Engineering Applications of Artificial Intelligence 167 (2026) 113756

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Research paper

# Iterative deep learning for cetacean whistle detection in the Strait of Gibraltar

Alba Márquez-Rodríguez [a,*,1], Neus Pérez-Gimeno [a,b,1], Daniel Benítez-Aragón [a,b], Gonzalo M. Arroyo [a,c], Andrés De la Cruz [a,c]

[a] University of Cádiz, Institute of Marine Research (INMAR), Puerto Real, Cádiz, Spain
[b] University of Cádiz, Laboratory of Acoustic Engineering, Institute of Marine Research (INMAR), Puerto Real, Cádiz, Spain
[c] University of Cádiz, Department of Biology, Puerto Real, Cádiz, Spain

ARTICLE INFO

Dataset link: https://github.com/SEANIMALMOVE/CetaceanWhistleDetection_StraitOfGibraltar

*Keywords:*
Convolutional neural networks
Deep learning
Ecoacoustics
Passive acoustic monitoring
Strait of Gibraltar
Underwater acoustics

ABSTRACT

Deep learning has shown remarkable potential for complex signal recognition, yet its deployment in noisy real-world environments remains a major challenge. This study presents an adaptive deep learning framework for acoustic signal detection, demonstrated through the identification of cetacean whistles in the Strait of Gibraltar, a mandatory migratory bottleneck where several species coexist with intense maritime traffic and complex acoustic conditions. Passive acoustic monitoring (PAM) enables long-term assessment of marine mammal presence and behavior, yet its application remains challenging due to overlapping anthropogenic and environmental sounds and the large volumes of data generated. The proposed framework combines transfer learning, semi-supervised iterative fine-tuning, and confidence-threshold calibration to automate the analysis of PAM data and improve robustness in acoustically variable conditions. Pre-trained bioacoustic models (BirdNET and Perch) were repurposed as feature extractors and coupled with custom classifiers to assess generalization across domains. Model performance was evaluated using standard deep learning metrics, including accuracy, recall, and F1-score. Models were validated using a clean benchmark dataset and a real-world deployment dataset characterized by substantial anthropogenic and geophonic noise. While baseline models achieved over 0.95 accuracy on clean data, their performance degraded under realistic noise conditions (whistle F1-score $\leq$ 0.10). The best-performing model, fine-tuned with local data and semi-supervised iterative validation, achieved an F1-score of 0.88 and improved recall across deployments. Confidence-threshold optimization further enhanced adaptability, supporting both automated and expert-assisted monitoring workflows. The final model was tested across three independent acoustic deployments spanning different seasons and locations. Results demonstrated robust whistle detection with minimal annotation effort. This work provides a reproducible, AI-driven framework for signal detection in noisy environments, demonstrating how adaptive learning and threshold calibration can extend the applicability of deep neural networks to real-world acoustic monitoring and contribute to the development of automated, ecologically meaningful soundscape analysis systems.

## 1. Introduction

Many marine organisms rely on sound to interpret their environment, using it for essential functions such as communication, mating, prey and predator detection, navigation, habitat selection, and echolocation (Peng et al., 2015; de Soto, 2016; Popper and Hawkins, 2018). A prominent example are cetaceans, whose marked sensitivity to acoustic variation makes them key biological indicators of the acoustic health of marine ecosystems. Shifts in their behavioral patterns and vocal emissions can reflect environmental disturbances across spatial and temporal scales, capturing ecosystem changes at both population and community levels (Hazen et al., 2019; Cartagena-Matos et al., 2021; Wang and Houser, 2023; Guan and Brookens, 2023).

For cetaceans, whose vital behaviors rely on acoustic cues, the marine soundscape represents a complex acoustic environment shaped by biological (biophony), environmental (geophony), and anthropogenic (anthrophony) sources, integrating their contributions across space and time (Pijanowski et al., 2011; Duarte et al., 2021). Within this approach, soundscape analysis enables the assessment of acoustic diversity, the characterization of habitat conditions, and the detection of alterations caused by increasing underwater noise (Havlik et al., 2022).

Passive acoustic monitoring (PAM) has become a key methodology for the non-invasive study of marine mammals, allowing the detection and analysis of their acoustic emissions in natural habitats (Van Parijs et al., 2009). This technique is particularly effective in

* Corresponding author.
*E-mail address:* alba.marquez@uca.es (A. Márquez-Rodríguez).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.engappai.2026.113756
Received 4 November 2025; Received in revised form 23 December 2025; Accepted 5 January 2026

low-visibility environments for continuous data collection, facilitating long-term ecosystems monitoring (Zimmer, 2011; Howe et al., 2019; Ross et al., 2023).

Despite its advantages, PAM techniques generate massive volumes of data that require intensive manual analysis due to the need for expert-driven inspection of spectrograms and signal characteristics to detect and classify biologically relevant sounds (Mellinger et al., 2007; Van Parijs et al., 2009; Roch et al., 2011). This process is labor intensive and limits scalability. To overcome this bottleneck, automated methods, including classification approaches based on Machine and Deep Learning, have been introduced to facilitate the processing of large acoustic datasets (Usman et al., 2020; Shiu et al., 2020; Tuia et al., 2022).

Early computational approaches for bioacoustic signal classification relied on classical Machine Learning algorithms such as Random Forest, Support Vector Machines (SVMs), and Hidden Markov Models (HMMs), which require manual feature extraction and expert-driven selection (Mellinger et al., 2007; Roch et al., 2011; Frederick et al., 2020). While these methods improved processing efficiency, their performance remained limited in terms of generalization, particularly in acoustically diverse or noisy environments (Bravo Sanchez et al., 2021; White et al., 2022).

In recent years, Deep Learning has emerged as a powerful alternative, capable of learning hierarchical representations directly from raw or minimally processed data, thereby reducing the dependency on handcrafted features and enhancing robustness to variable acoustic conditions (Frederick et al., 2020; Stowell, 2022). Among these, Convolutional Neural Networks (CNNs) have shown particular promise for bioacoustic applications, especially in the classification of time-series and spectrogram-based representation of audio data (LeCun et al., 2015). CNNs automatically extract relevant acoustic features visible in the spectrogram, making them well suited for the detection and classification of bioacoustic signals, including cetacean whistles (Gibb et al., 2019; Padovese et al., 2023; Márquez-Rodríguez et al., 2025). Their effectiveness has been demonstrated in dolphin whistle detection (Nur Korkmaz et al., 2023) and in generalizing across complex marine soundscapes (Gibb et al., 2024), making them particularly suitable for large acoustic datasets, which often contain high levels of ambient noise (Bergler et al., 2019).

Recent studies have highlighted the advantages of cross-domain transfer learning strategies, in which pretrained bioacoustic models are adapted to a different target domain that may exhibit distinct acoustic and ecological characteristics. In Deep Learning, a *domain* typically refers to a combination of data distribution and feature space, such as those associated with avian versus marine soundscapes. In bioacoustics, cross-domain adaptation often involves repurposing models trained on terrestrial or bird vocalizations for underwater applications, despite differences in frequency range, background noise levels, and recording conditions (Stowell, 2022; Ghani et al., 2023; Williams et al., 2024).

Two notable bioacoustic models leveraging CNNs are BirdNET (Kahl et al., 2021) and Perch (Williams et al., 2024), both originally developed for the classification of bird vocalizations. These models were trained on large-scale, heterogeneous avian datasets and have demonstrated the ability to generalize across diverse acoustic contexts. Their internal feature representations (known as *embeddings*) capture high-dimensional acoustic patterns extracted from input spectrograms. These embeddings can be reused as input features for new classification tasks, offering a computationally efficient way to transfer knowledge without retraining the entire network (Ghani et al., 2023). This strategy has shown particular promise in the detection of marine mammals, including cetacean whistles, particularly when combined with custom classifiers or fine-tuning approaches (Padovese et al., 2023; Licciardi and Carbone, 2024).

This capability becomes particularly relevant in the context of marine environments characterized by high levels of background noise. A significant percentage of the world's oceans, estimated between 50% and 70%, exhibit ambient noise levels exceeding 90 dB re 1 μPa, particularly in regions heavily impacted by anthropogenic activities (McKenna et al., 2012; Stanley et al., 2017; Wisniewska et al., 2018). Specifically, in marine areas with high-density shipping corridors or industrial activity zones, the average sound pressure levels can reach critical values, with measurements exceeding 120 dB (Wisniewska et al., 2018; Javier et al., 2023). In light of this scenario, the application of CNN-based detection methods becomes especially promising in high-impact acoustic zones that are ecologically significant for cetaceans. The Strait of Gibraltar serves as a prime example of such a noisy environment, where noise levels are notably higher than the average for marine areas, primarily due to heavy maritime traffic and natural oceanographic conditions (Aranda et al., 2013; Criado-Aldeanueva et al., 2012).

The Strait of Gibraltar is a crucial connection zone between the Atlantic Ocean and the Mediterranean Sea, serving as the only natural link between these two water masses. The Strait spans approximately 60 km in length and is only 15 km wide at its narrowest point, separating the Spanish and Moroccan coasts. Its geographic and oceanographic characteristics play a fundamental role in the migratory movements of various marine species (Bruno et al., 2013; Arroyo et al., 2016; Gauffier et al., 2018; Ramos, 2019).

Throughout the annual cycle, the Strait of Gibraltar hosts several marine mammal species (Esteban et al., 2016; de Stephanis et al., 2008; Pons et al., 2022; Rojo-Nieto et al., 2011), including large cetaceans such as the sperm whale (*Physeter macrocephalus*) and the fin whale (*Balaenoptera physalus*) (Herr et al., 2020). The classification of cetacean signals in this context poses several challenges, including high ambient noise, overlapping anthropogenic sound sources, and the difficulty of species-level identification in whistle-rich recordings.

This study presents a Deep Learning pipeline for the detection of cetacean whistles in the Strait of Gibraltar, leveraging embeddings from bioacoustically pretrained CNNs (BirdNET and Perch) and a custom neural classifier. The method integrates model-assisted annotation with iterative expert validation to construct an increasingly robust training dataset from unlabeled recordings. The final model is evaluated across three independent acoustic deployments characterized by seasonal and spatial variability in background noise. This approach aims to optimize detection performance while minimizing manual effort, contributing to the development of scalable and generalizable tools for marine mammal monitoring. The results provide insights into the design of bioacoustic classifiers capable of operating under the challenging acoustic conditions typical of strategic marine corridors like the Strait of Gibraltar.

## 2. Material and methods

### 2.1. Data preparation

#### 2.1.1. Acoustic data acquisition

Acoustic data were collected using a single-channel recorder (Sylence RTSys), configured with a sampling rate of 256 kHz and capable of continuous recording for over a month.

Three deployments were conducted between May 2024 and March 2025 (Table 1) in the vicinity of Tarifa Island (Fig. 1) at a depth of 10 m. Each deployment was conducted under distinct seasonal conditions, aligned with the tourist seasons of the study area, considering that increased tourism correlates with higher levels of recreational vessel activity, including whale watching and diving (Ayuntamiento de Tarifa, 2016).

