Check for updates

Remote Sensing in Ecology and Conservation
Open Access
ZSL

RESEARCH ARTICLE

Deep learning in marine bioacoustics: a benchmark for baleen whale detection

Elena Schall[1], Idil Ilgaz Kaya[2], Elisabeth Debusschere[3], Paul Devos[4] & Clea Parcerisas[3,4]

[1]Alfred-Wegener-Institute Helmhotz Center for Polar and Marine Research (AWI), Klußmannstr. 3d, 27570, Bremerhaven, Germany
[2]Middle East Technical University, Üniversiteler Mahallesi, Dumlupınar Bulvarı No: 1, 06800, Çankaya, Ankara, Turkey
[3]Flanders Marine Institute (VLIZ), InnovOcean Campus, Jacobsenstraat 1, 8400, Oostende, Belgium
[4]Ghent University, WAVES, Technologiepark-Zwijnaarde 126, 9052, Gent, Belgium

Keywords
Baleen whales, big data, deep learning, marine bioacoustics, passive acoustic monitoring (PAM), sound detection

Correspondence
Elena Schall, Alfred-Wegener-Institute Helmhotz Center for Polar and Marine Research (AWI), Klußmannstr. 3d, 27570 Bremerhaven, Germany.
Tel: +49 471 4831 2157;
E-mail: elena.schall@awi.de

Funding Information
This research was possible thanks to the Research Foundation – Flanders (Belgium) (FWO) under the research grant V509723N.

Editor: Vincent Lecours
Associate Editor: Alice Jones

Received: 30 November 2023; Revised: 6 February 2024; Accepted: 2 April 2024

doi: 10.1002/rse2.392

Abstract

Passive acoustic monitoring (PAM) is commonly used to obtain year-round continuous data on marine soundscapes harboring valuable information on species distributions or ecosystem dynamics. This continuously increasing amount of data requires highly efficient automated analysis techniques in order to exploit the full potential of the available data. Here, we propose a benchmark, which consists of a public dataset, a well-defined task and evaluation procedure to develop and test automated analysis techniques. This benchmark focuses on the special case of detecting animal vocalizations in a real-world dataset from the marine realm. We believe that such a benchmark is necessary to monitor the progress in the development of new detection algorithms in the field of marine bioacoustics. We ultimately use the proposed benchmark to test three detection approaches, namely ANIMAL-SPOT, Koogu and a simple custom sequential convolutional neural network (CNN), and report performances. We report the performance of the three detection approaches in a blocked cross-validation fashion with 11 site-year blocks for a multi-species detection scenario in a large marine passive acoustic dataset. Performance was measured with three simple metrics (i.e., true classification rate, noise misclassification rate and call misclassification rate) and one combined fitness metric, which allocates more weight to the minimization of false positives created by noise. Overall, ANIMAL-SPOT performed the best with an average fitness metric of 0.6, followed by the custom CNN with an average fitness metric of 0.57 and finally Koogu with an average fitness metric of 0.42. The presented benchmark is an important step to advance in the automatic processing of the continuously growing amount of PAM data that are collected throughout the world's oceans. To ultimately achieve usability of developed algorithms, the focus of future work should be laid on the reduction of the false positives created by noise.

Introduction

Passive acoustic monitoring (PAM) creates massive amounts of valuable data to monitor fauna with bioacoustic methods. PAM has the great advantage of being able to collect continuous data at logistically challenging locations. Therefore, this method of data collection plays an especially large role in the investigation of marine habitats and species. The continuously increasing amounts of PAM data make the manual review of these data by human experts more and more impractical and require automatized approaches to process and analyze data to gain knowledge on spatio-temporal patterns of soundscapes, species presence and behavior. While numerous statistical and threshold-based methods are available to solve acoustic detection and classification problems automatically, these methods often suffer from either low sensitivity (i.e., high false-negative rate) or low selectivity (i.e., high false-positive rate), not achieving the accuracy of human classification abilities (Baumgartner & Mussoline, 2011; Kowarski & Moors-Murphy, 2021; Roca &

© 2024 The Authors. *Remote Sensing in Ecology and Conservation* published by John Wiley & Sons Ltd on behalf of Zoological Society of London. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Van Opzeeland, 2019; Schall et al., 2021; Thomisch et al., 2016). Common statistical methods extract representative measurements from the time, frequency and amplitude domain of given audio files which are then supplied to machine learning algorithms or filtered by applying previously tested threshold values (Bittle & Duncan, 2013; Usman et al., 2020). These methods often still require a considerable amount of control through manual post-processing of results (Kowarski & Moors-Murphy, 2021; Schall et al., 2021) The biggest challenges when it comes to automatic detection and classification methods are coping with low signal-to-noise ratios, simultaneous species and noise presence, sparsity of signals of interest, variability in vocalizations (e.g., inter- and intra-populational or interindividual) and overlapping vocalizations (Stowell, 2022).

The analysis of acoustic data is usually based on the use of spectrograms featuring high temporal and spectral resolutions. Computer vision techniques have delivered promising results in recent years with regard to automated spectrogram classification mainly implementing deep learning techniques such as convolutional neural networks (CNNs; Allen et al., 2021; Bergler et al., 2019; Dugan et al., 2014; Halkias et al., 2013; Poupard et al., 2021; Stowell et al., 2019). A main requirement for achieving good results with deep learning techniques is the availability of large annotated datasets to train, validate and test models, which contain as much of the existing variability of soundscapes as possible. This is beneficial so that the model learns to generalize over different spatial and temporal scales to ensure a broad applicability of the developed models in the light of increasing data availability. Numerous bioacoustic software tools provide the functionality to train your own CNNs or even predict your data with pre-trained models, such as Ketos (Kirsebom et al., 2021), Koogu (Madhusudhana, 2022), aviaNZ (Marsland et al., 2019), ANIMAL-SPOT (Bergler et al., 2022), gibbonfindR (Clink & Klinck, 2019), soundClass (Silva et al., 2022), OpenSoundscape (Lapp et al., 2023) and Raven Pro 1.6.5 (K. Lisa Yang Center for Conservation Bioacoustics, 2023). Even more publications report on developed and tested models that can be used to analyze marine passive acoustic data to detect marine animal vocalizations (Allen et al., 2021; Belghith et al., 2018; Bergler et al., 2019; Best et al., 2020, 2022; Bohnenstiehl, 2023; Kirsebom et al., 2020; Madhusudhana et al., 2021; Miller et al., 2023; Rasmussen & Širović, 2021; Rycyk et al., 2022; Shiu et al., 2020; Vickers et al., 2021; White et al., 2022; Zhong et al., 2020, 2021). However, there is only a small number of actual applications of these models to long-term data (Allen et al., 2021; Best et al., 2022; Bohnenstiehl, 2023; Lammers et al., 2023; Rycyk et al., 2022). We believe that this is because most published approaches were evaluated on subsets of data that do not necessarily represent a real-world detection scenario (e.g., Belghith et al., 2018; Vickers et al., 2021; White et al., 2022). A real-world detection scenario in the marine realm is almost always characterized by a large imbalance between shorter time periods when animal vocalizations are present and longer time periods when ‘only’ environmental (e.g., rain, earthquakes, currents) or anthropogenic noise is present (e.g., shipping). Additionally, not only should the dataset be imbalanced, but the soundscape diversity of training, validation and testing datasets should be representative of long-term and large-scale PAM data. These two factors are of key importance when designing a real-world dataset for the evaluation of broadly applicable detection algorithms for marine sounds, posing additional challenges for the development of detection algorithms.

Therefore, to measure the performance of available approaches in a controlled and standardized way, there is the need for a neutral benchmark (Weber et al., 2019), which consists of a defined task specifically designed to measure the performance of different algorithms on a real-world dataset of marine passive acoustic data. For this special case of highly imbalanced datasets from marine PAM (in comparison to terrestrial recordings), we need to make sure that the dataset on which performance is reported on is representative of real-world marine PAM datasets. Here we describe a benchmark for the detection of baleen whale vocalizations which is based on a large annotated dataset (1880.25 h) from seven Antarctic locations, 4 years and all months of the year. This dataset is considered to be representative of real-world marine PAM data. We highlight the need to document performance with three simple and one combined performance metrics and already evaluated three CNN-based detection algorithms on the basis of the presented benchmark.

