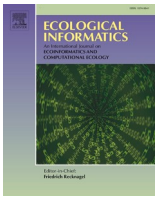
Ecological Informatics 84 (2024) 102906

Contents lists available at ScienceDirect

Ecological Informatics

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



A deep learning model for detecting and classifying multiple marine mammal species from passive acoustic data

Quentin Hamard [a], Minh-Tan Pham [b], Dorian Cazau [c], Karine Heerah [a,*]

[a] France Energies Marines, 29280 Plouzané, France
[b] Université Bretagne Sud, IRISA, UMR 6074, 56000 Vannes, France
[c] ENSTA Bretagne, Lab-STICC, UMR CNRS 6285, 29200 Brest, France

ARTICLE INFO

Keywords:
Ecoacoustics
Object detection
Offshore wind farm
Passive acoustic monitoring
Marine mammal

ABSTRACT

Underwater passive acoustics is used worldwide for multi-year monitoring of marine mammals. Yet, the large amount of audio recordings raises the need to automate the detection of acoustic events. For instance, the increasing number of Offshore Wind Farms (OWF) raises key environmental and societal issues relating to their impacts on wildlife. In this context, monitoring marine mammals along with information on their acoustic environment throughout the OWF life cycle is crucial. The objective of this study is to evaluate the ability of a single deep learning model to precisely detect and localize, in time and in frequency, the marine mammal sounds over a wide frequency range and classify them by species and sound types.

A broadband hydrophone, deployed at the Fécamp OWF (Normandy, France), recorded the underwater soundscape including sounds from marine mammals occurring in the area. To visualize these sounds, 15-s spectrograms were computed. From these images, dolphin (D) and porpoise (P) sounds were manually annotated, including different types of sounds: Click-Trains (D_{CT}, P_{CT}), Buzzes (D_B, P_B) and Whistles (D_W). The spectrograms were then split into five-fold cross-validation datasets, each containing one half of manual annotations and one half of only background noise. A Faster R-CNN model was trained to precisely detect and classify the marine mammal sounds in the spectrograms.

Three model output configurations were used: (1) overall detection of marine mammals (presence vs. absence), (2) detection and classification of species (two classes: dolphin, porpoise) and (3) sound types (five classes: D_{CT}, D_B, D_W, P_{CT}, P_B). For the simplest configuration (1) 15.4 % of the spectrogram dataset had detections while missing only 6.6 % of annotated spectrograms. For the more precise configurations, (2) and (3), the mean Average Precision (mAP) achieved were 92.3 % (2) and 84.3 % (3), and the macro average Area under the curve (AUC) 95.7 % (2) and 94.9 % (3).

This model will help to speed up the annotation processes, by reducing the spectrogram quantity to be manually analyzed and having time-frequency boxes already drawn. Several model parameters can be adjusted to trade off missed detections and false positives which need to be carefully considered and adapted to the problem. For instance, these adjustments would be particularly relevant depending on the human resources available to manually check the model detections and the criticality of missing marine mammal sounds. These models are promising, ranging from the simple detection of marine mammal presence to precise ecological inferences over the long term.

1. Introduction

The modification of marine habitats resulting from offshore human activities raises key environmental issues relating to their impact on wildlife. Marine megafauna species, such as marine mammals, are influenced by anthropogenic activities, in their distribution, abundance and behavior in relation to potential direct effects – including mortalities, injuries caused by fisheries bycatch, collisions with ships, pollutants - but also a large array of underlying ecological effects - including loss of habitat, acoustic disturbances, changes in prey distribution and availability - which can ultimately affect population dynamics. For instance, the development of offshore wind farms generates effects and

* Corresponding author.
E-mail address: karine.heerah@france-energies-marines.org (K. Heerah).

https://doi.org/10.1016/j.ecoinf.2024.102906
Received 16 January 2024; Received in revised form 17 October 2024; Accepted 18 November 2024
Available online 22 November 2024

raises concerns about potential impacts on the environment, which need to be assessed. The construction or operation of an offshore wind farm (i.e., activity) increases ambient noise (i.e., effect) which could, for marine mammals (i.e., ecological receptor), mask biological communication signals, such as the echolocation used to locate and forage (i.e., impact). These effects and impacts need to be identified and quantified through comprehensive and long-term monitoring as required by the Marine Strategy Framework Directive (MSFD) to ensure the good ecological status of European waters. Marine mammals are at the heart of concerns considering their environmental importance. These species represent the high end of food webs and play key roles in ecosystem functioning (Mariano-Jelicich et al., 2021). Their ecologies and distributions integrate spatiotemporal variations of their food resources. Thus, such species are generally considered as ecological indicators of ecosystem health (Hazen et al., 2019).

While long-term monitoring of marine mammals is crucial to capture the drivers of potential changes in population dynamics, it remains particularly challenging to observe them as they live in the open ocean, where they can cover large distances and spend most of their time underwater. Several methods have been developed to study their distribution at sea (Frasier et al., 2021; Hammond et al., 2021), habitat use (Lambert et al., 2017; Virgili et al., 2017), foraging habits (Bowen and Iverson, 2013) and behavior at sea (Todd et al., 2020). Each method has its advantages and drawbacks. For instance, studies on the distribution and abundance of marine megafauna species are traditionally based on direct observation data obtained from aerial and/or ship-based surveys (Hammond et al., 2021; Waggitt et al., 2019). While these methods offer the opportunity to sample large areas, they only provide a short-term snapshot of the ecological scene and thus, do not fully integrate the temporal variability in species abundance and distribution (Virgili et al., 2018), driven either by natural environmental variability and/or by human activities. Biologging of marine mammals equipped with movement recording tags allows for mid/long term monitoring of individuals at large spatial and temporal scales. While biologging methods offer invaluable information on the movements and behavior of tagged individuals, their efficiency to accurately characterize species ecology is drastically dependent on sample sizes (Sequeira et al., 2019). Underwater, passive acoustic monitoring from a fixed point (i.e., restricted spatial scale) is another approach used to study marine mammals (Mellinger et al., 2007). It is classically used for mid−/long-term monitoring of species emitting an acoustic signal (Gervaise et al., 2021; Mellinger et al., 2007; Nowacek et al., 2016), used to locate, forage and/or interact with each other. Each species possesses specific acoustic signatures and a specific vocal repertoire associated with different activities. As such, passive acoustic monitoring can be used to identify the presence of a given species in an area and to characterize its behavior. Moreover, this solution is able to characterize the environment by measuring the ambient noise.

The analysis of long-term acoustic recordings can be performed manually by labelling sound events. Advances in technologies, including energy and storage capacities, mean that acoustic data can be collected over longer durations with higher frequency sampling (up to more than 300 kHz). However, these represent large and complex datasets, rendering manual analyses time-consuming. To address this challenge and achieve cost-effective processing, alternative methods have emerged. From acoustic signal processing to artificial intelligence, these methods have proved their effectiveness over the last decades in automatically analyzing large amounts of audio data for acoustic environment monitoring (Morgan and Braasch, 2021; Towsey et al., 2014). More specifically for underwater bioacoustics, solutions based on signal processing such as the C-pod and F-pod detectors (Todd et al., 2023) or the Pamguard software (Gillespie et al., 2009) are used in particular to detect marine mammal clicks. However, these algorithms have shown a certain sensitivity to background noise (Clausen et al., 2019) and a lack of diversity in the sound types and species detected and classified. More recently, the use of artificial intelligence (machine learning, and especially deep learning), has demonstrated its efficiency across various domains, including image and speech processing (LeCun et al., 2015; Shinde and Shah, 2018). With the significant breakthrough of deep learning, the automation of monitoring processes in bioacoustics has become a realistic objective (Goodwin et al., 2022; Parsons et al., 2022). Indeed, deep learning has proven essential in making the wealth of available data beneficial for learning how to extract relevant features from the sensed information (Heaton et al., 2018). Within the context of marine fauna monitoring, deep learning has been adopted to tackle various tasks including the automatic identification of fishes from echosounding data (Brautaset et al., 2020), the classification of marine mammal sounds from passive acoustic recordings (Mutanu et al., 2022; Shiu et al., 2020), and automatic object recognition and tracking in underwater videos (Beyan and Browman, 2020; Malde et al., 2020).

In recent reviews, Mutanu et al. (2022) and Stowell (2022) discussed the current state-of-the-art methods employed for detection and classification in the field of bioacoustics. Typically, detection in bioacoustics involves a binary classification of spectrograms generated from acoustic signals to determine the presence or absence of target sounds. Subsequently, a more detailed classification task is applied to positive spectrograms, involving multiple classes (i.e., one label is affected per spectrogram) and, in some cases, multiple labels (i.e. multiple labels can be affected per spectrogram). Traditional machine learning approaches, such as Support Vector Machines (SVM), Bayesian Hidden Markov and Random Forest can be trained with relatively small datasets. However, due to their limited use in only a few studies, their capabilities cannot be assessed in a wider range of scenarios (Mutanu et al., 2022). On the other hand, deep learning models, particularly Convolutional Neural Networks (CNNs) have emerged as the predominant algorithms for acoustic classification, mainly due to their ability to offer superior performance when dealing with raw data or spectrograms (Shiu et al., 2020; Stowell, 2022). Nevertheless, they necessitate a substantial amount of labeled acoustic data for effective training. For detection tasks, recent deep learning-based methods have introduced a new dimension to the field, allowing for precise localization and classification of sound events in time and in frequency. Popular object detection models, such as Faster R-CNN or YOLO architectures, have already been applied in wireless signal detection (Prasad et al., 2020) and temporal localization of sound events (Pham et al., 2018). Specifically, bioacoustics studies have used these models to detect single-class vocalizations (Coffey et al., 2019; Ferguson et al., 2022; Romero Mujalli et al., 2021) and multispecies or multitype sounds in a low frequency range (Escobar-Amado et al., 2024, Wu et al., 2021). Such approaches are particularly useful in scenarios where occurrences of sounds are overlapping or when different sound classes are present. Despite being demanding in the annotation process, this is a promising method to precisely detect and localize (in time and in frequency) several marine mammal species and the different sound types of their broadband vocal repertoire using a single deep learning model.

In this study, we exploited the Faster R-CNN model to detect five sound types of two marine mammal species in broadband underwater audio recorded in an OWF: dolphin click-train, buzz and whistle, and porpoise click-train and buzz. The precise localization (in time and in frequency) and classification of these sounds in the spectrograms enabled us to aggregate the detections at the spectrogram scale and evaluate our model at several precision levels: marine mammal (presence/absence), species (dolphin/porpoise) and sound type (dolphin click-train, buzz and whistle, and porpoise click-train and buzz). Furthermore, our study has confirmed the model's ability to expedite the annotation process compared to a manual annotation, while effectively detecting sparse marine mammal sounds. This was achieved by minimizing the number of spectrograms requiring manual analysis and providing pre-drawn time-frequency boxes.

The remainder of the paper is organized as follows. Section 2 first describes the data collection and processing, as well as the dataset preparation. Then, details about the selection, configuration and

evaluation of our deep learning model are presented. Section 3 provides the experimental results including the analysis of the trade-off between false positives and missed detections, the performance of the three classification levels and the results on the generalization dataset. Next, Section 4 provides further discussions with an assessment of the model's performance and relevance in the context of marine mammal monitoring. In addition, potential future directions for research and development are described, focusing on strategies to improve input data and the integration of information from various sensors to enhance monitoring capabilities. Finally, we conclude the paper in Section 5.

2. Material and methods

2.1. Data collection

Underwater acoustic data were recorded at the Fécamp (Normandy, France) offshore wind farm (OWF) with instruments deployed at the south-east end of the OWF (49°51.3759′N, 0°14.2259′E) (Fig. 1). In this area of the English Channel, permanent species include dolphins (Bottlenose dolphin (*Tursiops truncates*), Common dolphin (*Delphinus delphis*), White-beaked dolphin (*Lagenorhyncus albirostris*)), porpoises (Harbor porpoise (*Phocoena Phocoena*)) and seals (Gray seal (*Halichoerus grypus*), Harbor seal (*Phoca vitulina*)). Other species can occur in this area but are considered rare. Very few seal sounds were present in the recordings, therefore this study focused on sounds from dolphins and porpoises. Dolphins emit whistles (frequency modulation around 1–40 kHz) and broadband clicks and buzzes (pulses up to 150 kHz) (Jones et al., 2019). Porpoises only emit clicks and buzzes (pulses around 100–150 kHz) (Verboom and Kastelein, 1995). Each sound type can be associated with a behavior as dolphins and porpoises use click-trains and buzzes mainly to echolocate and feed, respectively. In addition, dolphins use whistles to communicate (Dudzinski et al., 2009).