**Table 1**
Recording deployments. Summary of deployment periods, total recording durations, and corresponding seasonal noise contexts collected using the Sylence PAM system in the Strait of Gibraltar.

| Deployment | Period | Total hours | Seasonal context |
|---|---|---|---|
| *Deployment 1* | May 27 – June 13, 2024 | 367.3 | Mid-season (Moderate noise) |
| *Deployment 2* | June 16 – July 15, 2024 | 686.7 | Peak season (High noise) |
| *Deployment 3* | February 21– March 3, 2025 | 264.9 | Low season (Reduced noise) |
| **Total** | | **1318.9** | |

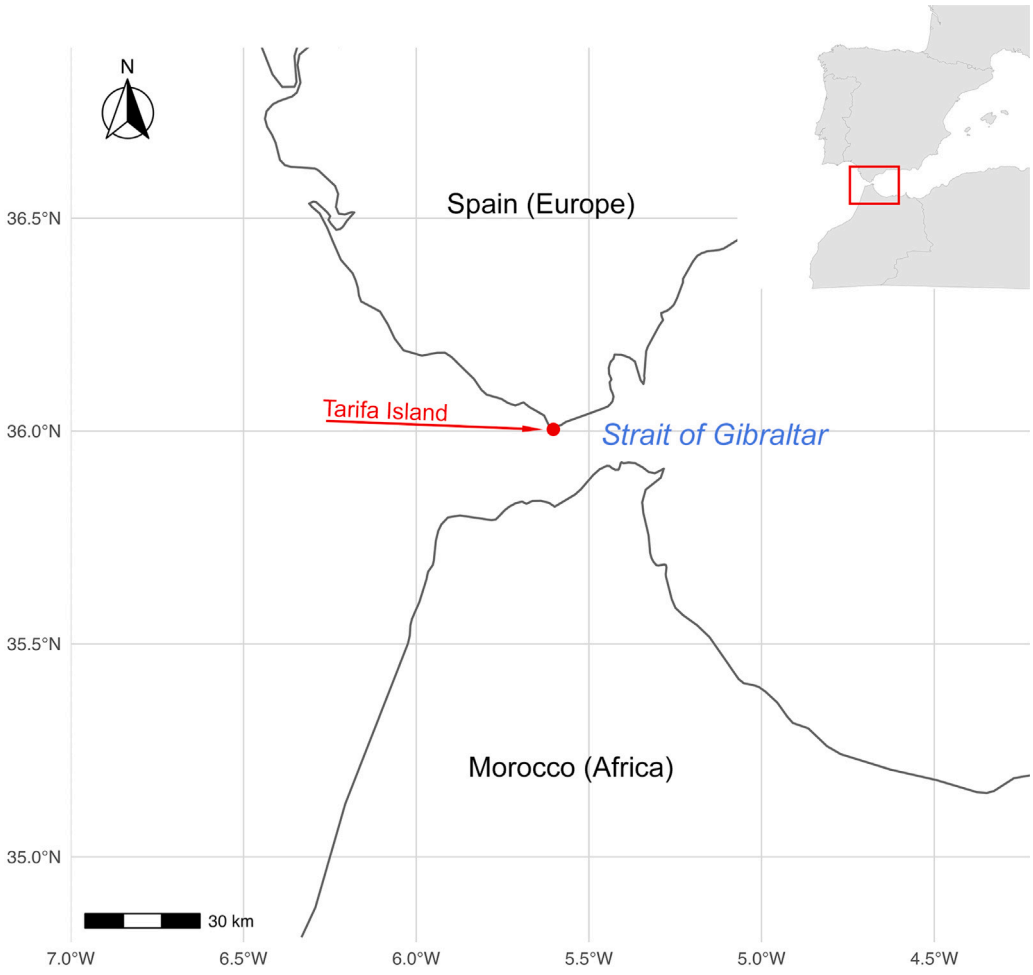


**Fig. 1.** Map of the study area. The area marks the southernmost point of mainland Spain, showing the location of the Sylence system deployments near the island of Tarifa, in the Strait of Gibraltar.

*2.1.2. Data preprocessing*

All audio recordings were segmented into fixed 3-second windows, and segments shorter than 3 s were zero-padded with silence to ensure uniform input length. This window duration reflects common practice in bioacoustic deep learning and was selected to satisfy the most restrictive input constraint among the models evaluated, namely BirdNET, which requires 3-second inputs (Kahl et al., 2021). Although other models used in this study (e.g., Perch) can accommodate longer segments, standardizing all inputs to 3 s ensured full compatibility across architectures and consistent feature extraction (Ghani et al., 2023). Padding preserves the complete temporal structure of detected events while maintaining a uniform input size, thereby reducing preprocessing-induced variability and ensuring reproducibility across models and deployments.

Sound events were then mapped into a binary classification framework: **Whistle** (any biophonic whistle-like vocalization, regardless of species) and **Background** (all non-whistle sounds, including anthropogenic noise, geophonic sources, and silence).

This preprocessing ensured compatibility with the fixed input requirements of the models and enabled precise control over the distribution of training, validation, and test subsets, avoiding data leakage or overestimation during evaluation. It also allowed for the integration of data augmentation techniques, applied dynamically during training, to improve generalization across diverse and noisy acoustic conditions.

Additional details on class balancing, dataset splitting, and data augmentation strategies are provided in the Appendix A. Detailed data preparation pipeline.

*2.1.3. Annotation process*

The raw recordings were initially annotated based on broad sound categories, including clicks, whistles, recreational boats and ferry noise.

**Table 2**
Summary of the datasets. Description of the origin, acoustic characteristics, and labeling methodology of each dataset used in model training and evaluation, including both curated global sound archives and field recordings from the Strait of Gibraltar.

| Dataset name | Data information | Labeling process |
|---|---|---|
| Watkins Marine Mammal Sound Archive (Sayigh et al., 2016) | High-quality, species-specific vocalizations with minimal background noise. | Provided pre-annotated vocalizations. Segments longer than 3 s were split, inheriting the original label. |
| SEANIMALMOVE - WOPAM | Originally annotated dataset based on three manually annotated days. Only non-biophonic sounds were retained. | Manually labeled sound events of varying durations. Segments exceeding 3 s were split while maintaining the original class label. |
| Laboratory of Acoustic Engineering of the University of Cadiz Marine Mammals Dataset | Locally recorded cetacean sounds with ambient noise, reflecting deep-sea conditions in the study area. | Manually reviewed and segmented into 3-second windows, ensuring exclusion of non-target noise. |
| Laboratory of Acoustic Engineering of the University of Cadiz Fast Ferries Dataset | Ambient noise recordings from deep-sea environments to improve model robustness against noise interference. | Manually reviewed and segmented into 3-second windows, ensuring exclusion of non-target noise. |
| SEANIMALMOVE HumanDivers, FastFerry and MotorBoat Dataset | Additional datasets created during model testing from classifier-detected positives. Later manually reviewed. Often contained anthropogenic noise such as divers, ferries, and boats. | A preliminary classifier identified potential positives. Segments were manually reviewed using the Validation Review application. Validated False Positives were relabeled as background noise and incorporated into the dataset. |

Only non-biophonic segments (such as anthropogenic and geophonic sounds) were classified as background noise. Annotations were performed by a single expert reviewer, and ambiguous segments were excluded to ensure data quality.

**Addressing domain mismatch: Domain adaptation & dataset integration.**

To train robust classification models and increase dataset size, annotated data from multiple sources were integrated to capture a wide range of acoustic conditions and signal types. These sources included both publicly available archives (Sayigh et al., 2016) and locally recorded datasets curated by the Acoustic Engineering Laboratory of the University of Cádiz and the SEANIMALMOVE project.

Table 2 summarizes the datasets used, detailing their acoustic characteristics and annotation methodologies. Notably, the training material combined high-quality vocalizations with minimal noise (e.g., Watkins Marine Mammals Dataset Sayigh et al., 2016) and complex, noisy recordings collected under field conditions in the Strait of Gibraltar. This integration introduced variability in signal-to-noise ratios, soundscape composition, and recording hardware, factors that can result in domain mismatch (Chotiros, 1995; Buckingham, 2000) and affect model generalization (Liang et al., 2024; Stowell, 2022; Van Merriënboer et al., 2024).

A key subset of locally recorded material, collected between June 7th and 9th, 2024, and initially annotated as part of the WOPAM initiative (WOPAM Project, 2025) was used as the foundation for a progressive, model-assisted annotation strategy. This subset from *Deployment 1* was essential for grounding the training dataset in the specific acoustic characteristics of the study site. The distribution of data origins across sound classes is shown in Figure C.1, illustrating the diversity of the sources and highlighting the potential for domain shifts requiring adaptation techniques during model development.

Since the target recordings were collected in very shallow waters around Tarifa Island, while some datasets originated from deep-sea environments and different regions, domain mismatch was expected due to differences in acoustic properties (Chotiros, 1995; Buckingham, 2000), affecting model generalization (Liang et al., 2024; Stowell, 2022; Van Merriënboer et al., 2024). To address these challenges, we applied targeted mitigation strategies during model training. First, a mixup-based data augmentation technique was employed with a fixed mixing ratio of 0.5, which provides a balanced blend (Abayomi-Alli et al., 2022; Guo et al., 2023). This method artificially combined clean cetacean whistles with noisy backgrounds, simulating how signals might appear in shallow, high-noise environments like the Strait of Gibraltar (Tokozume et al., 2017; Ghani et al., 2023; Nshimiyimana, 2024). Second, the training set was restricted to cetacean species known to produce whistle vocalizations, as advised by expert annotators. All other sounds, including geophonic noise, vessel traffic, and diver activity, were grouped under the **Background** class.

The resulting dataset covered a range of species and noise sources, enabling the model to learn distinguishing features of whistle vocalizations while improving robustness to background variability. Fig. 2 presents the final class distribution, showing the number of recordings and total accumulated duration per class.

**Iterative annotation strategy using *Deployment 1*.**

To improve model generalization under real-world acoustic conditions, we implemented an iterative annotation strategy that progressively expanded the training dataset using *Deployment 1* recordings. This approach aimed to address domain mismatch by exposing the model to the specific soundscape characteristics of the target environment, such as vessel traffic, geophonic noise, and shallow-water propagation effects, factors that were underrepresented in the initial training datasets. Similar semi-supervised or active, learning strategies have been shown to reduce labeling effort while improving performance in underwater acoustic detection tasks (Shiu et al., 2020; Bergler et al., 2019)

Rather than relying on random sampling or static annotations, the iterative strategy used model predictions to guide expert validation, prioritizing segments likely to contain cetacean whistles. The process began with model inference on a three-day subset (June 7–9, 2024) of *Deployment 1*, producing an initial annotated subset (Deployment-1-subset1, designated as *SEANIMALMOVE-WOPAM*). This subset, characterized by elevated anthropogenic noise, was reserved exclusively as an independent test set. Two additional three-day subsets (Subsets 2 and 3), representing comparatively cleaner acoustic conditions, were subsequently selected, annotated based on model-inferred detections, and incorporated into the training dataset. After each annotation cycle, the model was retrained and reapplied to a new unlabeled data subset, enabling progressive domain adaptation to local soundscape characteristics while substantially reducing manual labeling effort. Although not a fully automated active learning system, this hybrid passive–active strategy leverages model-guided annotation combined with expert validation and has been shown to be effective in underwater acoustic detection tasks (Bergler et al., 2019; Shiu et al., 2020; Kath et al., 2024). The complete workflow is illustrated in Fig. 3.