The evaluation code for this benchmark is available at https://gitlab.awi.de/oza-sound-detectors/cnn_sound_detection.

Materials and Methods

Showcase for choice of dataset and metrics

Dealing with imbalance in the data is a well-known challenge in machine learning. Some metrics commonly used in classification tasks can be misleading when evaluating highly imbalanced data (Johnson & Khoshgoftaar, 2019). Furthermore, in marine bioacoustics, there has been a challenge with proper generalization of the positive class, due to the aforementioned imbalance of the signal of interests and the non-gaussian distribution of ocean noise. Properly evaluating algorithms dealing with such

© 2024 The Authors. *Remote Sensing in Ecology and Conservation* published by John Wiley & Sons Ltd on behalf of Zoological Society of London.

20563485, 0, Downloaded from https://zslpublications.onlinelibrary.wiley.com/doi/10.1002/rse2.392 by Bibliothèque de Sorbonne Université, Wiley Online Library on [30/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

challenges depends on the specific task at hand to, for example, avoid that algorithms trained in one single location do not always perform well on unseen locations. We showcase the importance of the dataset and evaluation metrics' choice by training and testing a simple CNN classifier (for the description of model setup, see "Experimental model setup", Appendix S1) on an annotated marine passive acoustic dataset with varying levels of imbalance: when the model is trained on a dataset in which noise is only represented by a small proportion, its performance is usually also good on a similar test dataset in which noise is represented by a small proportion, but its performance can drop when tested on more and more imbalanced datasets (Fig. 1). Second, we need to make sure that performance is monitored in a way that the used metric is also representative of the detection goal, namely a high recall for all the vocalization classes and a low false-positive rate caused by noise. The importance of this can be illustrated by the results of the simple CNN classifier trained on a high noise percentage (i.e., 90%) and tested on varying levels of imbalance: its accuracy increases when tested on datasets with increasing levels of imbalance; however, this is only due to its ability to correctly classify noise as the recall for the vocalization classes remains at moderate values for all tests (Fig. 1).

Dataset

The library of annotated circum-Antarctic recordings for Antarctic blue and fin whale vocalizations published by

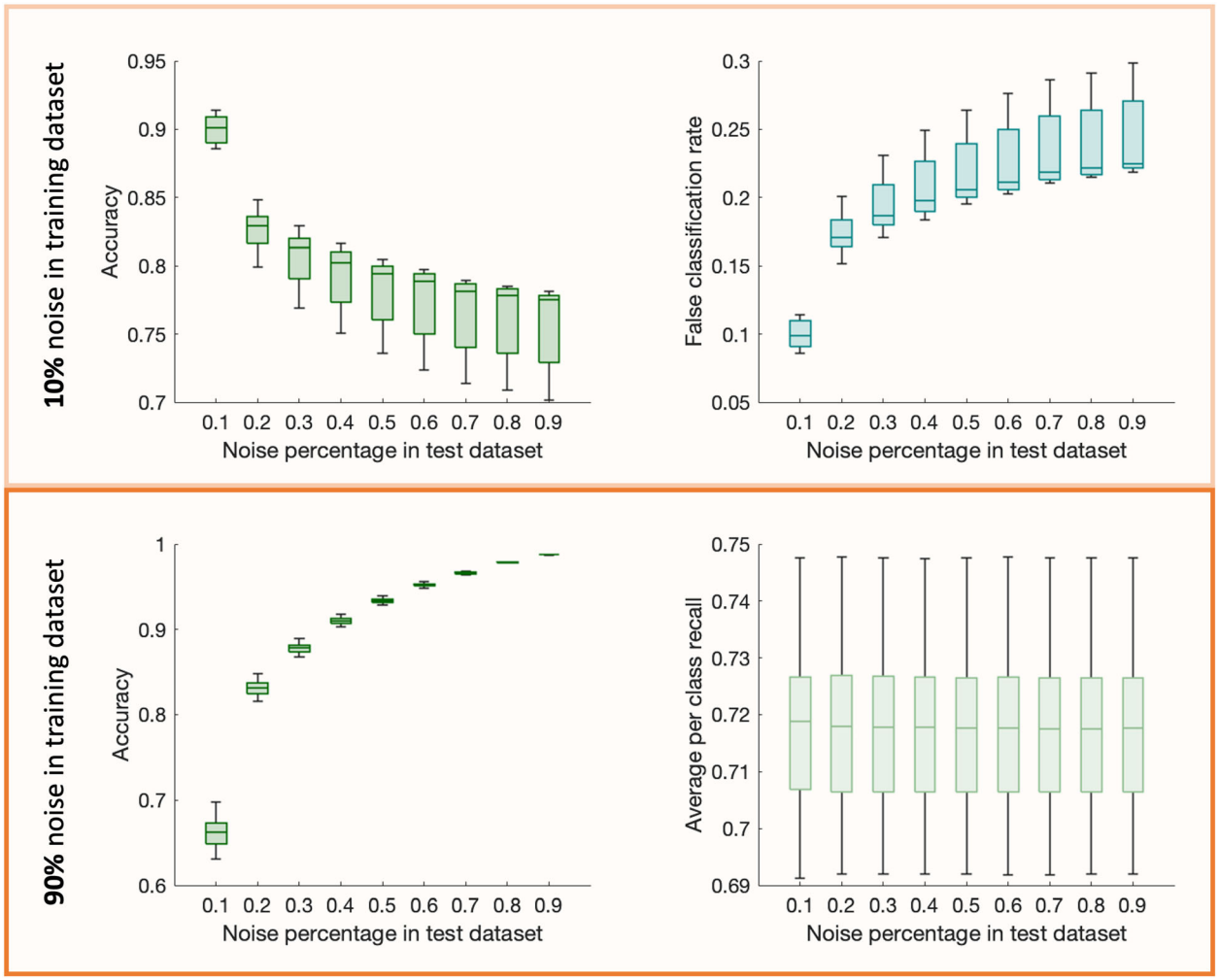

Figure 1. Model evaluation for two different training conditions: 10% noise in the training set versus 90% noise in the training set. Accuracy represents the number of correct predictions divided by the total number of predictions; the false classification rate represents the number of false predictions divided by the total number of predictions and the average of the recalls per class (per class true predictions divided by per class actual positives).

© 2024 The Authors. *Remote Sensing in Ecology and Conservation* published by John Wiley & Sons Ltd on behalf of Zoological Society of London.

20563485, 0, Downloaded from https://zslpublications.onlinelibrary.wiley.com/doi/10.1002/rse2.392 by Bibliotheque de Sorbonne Universite, Wiley Online Library on [30/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Miller et al. (2021) was used for this benchmark. This library consists of 1880.25 h of annotated audio recordings across seven Antarctic sites and 11 site-year combinations, covering multiple months per year.

This dataset was chosen for the benchmark because it represents a multi-class detection scenario for baleen whale vocalizations in a real-world long-term dataset with a realistic imbalance toward the presence of noise with 73–99.9% of the time having only noise presence (average 91%, median 94%). The annotations comprise seven vocalization categories from Antarctic blue whales and fin whales: the Antarctic blue whale Z-call represented by the A-call (A), B-call (B) and the entire Z-call (Z), the Antarctic blue whale D-call (D), the fin whale 40Hz-downsweep (Dswp) and the fin whale 20Hz-pulse represented by the 20Hz-pulse (20Hz) and the 20Hz-pulse plus overtone (20Plus). The representation of these seven vocalization categories within the entire dataset and the different site-year combinations is highly imbalanced (Table 1).