To record the sounds emitted by these species, the instrumentation included a broadband hydrophone (HTI-99, Sensitivity: −164 dB) and a recorder able to record the soundscape up to 156 kHz (Rtsys – EA-SD14, Fs: 312.5 kHz, Bit depth: 24 bits, Bandwidth: 3 Hz-150 kHz, Gain: 14.7

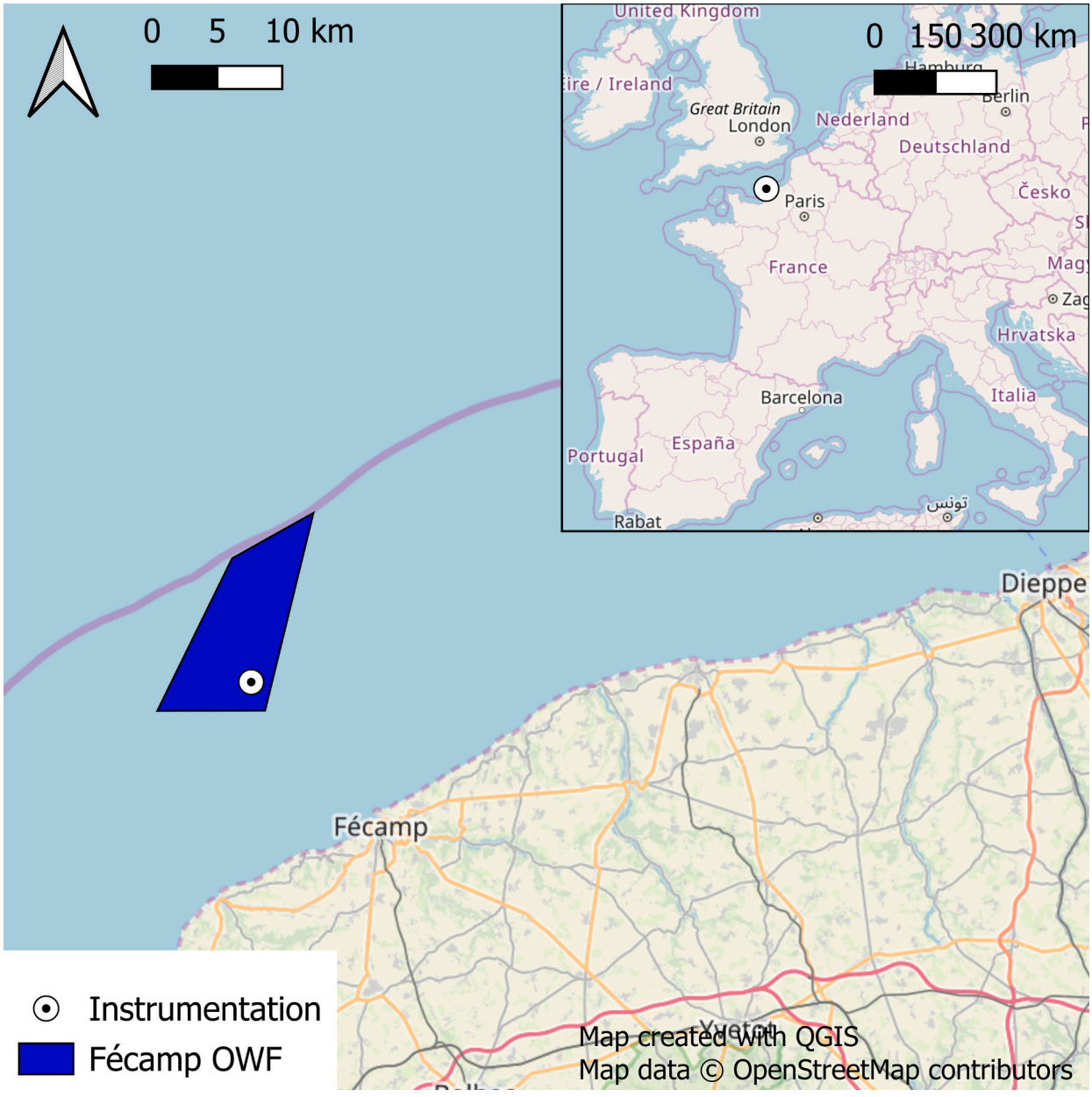


Fig. 1. Location of the Fécamp OWF and instrumentation.

dB, V_{ADC}=V_{PP}/2 = 2.5 V). To save energy and storage space, recording occurred 50 % of the time on a duty cycle of 5 min on/5 min off. The instrumentation was attached to a mooring line consisting of a stationary chain (on the seabed), a tidal chain (in the water column) and a buoy (at the surface). The hydrophone was placed along the stationary chain a few meters above the seabed (bathymetry of around 30 m depth). The marine mammal sounds could be detected if the instrumentation is less than a few hundred meters or a few kilometers from the sound source (depending on the ambient noise level and the type of signal and its characteristics: orientation, amplitude) (Nuuttila et al., 2013; Nuuttila et al., 2018; Quintana-Rizzo et al., 2006).

Two sound recording datasets were available. Both were recorded before the beginning of the construction works of the OWF: one in November/December 2020, hereafter referred to as Dataset 1 (305 h, 961 Go) and, one in January/February 2020, hereafter referred to as Dataset 2 (288 h, 905 Go). Dataset 1 was manually annotated (§2.2 Data processing) and used for training and validation of the object detection model. For Dataset 2, five consecutive days were manually annotated and used to assess the generalization of model performances.

2.2. Data processing and dataset preparation

Fig. 2 provides an overview of the workflow for data processing, model training and evaluation.

2.2.1. Sound processing

Firstly, from the raw acoustic data (.wav files), sounds were transformed into 15-s-long spectrograms to visualize their frequency components and thus marine mammal vocalizations. Two different sets of resolution and frequency intervals were used to compute the spectrograms, resulting in: (i) one broadband spectrogram (0–156 kHz; resolution: 8.3 ms, 305.2 Hz) and; (ii) one low frequency spectrogram (0–25 kHz; resolution: 8.3 ms, 48.8 Hz) to focus on low frequency vocalizations, such as dolphin whistles (FFT size: 2048, window size: 2048 (1024 for low frequency spectrogram), overlap: 50 % (70 % for low frequency spectrogram), type: power spectrum density). Secondly, each

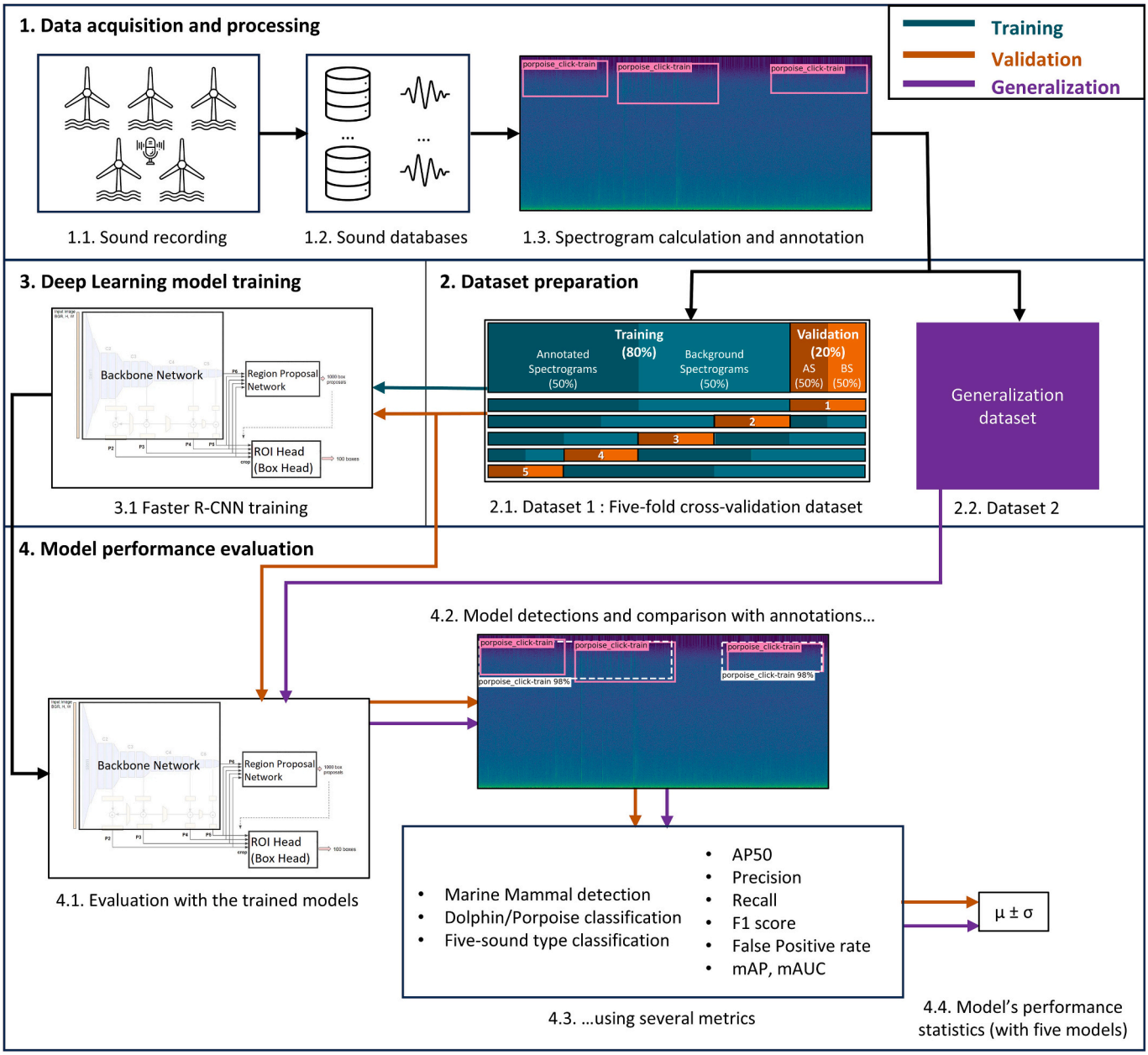


Fig. 2. Workflow of the marine mammal sound detection and classification model. Icons from www.flaticon.com, Faster R-CNN architecture from (Honda, 2020).

spectrogram matrix was converted into an image to reduce spectrogram size (color value range: 0–120 dB, color: viridis, image size (width, height): 1813, 512), allowing faster model training. Examples of spectrograms are shown in Fig. 3. Spectrograms and images were generated using the open source OSmOSE python package (https://github.com/Project-OSmOSE/osmose-toolkit).

2.2.2. Annotation

The entire datasets 1 and 2 were manually annotated by an expert using the temporal and spectrogram visualizations of the Raven Pro 1.6 software in order to identify marine mammal sounds in the recordings. The resulting time-frequency boxes drawn around the sounds of interest were then transferred to our 15-s spectrograms generated with the OSmOSE pyhton package to obtain standardized spectrograms. Each box was annotated with one of these labels: dolphin click-train, dolphin buzz, dolphin whistle, porpoise click-train and porpoise buzz. In Dataset 1, 4288 sounds were annotated, including: 2599 dolphin sounds (2028 click-trains, 254 buzzes and 317 whistles) and 1689 porpoise sounds (1613 click-trains and 76 buzzes). Buzzes and whistles were the least represented in the dataset, especially porpoise buzzes. Out of 305 h of sound recordings, annotations represented around 105 min (0.6 % of the time). In terms of 15-s spectrograms, 1959 segments included annotation(s) out of 75864 segments. Each annotation was associated with a certainty level, ranging from low (1) to high (3). For a high certainty level, there was no doubt about the origin of the signal. For a medium certainty level, it was not possible to determine its origin with certainty. Finally, for a low certainty level, the signal resembled a signal of interest but did not appear to be emitted by a biological source. The average certainty level of all annotations was 2.67 (SD: 0.54).

In Dataset 2415 sounds were annotated, including: 137 dolphin

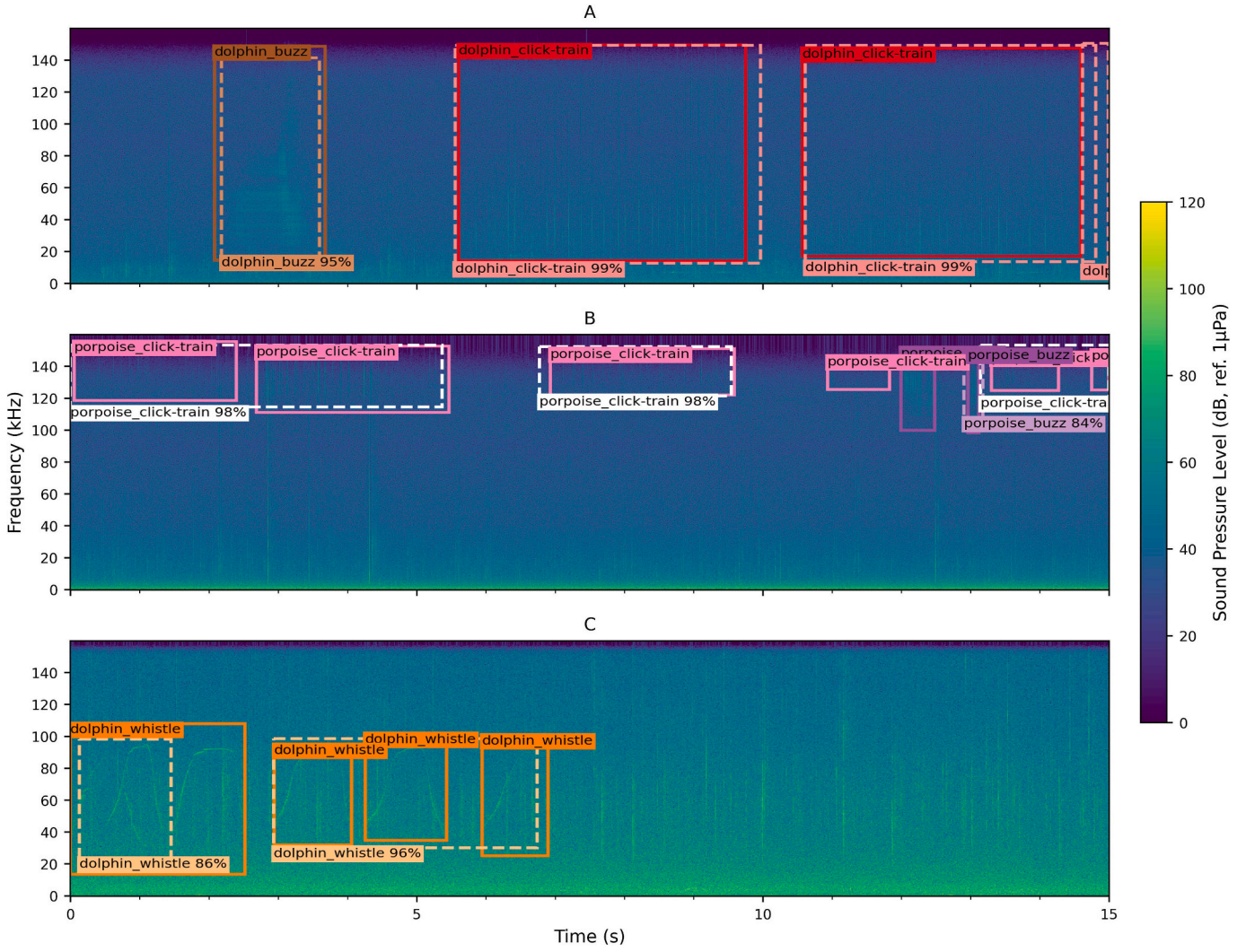


Fig. 3. Examples of annotations and detections (spectrogram parameters detailed in 2.2.1).
A) Detections matching annotations.
B) Detections matching annotations, however some annotations are comprised in one detection. With the common metric for object detection (AP50), these annotations would have been considered as missed detections. Using our customized metrics at the 15-s scale, these detections are considered correct. One annotation of porpoise click-train and one of buzz are missed due to the low signal-to-noise ratio. At the 15-s scale, the missed detections are not considered due to the good detections of the other porpoise click-trains and buzz.
C) Good detections of dolphin whistles. As in B), three whistle annotations are comprised in one detection.
D) No signal to detect and no false detections despite the high background noise.
E) Two good detections and one missed detection of porpoise click-trains, due to the low signal-to-noise ratio. And one false detection of a dolphin click-train, due to impulsive noise similar to dolphin clicks.
F) False positives probably due to acoustic deterrent devices similar to dolphin and porpoise buzzes.
The annotations are in solid lines and detections in dashed lines. To each detection the model attributes a confidence score.