### 2.2. Models

The Deep Learning models used in this study, specifically BirdNET and Perch, are detailed below. Additionally, the integration of

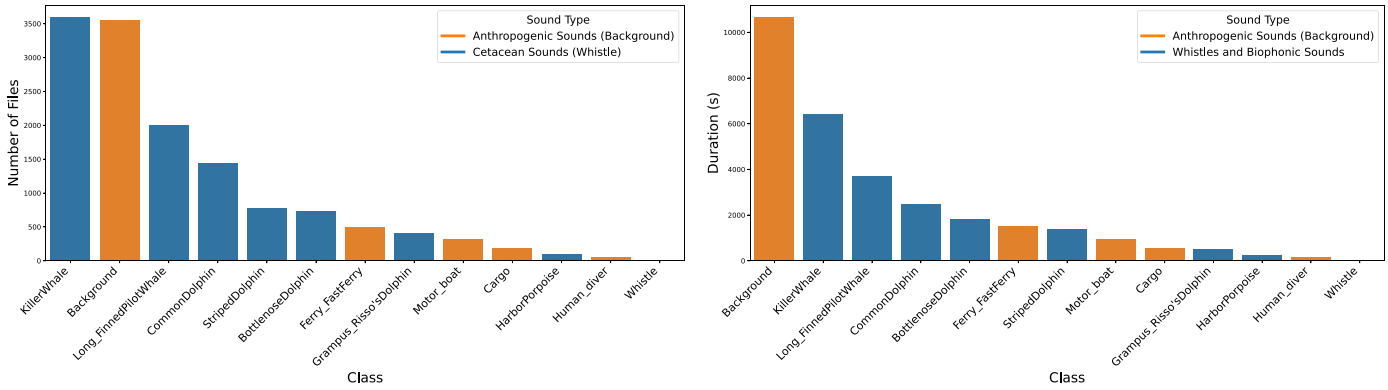


**Fig. 2.** Dataset composition. **Left:** Number of recordings per class. **Right:** Total accumulated duration per class (seconds). Cetacean whistles are in blue, and anthropogenic noise sources (e.g., vessel traffic, human divers) are in orange.

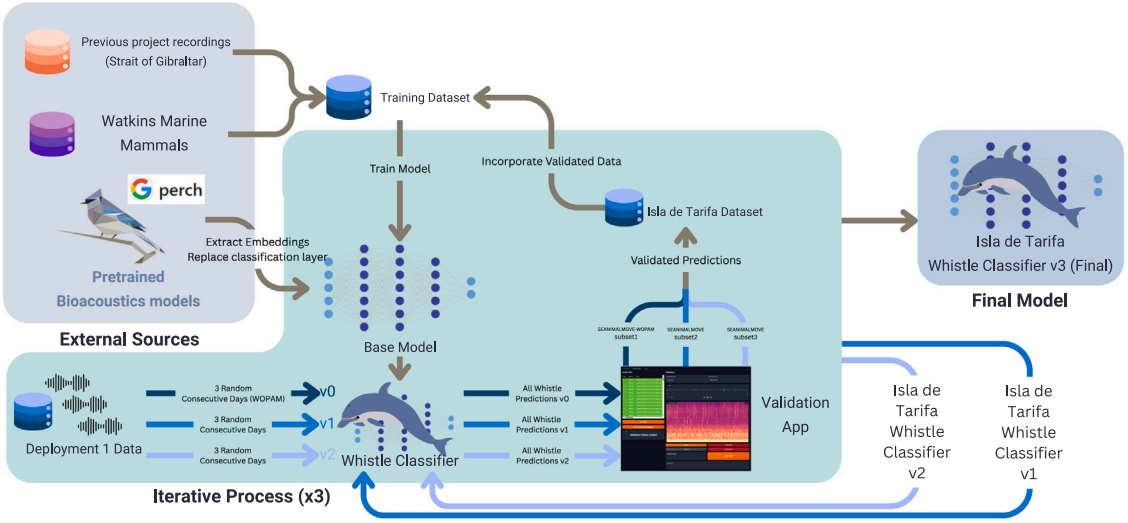


**Fig. 3.** Model development workflow. Diagram of the iterative training process for the final whistle classifier. A base model (v0) was first trained using public datasets (Watkins Marine Mammals) and recordings from prior local projects. This model was applied to three randomly selected days from *Deployment 1*, and all whistle predictions were manually validated through the custom web-based Validation App, forming subset 1 (*SEANIMALMOVE-WOPAM*). This data was incorporated into the Isla de Tarifa Dataset (dark blue arrow), and a new model (v1) was trained. The process was repeated with two more 3-day segments (middle and light blue arrows), producing Subsets 2 and 3 and leading to Whistle Classifiers v2 and v3. Each iteration expanded the training dataset with newly validated predictions, while the *SEANIMALMOVE-WOPAM* subset was retained as an independent test set. The final model (v3) was trained on all validated subsets and evaluated in real-world conditions, *SEANIMALMOVE-WOPAM*.

embeddings generated by these models with traditional Machine Learning algorithms, such as Random Forest, SVC-RBF, XGBoost and BirdNET custom neural network, are discussed with the aim of improving classification performance

#### 2.2.1. Feature extractors

- **BirdNET** is a Deep Learning model designed to classify bird species from audio recordings. It segments audio inputs into 3-second clips, converts them into spectrograms, and utilizes a CNN for classification. The model employs an EfficientNetB0-like architecture, producing 1024-dimensional embeddings that capture essential audio features. It covers frequencies from 0 Hz to 15 kHz, and it is trained on over 6500 classes, including 10 non-event categories. Non-event classes refer to categories that represent sounds or signals that are not related to bird vocalizations, such as background noise, anthropogenic sounds, or other environmental noises (Kahl et al., 2021).
- **Perch** is a Deep Learning model developed by Google Research for bird species classification. It utilizes an EfficientNet-B1 architecture and is trained on audio recordings from Xeno-Canto, encompassing over 10,000 bird species. The model generates 1280-dimensional vector embeddings from 5-second audio inputs, facilitating tasks such as similarity searches and classification (Ghani et al., 2023; Williams et al., 2024).

#### 2.2.2. Embedding-based classification

The embeddings extracted from BirdNET and Perch can serve as input features for traditional Machine Learning classifiers. Using the `OpenSoundscape` Python library (Lapp et al., 2023), embeddings from these models can be integrated with algorithms like Random Forest, SVM classifier with Radial Basis Function (SVC-RBF) kernels, and Gradient Boosting (XGBoost). This dual comparison between feature extractors and classification algorithms was designed to evaluate how different embedding representations and decision boundaries influence whistle detection performance. While BirdNET and Perch provide distinct feature spaces learned from large-scale bioacoustics datasets, traditional classifiers such as Random Forest, SVC-RBF, and XGBoost

**Table 3**
Best-performing classifier configurations for Perch and BirdNET embeddings. These configurations yielded the highest classification performance in distinguishing whistles from background noise.

| Classifier | Perch embeddings configuration | BirdNET embeddings configuration |
|---|---|---|
| Random Forest | max_depth = 50, n_estimators = 50 | max_depth = 10, n_estimators =50 |
| SVC-RBF | C = 100, gamma = 0.001, probability = True | C = 1, gamma = 0.001, probability = True |
| Gradient boosting classifier | max_depth = 10 | max_depth = 50 |

have demonstrated robust performance in bioacoustic classification tasks (Frederick et al., 2020; Henderson et al., 2012; Serra et al., 2019). Testing their combination therefore allowed assessing whether model performance was primarily constrained by the embedding representation or by the classifier's capacity to separate overlapping acoustic classes.

- **Random Forest** is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of their predictions for classification tasks (Breiman, 2001). This approach has already been used for ecological signal classification and biodiversity monitoring (Haider et al., 2023; Mohebbi-Kalkhoran, 2022; Márquez-Rodríguez et al., 2025).
- **SVC-RBF** kernels are well-established for handling non-linearly separable data through implicit feature transformation (Cortes and Vapnik, 1995; Han et al., 2012). The RBF kernel has demonstrated reliable performance in various bioacoustic applications, especially when paired with dense feature embeddings (de Stephanis et al., 2008; Mohebbi-Kalkhoran, 2022).
- **XGBoost** is a high-performance gradient boosting algorithm that constructs an ensemble of shallow decision trees with built-in regularization to prevent overfitting (Chen and Guestrin, 2016). It has shown advantages over Deep Learning methods in settings with limited labeled data or transient marine signals (Zhou et al., 2024; Nadir et al., 2020).

All classifiers were trained with grid search for hyperparameter optimization. Table 3 summarizes the best-performing configurations for each combination, selected based on the highest F1-Score obtained.

### 2.2.3. BirdNET-custom neural network

In addition to using BirdNET embeddings with traditional classifiers, a custom Deep Learning model, hereafter referred to as *BirdNET-custom neural network*, was developed to perform direct binary classification between whistle and background segments. The model was implemented using the `TensorFlow` `Keras` framework (Abadi et al., 2016; Gulli and Pal, 2017), which allows full control over the training process and enables customization of architectural components, optimization strategies, and evaluation metrics.

This approach facilitated a more customized and controlled adaptation of the general-purpose BirdNET feature extractor to the specific ecological and acoustic context of the study area, while maintaining flexibility to adjust training strategies as needed. A detailed description of the model architecture, training configuration, and optimization procedure is provided in the Appendix B. BirdNET-custom neural network details.

### 2.2.4. Model evaluation

To assess the classification performance of the developed models, we employed standard evaluation metrics using the `scikit-learn` Python library (Pedregosa et al., 2011). These metrics provide both class-specific and global indicators of model behavior under varying conditions, particularly in the presence of class imbalance.

- **True Positives (TP)**: Audio segments that contain a cetacean whistle and are correctly predicted as whistle segments.
- **False Positives (FP)**: Audio segments that do not contain a whistle and are incorrectly predicted as containing one.
- **True Negatives (TN)**: Audio segments that do not contain a whistle and are correctly predicted as background.
- **False Negatives (FN)**: Audio segments that contain a cetacean whistle and are incorrectly predicted as background.

Together, these outcomes form the basis of the **confusion matrix**, a tabular summary of a classifier's predictions. This matrix is typically represented as a $2 \times 2$ grid, where rows correspond to the actual labels and columns represent the predicted labels. It provides a comprehensive view of how the model differentiates between the target class (whistles) and the background class, enabling visual identification of common misclassifications and biases in prediction.

From these values, the following core performance metrics are derived:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

These metrics were calculated for both the whistle class, indicating the model's ability to correctly detect cetacean vocalizations, and the background class, which includes non-biophonic sounds such as anthropogenic and geophonic sounds. This class-level evaluation allows for a detailed understanding of model behavior in unbalanced datasets where one class (background) is more prevalent than the other (whistle).