For the usage of this dataset in this benchmark, we consider the possibility to join certain vocalization categories into one category. For example, the Antarctic blue whale 'A', 'B' and 'Z' vocalization categories are all part of the same vocalization, namely the Antarctic blue whale Z-call, and they occur rather as a continuum in the dataset than as separable categories (Fig. 2; Ljungblad et al., 1998; Rankin et al., 2005). Second, the fin whale 20Hz-pulse is represented by both only the 20 Hz component and the 20 Hz component plus its overtone (Fig. 2; Širović et al., 2004). Finally, the Antarctic blue whale D-Call and the fin whale 40Hz-downsweep are two vocalization categories for which differentiation criteria are not fully understood up to date (Fig. 2; Ou et al., 2015) which makes it likely to have confusions between these two categories in the ground truth annotations. Therefore, we consider the possibility to join 'A', 'B' and 'Z', '20Hz' and '20Plus', as well as 'D' and ' Dswp' into each one category, which consequently allows for five different cases of class separation (Table 2). Preliminary model training and following performance analyses (for the description of model setup, see "Experimental model setup", Appendix S1) including all five cases showed a considerable increase in classification performance when joining all three above described (sub-) categories as in case 5 (Table 2). Therefore, all below-described experiments were conducted concerning the detection/classification task of this case and future uses of this benchmark can consider joining categories as in any of the presented cases in order to increase detection feasibility.

Data split

For the purpose of a neutral benchmark, the focus should be on creating a realistic scenario of using published algorithms or even pre-trained models on someone's (newly collected and therefore) unseen data (Weber et al., 2019). The test data for this benchmark should therefore be split in a blocked-cross-validation approach for which each one site-year combination is excluded from the training and validation procedure and only used for the test. This way our benchmark makes sure that different algorithms are tested thoroughly on different independent unseen (in this case 11) datasets. The test dataset from each site-year combination should then be sequenced with a sliding-window approach to imitate a realistic detection task in an unknown dataset. The model evaluation results

Table 1. Dataset composition from Miller et al. (2021) for the 11 site-years and seven vocalization categories ('A', 'B', 'Z', 'D', 'Dswp', '20Hz', '20Plus').

	Vocalization categories						
Site-year	A	B	Z	D	20Hz	20Plus	Dswp
Maud Rise 2014	2191	37	28	70	23	5	6
Greenwich 2015	827	157	29	66	2	1	46
Kerguelen 2005	812	237	166	435	788	78	444
Kerguelen 2014	2557	1177	563	435	1920	1826	344
Kerguelen 2015	1970	542	236	1180	552	718	344
Casey 2014	3681	1398	1091	679	17	0	0
Casey 2017	1741	558	119	553	78	214	0
Ross Sea 2014	104	0	0	0	0	0	0
Balleny Islands 2015	923	44	31	46	951	148	78
Elephant Island 2013	2447	1672	141	10 600	3266	1599	965
Elephant Island 2014	6934	967	100	1034	4940	2912	4077
Total	24 189	6791	2506	15 100	12 539	7503	6306

Note the natural imbalance in the dataset.

© 2024 The Authors. *Remote Sensing in Ecology and Conservation* published by John Wiley & Sons Ltd on behalf of Zoological Society of London.

20563485, 0, Downloaded from https://zslpublications.onlinelibrary.wiley.com/doi/10.1002/rse2.392 by Bibliothèque de Sorbonne Université, Wiley Online Library on [30/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

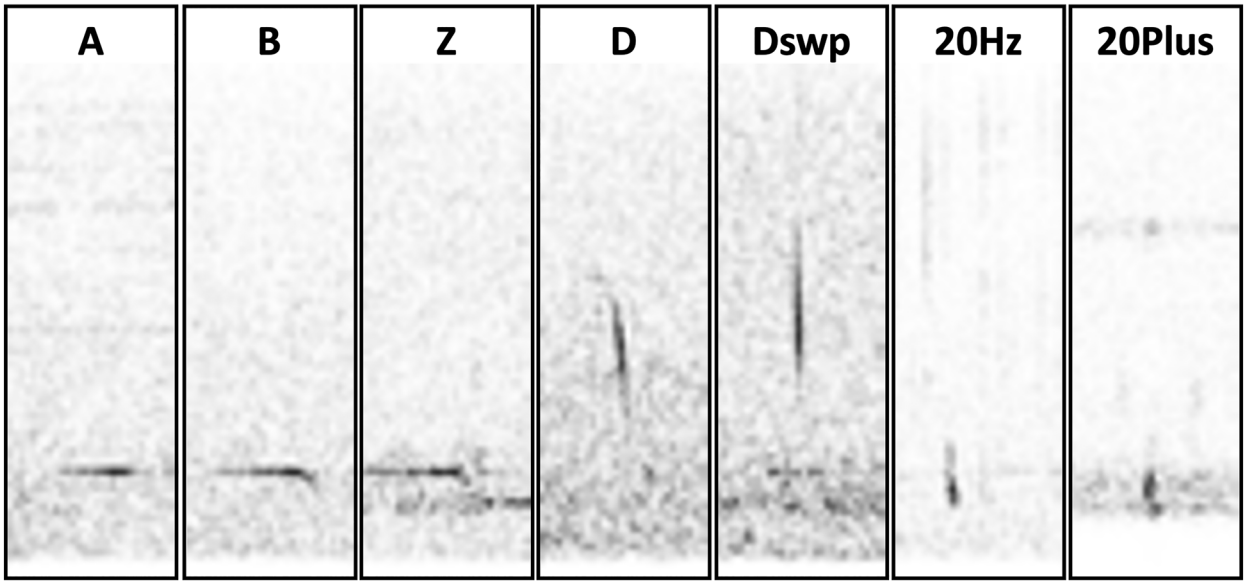

Figure 2. Exemplary spectrograms of the seven vocalization categories. All displayed sounds have a sampling rate of 250 Hz and are bandpass-filtered between 5 and 124 Hz. The spectrograms have 15 s duration and 125-Hz bandwidth and were created with 256 window size, 98% overlap and 3570 FFT (Fast Fourier Transform) size, which is representative of the input to the models in the experiments.

Table 2. Cases for category combinations (original categories: 'A', 'B', 'Z', 'D', 'Dswp', '20Hz', '20Plus', 'Noise').

Case								
1)	A	B	Z	D	Dswp	20Hz	20Plus	Noise
2)	A	B	Z	D	Dswp	20Hz	20Plus	Noise
3)	A	B	Z	D	Dswp	20Hz	20Plus	Noise
4)	A	B	Z	D	Dswp	20Hz	20Plus	Noise
5)	A	B	Z	D	Dswp	20Hz	20Plus	Noise

accordingly have to be reported on all 11 site-year combinations.

The training validation split of the benchmark is up to the model developer. For the three runs presented here, the training and validation sets consisted of a total of 30 000 samples for each of the four classes ('ABZ', '20Hz20Plus', 'DDswp' and 'Noise') which were split randomly into training and validation sets, with slightly different training–validation ratio proportions (we consider these proportions as part of the approaches that are to be compared based on the presented benchmark as different proportions have both advantages and disadvantages).

Model evaluation

For the model evaluation within the framework of this benchmark for baleen whale detection, we want to put emphasis on the reduction of false positives created by noise (e.g., electronic noise, strumming noise, ambient noise caused by wind, rain, sea ice or earthquakes, vocalizations from other species) as this is the biggest challenge when it comes to detecting baleen whale sounds in large passive acoustic datasets. The focus of this benchmark lays on finding solutions for the continuously increasing (large) passive acoustic datasets that require highly

© 2024 The Authors. *Remote Sensing in Ecology and Conservation* published by John Wiley & Sons Ltd on behalf of Zoological Society of London.