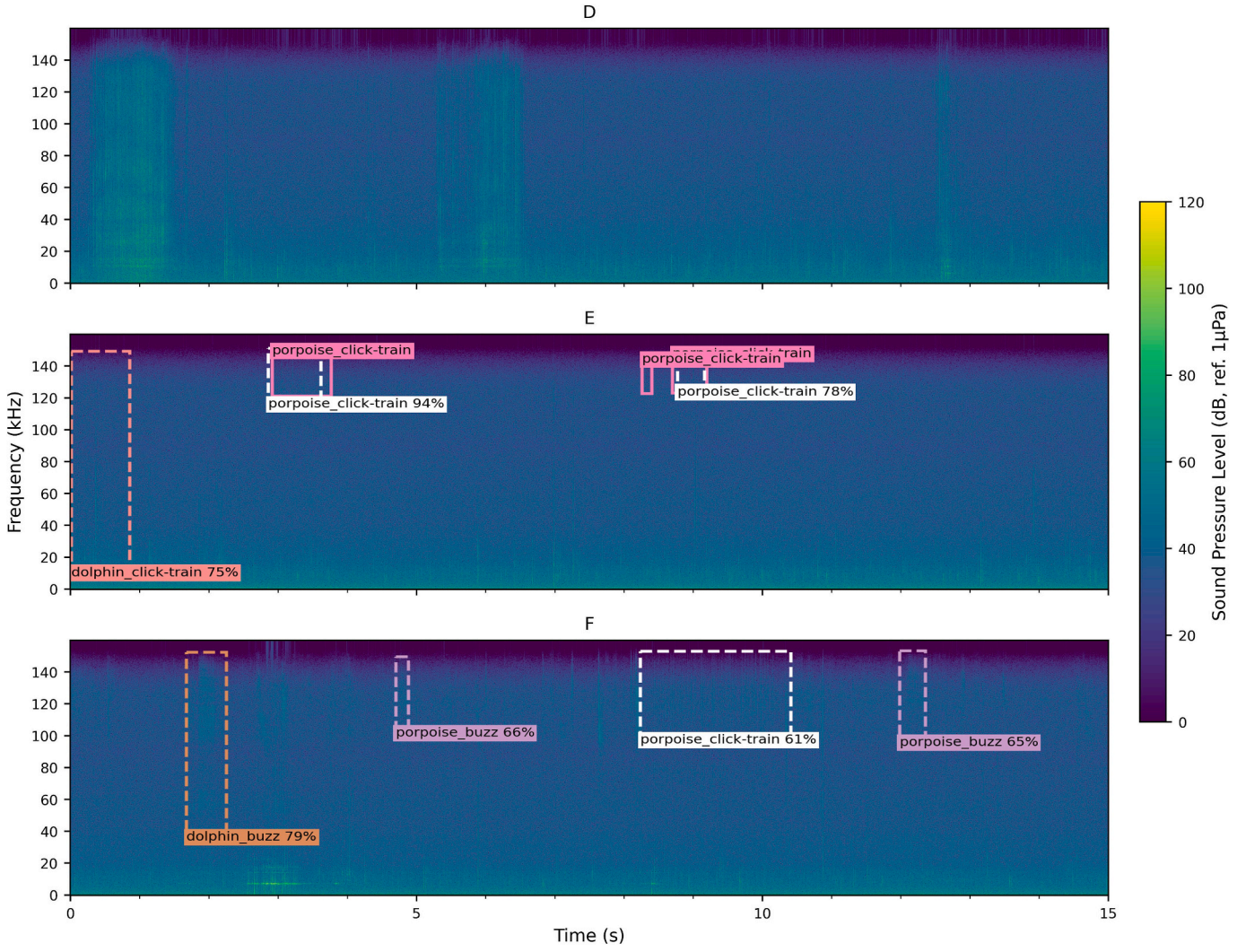


Fig. 3. (*continued*).

click-trains and 278 porpoise click-trains. No buzzes or whistles were found in these recordings. Annotations represented 9 min of sound recording out of 61 h (0.3 % of the time). The average certainty level of all annotations was 2.16 (SD: 0.87).

2.2.3. Dataset preparation

From Dataset 1, training and validation datasets were built to train and validate the deep learning model, respectively. Both datasets were composed of 15-s spectrograms, 50 % of them with annotations (time-frequency boxes) and 50 % without annotations (these datasets will be referred to as annotated and background spectrograms respectively in the following). Although the deep learning model could have been trained without background spectrograms (including any type of sounds different from the target sounds), preliminary numerical experiments showed that adding background spectrograms made it more robust to false detections. Out of the 147810 spectrograms available in the dataset, 1959 background spectrograms were selected randomly. All the annotated spectrograms were included in these training and validation datasets. Moreover, to evaluate the performance statistics of the deep learning model and its generalization to different datasets, the data were prepared for a five-fold cross validation. The dataset was split randomly into five folds allowing the model to be trained and validated five times, with 80 % of training data and 20 % of validation data (Fig. 1–2.1 and Table A.1).

For Dataset 2, all spectrograms from the five consecutive annotated days were used to assess the generalization of model performances.

2.3. Deep learning model

2.3.1. Model selection and setup

The aim of this study was to investigate and assess the ability of a single deep learning model to precisely detect (in time and in frequency) and classify marine mammal sounds over a wide frequency range. In the literature, several object detection models exist and could be employed (Zou et al., 2023). Here, the Faster R-CNN model equipped with the Feature Pyramid Network (FPN) was adopted, taking advantage of the multi-scale features to yield better detection performance (Lin et al., 2017). This two-stage detector has been proven to provide better detection performance than single-stage detection models which focus more on the prediction speed (Beyan and Browman, 2020). Indeed, Faster R-CNN + FPN has been one of the most widely used models in various application domains (Zou et al., 2023). The Pytorch implementation from Detectron2 (Wu et al., 2019) version 0.6, developed by Facebook AI Research, was used. The model can be adjusted by many hyper-parameters in its different components. In this study, most of them remained unchanged and were set using their default values. The input image (original size of 512 × 1813) was resized to (377 × 1333) (default max length = 1333). The ResNet50 architecture with ImageNet pretrained weights was adopted as a backbone.

For model training, the following parameters were configured: batch

size of 16, learning rate of 10^{-4}, approximately 50 epochs (equivalent to 10000 iterations), and stochastic gradient descent (SGD) optimizer. During the training process, using the input data (2.2.3 Dataset preparation) consisting of 15-s spectrograms with time-frequency boxes indicating the sounds of interest (annotations), the model learned to precisely localize (in time and in frequency) and classify the five sound-types (if any) in the spectrograms. For each cross-validation fold, the model providing the best performance with the validation set was selected for further evaluation and analysis.

When using the model, each predicted box was associated with a class and a confidence index. Two parameters could be adjusted to trade off between missed detections and false positives: the confidence threshold score and the Non-Maximum Suppression (NMS) threshold. The confidence threshold retained predictions whose confidence score was greater than this value. The lower this value, the higher the number of predictions. The NMS threshold filtered out overlapping predicted boxes. Based on the Intersection over Union (IoU) value (Fig. 4), the lower the threshold, the fewer overlapping boxes there were. To avoid too many detections on the same objects, this value was set to 0.1. Data were processed using a PC with Intel® Core™ i9-10900KF CPU @ 3.70GHz, 32 Go RAM, NVIDIA GeForce RTX 3080.

2.3.2. Scenario setting

The model was used under three different scenarios: (1) Detecting marine mammal sounds without classification to assess the presence/absence of individuals; (2) Detecting and classifying at species level, either dolphin or porpoise class; (3) Detecting and classifying the five sound types, including dolphin click-trains (D_{CT}), dolphin buzzes (D_B), dolphin whistles (D_W), porpoise click-trains (P_{CT}) and porpoise buzzes (P_B). The number of model detections could be adjusted according to whether a conservative model (i.e., one that is insensitive to noise, with few false positives) or a sensitive model (i.e., one that misses few sounds of interest) was wanted. To achieve this, the confidence threshold was either raised or lowered. Then, for each scenario, the detections were aggregated at the spectrogram scale (15 s) to evaluate the model using a single-class classification (i.e., each spectrogram can be associated with one label: (1) marine mammal) or a multilabel classification (i.e., each spectrogram can be associated with several labels: (2) dolphin/porpoise, (3) D_{CT}/DB/D_W/P_{CT}/P_B).

2.3.3. Model evaluation

As commonly used in deep learning-based object detection, the main metric used for model evaluation during the training process was the mean average precision (mAP). This metric was computed using the average precision (AP) of each class based on the Intersection over Union (IoU) (1) between the predicted box (P_{box}) and grounding truth box (GT_{box}) (Fig. 4). The IoU threshold was set to 0.5, meaning that predicted boxes with an IoU above 50 % were considered to be correct.

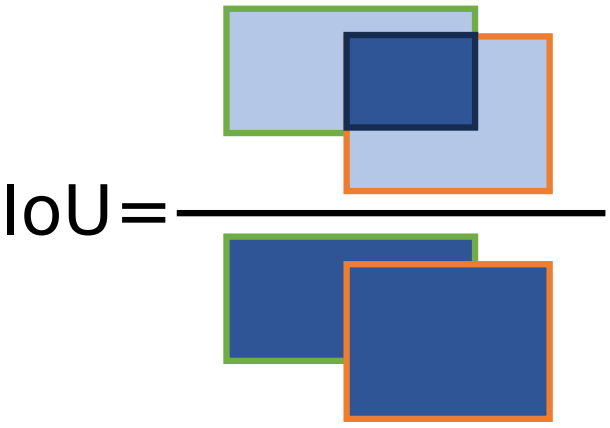


Fig. 4. Intersection over Union (IoU).

Consequently, the True Positives (TP, i.e. a box has been predicted and was expected), False Positives (FP, i.e. a box has been predicted but was not expected) and False Negatives (FN, i.e., no box has been predicted but one was expected) were determined for each confidence threshold, enabling the calculation of the precisions (2) and the recalls (3). The precision-recall curve was computed using the confidence scores of the detections, and the AP corresponded to the area under this curve. The APs of each class are then averaged to give the mean average precision (mAP).

It should be noted that in this study, a precise measurement at the bounding box scale was not necessary. Metrics at the spectrogram scale (15 s in this study) were sufficient and helped to reduce the false positives. For the calculation of classification metrics, at a given confidence threshold, each spectrogram was associated with the grounding truth and predicted labels of the selected scenario (2.3.2. Scenario setting) from the grounding truth boxes labels and those of the predicted boxes. Thus, for each scenario setup and each class, the number of TP, FP, FN and True Negatives (TN) enabled the calculation of the precision, the recall, the F1-score (4) and the false positive rate (FPR) (5). To demonstrate the possibilities of using the model with different trade-offs between false positives and missed detections, precision-recall curves were produced for the three classification levels. Additionally, the areas under the precision-recall curve (AP) and the ROC curve (FPR-recall curve) (AUC) were computed (Hildebrand et al., 2022). Finally, the models were evaluated on the five-fold cross-validation datasets, with statistics (mean and standard deviation) computed for each metric mentioned above.

$$IoU = \frac{P_{box} \cap GT_{box}}{P_{box} \cup GT_{box}} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1\ score = \frac{2{*}Precsion{*}Recall}{Precision + Recall} \quad (4)$$

$$FP\ rate = \frac{FP}{(FP + TN)} {*} \frac{3600}{Ls} \quad (5)$$

With *Ls* the length of the spectrogram (here 15 s) to calculate the number of false positives per hour.

Moreover, to assess the generalization of the model to a new dataset, the previous metrics were calculated on Dataset 2 using the five cross-validation models. This allowed the comparison of means and standard deviations with the performances on the cross-validation datasets.

Finally, to explain the variability of model performance across datasets, Pearson correlation coefficients were calculated between performance metrics (i.e., the presence of false positives and true positives) and annotation certainty levels as well as noise levels. The latter were calculated with the python package scikit-maad (Ulloa et al., 2021) and correspond to the equivalent continuous sound pressure level (Leq) calculated over the 15-s spectrograms.

3. Results

This section analyzes the performance and adaptability of the object detection model on the cross-validation dataset for different classification tasks, including marine mammal detection (presence/absence), species classification and five-sound type classification. It will also highlight the model effectiveness in achieving a trade-off between false positives and missed detections, by presenting its performance at different confidence thresholds. Additionally, the evaluation of the model metrics for the three classification levels will highlight the model's accuracy in detecting and classifying different marine mammal

sounds. Furthermore, the evaluation of the model's performance on a generalization dataset reveals its resilience to different environmental conditions. These results open prospects for optimizing the model's effectiveness and applicability in other real-world scenarios.