**Accuracy** represents the overall proportion of correctly classified segments, encompassing, in this case, both, whistle and background detections. It provides a general indicator of model correctness by measuring how frequently predictions match the true labels. However, in highly unbalanced datasets, accuracy can be misleading, as a model may achieve a high value simply by favoring the majority class. For this reason, while accuracy offers a useful global measure, it is complemented by precision, recall, and F1-score, which better capture performance on the minority class (Juba and Le, 2019; Ghanem et al., 2023).

**Precision** quantifies the proportion of predicted positive segments (whistles) that are actually correct, while **recall** (or True Positive Rate, TPR) measures the proportion of actual whistle segments that the model successfully identified. These metrics are especially important in ecological acoustic monitoring, where the cost of FN or excessive FP can bias population estimates or effort allocation.

**F1-score** is the harmonic mean of precision and recall, offering a balanced measure when both metrics are equally important. F1-score assumes a symmetrical importance between detecting as many true whistles as possible and avoiding incorrect detections. This makes it a particularly suitable choice for the current study, where both sensitivity and specificity are valued in modeling cetacean acoustic presence.

In addition to reporting metrics for individual classes, two averaging strategies were used to summarize overall performance across classes:

- **Macro Average**: Arithmetic mean of a metric computed independently for each class (treats all classes equally).
- **Weighted Average**: Computes a class-wise average weighted by the support (number of true instances per class), thus reflecting class imbalance.

To further assess classifier behavior across varying decision thresholds, curve-based evaluation methods were employed:

- **Receiver Operating Characteristic (ROC) Curve**: Plots Recall vs. False Positive Rate (FPR), where

$$\text{FPR} = \frac{FP}{FP + TN} \tag{5}$$

  The **Area Under the ROC Curve (AUC-ROC)** quantifies overall classification performance across thresholds. A value of 1 indicates perfect classification, 0.5 implies random guessing.
- **Precision-Recall (PR) Curve**: Especially useful in class-imbalanced settings, it plots Precision vs. Recall. The **Area Under the PR Curve (AUC-PR)** indicates how well the classifier maintains high precision across recall levels.

To test generalization, we excluded annotations from the *SEANIMALMOVE-WOPAM* dataset during training and used them for final tests. This set, comprising 150 background segments and 6 whistle segments, is a real-world deployment scenario. Given the low occurrence of true whistle events, this test is particularly well-suited to evaluate FP behavior in response to common confounding sources in PAM data, such as diver and vessel noise.

**Confidence threshold analysis.**

To enable practical deployment of the whistle identification model, a confidence threshold analysis was performed to balance sensitivity and specificity. This analysis was based on the confidence scores produced by the BirdNET-custom neural network using data from *Deployment 1*. The objective was to identify a decision threshold that maximizes TP detections while minimizing FPs, particularly under the complex and noisy conditions typical of field deployments.

The first step involved the analysis of the ROC and PR curves. The ROC–AUC curve provided the statistically optimal threshold, defined as the point that maximizes the TPR for the lowest possible FPR. However, this discrimination-oriented criterion does not necessarily align with the operational priorities of ecological monitoring, where missing whistle events (FNs) can be more detrimental than including additional false detections.

To account for these practical considerations, a complementary threshold exploration was carried out. In this analysis, classification metrics (precision, recall, and F1-score) were computed across a continuous range of confidence scores, together with a logistic regression model fitted to prediction correctness. Thresholds that yielded higher F1-scores than the ROC–AUC–derived value were identified, as they represented a more favorable balance between precision and recall. Among these, the threshold achieving the highest recall was selected, ensuring greater sensitivity to whistle detections while maintaining acceptable levels of FPs.

This approach resulted in the identification of an alternative, operational confidence threshold better suited to our study objective. By explicitly prioritizing recall, it allowed the model to recover a larger number of true detections, while still providing flexibility for adaptive calibration depending on monitoring goals—either conservative, automated detection or expert-assisted validation workflows.

In addition to threshold selection based on summary performance metrics, the confidence scores produced by the BirdNET-custom neural network model were retained for all expert-validated detections across deployments. These scores were subsequently grouped by validation outcome and acoustic class to enable a post hoc, threshold-consistent characterization of model decisions. For a given confidence threshold, detections validated as cetacean whistles were labeled as TPs when their confidence exceeded the threshold and as FNs otherwise, whereas detections corresponding to non-cetacean sound classes were labeled as FPs when exceeding the threshold and as TNs when below it. This formulation provides a unified, threshold-dependent framework for categorizing detection outcomes directly from model confidence scores and expert annotations, ensuring consistency between performance metrics, operating-point selection, and subsequent error analysis.

## 3. Results

### 3.1. Classifier performance

Evaluation outcomes of the proposed "Cetacean Whistle Detector" focus on classifier performance, robustness under noisy conditions, and confidence threshold calibration. Table 4 summarizes the classification metrics for all tested models on both the internal Test Dataset and the external *SEANIMALMOVE-WOPAM* dataset. All models achieved high accuracy on the Test Dataset, ranging from 0.96 to 1.00, with F1-scores exceeding 0.95 across all model–classifier configurations. For instance, the Perch-SVC-RBF and BirdNET-SVC-RBF models reached F1-scores of 1.00, indicating both high precision and high recall when classifying whistle and background segments in low-noise, well-balanced scenarios.

However, performance declined when evaluated on the *SEANIMALMOVE-WOPAM* dataset. Although accuracy remained high in some configurations (e.g., 0.88 for Perch-SVC-RBF), whistle-specific F1-scores dropped substantially, ranging from 0.04 to 0.10 in most models. This performance degradation was not reflected in global metrics such as accuracy or macro-averaged F1-score, which remained moderately high due to a class imbalance favoring background segments. Most classifiers maintained background F1-scores above 0.80, confirming their ability to correctly label negative segments even under noisy conditions.

A comparison between Perch- and BirdNET-based classifiers showed similar behavior across datasets. On the Test Dataset, BirdNET-based classifiers slightly outperformed their Perch counterparts, often increasing metric scores from 0.98 to 0.99 or 1.00. In contrast, on *SEANIMALMOVE-WOPAM*, Perch-based classifiers generally exhibited better average performance. For example, the Perch-SVC-RBF model achieved an overall F1-score of 0.52, compared to 0.45 for BirdNET-SVC-RBF. However, this difference was largely driven by background classification, F1-scores for the background class were higher in Perch models (up to 0.94) than in BirdNET models (as low as 0.64). In contrast, whistle-specific F1-scores were similarly low across both model types, rarely exceeding 0.10.

The BirdNET-custom neural network achieved the best performance across all scenarios. On the Test Dataset, it obtained perfect scores in all metrics, including accuracy, precision, recall, and F1-score for both classes, indicating a correct classification of all whistle and background segments in a controlled conditions. It is important to note that this dataset represents clean and balanced data, where all models generally perform well, allowing this network to reach the upper limit of the evaluation metrics. When evaluated on the *SEANIMALMOVE-WOPAM* dataset, which contains substantial background noise and greater acoustic variability, performance naturally decreased but remained markedly superior to all other configurations. The model achieved an overall F1-score of 0.88, with 0.99 for background and 0.77 for whistle, and a recall of 0.91.

### 3.2. Confidence threshold analysis

Confidence score analysis revealed variation in detection behavior across thresholds. On *Deployment 1*, the ROC curve yielded an AUC of 0.87 and indicated 0.84 as an optimal threshold value, located near the elbow of the curve (Fig. 4). At this threshold, the model retained 63 TPs and 6 FPs, while 19 whistle events were missed (FNs). The PR curve corroborated this selection, showing a favorable balance between precision and recall at the 0.84 threshold (Fig. 5). However, this optimal point from a discrimination perspective does not necessarily maximize performance for the specific objectives of this study.

The logistic regression model fitted on prediction correctness (Figure C.2) supported this choice but also revealed the potential utility of adopting a less stringent threshold. Table C.1 shows the variation in classification metrics (precision, recall, and F1-score) across confidence

**Table 4**
Performance comparison of different models and classifiers on the Test and *SEANIMALMOVE-WOPAM* datasets. The results highlight the performance across various metrics, including accuracy, precision, recall, and F1-score for both background and whistle detection. Highlighted in gray, the best performing model in both datasets.

| Base model | Classification layer | Test dataset | | | | *SEANIMALMOVE-WOPAM dataset* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | Precision | Recall | F1-Score | Accuracy | Precision | Recall | F1-Score | F1-Score Background | F1-Score Whistle |
| Perch | Basic Neural Network | 0.98 | 0.98 | 0.98 | 0.98 | 0.86 | 0.51 | 0.53 | 0.50 | 0.92 | 0.08 |
| Perch | Random Forest | 0.97 | 0.98 | 0.97 | 0.97 | 0.71 | 0.49 | 0.45 | 0.43 | 0.83 | 0.04 |
| Perch | SVC-RBF | 1.00 | 1.00 | 1.00 | 1.00 | 0.88 | 0.52 | 0.54 | 0.52 | 0.94 | 0.10 |
| Perch | XGBoost | 0.96 | 0.96 | 0.95 | 0.95 | 0.72 | 0.49 | 0.45 | 0.44 | 0.83 | 0.04 |
| BirdNET | Basic Neural Network | 1.00 | 1.00 | 1.00 | 1.00 | 0.65 | 0.50 | 0.50 | 0.43 | 0.79 | 0.07 |
| BirdNET | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 0.48 | 0.49 | 0.41 | 0.35 | 0.64 | 0.05 |
| BirdNET | SVC-RBF | 1.00 | 1.00 | 1.00 | 1.00 | 0.71 | 0.50 | 0.53 | 0.45 | 0.82 | 0.08 |
| BirdNET | XGBoost | 0.99 | 0.99 | 0.98 | 0.99 | 0.51 | 0.49 | 0.42 | 0.36 | 0.67 | 0.05 |
| BirdNET | Custom neural network | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 | 0.85 | 0.91 | 0.88 | 0.99 | 0.77 |