20563485, 0, Downloaded from https://zslpublications.onlinelibrary.wiley.com/doi/10.1002/rse2.392 by Bibliothèque de Sorbonne Université, Wiley Online Library on [30/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

efficient automated analysis techniques in order to exploit the full potential of the available data. Therefore, we consider three different metrics for model evaluation: (1) the vocalizations' true classification rate (TCR) that describes the average TCR or the recall from all the vocalization categories (i.e., excluding the Noise category), (2) the noise misclassification rate (NMR) that describes the false-positive rate generated by the noise category for the vocalization categories and (3) the call misclassification rate (CMR) that describes the average misclassification rate of the vocalization categories among each other. True positives, false positives and false negatives are counted from the confusion matrices that are based on the count of all input segments provided to a specific approach. The exemplary confusion matrix in Table 3 and the following calculations show how to obtain the three metrics in the case of the 4-class detection task (i.e., case 5, Table 2):

$$\mathrm{TCR} = \overline{\left(\frac{\mathrm{TP}_1}{\mathrm{ABZ}_{\mathrm{total}}}, \frac{\mathrm{TP}_2}{20\mathrm{Hz20Plus}_{\mathrm{total}}}, \frac{\mathrm{TP}_3}{\mathrm{DDswp}_{\mathrm{total}}}\right)};$$

$$\mathrm{NMR} = \frac{\mathrm{FP}n_1 + \mathrm{FP}n_2 + \mathrm{FP}n_3}{\mathrm{Noise}_{\mathrm{total}}};$$

$$\mathrm{CMR} = \overline{\left(\frac{\mathrm{FPc}_1}{\mathrm{ABZ}_{\mathrm{total}}}, \frac{\mathrm{FPc}_2}{\mathrm{ABZ}_{\mathrm{total}}}, \frac{\mathrm{FPc}_3}{20\mathrm{Hz20Plus}_{\mathrm{total}}}, \frac{\mathrm{FPc}_4}{20\mathrm{Hz20Plus}_{\mathrm{total}}}, \frac{\mathrm{FPc}_5}{\mathrm{DDswp}_{\mathrm{total}}}, \frac{\mathrm{FPc}_6}{\mathrm{DDswp}_{\mathrm{total}}}\right)}.$$

True classification rate has to be maximized and NMR and CMR have to be minimized. As a fourth metric, we combine these three metrics into one fitness metric (F) which will be used for model comparison. We propose the following in order to give more weight to the minimization of false positives created by noise

$$F = \overline{(\mathrm{TCR}, (1-\mathrm{NMR}), (1-\mathrm{NMR}), (1-\mathrm{CMR}))}.$$

The alternative confusion matrices and corresponding formulas for the four metrics in case of an evaluation based on the other four cases listed in Table 2 are provided in Appendix S1 (Case 1, Case 2, Case 3, Case 4).

Experiments

In order to provide the first model performance results alongside this benchmark, we tested three CNN-based classifiers. Two of these CNN models are published models that were designed for bioacoustic detection and classification tasks, namely ANIMAL-SPOT (Bergler et al., 2022) and the Python package Koogu (Madhusudhana, 2022). The third CNN model is a custom-made sequential CNN inspired by AlexNet (Krizhevsky et al., 2017) tuned to handle the current benchmark. All three models were trained, validated and tested on the task of classifying the three joined vocalization categories 'ABZ', '20Hz20Plus' and 'DDswp' and a separate 'Noise' category. Since all three approaches needed different data inputs and slightly different configurations, we will provide the information on how data were preprocessed and training, validation and tests were performed in separate sections below. Details are summarized in Table 4.

ANIMAL-SPOT

For ANIMAL-SPOT, the entire dataset was decimated to 250 Hz, bandpass-filtered between 5 and 124 Hz and then sequenced into 15 s wave files with 12.5 s overlap. If the time limits of a wave file overlapped to 100% with a corresponding vocalization annotation, we assigned the corresponding label, otherwise we assigned the label 'Noise'. If a wave file overlapped with more than one annotation of the different vocalization categories, the wave file got assigned one random choice of the labels from the annotations with overlap. The training-validation ratio was 82:18, which is the default. ANIMAL-SPOT creates spectrograms from all input wav files on the fly for which we set the NFFT (Nonuniform Fast Fourier Transform) size to 256 samples and the hop length to 62 samples, a minimum frequency of 5 Hz and a maximum frequency of

Table 3. Exemplary confusion matrix for the 4-class detection task.

	ABZ	20Hz20Plus	DDswp	Noise	Total
ABZ	TP_1	FPc_1	FPc_2	FN_1	$\mathrm{ABZ}_{\mathrm{total}}$
20Hz20Plus	FPc_3	TP_2	FPc_4	FN_2	$20\mathrm{Hz20Plus}_{\mathrm{total}}$
DDswp	FPc_5	FPc_6	TP_3	FN_3	$\mathrm{DDswp}_{\mathrm{total}}$
Noise	$\mathrm{FP}n_1$	$\mathrm{FP}n_2$	$\mathrm{FP}n_3$	TN	$\mathrm{Noise}_{\mathrm{total}}$

Green cells mark the true positive (TP) counts that are considered in the true classification rate (TCR) calculation, orange cells mark the false positive (FP) counts that are considered in the noise misclassification rate (NMR) calculation, and the yellow cells mark the FP counts that are considered in the call missclassification rate (CMR) calculation.

© 2024 The Authors. *Remote Sensing in Ecology and Conservation* published by John Wiley & Sons Ltd on behalf of Zoological Society of London.

20563485, 0, Downloaded from https://zslpublications.onlinelibrary.wiley.com/doi/10.1002/rse2.392 by Bibliothèque de Sorbonne Université, Wiley Online Library on [30/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 4. Settings for the three CNN-based classifiers tested on the presented benchmark.

Setting	ANIMAL-SPOT	Koogu	Custom CNN
Architecture	ResNet18 with initial kernel size of 7 and initial max-pooling disabled	DenseNet of 4 quasi-dense blocks with 4,4,4 and 2 layers and growth rate of $k = 8$	Sequential CNN inspired by AlexNet
Max training epochs	100	60	100
Batch size	16	64	16
Validation interval	Every 2 epochs	Every 5 epochs	Every epoch
Learning rate	0.00001	0.01	0.0001
Patience learning rate change	8	Epochs 10, 30, 50	10
Learning rate decay factor	0.5	0.1, 0.01, 0.001	0.5
Patience early-stopping	20	NA	20
Adam β1	0.5	0.9	0.9

CNN, convolutional neural network.

124 Hz, minimum/maximum normalization, and an additional linear frequency compression to 90 frequency bins. Additionally, we allowed ANIMAL-SPOT to apply its default data augmentation (i.e., random scaling of frequency, time and intensity; Bergler et al., 2022) within the training dataset. With this respective input, we trained, validated and tested ANIMAL-SPOT for each of the blocked site-year combinations with the default settings recommended in Bergler et al. (2022) (Table 4).

Koogu

Koogu comes with a data sequencer which we used to decimate the data to 250 Hz and then sequence it into 15 s audio windows with 12.5 s overlap which were stored as NumPy arrays as NPZ files. If the time limits of an audio window overlapped to 100% with a corresponding annotation, we assigned the corresponding label, otherwise we assigned the label 'Noise'. If an audio window overlapped with more than one annotation of the different vocalization categories, Koogu assigns multiple labels due to its multi-label training capability. The training–validation ratio was 85:15. As for Koogu, there are no recommended settings, we used the model training and validation specifications described in Miller et al. (2023), where Koogu was configured to detect the Antarctic blue whale D-call in the same dataset used for this benchmark. Koogu also creates input spectrograms on the fly for which we set the window length to 250 samples and the window overlap to 97%, again only considering the frequency range between 5 and 124 Hz. With this respective input, we trained, validated and tested Koogu for each of the blocked site-year combinations with the settings described in Miller et al. (2023) (Table 4). For one of the site-year combinations, we also evaluated the performance of Koogu disabling the multi-label setting in order to see if this would increase performance (but performance was worse than with multi-labeling enabled so that this setting was not further evaluated).