3.1. Performance on the cross-validation dataset

3.1.1. False positives/missed detections trade-off

With the confidence scores for each detection, the precision-recall curves were calculated (Fig. 5) for the three scenarios. The confidence threshold was adjusted to obtain a trade-off between false positives and missed detections. To reduce the number of missed detections (e.g. recall around 90 %), a low confidence threshold was selected (e.g. 40 %). Conversely, to reduce the number of false positives (e.g. precision around 90 %), a high confidence threshold was selected (e.g. 90 %). Finally, to achieve a balance between missed detections and false positives (precision roughly equivalent to recall), an intermediate threshold was selected (70 %). For instance, in the case of species classification, selecting a low confidence threshold (40 %) resulted in a reduction in missed detections (recall = 90.9 %). On the other hand, selecting a high confidence threshold (90 %) reduced the false positives (precision = 96.4 %). Finally, the best trade-off between these two was obtained with a confidence threshold of 70 %, giving precision = 86.5 % and recall = 83.6 %. The figure also shows that the more precise the classification, the smaller the area under the curve. In other words, the more complex the model, the lower the model's performance.

The count of spectrograms with detections was also retrieved to assess their proportion within the entire Dataset 1, i.e., the validation dataset plus the remaining background spectrograms (for a total of 148641 spectrograms including 391 with annotations) (Table 1 – Dataset 1). In particular, using a low confidence threshold (40 %), 15.4 % of spectrograms had detections and 6.6 % of annotated spectrograms were missed. As the confidence threshold increased, the number of spectrograms with detections decreased, while the number of missed detections increased.

3.1.2. Performance of the three classification levels

The detection metrics were calculated for the three scenarios (marine mammal detection, species classification, and five-sound type classification) at the three selected confidence thresholds (40 %, 70 %, 90 %). Metrics for the 70 % threshold are detailed in Table 2 – Dataset 1, and those for 40 % and 90 % in Table B.1 and B.2.

As observed in Table 2 – Dataset 1, the model achieved a good F1-score of 88.0 % for marine mammal detection, 84.7 % for species

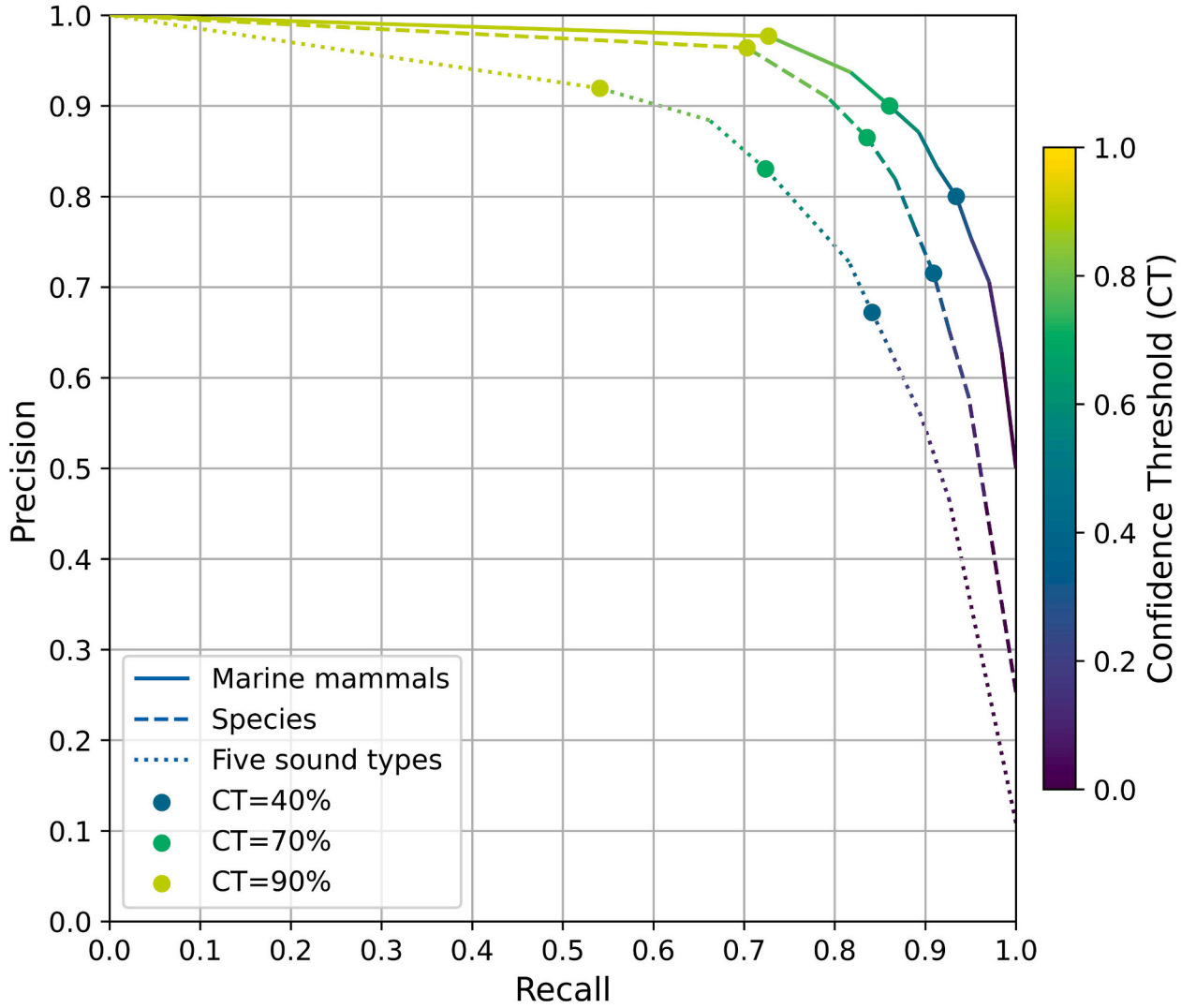


Fig. 5. Precision-recall curves. Calculated using a confidence threshold (CT) from 0 to 1 with a step of 0.1 and for the three classification levels (Marine mammal detection, Species classification, Five-sound type classification). Three thresholds selected: 40 %, 70 % and 90 %. (Recall, Precision) for Marine mammal detection, Species classification (Dolphin/Porpoise), Five-sound type classification (Dolphin Click-Train, Porpoise Click-Train, Dolphin Buzz, Porpoise Buzz, Dolphin Whistle): 40 %: (93.4, 80.0), (90.9, 71.5), (84.1, 67.2); 70 %: (86.1, 90.0), (83.6, 86.5), (72.4, 83.0); 90 %: (72.7, 97.7), (70.3, 96.4), (54.1, 91.9).

Table 1
Number of spectrograms to be checked manually, and number of missed annotated spectrograms for the cross-validation (Dataset 1) and generalization (Dataset 2) datasets. Percentages were calculated on the total number of spectrograms in each dataset (Dataset 1: 148641; Dataset 2: 30240) and on the number of annotated spectrograms (Dataset 1: 391 per fold; Dataset 2: 213). Percentage (number of spectrograms).

	Spectrograms with detections		Missed annotated spectrograms	
Confidence threshold	Dataset 1	Dataset 2	Dataset 1	Dataset 2
40 %	15.4 % (22939)	49.0 % (14822)	6.6 % (26)	3.7 % (8)
70 %	5.8 % (8606)	31.4 % (9483)	13.9 % (55)	22.8 % (49)
90 %	1.5 % (2196)	7.9 % (2381)	27.3 % (107)	57.7 % (123)

classification and 75.8 % for the five-sound type classification. In more detail, dolphin and porpoise click-trains achieved 87.7 (SD: 1.6) % and 82.0 (SD: 2.5) % respectively. Dolphin and porpoise buzzes achieved lower scores with a higher standard deviation: 64.8 (SD: 5.3) % and 53.0 (SD: 12.6) % respectively. Finally, the best score was obtained for dolphin whistles: 91.7 (SD: 5.2) %. Globally, the more precise the classification, the lower the mean values (precision (90.0 % > 83.0 %), recall (86.1 % > 72.4 %) and F1-score (88.0 % > 75.8 %)) and the higher the false positive rate (23.1 FP/H < 24.9 FP/H).

To obtain a more representative value for the false positive rate, the metric was also calculated using all the background spectrograms in Dataset 1 (validation dataset plus the remaining background spectrograms). This yielded lower false positive rates, from 13.4 FP/H (marine mammal detection) to 14.2 FP/H (five-sound type classification).

Finally, areas under the precision-recall curves (AP) and the ROC curves (AUC-ROC) were macro-averaged between values per classes (Table B.3 and Table B.4). For species classification, this yielded a mAP of 92.3 (SD: 1.1) % and a macro-average AUC of 95.7 (SD: 0.9) %. And for five-sound type classification, this yielded a mAP of 84.3 (SD: 1.7) % and a macro-average AUC of 94.9 (SD: 0.6) %.

3.2. Performance on the generalization dataset

In this subsection, we investigated the generalization capacity of the model, trained on Dataset 1, by evaluating its performance on Dataset 2. Firstly, the count of spectrograms with detections was obtained to assess their proportion within the five-day dataset (totaling 30240 spectrograms, including 213 annotated ones) (Table 1 – Dataset 2). In particular, by using a low confidence threshold (40 %), 49.0 % of spectrograms had detections and 3.7 % of annotated spectrograms were missed.

Then, the recall and false positive rate were calculated for the three scenarios (marine mammal detection, species classification and five-sound type classification) with a confidence threshold of 70 % (Table 2 – Dataset 2).

Compared to Dataset 1, the model achieved a good recall of 77.2 % for marine mammal detection. However, the macro-average recall is lower for the five-sound type classification with a value of 59.1 %, particularly due to the low value of 41.4 % for dolphin click-train. Additionally, false positive rates are much higher than those of Dataset 1 (100.6 > 24.9 FP/H).

Finally, areas under the ROC curves (AUC-ROC) were macro-averaged between values per classes (Table B.5) yielding a macro-average AUC of 87.0 (SD: 1.2) %.

3.3. Explanatory analysis of results

Fig. 3 illustrates annotated spectrograms showing the detections provided by the model and highlighting the diversity of marine mammal sounds and background noise present in the datasets. These outputs emphasized the model's ability to accurately detect a variety of marine mammal sounds, while illustrating specific instances of missed detections and false positives. In particular, Fig. 3-F illustrates the model detections of non-biological sounds resembling dolphin and porpoise buzzes. Finally, this figure highlights the importance of calculating metrics on a 15-s scale, demonstrating a reduction in false positives and

Table 2
Evaluation metrics (Precision, Recall, F1-score and False Positive Rate) for the three scenarios with a 70 % confidence threshold (best trade-off between false positives and missed detections) for Datasets 1 and 2. Precision, Recall and F1-score in percent, False Positive Rate in False Positives/Hour. Mean (standard deviation).

A) Marine mammal detection

	Precision (%)	Recall (%)		F1-score (%)	False Positive Rate (FP/H)	
Class	Dataset 1	Dataset 1	Dataset 2	Dataset 1	Dataset 1	Dataset 2
Marine Mammal	90.0 (1.7)	86.1 (1.6)	77.2 (5.5)	88.0 (1.0)	23.1 (4.6)	74.5 (10.2)

B) Species classification

	Precision (%)	Recall (%)		F1-score (%)	False Positive Rate (FP/H)	
Class	Dataset 1	Dataset 1	Dataset2	Dataset 1	Dataset 1	Dataset 2
Dolphin	84.8 (2.6)	90.5 (1.8)	48.6 (5.8)	84.7 (1.5)	15.7 (3.1)	30.0 (7.2)
Porpoise	88.2 (3.6)	76.7 (3.4)	83.1 (5.3)	82.0 (2.4)	6.9 (2.0)	58.6 (10.1)
Macro Average (Sum for FPR)	86.5 (1.9)	83.6 (1.9)	65.9 (5.2)	84.7 (1.5)	22.6 (3.5)	88.5 (13.5)

C) Five-sound type classification

	Precision (%)	Recall (%)		F1-score (%)	False Positive Rate (FP/H)	
Class	Dataset 1	Dataset 1	Dataset 2	Dataset 1	Dataset 1	Dataset 2
Dolphin Click-Train	84.5 (3.8)	91.4 (2.4)	41.4 (4.0)	87.7 (1.6)	12.8 (3.5)	9.8 (1.7)
Porpoise Click-Train	87.6 (4.4)	77.2 (3.9)	76.9 (5.6)	82.0 (2.5)	7.0 (2.5)	30.5 (10.4)
Dolphin Buzz	64.5 (9.0)	66.7 (10.5)	N/A	64.8 (5.3)	4.5 (2.5)	20.1 (5.1)
Porpoise Buzz	79.4 (27.7)	40.8 (7.9)	N/A	53.0 (12.6)	0.6 (0.8)	38.4 (5.8)
Dolphin Whistle	99.1 (2.0)	85.7 (8.5)	N/A	91.7 (5.2)	0.1 (0.1)	1.8 (1.6)
Macro Average (Sum for FPR)	83.0 (6.7)	72.4 (2.7)	59.1 (4.3)	75.8 (4.1)	24.9 (4.6)	100.6 (17.4)

missed detections.

Fig. 6 - Dataset 1 gives information about the temporal evolution of the ambient soundscape of Dataset 1. The presence of mooring-related harmonic noise (5, 10 and 15 kHz) was noticed most of the time in the dataset. Particularly between November 29th and December 5th, and from December 12th, with high frequency components, also from ships passing close by and potentially from Acoustic Deterrent Devices (ADDs).

The sound pressure level and the certainty level of the annotations give some explanation of the model's performance. The Pearson correlation coefficient between the sound pressure level at high frequency [80–156 kHz] and the presence of a false positive (0/1) per spectrogram was 0.20 ($p < 0.001$). Thus, the increases in the false positive rate tend to coincide with periods of high-frequency noise. However, the false positive rate was acceptable: averaging less than 20 FP/H for most of the time over the entire dataset (Fig. B.1). Additionally, the correlation coefficient between the mean level of certainty of the annotations and the presence of a true positive (0/1) per spectrogram was 0.28 ($p < 0.001$). Thus, the presence of annotations with low certainty levels and/or the absence of annotations with high certainty levels tend to decrease the recall.

The certainty level of the annotations also provides information on the causes of performance differences between the cross-validation and the generalization datasets. A higher proportion of annotations with low certainty levels was noticed compared to Dataset 1 (Fig. B.2) and the Pearson correlation coefficient between the mean level of certainty of the annotations and the presence of a true positive (0/1) per spectrogram was 0.39 ($p < 0.001$). Moreover, some acoustic differences were observed between these two datasets for the dolphin click-trains. The average value of the upper frequency limit was 101 (SD: 35) kHz for the generalization dataset as opposed to 139 (SD: 21) kHz for the cross-validation dataset, and a first quartile at 70 kHz as opposed to 142 kHz.