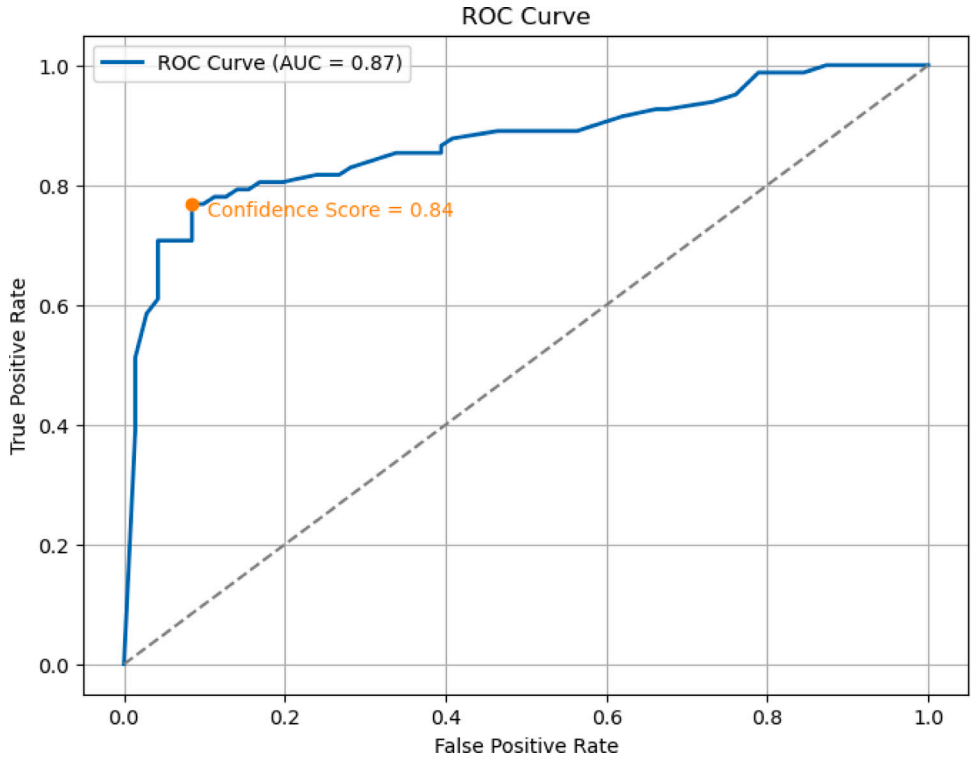


**Fig. 4.** ROC curve for the BirdNET-custom neural network evaluated on *Deployment 1*. AUC = 0.87. The $x$-axis shows the FPR, and the $y$-axis shows the TPR. The optimal operating threshold (0.84) is marked at the curve elbow.

score thresholds ranging from 0.70 to 0.90. As expected, precision increased progressively with higher thresholds, while recall decreased as the model became more conservative in labeling whistle detections. The ROC-AUC analysis identified 0.84 as the optimal discrimination point. However, multiple thresholds between 0.79 and 0.83 achieved comparable or slightly higher F1-scores, indicating stable classification performance within this interval. Among these, a threshold of 0.79 (logit $\approx$ 1.30) yielded the highest recall (0.793) with a precision of 0.867. This threshold recovered 82 true positives, 19 more than the 0.84 setting, at the cost of only four additional false positives. Both thresholds (0.79 and 0.84) were retained for subsequent analyses to evaluate their performance across deployments.

### 3.3. Performance in field deployments

The BirdNET-custom neural network was applied to two independent acoustic datasets (*Deployment 2* and *Deployment 3*). These recordings, which differed in background noise profiles and seasonal context, had not been used for training, validation or testing yet. Across these deployments, the classifier produced 375 whistle-labeled predictions in *Deployment 2* and 130 in *Deployment 3*, compared to 156 in *Deployment 1*.

Classification metrics were computed using three operating thresholds across the three different deployment scenarios: no confidence score threshold, a threshold of 0.84 selected based on ROC and PR curves, and a lower threshold of 0.79, which prioritizes recall (Table 5).

Without thresholding, all deployments exhibited whistle recall values of 1.00, as all whistle segments were correctly identified. In this configuration, the classification relied solely on the default decision rule of the binary model, where the predicted class corresponds to the one with the highest confidence score, effectively setting the decision threshold for the *Whistle* class at 0.5. Under these conditions, all segments, regardless of their actual label, were classified as whistles. As a result, although the model correctly recovered all true whistle events, it completely failed to identify background segments, leading to a background recall of 0.00 across all deployments. This imbalance resulted in a large number of FPs and poor overall performance. Macro

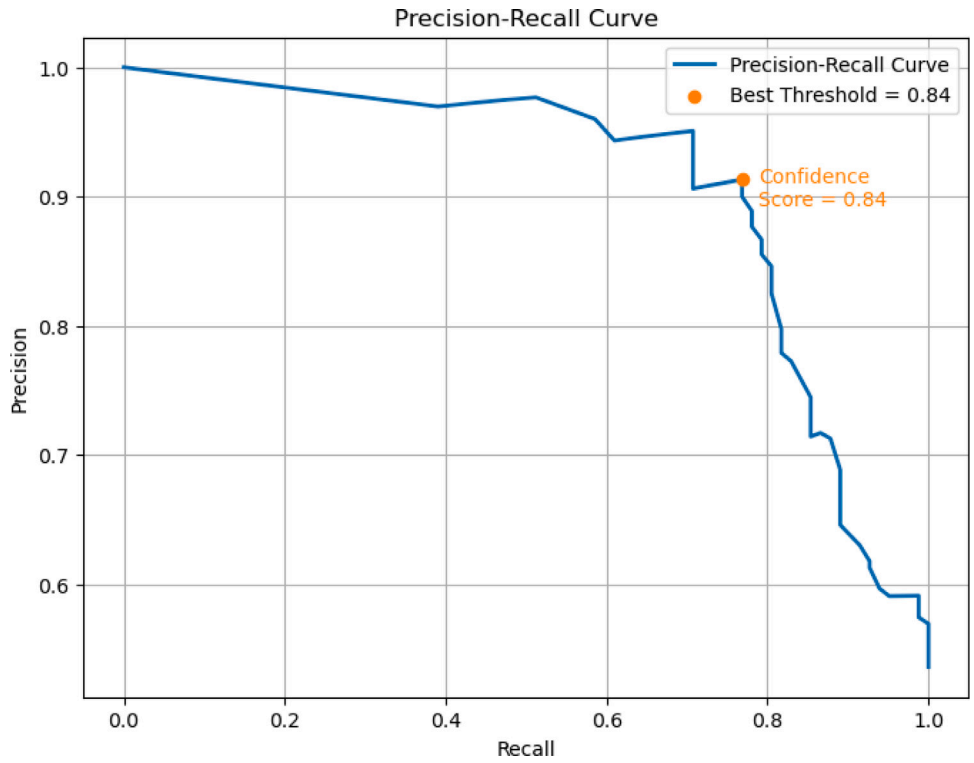


**Fig. 5.** PR curve for whistle detection on *Deployment 1*. The *x*-axis shows recall and the *y*-axis shows precision. The selected threshold at confidence score 0.84 offers a favorable trade-off between sensitivity and precision.

**Table 5**
Performance metrics across *Deployments 1*, *2* and *3* using three confidence score thresholds. Results are shown for both macro-averaged and weighted-averaged precision, recall, and F1-score. Macro averaging equally weights both classes, highlighting sensitivity to minority class (whistles), while weighted averages reflect performance based on class prevalence. Final F1-score values are also provided for each class (Background and Whistle). ✗ indicates no confidence score threshold was applied. Best performing confidence scores are highlighted in gray for each *Deployment*.

| Deployment | Threshold | Accuracy | Macro average | | | Weighted average | | | Recall | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Precision | Recall | F1 | Precision | Recall | F1 | Background | Whistle |
| *Deployment 1* | ✗ | 0.54 | 0.27 | 0.50 | 0.35 | 0.29 | 0.54 | 0.37 | 0.00 | 1.00 |
| | 0.79 | 0.82 | 0.82 | 0.83 | 0.82 | 0.83 | 0.82 | 0.82 | 0.86 | 0.79 |
| | 0.84 | 0.84 | 0.84 | 0.84 | 0.84 | 0.85 | 0.84 | 0.84 | 0.92 | 0.77 |
| *Deployment 2* | ✗ | 0.52 | 0.26 | 0.50 | 0.34 | 0.27 | 0.52 | 0.35 | 0.00 | 1.00 |
| | 0.79 | 0.76 | 0.76 | 0.76 | 0.76 | 0.76 | 0.76 | 0.76 | 0.73 | 0.79 |
| | 0.84 | 0.76 | 0.76 | 0.76 | 0.76 | 0.76 | 0.76 | 0.76 | 0.77 | 0.75 |
| *Deployment 3* | ✗ | 0.27 | 0.13 | 0.50 | 0.21 | 0.07 | 0.27 | 0.11 | 0.00 | 1.00 |
| | 0.79 | 0.81 | 0.76 | 0.76 | 0.76 | 0.81 | 0.81 | 0.81 | 0.87 | 0.65 |
| | 0.84 | 0.80 | 0.75 | 0.72 | 0.73 | 0.80 | 0.80 | 0.80 | 0.90 | 0.53 |

precision values were notably low, 0.27 in *Deployment 1*, 0.26 in *Deployment 2*, and 0.13 in *Deployment 3*, highlighting the low proportion of correct whistle predictions among all predicted whistles. Accuracy also dropped substantially, with values between 0.27 and 0.54 across deployments. The class-specific F1-scores reflected this imbalance: whistle F1-score remained high (1.00) due to perfect recall, while background F1-score dropped to 0.00, indicating a complete failure to identify background segments.

The intermediate threshold of 0.79 improved this imbalance by filtering out low-confidence predictions. Whistle recall values decreased slightly across deployments (0.79 in *Deployment 1*, 0.79 in *Deployment 2*, and 0.65 in *Deployment 3*), but background recall improved markedly, reaching 0.86, 0.73, and 0.87, respectively. This resulted in a more balanced distribution of predictions across the two classes. Accuracy improved to over 0.80 in *Deployments 1* and *3*, and reached 0.76 in *Deployment 2*. Macro F1-scores stabilized at 0.76–0.82, reflecting a fair compromise between precision and recall for both whistle and background events. Class-specific F1-scores showed that both classes were now reliably detected, and the reduction in FPs contributed to overall metric improvement.

The more conservative threshold of 0.84, identified via ROC and PR curve analysis, prioritized precision by further restricting detections to higher-confidence predictions. As a result, whistle recall dropped slightly (0.77 in *Deployment 1*, 0.75 in *Deployment 2*, and 0.53 in *Deployment 3*), but background recall remained high (0.92, 0.77, and 0.90, respectively). This led to improved classification of background events and a reduction in FPs. Overall accuracy remained stable or slightly increased compared to the 0.79 threshold, reaching 0.84 in *Deployment 1*, 0.76 in *Deployment 2*, and 0.80 in *Deployment 3*. Macro and weighted F1-scores remained in the 0.73–0.84 range, confirming that the 0.84 threshold reduces FPs while maintaining acceptable recall for whistle detection across varying acoustic conditions.