Custom-made sequential CNN

The custom-made sequential CNN has a much simpler architecture than the implemented models from ANIMAL-SPOT and Koogu as can be seen in Figure 3. As input for the model, we created spectrograms from the 15 s wave files with 12.5 s overlap (created for ANIMAL-SPOT with the same corresponding labels) with a window length of 256 samples, a window overlap of 250 samples, a NFFT size of 3570 samples and using the magnitude mode for the short-time Fourier transform, finally displaying the power spectral density. These spectrograms were compressed to 90×30 pixel 8-bit unsigned integer images in grayscale (where the 98% percentile was set to the maximum value for the grayscale in order to avoid loud outliers dominating the coloration). The training–validation ratio was 70:30. With this respective input, we trained, validated and tested the custom CNN for each of the blocked site-year combinations for 100 epochs with a batch size of 16, an initial learning rate of 0.0001, a patience to change the learning rate of 10 epochs, a learning rate decay factor of 0.5 and a patience for early-stopping of 20 epochs (Table 4).

Experimental Results

Three initial experiments have been conducted on the presented benchmark and provide first insights into the performance of available tools for sound detection in marine passive acoustic data. The three CNN models ANIMAL-SPOT, Koogu and the custom CNN were successfully trained to detect the four vocalization categories 'ABZ', '20Hz20Plus' and 'DDswp' in the open access dataset published by Miller et al. (2021) in a blocked

© 2024 The Authors. *Remote Sensing in Ecology and Conservation* published by John Wiley & Sons Ltd on behalf of Zoological Society of London.

20563485, 0, Downloaded from https://zslpublications.onlinelibrary.wiley.com/doi/10.1002/rse2.392 by Bibliothèque de Sorbonne Université, Wiley Online Library on [30/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

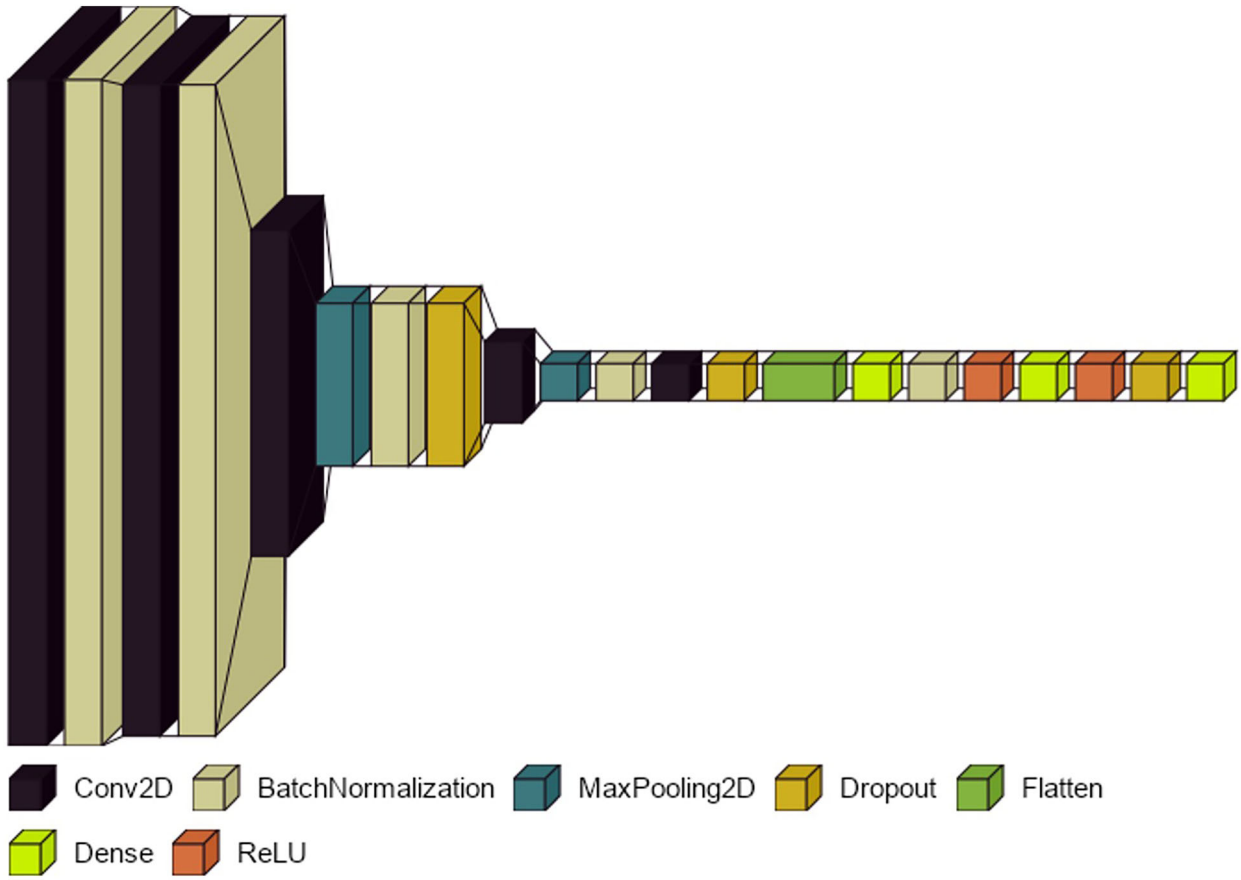

Figure 3. Custom sequential convolutional neural network model architecture.

cross-validation fashion with very diverse performances. The TCR metric was on average highest for the custom CNN (0.73), even though ANIMAL-SPOT achieved the maximum TCR for a single location (i.e., average = 0.67, maximum = 0.91; Fig. 4A; Table 5). Koogu's TCRs were generally quite low with two outlier site-year combinations and its NMRs and CMRs were represented by a wide spread of values (Fig. 4A–C; Table 5). ANIMAL-SPOT and the custom CNN had very similarly good performance in terms of the CMR metric with averages of 0.04 and 0.06, respectively (Fig. 4C; Table 5). Finally, ANIMAL-SPOT's high F metric results from its low average NMR of 0.16, while the custom CNN only achieves an average of 0.24 (Fig. 4B; Table 5). Overall, ANIMAL-SPOT performed the best with an average F metric of 0.6, followed by the custom CNN with an average fitness metric of 0.57 and finally Koogu with an average fitness metric of 0.42 (Fig. 4D; Table 5).

The inspection of the timelines of predictions of the three models in comparison to the timelines of the underlying ground truth data for the four classes '20Hz20Plus', 'ABZ', 'DDswp' and 'Noise' provides insights into the usability of the three models to analyze real-world (long-term) data. The predictions of ANIMAL-SPOT and the custom CNN in most cases resemble closely the real distribution of the four classes within the ground truth (Figure S1). Exceptions are, for example, that ANIMAL-SPOT predicted for Elephant Island 2014 an underrepresented occurrence for the '20Hz20Plus' class and that the custom CNN predicted for Kerguelen 2005 an overrepresented occurrence for the 'DDswp' class. Koogu's predictions, on the other hand, in many cases were biased toward the 'ABZ' or 'DDswp' classes. This translated into high proportions of the time the model predicted baleen whale acoustic presence when actually only noise was present in the ground truth (Figure S1). Exceptions of better resemblance of the ground truth by Koogu were for both years of Elephant Island and the Ross Sea, which represent special cases within the entire dataset due to the high density of vocalizations within the Elephant Island data and very homogeneous noise conditions in the Ross Sea data.

© 2024 The Authors. *Remote Sensing in Ecology and Conservation* published by John Wiley & Sons Ltd on behalf of Zoological Society of London.