Moreover, the presence of sounds potentially generated by acoustic deterrent devices (ADDs) resembling marine mammal sounds was noted in the recordings, particularly in Dataset 2. Indeed, out of 100 randomly selected false positives, 68 were likely to be sounds of ADDs and occurred regularly in Dataset 2. On the other hand, in Dataset 1, only 19 % of false positives were due to ADDs, particularly at the end of the dataset (from November 13th). Moreover, Fig. 6 – Dataset 2 highlights the presence of high-frequency noise throughout Dataset 2.

4. Discussion

In this study, the development and evaluation of an object detection model adapted to the detection and classification of marine mammal sounds from spectrograms over a wide frequency range were investigated. Thanks to the model's capabilities, it was possible to accurately localize (in time and frequency) and classify marine mammal sounds in large datasets, demonstrating its potential to facilitate manual annotation processes in the context of environmental monitoring of underwater fauna in OWFs. This section presents an in-depth discussion of the results, focusing on the model's performance across three levels of classification (marine mammal detection, species classification, five-sound type classification), its ability to adapt to find a compromise between false positives and missed detections, and its relevance to marine mammal monitoring. In addition, the study of the model's performance on a generalization dataset revealed information about its resilience to new environmental conditions. These results have opened up prospects for improving the model and broadening its possible uses, in order to contribute to the advancement of marine mammal research (e.g. acoustic behavior) and the enhancement of passive acoustic monitoring techniques.

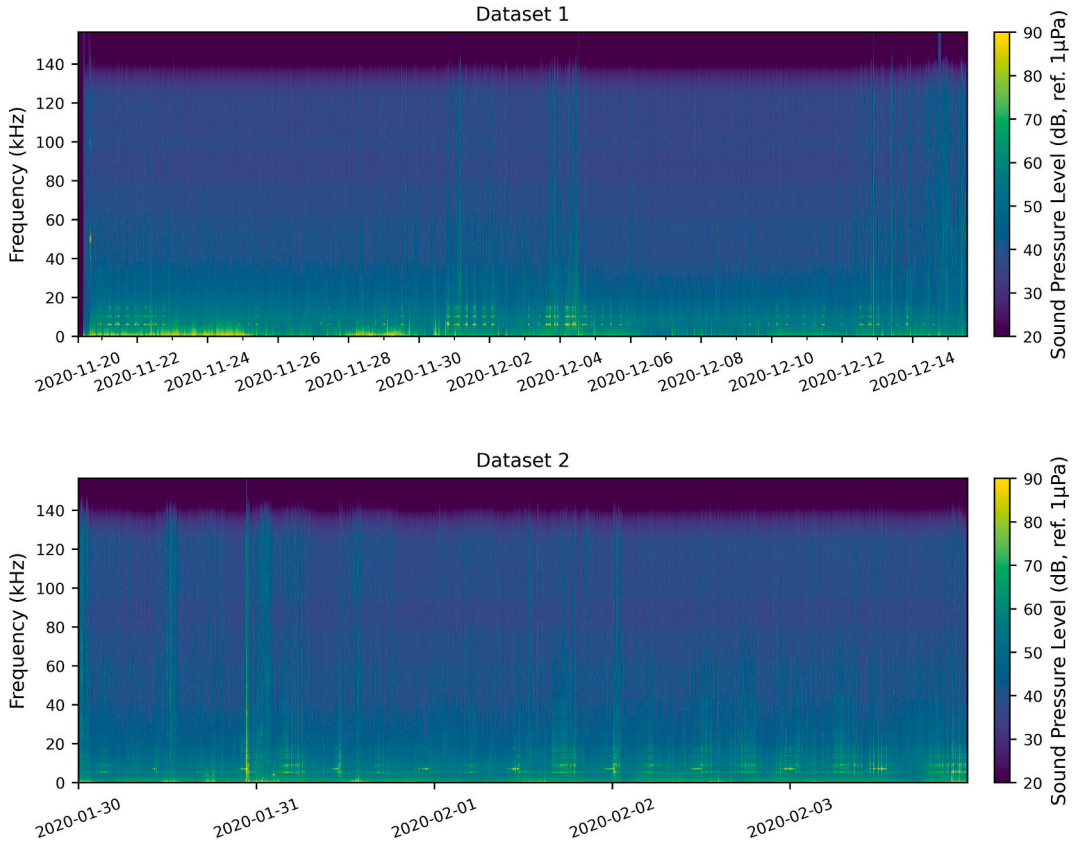


Fig. 6. Long-term average spectrum (LTAS) of Dataset 1 (upper) and Dataset 2 (lower). Sound pressure level in dB (ref. 1 µPa).

4.1. Qualitative evaluation

4.1.1. Model performance

As one of the first applications of an object detection model to underwater acoustic data, the performances obtained in this study are encouraging. When applied to the whole dataset, the model efficiently identified the 15-s spectrograms containing marine mammal sounds while only missing 7 % of the spectrograms with annotations. Detections from the model were spread along one sixth of the spectrograms (22939 spectrograms with detections out of 148641 spectrograms) meaning that the model can greatly ease manual verification by: (i) removing more than three-quarters of the dataset without sounds of interest, and (ii) providing spectrograms where detection boxes are drawn, guiding annotators toward potential marine mammal sounds.

When comparing model performances for the three classification levels (presence/absence, dolphin/ porpoise, five sounds), the most complex task, the five-sound type classification, exhibited a macro average F1-score lower than solely identifying the presence or absence of a marine mammal sound. This loss of performance when increasing the complexity of the classification task is likely due to low occurrences of some sound classes, such as buzzes, for which the model did not have enough examples to train and then detect these rare sound events. Other methods like few-shot learning (Nolasco et al., 2022; Xu et al., 2021) may be more relevant for this type of rare sounds. Increasing the data quantity using data augmentation (Li et al., 2021; Padovese et al., 2021) could also be tested to improve the results (i.e., raise mean values and reduce standard deviations). This solution is based on the creation of new artificial data to have a larger training dataset. These new data can be created by adding, for instance, background noise to known sounds of interest. In the case of passive acoustic monitoring this would be interesting, as the datasets mainly contain recordings without sounds of interest, constituting a large dataset for data augmentation. Finally, in this study, the Faster R-CNN model was used with few modifications. To enhance the model's performance, hyperparameters could be fine-tuned according to the input data. For instance, this might involve shaping the anchors (time-frequency box proposals) or adjusting the pooling size (ensuring identical resolution for all region proposals).

As a last general comment on the results, despite the limited instances of dolphin whistles, the mean metrics and standard deviations remained acceptable. The accurate detection of dolphin whistles may be attributed to their distinctiveness, making them less susceptible to confusion with other background noises. Impulsive sounds like click-trains and buzzes are more likely to be confused with other impulsive noises, such as impacts, mooring noise or acoustic deterrent devices. We noted that a major strength of this type of model, compared with simple classification models, is its ability to precisely localize different types of sound in time and particularly in frequency, enabling it to easily distinguish between the clicks of the two species with very distinct frequency ranges (dolphins: 20–100 kHz and porpoises: 100–150 kHz).

Regarding the results on the generalization dataset, although not all the training sound classes were found in these recordings, the evaluation of the model's performance yielded insights. Firstly, the model effectively reduced the number of spectrograms to be manually checked: half of the dataset (49 %) included spectrograms with detections, with only 4 % of spectrograms with annotations missed. Then, an increase in false positives was noticed compared to the previous dataset. Indeed, during the annotation process, it was observed that this dataset contained many more high frequency noise events. Further investigation suggested that these sounds may originate from acoustic pingers emitting signals resembling dolphin and porpoise buzzes (McGarry et al., 2022; Schaffeld, 2016). This emphasized the need for deeper acoustic studies for a better distinction between marine mammal buzzes and acoustic pingers. One possible approach to address this question is to calculate low-level acoustic features (such as the inter-click interval) within the time-frequency boxes of the sounds detected by the model, followed by dimensionality reduction and clustering techniques. For instance, techniques like UMAP (McInnes et al., 2020) and DBSCAN (Ester et al., 1996) could be employed to identify clusters of click-trains, buzzes and anthropogenic noises. This model is also promising for advancing bioacoustics research, enabling more in-depth studies of vocalizations. It could be useful for studying the existence of different types of buzzes and click-trains. Indeed, click-trains and buzzes are not necessarily used only for echolocation and foraging. Some studies (Clausen et al., 2011; Sørensen et al., 2018) have found that click-trains and buzzes may also be used for communication between individuals.

Moreover, dolphin click-trains were less well detected in the generalization dataset probably due to the changes in their acoustic signatures, and especially to the decrease in energy in the high frequencies. Since high frequencies propagate less well than low frequencies, these results may be explained by a non-optimal source-receiver orientation, or by changes in environmental parameters (such as temperature or salinity) (Urick, 1983). In addition, even for the expert annotator it was difficult to identify the marine mammal sounds. Indeed, a higher proportion of annotations with a low certainty level was noticed compared to Dataset 1.

4.1.2. Model relevance for marine mammal monitoring

Regulatory studies are currently required throughout the life cycle of OWF. Most are carried out using a hydrophone to measure ambient noise, accompanied by a C-Pod or F-Pod to detect marine mammal clicks. The latter are quick and inexpensive to set up, and can be used to monitor site frequentation, but the output data are limited and allows only limited post-processing. With a broadband hydrophone and an object detection model, it is possible to obtain high-quality data, giving more precise temporal and frequency information on the various vocalizations of marine mammals, enabling us to study their behavior and carry out ecological inferences over the long term.

Using a model to detect and classify marine mammal sounds is more complex than merely detecting them without classification. However, achieving a perfect model (100 % precision and recall) in both cases is challenging. Thus, different confidence thresholds can be used to adapt the model to favor either missed detections or false detections, making it versatile for diverse situations. For instance, if the model is employed to simply detect marine mammal sounds in the vicinity of OWFs during periods requiring attention (e.g. construction works), the threshold could be lowered to ensure that marine mammal sounds are not missed. On the other hand, if the model is used for long-term monitoring of the presence and behavior of marine mammals around the OWF, it may be interesting to raise the confidence threshold to reduce the number of spectrograms that have to be checked manually. Thus, using a low confidence threshold, 6.6 % of missed detections and 15.4 % of false detections were reached for the detection of marine mammals. With a high confidence threshold, 45.9 % of missed detections and 8.1 % of false detections were reached for the detection and classification of the five sound types. In real applications, these values could be used to adjust the results of detections (i.e., estimating the real quantity of detections by adding the probability of missed detections and subtracting the probability of false detections).

Additionally, Shiu et al., 2020 estimated that to be efficient (i.e., to obtain an acceptable quantity of detections to be manually checked) the model should not exceed 20 FP/H. In this study, with a confidence threshold of 40 % and 90 %, the false positive rate was 68 and 5 FP/H respectively. Improvements have to be made to reduce false positives and to obtain higher recall and precision scores. Implementing iterative learning and hard negative mining techniques (Allen et al., 2021; Shiu et al., 2020) could be a solution. Through strategic selection of background spectrograms rather than random selection, as is the case here the model could be exposed to more challenging spectrograms, particularly those for which it is prone to false detections (with ADDs for instance). This deliberate exposure enhances the model robustness to diverse ambient background noises, by reducing the occurrence of false positives. However, it is essential to note that this approach may result in

a higher rate of missed detections.

Although it was difficult to compare model performance across different research studies due to a lack of standardization in evaluation protocols, the mAP (84.3 %) and AUC-ROC (94.9 %) of the validation datasets were within a similar order of magnitude to other studies using CNNs to classify animal sounds (Shiu et al. (2020), LeBien et al. (2020), Ruff et al. (2021)). In particular Shiu et al. (2020), who compared two CNN models for binary classification of a single marine mammal sound (AP of 90 % and 83 %). The advantage of our model over conventional CNN classification models was the precise temporal and frequency localization of sounds in the spectrogram, making it easier to distinguish sounds of multiple species over a wide frequency range and enabling possible future analyses on the sounds of interest.

Despite the different scales of analysis of object detection model performance (spectrogram or box scale), the performance of our model could be considered to be of the same order of magnitude as other object detection models. Ferguson et al. (2022) trained three separate models to detect three marine mammal species (without call classification). They obtained an f1 score between 45 % and 63 %, compared with 82 % and 85 % for dolphin and porpoise in our study. Wu et al. (2021) trained a single model to detect the sounds of six owl species in a narrow frequency band (100 Hz-16 kHz). They obtained a mAP and AUC-ROC of 83 % and 89 % respectively. By aggregating our detections at spectrogram scale, we improved our results, but by having a single model with more species and sound types to classify, we achieved performance of the same order of magnitude as these studies. Compared with these models, we demonstrated the ability of a single object detection model to detect and classify multiple species and types of sound over a wide frequency range.

Thus, our model could find interesting use in many areas where species are numerous and have a wide vocal repertoire, such as in tropical forests (Hart et al., 2021) or coral reefs (Noble et al., 2024). It is also promising for use on a larger scale (other locations and time scales) using transfer learning. The latter has been successfully applied to marine acoustic environments using CNN (White et al., 2023).

Finally, this study demonstrated the ability of a single model to detect and classify several marine mammal sounds over a wide frequency range. In its current form, the model is not applicable to real-time but is a proof of concept that could be transferred to real-time in future work and can currently be used for post hoc studies.

4.2. Future work

In addition to the avenues for improvement identified in the previous paragraphs, some ideas for future work are proposed in this section.