*Deployment 2*, recorded under similar seasonal and environmental conditions to *Deployment 1*, during late spring and summer, with high

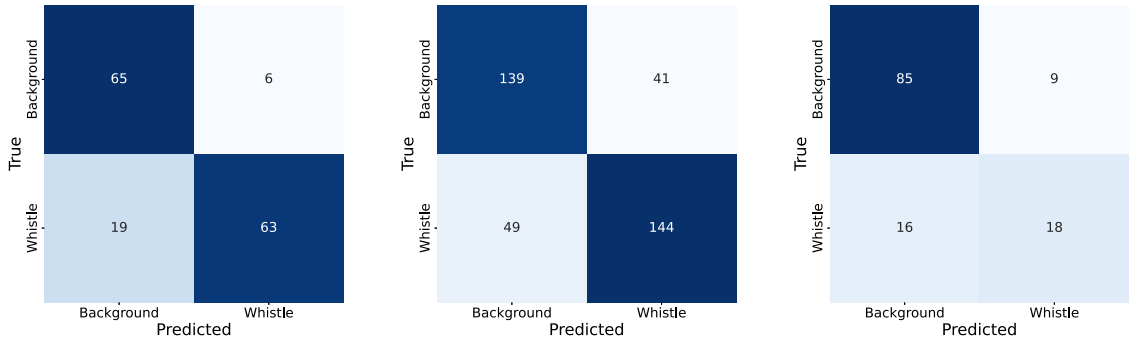


**Fig. 6.** Confusion matrices for *Deployments 1* (left), *2* (center), and *3* (right) at the operating confidence threshold of 0.79. Each matrix shows the number of TPs, TNs, FPs, and FNs, providing insight into detection quality across environments.

levels of anthropogenic presence, achieved similar performance, with whistle F1-scores ranging from 0.75 to 0.79 depending on the confidence threshold. This deployment had not been seen during model development, yet the model maintained stable performance. As in *Deployment 1*, the highest overall performance was observed at the 0.84 confidence threshold. In contrast, performance in *Deployment 3* was lower, with whistle F1-score dropping to 0.53 at the 0.84 threshold. This deployment was recorded approximately six months later, during winter, at a slightly different location.

Fig. 6 shows the confusion matrices for each deployment using the threshold of 0.84. These matrices confirm the trade-offs observed in the quantitative metrics, highlighting an improved balance between TPs and FPs across varying environmental conditions.

In *Deployment 1*, 63 TPs were retained with only 6 FPs and 19 FNs, this yielded a whistle recall of 77%, confirming strong whistle sensitivity in the deployment used for iterative annotation. In this case, FPs represented just 8.7% of all positive predictions, indicating high precision. This aligns with the high overall accuracy and whistle recall previously reported.

*Deployment 2* yielded 144 TPs and 41 FPs, with 49 missed whistles. Despite the increased number of FPs, the model maintained comparable whistle sensitivity, with a whistle recall of 75%. However, the FP rate also increased (22.2% of positive predictions). The confusion matrix reflects this trade-off, with a broader spread of predicted positives compared to *Deployment 1*.

In contrast, *Deployment 3* showed a notable drop in whistle detections, with only 18 TPs and 16 FNs. This corresponds to a whistle recall of 53%, significantly lower than in the other two deployments. However, the total number of whistle events in this deployment was lower, and when adjusted for total whistle presence, the TP-to-FN ratio still reflects some level of consistency. The background classification performance remained strong, with TNs far outnumbering FPs, indicating that the model continued to reject non-biophonic segments effectively. The FP rate was 25.0% of all positive predictions, slightly higher but within the same order of magnitude as in *Deployment 2*.

To further investigate the causes of recall degradation and threshold sensitivity, we analyzed the distribution of model confidence scores across all expert-validated detections from the three deployments, grouped by acoustic class Fig. 7. For detections validated as cetacean whistles (blue boxplot), confidence scores above a given threshold correspond to TPs, while scores below the threshold represent FNs. Conversely, for all non-cetacean classes, detections with confidence scores above the threshold constitute FPs, whereas those below the threshold are correctly rejected as true negatives TNs.

Under both thresholds, the majority of cetacean whistle detections fall above the decision boundary, indicating high separability from background sounds. The more conservative threshold (0.84) excludes a subset of lower-confidence whistle detections, increasing FN counts, while the more permissive threshold (0.79) retains a larger fraction of whistle detections at the cost of admitting additional FPs from background classes.

Background sound classes exhibit confidence distributions largely concentrated below both thresholds, although some classes show extended upper tails crossing the decision boundary. These instances correspond to high-confidence FP events and directly contribute to the threshold-dependent trade-off between recall and precision quantified in the previous sections.

## 4. Discussion

In this study, a multi-season underwater PAM campaign was conducted near Tarifa Island, in the Strait of Gibraltar, an ecologically significant marine area characterized by intense anthropogenic and environmental acoustic noise (André et al., 2011; Contreras Merida et al., 2024). To address the challenges of analyzing large, unlabeled, and acoustically complex datasets, Deep Learning-based PAM techniques were applied, particularly using BirdNET and Perch-based feature extraction with different classifiers (Kahl et al., 2021; Williams et al., 2024) and iterative, model-assisted annotation (Kath et al., 2024). While originally trained on avian sounds, the models were adapted to cetacean vocalizations through transfer learning, showing improved detection performance despite domain shifts (Padovese et al., 2023; Ghani et al., 2023; Williams et al., 2025). This approach offers a scalable and consistent alternative to traditional, expert-driven methods, although challenges related to generalization across species and conditions remain (Usman et al., 2020; Padovese et al., 2023; Nur Korkmaz et al., 2023).

**Model comparison.**

The pronounced contrast between the near-perfect performance achieved on the clean *Test dataset* (F1-score > 0.95 for all tested models) and the degradation observed in the *SEANIMALMOVE-WOPAM* dataset aligns with previous studies reporting similar domain-shift effects in bioacoustic applications (Stowell, 2022; Shiu et al., 2020; Ghani et al., 2023). Under controlled conditions, the models performed optimally because the benchmark dataset contained balanced, high-quality recordings representative of the training domain. However, real-world deployments introduced substantial acoustic variability, including changes in background composition, noise levels, and signal propagation characteristics. These discrepancies caused a marked decline in whistle F1-scores, often below 0.10 in most models, despite globally acceptable accuracy values. For example, the Perch feature extractor and Random Forest classifier model reached 97% accuracy on the *Test dataset* but only 43% F1-score on *SEANIMALMOVE-WOPAM*, with whistle-specific F1-score as low as 0.04. Such performance degradation reflects the models' bias toward the dominant background class and the limited capacity of some architectures to generalize to acoustically divergent environments (Shiu et al., 2020; Stowell, 2022).

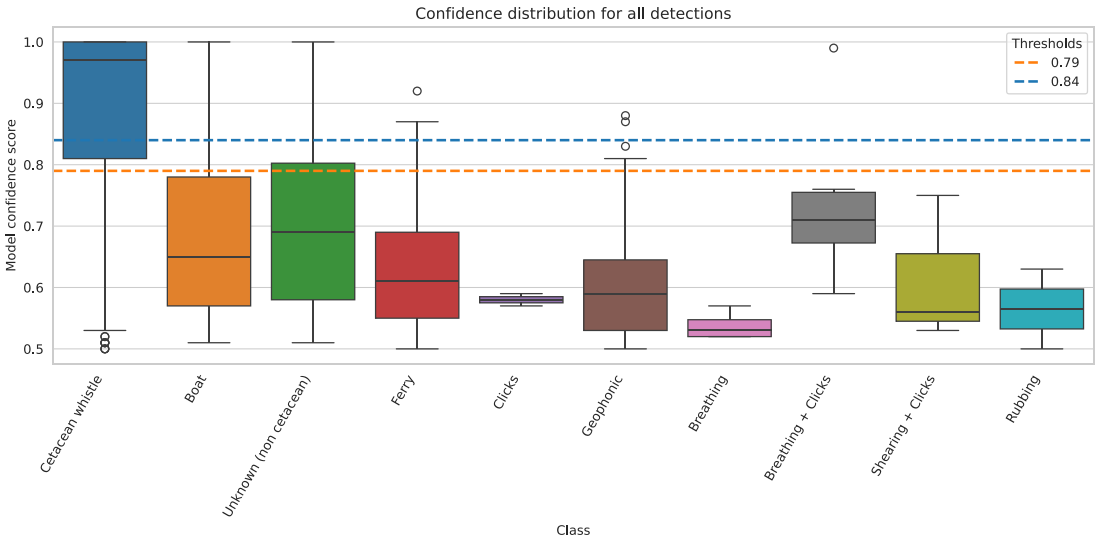


**Fig. 7.** Confidence score distribution across expert-validated acoustic classes for all deployments. Boxplots show the distribution of model confidence scores assigned by the BirdNET-custom detector to expert-validated detections pooled across Deployments 1–3. Classes correspond to cetacean whistles and major background sound categories identified during manual validation. The dashed horizontal lines indicate the two operating thresholds evaluated in this study (0.79 (orange) and 0.84 (blue)). True cetacean whistle detections (blue box-plot) concentrate at high confidence values well above both thresholds, whereas non-cetacean and acoustically ambiguous contexts (e.g., boat, unknown non cetacean) cluster closer to the decision boundary because of similarities to the target whistle cetacean class.

The BirdNET-custom neural network was the only configuration that maintained high detection performance across conditions, with an overall F1-score of 0.88 and a whistle-specific F1-score of 0.77. This robustness stems from its capacity to jointly adapt feature representations and decision boundaries, indicating strong generalization to real-world, noisy conditions (Padovese et al., 2023; Ghani et al., 2023; Licciardi and Carbone, 2024). The comparative evaluation of multiple architectures thus provides a broader view of how feature extractors and classifiers interact, and supports the use of fine-tuned neural heads when representative local data are available. Although simpler models such as Random Forests can yield high performance on controlled datasets, they tend to overfit dominant acoustic structures and fail under the complex noise conditions typical of coastal PAM deployments. In contrast, fine-tuned CNNs appear better suited to learn the subtle discriminative cues that separate cetacean whistles from overlapping anthropogenic and geophonic noise.

Importantly, the sharp degradation observed in baseline models under real-world conditions should not be interpreted as a failure of the proposed framework, but rather as an explicit demonstration of the domain-shift problem that motivates this work. All baseline models were trained using the same data and procedures, allowing for a controlled comparison across architectures. Within this setting, the BirdNET-custom neural network emerged as the only configuration capable of substantially mitigating the effects of noise and acoustic mismatch, achieving a whistle F1-score of 0.77 and an overall F1-score of 0.88 on the SEANIMALMOVE-WOPAM dataset. These results guided the subsequent focus on this model, which was then further analyzed across deployments and threshold configurations.

**Thresholding strategy.**

Confidence thresholding played a decisive role in defining model behavior during field deployment (Wenkel et al., 2021; Wood and Kahl, 2024; Tseng et al., 2025). The threshold derived from the ROC–AUC analysis (0.84) maximized discrimination between classes but did not necessarily correspond to the optimal operational balance for this study objective. In field scenarios, particularly where missed detections carry greater cost than FPs, maximizing recall becomes a priority. For this reason, a complementary threshold exploration was performed to assess the evolution of precision, recall, and F1-score across a range of confidence scores. The analysis showed that several thresholds produced comparable or higher F1-scores than the ROC-optimal one, with those between 0.79 and 0.83 offering a more favorable trade-off between precision and recall. Among them, 0.79 provided the highest recall with a modest increase in FPs. This finding supports the adoption of adaptive threshold calibration rather than rigid thresholding, as it enables the detector to be tuned according to the operational objectives of a specific deployment.

In automated monitoring systems where human verification is not possible, conservative thresholds such as 0.84 remain appropriate to minimize false alarms. Conversely, in expert-assisted workflows where detections can be manually validated, a more permissive threshold (e.g., 0.79) allows the recovery of additional true detections while maintaining manageable validation effort. This conclusion is consistent with previous studies emphasizing the need for context-aware threshold calibration in passive acoustic detection systems (Bergler et al., 2019; Shiu et al., 2020; Navine et al., 2024).

