W., 2022. Research on plant disease identification based on CNN. Cognitive Robot. 2, 155–163.

Tan, M., Le, Q., 2019, May. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.

Tyack, P.L., Miller, E.H., 2002. Vocal anatomy, acoustic communication and echolocation. Mar. Mammal Biolo. Evol. Approach 59, 142–184.

Wall, C.C., Haver, S.M., Hatch, L.T., Miksis-Olds, J., Bochenek, R., Dziak, R.P., Gedamke, J., 2021. The next wave of passive acoustic data management: how centralized access can enhance science. Front. Mar. Sci. 8, 703682.

Wang, Z.A., Moustahfid, H.A., Mueller, A., Mowlem, M.C., Michel, A.P.M., Glazer, B.T., Brehmer, P., Mooney, A., Friedman, K.J., Michaels, W., McQuillan, J., 2019. Advancing observation of ocean biogeochemistry, biology, and ecosystems with cost-effective in situ sensing technologies. Front. Mar. Sci. 6, 519.

White, E.L., White, P.R., Bull, J.M., Risch, D., Beck, S., Edwards, E.W.J., 2022. More than a whistle: automated detection of marine sound sources with a convolutional neural network. Front. Mar. Sci. 9 https://doi.org/10.3389/fmars.2022.879145.

Yang, W., Luo, W., Zhang, Y., 2020. Classification of odontocete echolocation clicks using convolutional neural network. J. Acoustical Soc. America 147 (1), 49–55.

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

International Organization for Standardization, 2014. ISO 12913-1: 2014 acoustics—Soundscape—part 1: definition and conceptual framework. ISO, Geneva.