4.2.1. Improvement of input data

Data-centric AI (Motamedi et al., 2021) reminds us of the need for good input data for model training. For this, the annotation process may be improved and further standardized. Currently, the annotation process is not necessarily performed by the same person or with the same acquisition and signal processing parameters, influencing the manner click-trains and buzzes are annotated and thus the learning of the model. Implementing an annotation platform (e.g. Aplose platform: https://github.com/Project-OSmOSE/osmose-app) that standardizes the spectrogram display eliminates the biases induced by the choices of signal processing parameters. Moreover, annotators are not perfect and may make some annotation errors (missed detections or false detections, as well as variation of the annotation method over time) (Nguyen Hong Duc et al., 2021). A solution could be to invite more people to analyze the same datasets to reinforce the validity of annotations. In our case, the presence of annotations with a low certainty level influenced model learning and performance. Grouping annotations from several annotators would enable the model to be trained with a greater number of high-certainty annotations, thus improving model performance. The study of inter-annotator variability and the development of methods to group and validate the annotations from multiple annotators is an area of interest that should continue to be explored (Dubus et al., 2024).

4.2.2. Toward the generalization of the model

In this study, the model was evaluated over a different time period from the training one to assess its robustness against other noise environments, with the aim of evaluating the relevance of precision and recall over time. A decline in recall and an increase in the false positive rate were observed in the generalization dataset. The ambient noise level and the degree of certainty of annotations affect the model's performance, as does the diversity of ambient background noise, which can vary significantly over time. Thus, it is important to obtain more diverse data (i.e., from other periods, locations and environmental conditions) for a more complete model training to achieve better performance for new environments. Moreover, in the case of cross-validation, having more diverse data would enable us to obtain folds that are more representative of reality and thus reduce uncertainties in the cross-validation results and give a better idea of the generalization capabilities of the model. This need of data has been mentioned by Parsons et al. (2022) but obtaining acoustic and annotated datasets is rare and costly.

Furthermore, in the future, as the model may need to be trained on acoustic data covering a wide spatial and temporal domain, it could be interesting to add contextual information (e.g. date and location) as an additional model input to improve detection and classification of vocalizations (Jeantet and Dufourq, 2023) which can vary over the seasons and between local species. Finally, the model has been trained on spectrograms with a particular sampling frequency. Further developments are necessary to obtain a model usable with data recorded with different sampling frequencies. The model needs to be adapted to better consider the ordinate axis of the spectrograms (frequencies). It may be interesting to study rainbow mapping (Wu et al., 2021) or coordinate convolution (Liu et al., 2018).

5. Conclusion

The advancements in deep learning have led to the development of numerous sound detection and classification models. These models play a crucial role in the monitoring of species based on their vocalizations. This study has introduced a promising object detection model specifically tailored to the detection and classification of marine mammal sounds over a wide frequency range. Compared with manual analysis, this model facilitates the monitoring of marine mammal species, especially in the assessment of their behavioral patterns around offshore wind farms. However, the disadvantage of passive acoustics is that it does not readily provides abundance estimates of marine mammals. One solution would be to merge the information from different survey methods (such as aerial surveys, cameras, biologging or eDNA) to compensate for the advantages and disadvantages of each method. This advancement would enhance the monitoring of marine mammals in the vicinity of OWFs. Further studies are required to develop methods for automating the processing and merging of the substantial quantity of data collected. The automation of data collection and processing also offers the unique opportunity to transmit ecological information in real-time which is crucial to optimize the cost-effectiveness of monitoring strategies for OWFs that are bound to be increasingly distant from the coast.

Funding

This work is part of the OWFSOMM (Offshore Wind Farm Surveys Of Marine Megafauna) project which receives funding from: France Energies Marines' members and partners; the French National Research Agency under the France 2030 Investment Plan (ANR10 IEED 0006 34); the French General Direction for Energy and Climate (DGEC); the French Office for Biodiversity (OFB); the Brittany region; the Provence Alps French Riviera (PACA) region; the Brittany (UBO) University; the Caen

(UniCaean) University.

CRediT authorship contribution statement

Quentin Hamard: Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Minh-Tan Pham:** Writing – review & editing, Validation, Supervision, Resources, Methodology, Funding acquisition, Conceptualization. **Dorian Cazau:** Writing – review & editing, Validation, Supervision, Resources, Methodology, Conceptualization. **Karine Heerah:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Acknowledgements

This study is part of the collaborative OWFSOMM (Offshore Wind Farm Surveys Of Marine Megafauna) project, carried out at France Energies Marines. As such, we would like to thank all 13 project partners (FEM, UBS, CNRS, LRU-PELAGIS, WIPSEA, EDR-RE, Ocean Winds, EOLFI, RWE, Ifremer, CE-Sigma, ENSTA Bretagne, GIS EMDT) for their contributions throughout the duration of the OWFSOMM project. More specifically, we thank EDF-RE and Eolienne en Mer des Hautes Falaises (EOHF) whom provided the underwater passive acoustic dataset used in this study. We also thank the OSmOSE team (ENSTA Bretagne, Lab-STICC, UMR CNRS 6285) for their advice on bioacoustics. Finally, thanks to "Alba Traduction" for proofreading the paper.

Appendix A. Dataset

Table A.1
Content of the five folds. Number of spectrograms (S) per class and per fold, and number of annotations (A) per class and per fold.

Fold	Dolphin Click-train		Porpoise Click-Train		Dolphin Buzz		Porpoise Buzz		Dolphin Whistle		S	A	Background Spectrograms	
	S	A	S	A	S	A	S	A	S	A			25 kHz	156 kHz
1	87 (19.9 %)	399 (19.7 %)	169 (20.4 %)	321 (19.9 %)	32 (17.9 %)	44 (17.3 %)	12 (19.4 %)	16 (21.1 %)	24 (20.3 %)	57 (18.0 %)	392 (20.0 %)	837 (19.5 %)	24 (20.3 %)	368 (20.0 %)
2	179 (19.1 %)	404 (19.9 %)	177 (21.4 %)	380 (23.6 %)	30 (16.8 %)	47 (18.5 %)	11 (17.7 %)	15 (19.7 %)	24 (20.3 %)	59 (18.6 %)	392 (20.0 %)	905 (21.1 %)	24 (20.3 %)	368 (20.0 %)
3	195 (20.8 %)	425 (21.0 %)	161 (19.5 %)	318 (19.7 %)	35 (19.6 %)	52 (20.5 %)	10 (16.1 %)	10 (13.2 %)	24 (20.3 %)	60 (18.9 %)	392 (20.0 %)	865 (20.2 %)	24 (20.3 %)	368 (20.0 %)
4	182 (19.4 %)	410 (20.2 %)	172 (20.8 %)	315 (19.5 %)	35 (19.6 %)	48 (18.9 %)	13 (21.0 %)	19 (25.0 %)	23 (19.5 %)	83 (26.2 %)	391 (20.0 %)	875 (20.4 %)	23 (19.5 %)	368 (20.0 %)
5	196 (20.9 %)	390 (19.2 %)	148 (17.9 %)	279 (17.3 %)	47 (26.3 %)	63 (24.8 %)	16 (25.8 %)	16 (21.1 %)	23 (19.5 %)	58 (18.3 %)	390 (19.9 %)	806 (18.8 %)	23 (19.5 %)	367 (20.0 %)
Sum	939	2028	827	1613	179	254	62	76	118	317	1957	4288	118	1839

Appendix B. Results

Table B.1
Evaluation metrics (Precision, Recall, F1-score and False Positive Rate) for the three scenarios with a 40 % confidence threshold for Dataset 1. Precision, Recall and F1-score in percent, False Positive Rate in False Positives / Hour. Mean (standard deviation).

A) Marine mammal detection

Class	Precision (%)	Recall (%)	F1-score (%)	False Positive Rate (FP/H)
Marine Mammal	80.0 (1.6)	93.4 (1.1)	86.2 (0.7)	56.2 (6.0)

B) Species classification

Class	Precision (%)	Recall (%)	F1-score (%)	False Positive Rate (FP/H)
Dolphin	70.2 (3.1)	95.4 (1.8)	80.8 (1.5)	39.3 (6.8)
Porpoise	72.9 (5.0)	86.4 (1.1)	79.0 (2.7)	21.9 (4.4)
Macro Average (Sum for FPR)	71.5 (2.5)	90.9 (0.6)	79.9 (1.5)	61.2 (7.3)

C) Five-sound type classification

(*continued on next page*)

Table B.1 (*continued*)

C) Five-sound type classification				
Class	Precision (%)	Recall (%)	F1-score (%)	False Positive Rate (FP/H)
Class	Precision (%)	Recall (%)	F1-score (%)	False Positive Rate (FP/H)
Dolphin Click-Train	68.8 (3.6)	94.9 (2.3)	79.7 (1.7)	32.9 (5.9)
Porpoise Click-Train	72.2 (6.9)	86.6 (1.2)	78.6 (4.0)	21.6 (5.7)
Dolphin Buzz	45.4 (4.8)	78.8 (6.8)	57.3 (3.3)	11.0 (2.6)
Porpoise Buzz	54.2 (20.7)	62.9 (13.5)	57.5 (17.2)	2.3 (1.5)
Dolphin Whistle	95.4 (6.7)	97.5 (2.3)	96.3 (3.7)	0.4 (0.6)
Macro Average (Sum for FPR)	67.2 (4.0)	84.1 (1.6)	73.9 (3.4)	68.2 (8.9)

Table B.2
Evaluation metrics (Precision, Recall, F1-score and False Positive Rate) for the three scenarios with a 90 % confidence threshold for Dataset 1. Precision, Recall and F1-score in percent, False Positive Rate in False Positives / Hour. Mean (standard deviation).

A) Marine mammal detection				
Class	Precision (%)	Recall (%)	F1-score (%)	False Positive Rate (FP/H)
Marine Mammal	97.7 (0.8)	72.7 (3.0)	83.3 (1.9)	4.2 (1.7)

B) Species classification				
Class	Precision (%)	Recall (%)	F1-score (%)	False Positive Rate (FP/H)
Dolphin	95.9 (1.1)	80.9 (3.3)	87.7 (1.7)	3.3 (0.9)
Porpoise	96.9 (2.7)	59.8 (7.0)	73.7 (5.5)	1.3 (1.1)
Macro Average (Sum for FPR)	96.4 (1.7)	70.3 (4.0)	80.7 (2.9)	4.7 (1.8)

C) Five-sound type classification				
Class	Precision (%)	Recall (%)	F1-score (%)	False Positive Rate (FP/H)
Dolphin Click-Train	95.6 (1.8)	84.3 (4.4)	89.5 (2.3)	3.0 (1.2)
Porpoise Click-Train	96.1 (2.7)	61.2 (7.6)	74.5 (5.8)	1.6 (1.1)
Dolphin Buzz	88.0 (5.4)	48.7 (11.0)	62.1 (8.9)	0.8 (0.5)
Porpoise Buzz	80.0 (44.7)	11.7 (8.1)	20.1 (13.5)	0.0 (0.0)
Dolphin Whistle	100.0 (0.0)	64.5 (10.1)	78.1 (7.6)	0.0 (0.0)
Macro Average (Sum for FPR)	91.9 (8.7)	54.1 (2.6)	64.9 (2.9)	5.4 (1.8)

Table B.3
Dataset 1 - Average Precision (AP) per fold for each class of each classification level. For each classification level, the APs are macro-averaged.

AP	MM	D	P	D/P	DCT	PCT	DB	PB	DW	Five classes
1	95.7 %	94.8 %	90.3 %	92.6 %	94.9 %	90.4 %	70.5 %	74.1 %	100.0 %	86.0 %
2	95.9 %	95.5 %	92.3 %	93.9 %	96.1 %	92.5 %	65.8 %	66.6 %	98.3 %	83.9 %
3	96.0 %	95.1 %	89.7 %	92.4 %	96.1 %	89.3 %	82.1 %	42.8 %	99.8 %	82.0 %
4	95.6 %	96.0 %	87.2 %	91.6 %	96.0 %	87.4 %	74.4 %	61.8 %	99.2 %	83.8 %
5	95.3 %	95.1 %	86.7 %	90.9 %	94.1 %	85.8 %	71.5 %	78.4 %	99.7 %	85.9 %
Mean	**95.7 %**	95.3 %	89.2 %	**92.3 %**	95.4 %	89.1 %	72.9 %	64.8 %	99.4 %	**84.3 %**
Std	0.3 %	0.5 %	2.3 %	1.1 %	0.9 %	2.6 %	6.0 %	13.9 %	0.7 %	1.7 %

Table B.4
Dataset 1 - Area under the receiver operating characteristic curve (AUC-ROC) per fold for each class of each classification level. For each classification level, the AUC-ROCs are macro-averaged.

AUC-ROC	MM	D	P	D/P	DCT	PCT	DB	PB	DW	5 classes
1	94.9 %	97.1 %	94.0 %	95.5 %	97.3 %	94.2 %	92.0 %	91.4 %	100.0 %	95.0 %
2	95.4 %	97.8 %	96.1 %	97.0 %	98.3 %	96.2 %	92.6 %	90.4 %	99.9 %	95.5 %
3	95.3 %	96.8 %	95.2 %	96.0 %	97.9 %	94.8 %	95.5 %	88.7 %	100.0 %	95.4 %
4	94.9 %	97.4 %	92.6 %	95.0 %	97.6 %	93.0 %	92.7 %	87.7 %	100.0 %	94.2 %
5	94.4 %	97.1 %	92.6 %	94.9 %	96.8 %	92.3 %	89.2 %	93.3 %	100.0 %	94.3 %
Mean	**95.0 %**	97.2 %	94.1 %	**95.7 %**	97.6 %	94.1 %	92.4 %	90.3 %	100.0 %	**94.9 %**
Std	0.4 %	0.4 %	1.5 %	0.9 %	0.6 %	1.6 %	2.3 %	2.2 %	0.0 %	0.6 %

Table B.5
Dataset 2 - Area under the receiver operating characteristic curve (AUC-ROC) for each class of each classification level. For each classification level, the AUC-ROCs are macro-averaged. For the five-sound type classification, only the two present classes were considered.