**Field deployments performance.**

Evaluation across independent field deployments confirmed that the BirdNET-custom model generalizes well when acoustic conditions remain comparable to those seen during fine-tuning (Kahl et al., 2021). In *Deployment 2*, recorded under similar seasonal and environmental conditions to the training data, the model achieved whistle F1-scores between 0.75 and 0.79, maintaining high recall and stable background discrimination. This indicates that the model can transfer effectively across deployments sharing comparable acoustic characteristics.

Rather than relying exclusively on clean or benchmark recordings, the training dataset incorporated data from *Deployment 1* that were iteratively validated and integrated through the semi-supervised annotation workflow. This process ensured that the model was fine-tuned on real-world acoustic variability, including representative anthropogenic noise, pronounced class imbalance, and environmental complexity typical of operational PAM conditions. As a result, the training data differ substantially from those used by many foundational and baseline models, which are often derived primarily from curated or controlled datasets (Tenan et al., 2020; Scuderi et al., 2024).

From a spatial and ecological perspective, all deployments analyzed in this study originate from the same geographic region, the Strait of Gibraltar, which also constitutes the source of the training and validation data. Consequently, the generalization assessed here reflects

the model's ability to transfer across seasons, locations, and highly variable acoustic conditions within this specific region, rather than across fundamentally different marine environments. Extending the framework to other marine regions would require additional region-specific data and ecological adaptation and therefore falls beyond the scope of the present work.

In contrast, performance in *Deployment 3* decreased substantially, with whistle F1-scores around 0.53 at the 0.84 threshold. This deployment differed from the others in both season (winter) and spatial location, producing distinct propagation conditions and noise regimes. While the model continued to classify background sounds accurately, the detection of whistles declined sharply, suggesting that unrepresented soundscapes and environmental shifts affected its generalization (Stowell, 2022; Kershenbaum et al., 2025).

Although the total number of predicted whistles scaled roughly with recording duration, the decline in recall and the increase in FNs during *Deployment 3* indicate that the model was less sensitive to whistle events under unfamiliar acoustic conditions. This does not necessarily imply that the reduced number of detections is solely attributable to model failure, as biological factors such as decreased cetacean presence during winter may also have contributed to lower whistle counts (Ross et al., 2023). Nevertheless, the observed reduction in recall suggests that both ecological variation and acoustic domain mismatch influenced detection performance. These findings are consistent with previous studies showing that fine-tuned models may overfit to site-specific acoustic conditions and lose sensitivity in unseen contexts (Ghani et al., 2023; Nur Korkmaz et al., 2023; Cominelli et al., 2024). Importantly, while the absolute number of detected whistles declined in *Deployment 3*, the ratio of true positives to total whistle events (recall) remained stable relative to each deployment's acoustic profile, indicating that the model's relative detection capacity was preserved despite changes in environmental conditions (Stowell, 2022).

These results highlights that the primary mitigation strategy for performance degradation in novel conditions is the progressive incorporation of multi-seasonal and multi-site data, from successive deployments, into the training dataset. In long-term PAM programs, such cumulative training allows the model to adapt incrementally to a broader range of acoustic variability, improving robustness in future deployments (Mcloughlin et al., 2019; Kershenbaum et al., 2025).

While the model retained consistent performance in background classification (background recall $\geq$ 0.90), the detection of whistle events was more affected. A lower threshold of 0.79 improved whistle recall and F1-score in *Deployment 3*, highlighting the benefit of threshold adaptation in unseen contexts (Navine et al., 2024; Kath et al., 2024). The variation in optimal threshold across deployments further supports this conclusion. While 0.84 was optimal in *Deployments 1* and *2*, *Deployment 3* benefited from a more permissive threshold (0.79), which improved recall with only a modest increase in FPs. In operational terms, this suggests that threshold tuning may be necessary when deploying the model in novel environments. When expert validation is possible, a lower threshold can be adopted to increase whistle recovery. In fully automated systems, a more conservative setting remains preferable to minimize false detections (Shiu et al., 2020; Kershenbaum et al., 2025).

**Confidence-based error structure across deployments.**

To further interpret the threshold-dependent performance observed across deployments, we analyzed the distribution of confidence scores assigned to all expert-validated detections, aggregated across the three deployments and grouped by acoustic class. This representation establishes a direct link between model confidence, decision thresholds, and resulting true and false detections, revealing systematic patterns underlying both FPs and FNs.

For cetacean whistles, confidence scores cluster predominantly above both operational thresholds, explaining the high recall observed in *Deployments 1 and 2*. The more conservative threshold (0.84) truncates the lower tail of this distribution, reducing FP incidence but increasing the number of FNs, whereas the 0.79 threshold retains nearly the entire central distribution while excluding only the lowest-confidence outliers. This distinction clarifies the functional implications of threshold selection in operational settings.

In contrast, most non-cetacean sound classes exhibit confidence distributions concentrated below both thresholds, resulting in stable background rejection across deployments. However, a subset of classes displays extended upper confidence tails that intersect the decision boundary, accounting for the majority of FPs under more permissive thresholding. These overlaps explain why threshold adaptation primarily affects recall while only moderately increasing FPR. High-confidence FPs arise predominantly from sound sources whose spectral or temporal structure partially overlaps with whistle characteristics, whereas other background sounds are consistently assigned low confidence scores. Similar confusion patterns have been documented in passive acoustic monitoring, particularly for tonal or quasi-harmonic anthropogenic signals that overlap odontocete whistles in frequency and modulation space (Zimmer, 2011; Mellinger et al., 2007; Shiu et al., 2020).

These results indicate that the recall degradation observed in acoustically distinct deployments reflects not only environmental variability but also shifts in confidence score distributions that move valid whistle detections closer to the decision boundary. This behavior is consistent with previous findings showing that fine-tuned deep learning models remain sensitive to unrepresented soundscapes and benefit from adaptive threshold calibration (Stowell et al., 2019; Ghani et al., 2023; Kershenbaum et al., 2025).

Importantly, this confidence-based analysis provides an interpretable, model-agnostic view of decision behavior without requiring architectural introspection. It enables practitioners to anticipate dominant error modes, select thresholds aligned with monitoring objectives, and assess deployment-specific trade-offs between recall and false alarm rates.

**Implications for model-assisted annotation and PAM workflows.**

The iterative model-assisted annotation approach implemented in this study proved highly effective for creating a representative and scalable training dataset. By combining model predictions with expert validation, the annotation process was accelerated while maintaining high labeling quality (Kath et al., 2024). This iterative strategy also facilitated fine-tuning, which significantly improved the model's capacity to generalize within the target region (Shiu et al., 2020; Ghani et al., 2023). Nevertheless, such workflows must be carefully managed to avoid potential biases, particularly the overrepresentation of high-confidence predictions and the under-sampling of rare or ambiguous whistle events (Mcloughlin et al., 2019; Kershenbaum et al., 2025). Periodic inclusion of randomly selected, low-confidence segments for manual inspection can mitigate these biases and provide a more balanced representation of the acoustic variability encountered in the field (McEwen et al., 2025).

Beyond the specific bioacoustic application addressed in this study, the proposed framework is inherently task-agnostic and may be transferable to other acoustic signal detection problems. Similar methodological challenges arise in engineering domains such as industrial condition monitoring and acoustic-based fault detection, where limited labeled data, changing operating conditions, and high background noise are common (Aslam et al., 2024). In this sense, the combination of transfer learning, iterative model-assisted annotation, and adaptive confidence-threshold calibration introduced here could be adapted to a range of industrial acoustic monitoring applications.

### 4.1. Limitations

While this study demonstrates promising results for automated cetacean whistle detection, certain limitations should be acknowledged to guide future improvements.

Although all detected whistle events were systematically validated by experts, the analysis did not include an exhaustive manual review of all segments classified as background. Performing a comprehensive validation of missed detections (FNs) across large-scale PAM datasets is often impractical due to the substantial annotation effort required (Mellinger et al., 2007; Christin et al., 2019; Stowell et al., 2019). Instead, the adopted semi-supervised workflow prioritized the validation of predicted whistle events across the full range of confidence scores, complemented by targeted inspection of selected low-confidence segments. This pragmatic trade-off reflects common operational constraints in real-world PAM applications. As a result, while recall estimates may be slightly conservative, the validation strategy provides a realistic assessment of detection performance under field conditions. Crucially, model predictions were never incorporated into the training set without expert verification, ensuring that labeling errors were not automatically propagated across iterations.

Another limitation of this study relates to the use of single-expert annotation during the iterative labeling process. While relying on a single expert ensures consistency in labeling criteria across iterations, it may introduce subjective bias, particularly for ambiguous or low signal-to-noise whistle events. Future work will aim to incorporate multiple annotators and inter-annotator agreement analyses to better quantify annotation uncertainty and improve the robustness of the training labels. Such multi-annotator validation would further strengthen the reliability and generalizability of the proposed framework.

Finally, temporal and spatial variability in the acoustic environment, particularly in *Deployment 3*, introduced differences in soundscape structure that were not fully represented in the training data. While the model performed well in familiar conditions, its ability to generalize to novel scenarios could be further enhanced by incorporating a broader range of acoustic contexts, seasons, and deployment locations during training.

### 4.2. Future directions

These findings open several promising directions for future research and development in automated cetacean whistle detection and marine soundscape analysis (Miksis-Olds and Nichols, 2016; Duarte et al., 2021). One clear priority is the continued enrichment and refinement of annotated datasets. The iterative model-assisted annotation process applied in this study has reduced dependency on fully manual labeling, which is often a bottleneck in acoustic ecology projects. And it has already produced a rapid and more comprehensive dataset, including not only cetacean whistles, but also human activity, vessel noise, echolocation clicks, and fish sounds. This enriched dataset offers future opportunities for multi-class classification, ecoacoustic research, and acoustic habitat assessment in the Strait of Gibraltar (Pijanowski et al., 2011; Aguzzi et al., 2019). In this context, future work will specifically aim to extend the current binary detection framework toward species-level and call-type classification, leveraging the progressively enriched annotations generated through the iterative workflow introduced in this study.

There is potential to further strengthen the model through architectural and training innovations. For example, transformer-based models (Vaswani et al., 2017) and self-supervised learning strategies have shown promise in audio analysis and could be explored to enhance transferability and reduce reliance on manual annotation (Mohamed et al., 2022; Gong et al., 2022). These approaches represent natural extensions to the current methodology and could help address any residual limitations in generalization and interpretability.

Model generalization across diverse acoustic environments also remains an important challenge. Incorporating training data from different seasons, background noise regimes, and deployment locations will help increase robustness to spatio-temporal variability. Also, benchmarking performance against established passive acoustic monitoring tools such as PAMGuard (Gillespie et al., 2008) and Chorus (Gavrilov and Parsons, 2014) would provide useful reference points and facilitate integration with current workflows. Expanding detection capabilities to include other types of cetacean vocalizations, particularly echolocation clicks could offer further insight into species presence and behavior. This will require specialized architectures and additional labeled data (Frasier, 2021), will enhance species coverage and behavioral analysis Complementary strategies incorporating advanced data augmentation and multi-modal representations, such as combining spectral and wavelet domain features, offers further potential to improve classification accuracy under challenging acoustic conditions (Mohamed et al., 2022; Gong et al., 2022).