20563485, 0, Downloaded from https://zslpublications.onlinelibrary.wiley.com/doi/10.1002/rse2.392 by Bibliothèque de Sorbonne Université, Wiley Online Library on [30/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

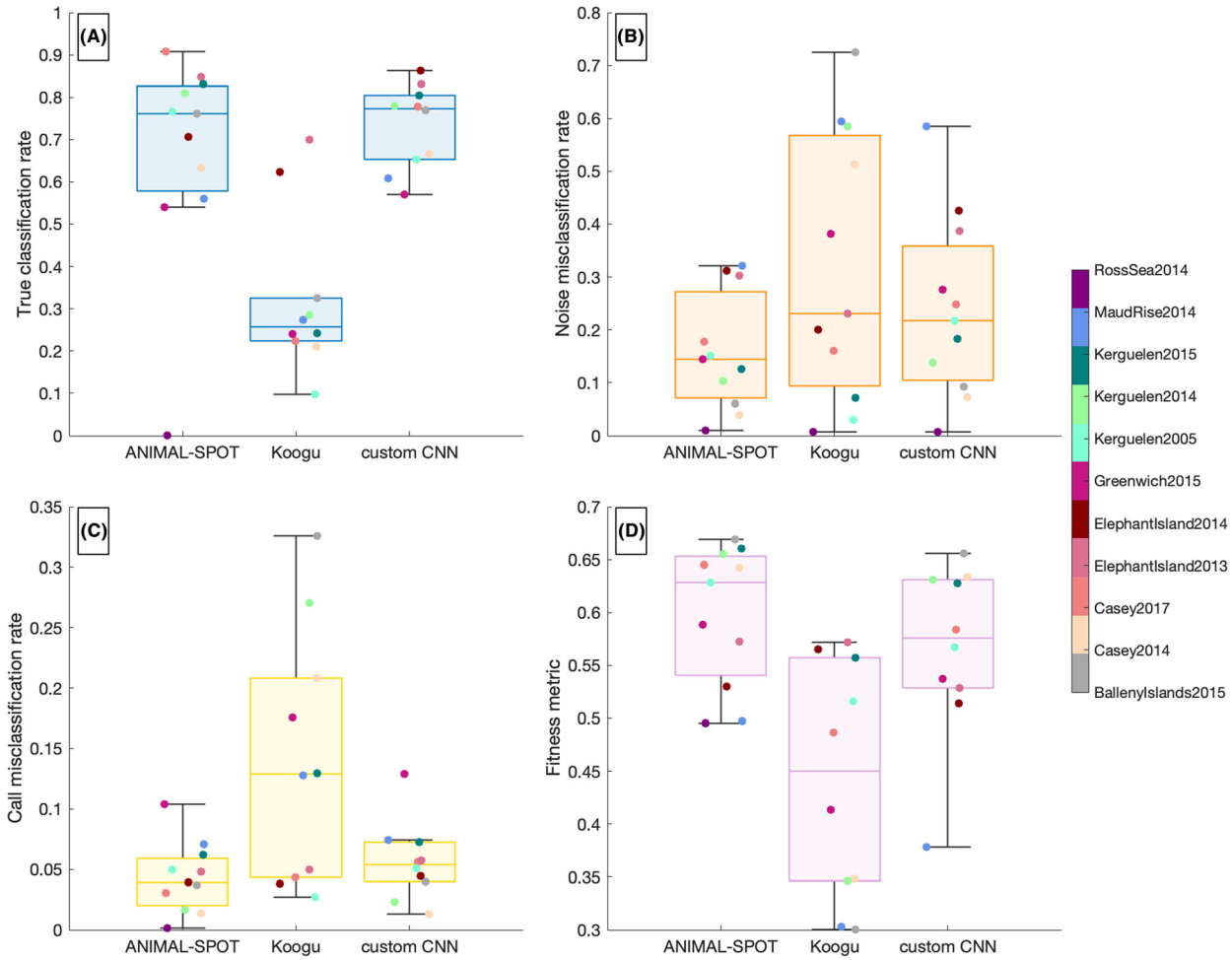

Figure 4. The four benchmark metrics [i.e., (A) true classification rate – TCR, (B) noise misclassification rate – NMR, (C) call misclassification rate – CMR and the (D) overall fitness metric – *F*] for the evaluation of the three models ANIMAL-SPOT, Koogu and the custom CNN in a blocked cross-validation fashion. Boxplots represent the overall distribution of the metrics per model and the colored dots are the single data points (note that only the vertical distribution of the dots has a meaning in terms of the metric value, but the horizontal distribution within each boxplot has been added arbitrarily to enhance visibility). CNN, convolutional neural network.

Table 5. Benchmark results (i.e., the four benchmark metrics *F*, TCR, NMR and CMR) for the three evaluated CNN-based classifiers.

Metric		ANIMAL-SPOT	Koogu	Custom CNN
TCR	Average	0.67	0.34	**0.73**
	Std	0.25	0.19	0.10
NMR	Average	**0.16**	0.38	0.24
	Std	0.11	0.31	0.17
CMR	Average	**0.04**	0.15	0.06
	Std	0.03	0.10	0.03
F	Average	**0.6**	0.42	0.57
	Std	0.07	0.13	0.08

CMR, call misclassification rate; CNN, convolutional neural network; NMR, noise misclassification rate; TCR, true misclassification rate.
Bold values indicate the highest achieved average value among the three tested models for each of the four metrics.

Discussion

In many fields of machine learning such as the classification of images or human speech, much of the recent progress has been driven by published benchmarks (e.g., Mehrish et al., 2023; Russakovsky et al., 2015; Yang et al., 2021). Similarly, the benchmark presented here sets a challenge in order to improve algorithms' performance for the sound detection task in marine long-term PAM data. The three approaches tested here delivered partly (i.e., ANIMAL-SPOT and custom CNN) promising results, while considerable improvement of performance is necessary to ensure the usability of approaches for the analysis of long-term data. Future applications of approaches to this benchmark should aim for better performances (i.e., higher *F* metrics) than presented here.

© 2024 The Authors. *Remote Sensing in Ecology and Conservation* published by John Wiley & Sons Ltd on behalf of Zoological Society of London.

20563485, 0, Downloaded from https://zslpublications.onlinelibrary.wiley.com/doi/10.1002/rse2.392 by Bibliotheque de Sorbonne Universite, Wiley Online Library on [30/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The focus for the development of algorithms to analyze long-term PAM data for the presence of marine animal vocalizations should be first laid on the reduction of the false positives created by noise while the recall of the vocalization classes should be kept reasonably high (i.e., ~0.8 TCR while minimization of the NMR and ultimately maximization of F). This is important to ultimately achieve usability of the developed algorithms without the need for manual post-processing of detections by human experts, which is often too time-consuming. Finally, we suggest to aim at a value of around 1% or lower for the false positives caused by noise (i.e., per vocalization class) in order to guarantee the direct applicability to real-world data for biological and ecological interpretation (Miller et al., 2021; Schall & Parcerisas, 2022; Širović et al., 2004; Thomisch et al., 2016). We are aware that the presented benchmark will, strictly speaking, only monitor the progress in algorithm development for the detection of baleen whale sounds (due to the lack of a published dataset including vocalizations of multiple taxa). However, we believe that approaches tuned for the detection task of vocalizations of a single taxon within a realistically imbalanced dataset, as proposed in this benchmark, will most likely also show comparable performances when re-trained and tested on real-world marine PAM data from other taxa.

Furthermore, this benchmark can be the basis to address additional challenges in the field of marine bioacoustics which go beyond the 'simple' detection of marine animal vocalization presence. Here, four additional challenges shall be named: (1) the separation of the joined vocalization classes such as 'ABZ' in 'A', 'B' and 'Z' and '20Hz20Plus' into '20Hz' and '20Plus', as well as the distinction of spectrally very similar vocalizations such as 'D' and 'Dswp' as the successful separate detection of these classes would provide potentially additional information on the distribution and behavior of the single species (e.g., Gedamke, 2009; Leroy et al., 2016; Oleson et al., 2007); (2) the successful detection of multiple vocalization classes within one time window (multi-labeling or the application of separate recognition tools per vocalization class); (3) the counting of cue rates in order to interpret behavior or estimate population densities (e.g., Schall et al., 2019; Thomas & Marques, 2012); (4) the determination of the exact time-stamps of the vocalizations in order to localize calling individuals from the time-difference-of-arrival with multiple recording devices (e.g., Helble et al., 2015; Warner et al., 2017). In the future, specifically defined tasks on the same dataset used in this benchmark or other similar datasets (representing the natural imbalance among classes, especially the noise class) should be provided as benchmarks to address these challenges.