AUC-ROC	MM	D	P	D/P	DCT	PCT	DB	PB	DW	5 classes
1	84.4 %	81.8 %	90.1 %	86.0 %	84.1 %	90.7 %	N/A	N/A	N/A	87.4 %
2	82.9 %	77.2 %	89.0 %	83.1 %	79.4 %	90.4 %	N/A	N/A	N/A	84.9 %
3	82.4 %	77.2 %	89.1 %	83.1 %	83.4 %	91.6 %	N/A	N/A	N/A	87.5 %
4	83.8 %	81.7 %	89.1 %	85.4 %	84.7 %	90.4 %	N/A	N/A	N/A	87.5 %
5	83.1 %	80.6 %	89.2 %	84.9 %	84.5 %	91.3 %	N/A	N/A	N/A	87.9 %
Mean	**83.3 %**	79.7 %	89.3 %	**84.5 %**	83.2 %	90.9 %	N/A	N/A	N/A	**87.0 %**
Std	0.8 %	2.3 %	0.5 %	1.3 %	2.2 %	0.6 %	N/A	N/A	N/A	1.2 %

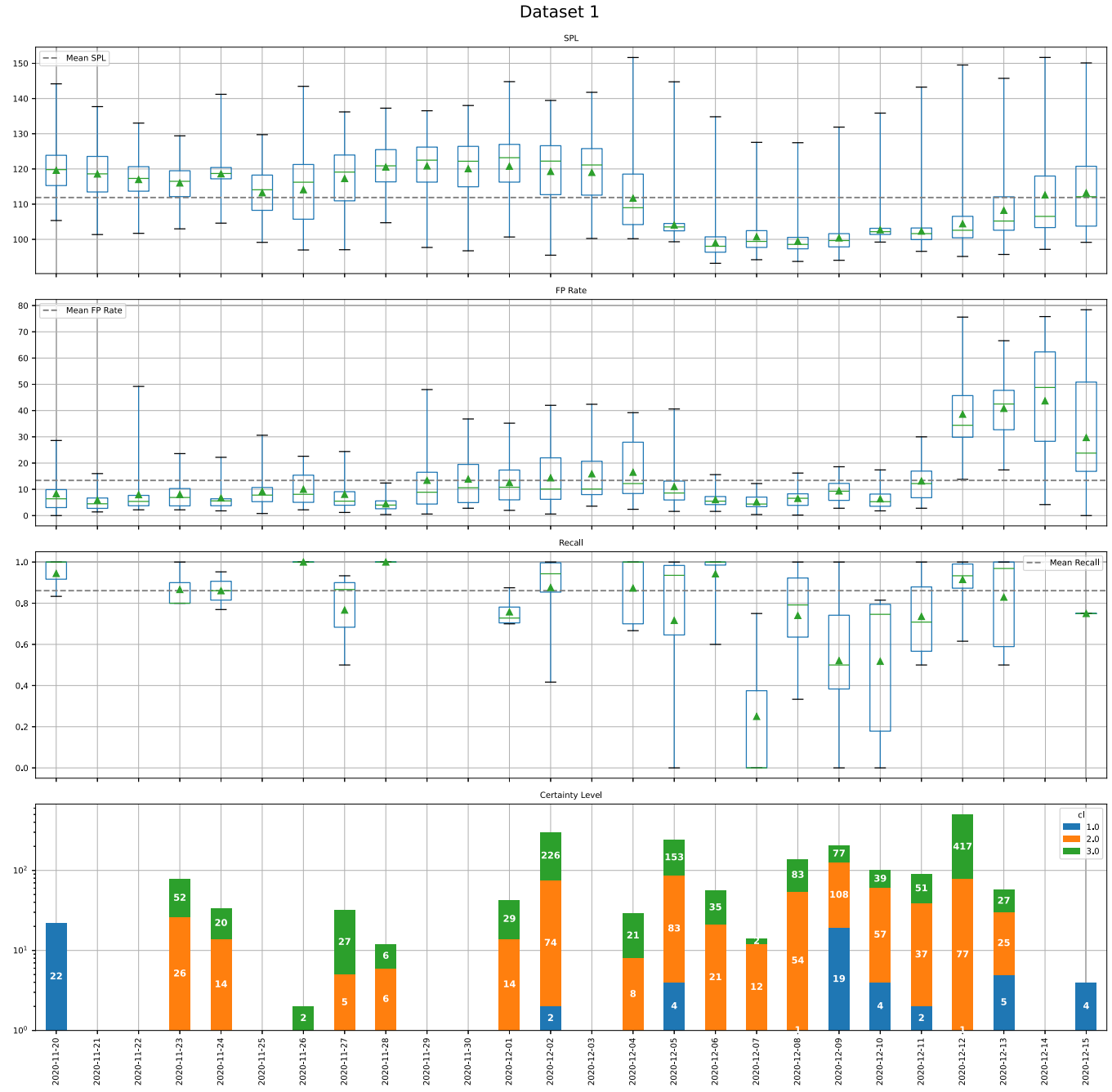


Fig. B.1. Dataset 1 – Temporal evolution of the sound pressure level (SPL, dB), the false positive (FP) rate (FP/H), the recall and the certainty level (CL) of the annotations. The sound pressure level was calculated with an integration time of one second and with a reference pressure of 1 µPa. The false positive rate and the recall were calculated for the marine mammal detection (without classification). The false positive rate corresponds to the number of false positives per hour. The recall was calculated per hour. The certainty level of the annotations was averaged and rounded per spectrograms (the y-axis was log-scaled to visualize the low presence of annotations with a low certainty level).

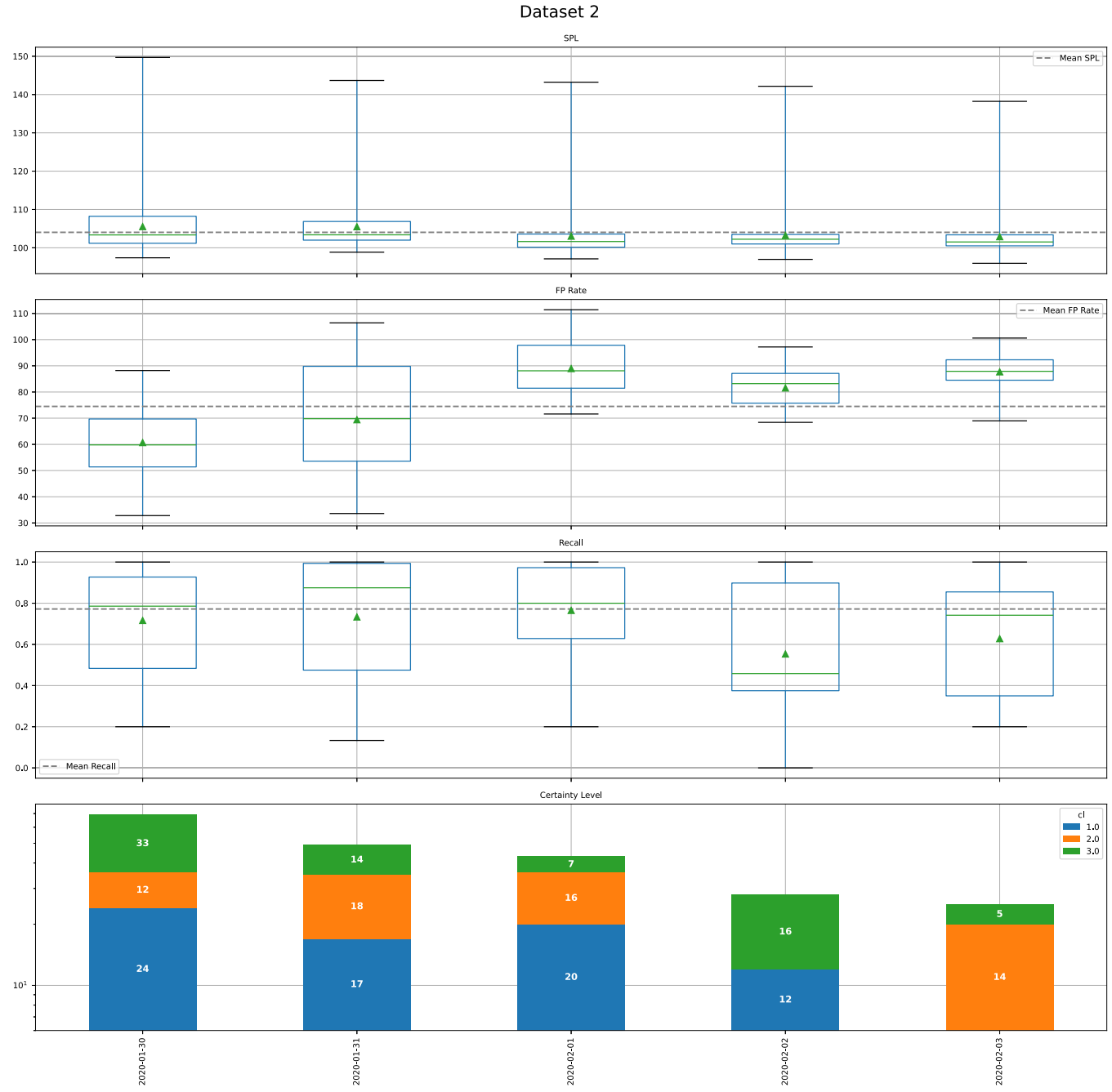


Fig. B.2. Dataset 2 – Temporal evolution of the sound pressure level (SPL, dB), the false positive (FP) rate (FP/H), the recall and the certainty level (CL) of the annotations. The sound pressure level was calculated with an integration time of one second and with a reference pressure of 1 µPa. The false positive rate and the recall were calculated for the marine mammal detection (without classification). The false positive rate corresponds to the number of false positives per hour. The recall was calculated per hour. The certainty level of the annotations was averaged and rounded per spectrograms.

Data availability

The source code of this work and the data are available on GitLab: https://gitlab.france-energies-marines.org/Quentin/owfsomm. The data are protected by a license CC BY-NC-ND 4.0. The full data are available upon request for academic research use.

References

Allen, A., Harvey, M., Harrell, L., Jansen, A., Merkens, K., Wall, C., Cattiau, J., Oleson, E., 2021. A convolutional neural network for automated detection of humpback whale song in a diverse, Long-term passive acoustic dataset. Front. Mar. Sci. 08, 607321. https://doi.org/10.3389/fmars.2021.607321.

Beyan, C., Browman, H., 2020. Setting the stage for the machine intelligence era in marine science. ICES J. Mar. Sci. 77, 1267–1273. https://doi.org/10.1093/icesjms/fsaa084.

Bowen, W., Iverson, S., 2013. Methods of estimating marine mammal diets: a review of validation experiments and sources of bias and uncertainty. Mar. Mamm. Sci. 29. https://doi.org/10.1111/j.1748-7692.2012.00604.x.

Brautaset, O., Waldeland, U., Johnsen, E., Malde, K., Eikvil, L., Salberg, A.-B., Handegard, N.O., 2020. Acoustic classification in multifrequency echosounder data using deep convolutional neural networks. ICES J. Mar. Sci. 77, 1391–1400. https://doi.org/10.1093/icesjms/fsz235.

Clausen, K., Wahlberg, M., Beedholm, K., Deruiter, S., Madsen, P., 2011. Click communication in harbor porpoises *Phocoena phocoena*. Bioacoustics Int. J. Anim. Sound Its Rec. Bioacoustics 20, 1–28. https://doi.org/10.1080/09524622.2011.9753630.

Clausen, K.T., Tougaard, J., Carstensen, J., Delefosse, M., Teilmann, J., 2019. Noise affects porpoise click detections – the magnitude of the effect depends on logger type and detection filter settings. Bioacoustics 28, 443–458. https://doi.org/10.1080/09524622.2018.1477071.

Coffey, K., Marx, R., Neumaier, J., 2019. DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations. Neuropsychopharmacology 44. https://doi.org/10.1038/s41386-018-0303-6.

Dubus, G., Cazau, D., Torterotot, M., Gros-Martial, A., Nguyen Hong Duc, P., Adam, O., 2024. From citizen science to AI models: advancing cetacean vocalization automatic detection through multi-annotator campaigns. Eco. Inform. 81, 102642. https://doi.org/10.1016/j.ecoinf.2024.102642.

Dudzinski, K., Thomas, J., Gregg, J., 2009. Communication in marine mammals. Encyclop. Mar. Mammals 260–269. https://doi.org/10.1016/B978-0-12-373553-9.00064-X.

Escobar-Amado, C., Badiey, M., Wan, L., 2024. Computer vision for bioacoustics: detection of bearded seal vocalizations in the Chukchi shelf using YOLOV5. IEEE J. Ocean. Eng. 49, 133–144. https://doi.org/10.1109/JOE.2023.3307175.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96. AAAI Press, Portland, Oregon, pp. 226–231.

Ferguson, E.L., Sugarman, P., Coffey, K.R., Pettis Schallert, J., Alongi, G.C., 2022. Development of deep neural networks for marine mammal call detection using an open-source, user friendly tool. J. Acoust. Soc. Am. 151, A28. https://doi.org/10.1121/10.0010547.

Frasier, K., Garrison, L., Soldevilla, M., Wiggins, S., Hildebrand, J., 2021. Cetacean distribution models based on visual and passive acoustic data. Sci. Rep. 11. https://doi.org/10.1038/s41598-021-87577-1.

Gervaise, C., Simard, Y., Aulanier, F., Roy, N., 2021. Optimizing passive acoustic systems for marine mammal detection and localization: application to real-time monitoring North Atlantic right whales in Gulf of St. Lawrence. Appl. Acoust. 178, 107949. https://doi.org/10.1016/j.apacoust.2021.107949.

Gillespie, D., Mellinger, D.K., Gordon, J., McLaren, D., Redmond, P., McHugh, R., Trinder, P., Deng, X., Thode, A., 2009. PAMGUARD: Semiautomated, open source software for real-time acoustic detection and localization of cetaceans. J. Acoustical Soc. Am. 125, 2547. https://doi.org/10.1121/1.4808713.

Goodwin, M., Halvorsen, K., Jiao, L., Knausgård, K., Martin, A., Moyano, M., Oomen, R., Rasmussen, J.H., Sørdalen, T., Thorbjørnsen, S., 2022. Unlocking the potential of deep learning for marine ecology: overview, applications, and outlook. ICES J. Mar. Sci. 79. https://doi.org/10.1093/icesjms/fsab255.

Hammond, P., Francis, T., Heinemann, D., Long, K., Moore, J., Punt, A., Reeves, R., Sepulveda, M., Sigurdsson, G., Siple, M., Víkingsson, G., Wade, P., Williams, R., Zerbini, A., 2021. Estimating the abundance of marine mammal populations. Front. Mar. Sci. 8, 735770. https://doi.org/10.3389/fmars.2021.735770.

Hart, P.J., Ibanez, T., Paxton, K., Tredinnick, G., Sebastián-González, E., Tanimoto-Johnson, A., 2021. Timing is everything: acoustic niche partitioning in two tropical wet Forest bird communities. Front. Ecol. Evol. 9. https://doi.org/10.3389/fevo.2021.753363.

Hazen, E., Abrahms, B., Brodie, S., Carroll, G., Jacox, M., Savoca, M., Scales, K., Sydeman, W., Bograd, S., 2019. Marine top predators as climate and ecosystem sentinels. Front. Ecol. Environ. 17. https://doi.org/10.1002/fee.2125.

Heaton, J., Goodfellow, I., Bengio, Y., Courville, A., 2018. Deep learning. Genet. Program Evolvable Mach. 19, 305–307. https://doi.org/10.1007/s10710-017-9314-z.

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

Honda, H., 2020. Digging into Detectron 2. Medium. URL. https://medium.com/@hirotoschwert/digging-into-detectron-2-47b2e794fabd (accessed 1.14.24).

Jeantet, L., Dufourq, E., 2023. Improving deep learning acoustic classifiers with contextual information for wildlife monitoring. Eco. Inform. 77, 102256. https://doi.org/10.1016/j.ecoinf.2023.102256.

Jones, B., Zapetis, M., Samuelson, M., Ridgway, S., 2019. Sounds produced by bottlenose dolphins (Tursiops): a review of the defining characteristics and acoustic criteria of the dolphin vocal repertoire. Bioacoustics 29. https://doi.org/10.1080/09524622.2019.1613265.

Lambert, C., Virgili, A., Pettex, E., Delavenne, J., Toison, V., Blanck, A., Ridoux, V., 2017. Habitat modelling predictions highlight seasonal relevance of marine protected areas for marine megafauna. Deep Sea Res. Part II Top. Stud. Oceanogr. 141. https://doi.org/10.1016/j.dsr2.2017.03.016.

LeBien, J., Zhong, M., Campos Cerqueira, M., Velev, J., Dodhia, R., Lavista Ferres, J., Aide, T.M., 2020. A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. Ecol. Inform. https://doi.org/10.1016/j.ecoinf.2020.101113, 101113.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539.

Li, L., Qiao, G., Liu, S., Qing, X., Zhang, H., Mazhar, S., Niu, F., 2021. Automated classification of *Tursiops aduncus* whistles based on a depth-wise separable convolutional neural network and data augmentation. J. Acoust. Soc. Am. 150, 3861–3873. https://doi.org/10.1121/10.0007291.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 936–944. https://doi.org/10.1109/CVPR.2017.106.

Liu, R., Lehman, J., Molino, P., Such, F.P., Frank, E., Sergeev, A., Yosinski, J., 2018. An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution. https://doi.org/10.48550/arXiv.1807.03247.

Malde, K., Handegard, N.O., Eikvil, L., Salberg, A.-B., 2020. Machine intelligence and the data-driven future of marine science. ICES J. Mar. Sci. 77, 1274–1285. https://doi.org/10.1093/icesjms/fsz057.

Mariano-Jelicich, R., Berón, P., Copello, S., Dellabianca, N., García, G., Labrada-Martagón, V., Paso Viola, M., Paz, J., Riccialdelli, L., San Martin, A., Seco Pon, J.P., Torres, M., Favero, M., 2021. Marine Megafauna Sea Turtles, Seabirds and Marine Mammals, pp. 297–324. https://doi.org/10.1201/9780429399244-14.

McGarry, T., De Silva, R., Canning, S., Mendes, S., Prior, A., Stephenson, S., Wilson, J., 2022. Evidence base for application of Acoustic Deterrent Devices (ADDs) as marine mammal mitigation (Version 4). JNCC Rep. No 615 JNCC Peterb.

McInnes, L., Healy, J., Melville, J., 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. https://doi.org/10.48550/arXiv.1802.03426.

Mellinger, D., Stafford, K., Moore, S., Dziak, B., Matsumoto, H., 2007. An overview of fixed passive acoustic observation methods for cetaceans. Oceanography 20, 36–45. https://doi.org/10.5670/oceanog.2007.03.

Morgan, M.M., Braasch, J., 2021. Long-term deep learning-facilitated environmental acoustic monitoring in the capital region of New York state. Eco. Inform. 61, 101242. https://doi.org/10.1016/j.ecoinf.2021.101242.

Motamedi, M., Sakharnykh, N., Kaldewey, T., 2021. A Data-Centric Approach for Training Deep Neural Networks with Less Data. https://doi.org/10.48550/arXiv.2110.03613.

Mutanu, L., Gohil, J., Gupta, K., Wagio, P., Kotonya, G., 2022. A review of automated bioacoustics and general acoustics classification research. Sensors 22, 8361. https://doi.org/10.3390/s22218361.

Nguyen Hong Duc, P., Torterotot, M., Samaran, F., White, P., Gerard, O., Adam, O., Cazau, D., 2021. Assessing inter-annotator agreement from collaborative annotation campaign in marine bioacoustics. Ecol. Inform. 61, 101185. https://doi.org/10.1016/j.ecoinf.2020.101185.

Noble, A.E., Jensen, F.H., Jarriel, S.D., Aoki, N., Ferguson, S., Hyer, M.D., Apprill, A., Mooney, T.A., 2024. Unsupervised clustering reveals acoustic diversity and niche differentiation in pulsed calls from a coral reef ecosystem. Front. Remote Sens. 5. https://doi.org/10.3389/frsen.2024.1429227.

Nolasco, I., Singh, S., Vidaña-Vila, E., Grout, E., Morford, J., Emmerson, M., Jensens, F., Whitehead, H., Kiskin, I., Strandburg-Peshkin, A., Gill, L., Pamula, H., Lostanlen, V., Morfi, V., Stowell, D., 2022. Few-shot Bioacoustic Event Detection at the DCASE 2022 Challenge. https://doi.org/10.48550/arXiv.2207.07911.

Nowacek, D., Christiansen, F., Bejder, L., Goldbogen, J., Friedlaender, A., 2016. Studying cetacean behaviour: new technological approaches and conservation applications. Anim. Behav. 120. https://doi.org/10.1016/j.anbehav.2016.07.019.

Nuuttila, H., Thomas, L., Hiddink, J., Austin, R., Turner, J., Bennell, J., Tregenza, N., Evans, P., 2013. Acoustic detection probability of bottlenose dolphins, *Tursiops truncatus*, with static acoustic dataloggers in Cardigan Bay, Wales. J. Acoust. Soc. Am. 134, 2596–2609. https://doi.org/10.1121/1.4816586.

Nuuttila, H., Brundiers, K., Dähne, M., Koblitz, J., Thomas, L., Courtene-Jones, W., Evans, P., Turner, J., Bennell, J., Hiddink, J., 2018. Estimating effective detection area of static passive acoustic data loggers from playback experiments with cetacean vocalisations. Methods Ecol. Evol. 9. https://doi.org/10.1111/2041-210x.13097.

Padovese, B., Frazao, F., Kirsebom, O., 2021. Data augmentation for the classification of North Atlantic right whales upcalls. J. Acoust. Soc. Am. 149, 2520–2530. https://doi.org/10.1121/10.0004258.

Parsons, M., Lin, T.-H., Mooney, A., Erbe, C., Juanes, F., Lammers, M., Li, S., Linke, S., Looby, A., Nedelec, S., Van Opzeeland, I., Radford, C., Rice, A., Sayigh, L., Stanley, J., Urban, E., Di Iorio, L., 2022. Sounding the call for a global library of underwater biological sounds. Front. Ecol. Evol. 10. https://doi.org/10.3389/fevo.2022.810156.

Pham, P., Li, J., Szurley, J., Das, S., 2018. Eventness: Object detection on spectrograms for temporal localization of audio events. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2491–2495. https://doi.org/10.1109/ICASSP.2018.8462062.

Prasad, K., Dsouza, K., Bhargava, V., 2020. A downscaled faster-RCNN framework for signal detection and time-frequency localization in wideband RF systems. In: IEEE Trans. Wirel. Commun., p. 1. https://doi.org/10.1109/TWC.2020.2987990.

Quintana-Rizzo, E., Mann, D., Wells, R., 2006. Estimated communication range of social sounds used by bottlenose dolphins (*Tursiops truncatus*). J. Acoust. Soc. Am. 120, 1671–1683. https://doi.org/10.1121/1.2226559.

Romero Mujalli, D., Bergmann, T., Zimmermann, A., Scheumann, M., 2021. Utilizing DeepSqueak for automatic detection and classification of mammalian vocalizations: a case study on primate vocalizations. Sci. Rep. 11. https://doi.org/10.1038/s41598-021-03941-1.

Ruff, Z., Lesmeister, D., Appel, C., Sullivan, C., 2021. Workflow and convolutional neural network for automated identification of animal sounds. Ecol. Indic. 124, 107419. https://doi.org/10.1016/j.ecolind.2021.107419.

Schaffeld, T., 2016. Reduction of Accidental Bycatches in Set-Net Fisheries by the Use of Acoustic Signals.

Sequeira, A., Heupel, M., Lea, M.-A., Eguíluz, V., Duarte, C., Meekan, M., Thums, M., Calich, H., Carmichael, R., Costa, D., Cerqueira Ferreira, L., Fernandéz-Gracia, J., Harcourt, R., Harrison, A.-L., Jonsen, I., McMahon, C., Sims, D., Wilson, R., Hays, G., 2019. The importance of sample size in marine megafauna tagging studies. Ecol. Appl. 29. https://doi.org/10.1002/eap.1947.

Shinde, P.P., Shah, S., 2018. A review of machine learning and deep learning applications. In: 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). Presented at the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), pp. 1–6. https://doi.org/10.1109/ICCUBEA.2018.8697857.

Shiu, Y., Palmer, K., Roch, M., 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, 607. https://doi.org/10.1038/s41598-020-57549-y.

Sørensen, P.M., Wisniewska, D., Jensen, F., Johnson, M., Teilmann, J., Madsen, P., 2018. Click communication in wild harbour porpoises (*Phocoena phocoena*). Sci. Rep. 8. https://doi.org/10.1038/s41598-018-28022-8.

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

Todd, V.L.G., Lazar, L., Williamson, L.D., Peters, I.T., Hoover, A.L., Cox, S.E., Todd Ian, B., Macreadie, P.I., McLean, D.L., 2020. Underwater visual records of marine megafauna around offshore anthropogenic structures. Front. Mar. Sci. 7. https://doi.org/10.3389/fmars.2020.00230.

Todd, N.R.E., Kavanagh, A.S., Rogan, E., Jessopp, M.J., 2023. What the F-POD? Comparing the F-POD and C-POD for monitoring of harbor porpoise (*Phocoena phocoena*). Ecol. Evol. 13, e10186. https://doi.org/10.1002/ece3.10186.

Towsey, M., Parsons, S., Sueur, J., 2014. Ecology and acoustics at a large scale. Ecol. Inform. Ecologic. Acoustics 21, 1–3. https://doi.org/10.1016/j.ecoinf.2014.02.002.

Ulloa, J.S., Haupert, S., Latorre, J.F., Aubin, T., Sueur, J., 2021. Scikit-maad: an open-source and modular toolbox for quantitative soundscape analysis in Python. Methods Ecol. Evol. 12, 2334–2340. https://doi.org/10.1111/2041-210X.13711.

Urick, R.J., 1983. Principles of Underwater Sound. McGraw-Hill.

Verboom, W., Kastelein, R.A., 1995. Acoustic signals by harbour porpoises (*Phocoena phocoena*). In: Harb. Porpoises Lab. Stud. Reduce Bycatch Eds Nachtigall PE Lien J Au WWL Read AJ Spil Publ. Woerden Neth, pp. 1–40.

Virgili, A., Racine, M., Authier, M., Monestiez, P., Ridoux, V., 2017. Comparison of habitat models for scarcely detected species. Ecol. Model. 346, 88–98. https://doi.org/10.1016/j.ecolmodel.2016.12.013.

Virgili, A., Authier, M., Monestiez, P., Ridoux, V., 2018. How many sightings to model rare marine species distributions. PLoS One 13, e0193231. https://doi.org/10.1371/journal.pone.0193231.

Waggitt, J., Evans, P., Andrade, J., Banks, A., Boisseau, O., Bolton, M., Bradbury, G., Brereton, T., Camphuysen, C., Durinck, J., Felce, T., Fijn, R., García-Barón, I., Garthe, S., Geelhoed, S., Gilles, A., Goodall, M., Haelters, J., Hamilton, S., Hiddink, J., 2019. Distribution maps of cetacean and seabird populations in the North-East Atlantic. J. Appl. Ecol. 57, 253–269. https://doi.org/10.1111/1365-2664.13525.

White, E.L., Klinck, H., Bull, J.M., White, P.R., Risch, D., 2023. One size fits all? Adaptation of trained CNNs to new marine acoustic environments. Eco. Inform. 78, 102363. https://doi.org/10.1016/j.ecoinf.2023.102363.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., Girshick, R., 2019. Detectron2.

Wu, S.-H., Chang, H.-W., Lin, R.-S., Tuanmu, M.-N., 2021. SILIC: a cross database framework for automatically extracting robust biodiversity information from soundscape recordings based on object detection and a tiny training dataset. Ecol. Inform. 68, 101534. https://doi.org/10.1016/j.ecoinf.2021.101534.

Xu, H., Wang, X., Shao, F., Duan, B., Zhang, P., 2021. Few-Shot Object Detection via Sample Processing. In: IEEE Access, 9, p. 1. https://doi.org/10.1109/ACCESS.2021.3059446.

Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J., 2023. Object detection in 20 years: a survey. In: Proc. IEEE, pp. 1–20. https://doi.org/10.1109/JPROC.2023.3238524.