In general, we hope that the presented benchmark will have an effect of rapid improvement in the field of marine bioacoustics monitoring and in particular for the detection of marine animal sounds in long-term passive acoustic data. We believe that a benchmark as the one described in this publication is necessary to be able to monitor progress and join forces worldwide in order to advance in the automatic processing of the continuously growing amount of PAM data that are collected throughout the world's oceans.

Summary of the Benchmark

In this publication, we present a benchmark for the detection of animal vocalizations in marine PAM data that aids at the facilitation of rapid progress within the field of marine bioacoustics and can be summarized as the following four conditions:

Dataset: The library of annotated circum-Antarctic recordings for Antarctic blue and fin whale vocalizations published in Miller et al. (2021) has to be used for this benchmark as it represents a multi-class detection scenario for baleen whale vocalizations in a real-world long-term dataset with a realistic imbalance toward the presence of noise.

Task: The dataset consists of annotations for in total seven vocalization classes of Antarctic blue and fin whales (i.e., 'A', 'B', 'Z', 'D', 'Dswp', '20Hz', '20Plus'), of which the temporal presence shall be detected within the entire dataset. In order to increase the feasibility of the detection task, we consider the possibility to join certain vocalization categories into one category (i.e., 'ABZ', 'DDswp', '20Hz20Plus').

Data split: For creating a realistic detection scenario the test data should be split in a blocked-cross-validation approach for which each one site-year combination is excluded from the training and validation procedure and only used for the test. The test dataset from each site-year combination should then be sequenced with a sliding-window approach to imitate a realistic detection task in an unknown dataset.

Metrics: Four different metrics should be finally reported for model evaluation: (1) the vocalizations' TCR, (2) the NMR, (3) the CMR and (4) one overall fitness metric (F) that is a combination of the first three metrics.

Acknowledgments

This research was possible thanks to the Research Foundation – Flanders (Belgium) (FWO) under the research grant V509723N. E.S. wants to thank the Flanders Marine Institute (VLIZ) and especially the Marine observation center (MOC) for the hospitality and the constant

© 2024 The Authors. *Remote Sensing in Ecology and Conservation* published by John Wiley & Sons Ltd on behalf of Zoological Society of London.

20563485, 0, Downloaded from https://zslpublications.onlinelibrary.wiley.com/doi/10.1002/rse2.392 by Bibliothèque de Sorbonne Université, Wiley Online Library on [30/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

support for making this project possible. Open Access funding enabled and organized by Projekt DEAL.

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. *Frontiers in Marine Science*, **8**, 165.

Baumgartner, M.F. & Mussoline, S.E. (2011) A generalized baleen whale call detection and classification system. *The Journal of the Acoustical Society of America*, **129**, 2889–2902.

Belghith, E.H., Rioult, F. & Bouzidi, M. (2018) Acoustic diversity classifier for automated marine big data analysis. *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, Volos, Greece, pp. 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. *Scientific Reports*, **9**, 1–17.

Bergler, C., Smeele, S.Q., Tyndel, S.A., Barnhill, A., Ortiz, S.T., Kalan, A.K. et al. (2022) ANIMAL-SPOT enables animal-independent signal detection and classification using deep learning. *Scientific Reports*, **12**, 21966.

Best, P., Ferrari, M., Poupard, M., Paris, S., Marxer, R., Symonds, H. et al. (2020) Deep learning and domain transfer for orca vocalization detection. *International Joint Conference on Neural Networks (IJCNN)*, Glasgow, UK, pp. 1–7.

Best, P., Marxer, R., Paris, S. & Glotin, H. (2022) Temporal evolution of the Mediterranean fin whale song. *Scientific Reports*, **12**, 13565.

Bittle, M. & Duncan, A. (2013) A review of current marine mammal detection and classification algorithms for use in automated passive acoustic monitoring. *Annual Conference of the Australian Acoustical Society 2013, Acoustics 2013: Science, Technology and Amenity*, Victor Harbor, Australia, pp. 208–215.

Bohnenstiehl, D.R. (2023) Automated cataloging of oyster toadfish (*Opsanus tau*) boatwhistle calls using template matching and machine learning. *Ecological Informatics*, **77**, 102268.

Clink, D.J. & Klinck, H. (2019) GIBBONFINDR: an R package for the detection and classification of acoustic signals. *arXiv*. Preprint arXiv:1906.02572.

Dugan, P.J., Zollweg, J., Glotin, H., Popescu, M., Risch, D., LeCun, Y. et al. (2014) High performance computer acoustic data accelerator (HPC-ADA): a new system for exploring marine mammal acoustics for big data applications. *Proceedings of ICML Unsupervised learning for Bioacoustics*, **1**, 1–8.

Gedamke, J. (2009) *Geographic variation in Southern Ocean fin whale song*. Impington: International Whaling Commission.

Halkias, X.C., Paris, S. & Glotin, H. (2013) Classification of mysticete sounds using machine learning techniques. *The Journal of the Acoustical Society of America*, **134**, 3496–3505.

Helble, T.A., Ierley, G.R., Spain, G.L. & Martin, S.W. (2015) Automated acoustic localization and call association for vocalizing humpback whales on the Navy's Pacific Missile Range Facility. *The Journal of the Acoustical Society of America*, **137**, 11–21.

Johnson, J.M. & Khoshgoftaar, T.M. (2019) Survey on deep learning with class imbalance. *Journal of Big Data*, **6**, 1–54.

K. Lisa Yang Center for Conservation Bioacoustics. (2023) *Raven pro: interactive sound analysis software*. Version 1.6.5. Ithaca, NY: The Cornell Lab of Ornithology.

Kirsebom, O.S., Frazao, F., Padovese, B., Sakib, S. & Matwin, S. (2021) Ketos—a deep learning package for creating acoustic detectors and classifiers. *The Journal of the Acoustical Society of America*, **150**, A164.

Kirsebom, O.S., Frazao, F., Simard, Y., Roy, N., Matwin, S. & Giard, S. (2020) Performance of a deep neural network at detecting North Atlantic right whale upcalls. *The Journal of the Acoustical Society of America*, **147**, 2636–2646.

Kowarski, K.A. & Moors-Murphy, H. (2021) A review of big data analysis methods for baleen whale passive acoustic monitoring. *Marine Mammal Science*, **37**, 652–673.

Krizhevsky, A., Sutskever, I. & Hinton, G.E. (2017) ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, **60**, 84–90.

Lammers, M.O., Goodwin, B., Kügler, A., Zang, E.J., Harvey, M., Margolina, T. et al. (2023) The occurrence of humpback whales across the Hawaiian archipelago revealed by fixed and mobile acoustic monitoring. *Frontiers in Marine Science*, **10**, 1083583.

Lapp, S., Rhinehart, T., Freeland-Haynes, L., Khilnani, J., Syunkova, A. & Kitzes, J. (2023) OpenSoundscape: an open-source bioacoustics analysis package for Python. *Methods in Ecology and Evolution*, **14**, 2321–2328.

Leroy, E.C., Samaran, F., Bonnel, J. & Royer, J.-Y. (2016) Seasonal and diel vocalization patterns of Antarctic blue whale (*Balaenoptera musculus intermedia*) in the Southern Indian Ocean: a multi-year and multi-site study. *PLoS One*, **11**, e0163587.

Ljungblad, D.K., Clark, C.W. & Shimada, H. (1998) A comparison of sounds attributed to pygmy blue whales (*Balaenoptera musculus brevicauda*) recorded south of the Madagascar Plateau and those attributed to 'true' blue whales (*Balaenoptera musculus*) recorded off Antarctica. *Reports of the International Whaling Commission*, **48**, 439–442.

Madhusudhana, S. (2022) Shyamblast/Koogu: V0.7.1. *Zenodo*.

Madhusudhana, S., Shiu, Y., Klinck, H., Fleishman, E., Liu, X., Nosal, E.-M. et al. (2021) Improve automatic detection of animal call sequences with temporal context. *The Journal of the Royal Society Interface*, **18**, 20210297.

© 2024 The Authors. *Remote Sensing in Ecology and Conservation* published by John Wiley & Sons Ltd on behalf of Zoological Society of London.

20563485, 0, Downloaded from https://zslpublications.onlinelibrary.wiley.com/doi/10.1002/rse2.392 by Bibliotheque de Sorbonne Universite, Wiley Online Library on [30/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Marsland, S., Priyadarshani, N., Juodakis, J. & Castro, I. (2019) AviaNZ: a future-proofed program for annotation and recognition of animal sounds in long-time field recordings. *Methods in Ecology and Evolution*, **10**, 1189–1195.

Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R. & Poria, S. (2023) A review of deep learning techniques for speech processing. *Information Fusion*, **99**, 101869.

Miller, B.S., Balcazar, N., Nieukirk, S., Leroy, E.C., Aulich, M., Shabangu, F.W. et al. (2021) An open access dataset for developing automated detectors of Antarctic baleen whale sounds and performance evaluation of two commonly used detectors. *Scientific Reports*, **11**, 1–18.

Miller, B.S., Madhusudhana, S., Aulich, M.G. & Kelly, N. (2023) Deep learning algorithm outperforms experienced human observer at detection of blue whale D-calls: a double-observer analysis. *Remote Sensing in Ecology and Conservation*, **9**, 104–116.

Oleson, E.M., Calambokidis, J., Burgess, W.C., McDonald, M.A., LeDuc, C.A. & Hildebrand, J.A. (2007) Behavioral context of call production by eastern North Pacific blue whales. *Marine Ecology Progress Series*, **330**, 269–284.

Ou, H., Au, W.W.L., Van Parijs, S., Oleson, E.M. & Rankin, S. (2015) Discrimination of frequency-modulated baleen whale downsweep calls with overlapping frequencies. *Journal of the Acoustical Society of America*, **137**, 3024–3032.

Poupard, M., Symonds, H., Spong, P. & Glotin, H. (2021) Intra-group orca call rate modulation estimation using compact four hydrophones array. *Frontiers in Marine Science*, **8**, 681036.

Rankin, S., Ljungblad, D., Clark, C. & Kato, H. (2005) Vocalizations of Antarctic blue whales, *Balaenoptera musculus intermedia*, recorded during the 2001/2002 and 2002/2003 IWC/SOWER circumpolar cruises, area V, Antarctica. *Journal of Cetacean Research and Management*, **7**, 13–20.

Rasmussen, J.H. & Širović, A. (2021) Automatic detection and classification of baleen whale social calls using convolutional neural networks. *The Journal of the Acoustical Society of America*, **149**, 3635–3644.

Roca, I.T. & Van Opzeeland, I. (2019) Using acoustic metrics to characterize underwater acoustic biodiversity in the Southern Ocean. *Remote Sensing in Ecology and Conservation*, **6**, 262–273.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S. et al. (2015) Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, **115**, 211–252.

Rycyk, A., Bolaji, D.A., Factheu, C. & Kamla Takoukam, A. (2022) Using transfer learning with a convolutional neural network to detect African manatee (*Trichechus senegalensis*) vocalizations. *JASA Express Letters*, **2**, 121201.

Schall, E., Di Iorio, L., Berchok, C., Filún, D., Bedriñana-Romano, L., Buchan, S.J. et al. (2019) Visual and passive acoustic observations of blue whale trios from two distinct populations. *Marine Mammal Science*, **36**, 365–374.

Schall, E. & Parcerisas, C. (2022) A robust method to automatically detect fin whale acoustic presence in large and diverse passive acoustic datasets. *Journal of Marine Science and Engineering*, **10**, 1831.

Schall, E., Thomisch, K., Boebel, O., Gerlach, G., Mangia Woods, S., El-Gabbas, A. et al. (2021) Humpback whale acoustic presence in the Atlantic sector of the Southern Ocean. *Dryad*. Available from: https://doi.org/10.5061/dryad.ncjsxkss0

Shiu, Y., Palmer, K., Roch, M.A., Fleishman, E., Liu, X., Nosal, E.-M. et al. (2020) Deep neural networks for automated detection of marine mammal species. *Scientific Reports*, **10**, 607.

Silva, B., Mestre, F., Barreiro, S., Alves, P.J. & Herrera, J.M. (2022) soundClass: an automatic sound classification tool for biodiversity monitoring using machine learning. *Methods in Ecology and Evolution*, **13**, 2356–2362.

Širović, A., Hildebrand, J.A., Wiggins, S.M., McDonald, M.A., Moore, S.E. & Thiele, D. (2004) Seasonality of blue and fin whale calls and the influence of sea ice in the Western Antarctic Peninsula. *Deep Sea Research Part II: Topical Studies in Oceanography*, **51**, 2327–2344.

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

Stowell, D., Wood, M.D., Pamuła, H., Stylianou, Y. & Glotin, H. (2019) Automatic acoustic detection of birds through deep learning: the first Bird Audio Detection challenge. *Methods in Ecology and Evolution*, **10**, 368–380.

Thomas, L. & Marques, T.A. (2012) Passive acoustic monitoring for estimating animal density. *Acoustics Today*, **8**, 35–44.

Thomisch, K., Boebel, O., Clark, C.W., Hagen, W., Spiesecke, S., Zitterbart, D.P. et al. (2016) Spatio-temporal patterns in acoustic presence and distribution of Antarctic blue whales *Balaenoptera musculus intermedia* in the Weddell Sea. *Endangered Species Research*, **30**, 239–253.

Usman, A.M., Ogundile, O.O. & Versfeld, D.J. (2020) Review of automatic detection and classification techniques for cetacean vocalization. *IEEE Access*, **8**, 105181–105206.

Vickers, W., Milner, B., Risch, D. & Lee, R. (2021) Robust North Atlantic right whale detection using deep learning models for denoising. *The Journal of the Acoustical Society of America*, **149**, 3797–3812.

Warner, G.A., Dosso, S.E. & Hannay, D.E. (2017) Bowhead whale localization using time-difference-of-arrival data from asynchronous recorders. *The Journal of the Acoustical Society of America*, **141**, 1921–1935.

Weber, L.M., Saelens, W., Cannoodt, R., Soneson, C., Hapfelmeier, A., Gardner, P.P. et al. (2019) Essential guidelines for computational method benchmarking. *Genome Biology*, **20**, 1–12.

© 2024 The Authors. *Remote Sensing in Ecology and Conservation* published by John Wiley & Sons Ltd on behalf of Zoological Society of London.

20563485, 0, Downloaded from https://zslpublications.onlinelibrary.wiley.com/doi/10.1002/rse2.392 by Bibliotheque de Sorbonne Universite, Wiley Online Library on [30/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

White, E.L., White, P.R., Bull, J.M., Risch, D., Beck, S. & Edwards, E.W. (2022) More than a whistle: automated detection of marine sound sources with a convolutional neural network. *Frontiers in Marine Science*, **9**, 879145.

Yang, S.-w., Chi, P.-H., Chuang, Y.-S., Lai, C.-I.J., Lakhotia, K., Lin, Y.Y. et al. (2021) Superb: speech processing universal performance benchmark. *arXiv*. Preprint arXiv:2105.01051.

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. *The Journal of the Acoustical Society of America*, **147**, 1834–1841.

Zhong, M., Torterotot, M., Branch, T.A., Stafford, K.M., Royer, J.-Y., Dodhia, R. et al. (2021) Detecting, classifying, and counting blue whale calls with Siamese neural networks. *The Journal of the Acoustical Society of America*, **149**, 3086–3094.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix S1. Experimental model set up.

Figure S1. The timelines of predictions from the three models (ANIMAL-SPOT, Koogu, and the custom CNN) and of the ground truth for the 11 site-year combinations and each of the four classes ('20Hz20Plus', 'ABZ', ' DDswp', and 'Noise').

© 2024 The Authors. *Remote Sensing in Ecology and Conservation* published by John Wiley & Sons Ltd on behalf of Zoological Society of London.

20563485, 0, Downloaded from https://zslpublications.onlinelibrary.wiley.com/doi/10.1002/rse2.392 by Bibliothèque de Sorbonne Université, Wiley Online Library on [30/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License