Finally, transitioning from offline analysis to real-time implementation on autonomous recorders or hydrophone networks represents a promising direction. Real-time detection could enable timely monitoring of cetacean presence, supporting conservation efforts and management decisions (Hazen et al., 2019; Cartagena-Matos et al., 2021), especially in areas with high vessel traffic. More broadly, the extension of this framework to include a wider range of biophonic and geophonic sources would contribute to more comprehensive marine soundscape assessments and provide valuable tools for biodiversity monitoring and ocean habitat protection (Chen et al., 2021; Corrias et al., 2023).

## 5. Conclusion

This study presents "Cetacean Whistle Detector", a domain-adapted Deep Learning approach capable of detecting cetacean whistles in the complex and noisy marine environment of the Strait of Gibraltar. By combining pretrained bioacoustic models with iterative annotation and custom classification, the approach proved effective in generating biologically meaningful detections under real-world acoustic conditions, reaching F1-scores of up to 0.88 after fine-tuning with locally curated data. The model's robustness was validated across multiple deployments, demonstrating that carefully calibrated thresholds and locally curated training data are key to achieving reliable performance in the absence of extensive labeled datasets.

Beyond detection accuracy, the methodology introduced here emphasizes practical solutions to common challenges in PAM, including limited annotations, domain shift, and high noise variability. Baseline models that performed well on clean benchmark data showed a marked degradation under realistic noise conditions (whistle F1-scores $\leq 0.10$), highlighting the necessity of domain adaptation and iterative validation. By reformulating the classification problem, focusing on scalable binary detection, and introducing a validation-informed thresholding strategy, we offer a replicable framework for similar underwater acoustic monitoring scenarios where labeled data are scarce.

In addition to cetacean detection, the integration of Deep Learning into marine soundscape analysis holds great promise. As acoustic monitoring continues to gain prominence in conservation efforts, automated classification techniques will be instrumental in managing the growing volume of passive acoustic data. This study serves as a foundation for further advancements in Deep Learning-based bioacoustics, paving the way for more accurate, scalable, and real-time monitoring solutions that can contribute to the protection and understanding of marine ecosystems.

## CRediT authorship contribution statement

**Alba Márquez-Rodríguez:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Neus Pérez-Gimeno:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Daniel Benítez-Aragón:** Writing – review & editing, Data curation. **Gonzalo M. Arroyo:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Andrés De la Cruz:** Writing – review & editing, Visualization, Supervision, Resources, Project administration, Funding acquisition.

## Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to enhance the writing quality and facilitate language translation. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This study was primarily funded by the SEANIMALMOVE project (PCM_00062_SEANIMALMOVE 2023-071/PAI/PC-CIEN-MARINAS23/PR): Monitoring the movement and population dynamics of marine and coastal vertebrates in response to anthropogenic impacts in a global change scenario. This project is supported by the I+D+i Grants within the framework of the Complementary Plan for Marine Sciences and the Recovery, Transformation, and Resilience Plan (2023 Call). The research is co-funded by the Regional Ministry of University, Research, and Innovation of the Junta de Andalucía and the European Union through Next Generation EU funds under the Recovery, Transformation, and Resilience Plan.

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.engappai.2026.113756.

## Data availability

The data supporting the results presented in this work will be made available upon request. To obtain access, please contact Neus Pérez at neus.perez@uca.es or servicioacustica.inmar@uca.exs.

All code used for this manuscript is presented here and is freely accessible at the following address: https://github.com/SEANIMALMOVE/CetaceanWhistleDetection_StraitOfGibraltar.

## References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al., 2016. TensorFlow: a system for large-scale machine learning. In: 12th USENIX Symposium on Operating Systems Design and Implementation. OSDI 16, pp. 265–283.

Abayomi-Alli, O.O., Damaševičius, R., Qazi, A., Adedoyin-Olowe, M., Misra, S., 2022. Data augmentation and deep learning methods in sound classification: A systematic review. Electronics 11 (22), 3795.

Aguzzi, J., Chatzievangelou, D., Marini, S., Fanelli, E., Danovaro, R., Flogel, S., Lebris, N., Juanes, F., De Leo, F.C., Del Rio, J., et al., 2019. New high-tech flexible networks for the monitoring of deep-sea ecosystems. Environ. Sci. Technol. 53 (12), 6616–6631.

André, M., Van Der Schaar, M., Zaugg, S., Houégnigan, L., Sánchez, A., Castell, J., 2011. Listening to the deep: live monitoring of ocean noise and cetacean acoustic signals. Marine Poll. Bull. 63 (1–4), 18–26.

Aranda, G., Abascal, F.J., Varela, J.L., Medina, A., 2013. Spawning behaviour and post-spawning migration patterns of Atlantic bluefin tuna (Thunnus thynnus) ascertained from satellite archival tags. PLoS One 8 (10), e76445.

Arroyo, G.M., Mateos-Rodriguez, M., Munoz, A.R., De la Cruz, A., Cuenca, D., Onrubia, A., 2016. New population estimates of a critically endangered species, the Balearic Shearwater Puffinus mauretanicus, based on coastal migration counts. Bird Conserv. Int. 26 (1), 87–99.

Aslam, M.A., Zhang, L., Liu, X., Irfan, M., Xu, Y., Li, N., Zhang, P., Jiangbin, Z., Yaan, L., 2024. Underwater sound classification using learning based methods: A review. Expert Syst. Appl. 255, 124498.

Ayuntamiento de Tarifa, 2016. Plan estratégico de turismo de tarifa (2016–2020). Tarifa, España: Ayuntamiento de Tarifa & Diputación de Cádiz. https://www.aytotarifa.com/wp-content/uploads/2018/04/Plan-Estrategico-de-Turismo-de-Tarifa.pdf.

Bergler, C., Schröter, H., Cheng, R.X., Barth, V., Weber, M., Nöth, E., Hofer, H., Maier, A., 2019. ORCA-SPOT: An automatic killer whale sound detection toolkit using deep learning. Sci. Rep. 9 (1), 10997.

Bravo Sanchez, F.J., Hossain, M.R., English, N.B., Moore, S.T., 2021. Bioacoustic classification of avian calls from raw sound waveforms with an open-source deep learning architecture. Sci. Rep. 11 (1), 15733.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Bruno, M., Chioua, J., Romero, J., Vázquez, A., Macías, D., Dastis, C., Ramírez-Romero, E., Echevarria, F., Reyes, J., García, C.M., 2013. The importance of sub-mesoscale processes for the exchange of properties through the strait of Gibraltar. Prog. Oceanogr. 116, 66–79. http://dx.doi.org/10.1016/j.pocean.2013.06.006.

Buckingham, M.J., 2000. Wave propagation, stress relaxation, and grain-to-grain shearing in saturated, unconsolidated marine sediments. J. Acoust. Soc. Am. 108 (6), 2796–2815.

Cartagena-Matos, B., Lugué, K., Fonseca, P., Marques, T.A., Prieto, R., Alves, F., 2021. Trends in cetacean research in the eastern north Atlantic. Mammal Rev. 51 (3), 436–453.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. pp. 785–794.

Chen, C., Lin, T.-H., Watanabe, H.K., Akamatsu, T., Kawagucci, S., 2021. Baseline soundscapes of deep-sea habitats reveal heterogeneity among ecosystems and sensitivity to anthropogenic impacts. Limnol. Oceanogr. 66 (10), 3714–3727.

Chotiros, N.P., 1995. Biot model of sound propagation in water-saturated sand. J. Acoust. Soc. Am. 97 (1), 199–214.

Christin, S., Hervet, É., Lecomte, N., 2019. Applications for deep learning in ecology. Methods Ecol. Evol. 10 (10), 1632–1644.

Cominelli, S., Bellin, N., Brown, C.D., Rossi, V., Lawson, J., 2024. Acoustic features as a tool to visualize and explore marine soundscapes: Applications illustrated using marine mammal passive acoustic monitoring datasets. Ecol. Evol. 14 (2), e10951.

Contreras Merida, M.R., Merchant, N.D., Warr, S., Dissanayake, A., 2024. Underwater noise in the bay of gibraltar. Available At SSRN 4711787.

Corrias, V., de Lucía, G.A., Filiciotto, F., 2023. Marine soundscape and its temporal acoustic characterisation in the Gulf of Oristano, Sardinia (western Mediterranean sea). Mediterr. Mar. Sci. 24 (3), 526–538.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.

Criado-Aldeanueva, F., Soto-Navarro, F.J., García-Lafuente, J., et al., 2012. Seasonal and interannual variability of surface heat and freshwater fluxes in the Mediterranean sea: Budgets and exchange through the strait of Gibraltar.

de Soto, N.A., 2016. Peer-reviewed studies on the effects of anthropogenic noise on marine invertebrates: from scallop larvae to giant squid. Eff. Noise Aquat. Life II 17–26.

de Stephanis, R., Cornulier, T., Verborgh, P., Sierra, J.S., Gimeno, N.P., Guinet, C., 2008. Summer spatial distribution of cetaceans in the strait of Gibraltar in relation to the oceanographic context. Mar. Ecol. Prog. Ser. 353, 275–288.

Duarte, C.M., Chapuis, L., Collin, S.P., Costa, D.P., Devassy, R.P., Eguiluz, V.M., Erbe, C., Gordon, T.A., Halpern, B.S., Harding, H.R., et al., 2021. The soundscape of the Anthropocene ocean. Science 371 (6529), eaba4658.

Esteban, R., Verborgh, P., Gauffier, P., Alarcón, D., Salazar-Sierra, J., Giménez, J., Foote, A., De Stephanis, R., 2016. Conservation status of killer whales, Orcinus orca, in the strait of Gibraltar. Adv. Mar. Biol. 75, 141–172.

Frasier, K.E., 2021. A machine learning pipeline for classification of cetacean echolocation clicks in large underwater acoustic datasets. PLoS Comput. Biol. 17 (12), e1009613.

Frederick, C., Villar, S., Michalopoulou, Z.-H., 2020. Seabed classification using physics-based modeling and machine learning. J. Acoust. Soc. Am. 148 (2), 859–872.

Gauffier, P., Verborgh, P., Giménez, J., Esteban, R., Sierra, J.M.S., de Stephanis, R., 2018. Contemporary migration of fin whales through the strait of Gibraltar. Mar. Ecol. Prog. Ser. 588, 215–228.

Gavrilov, A.N., Parsons, M.J., 2014. A Matlab tool for the characterisation of recorded underwater sound (CHORUS). Acoust. Aust. 42 (3).

Ghanem, M., Ghaith, A.K., El-Hajj, V.G., Bhandarkar, A., De Giorgio, A., Elmi-Terander, A., Bydon, M., 2023. Limitations in evaluating machine learning models for imbalanced binary outcome classification in spine surgery: a systematic review. Brain Sci. 13 (12), 1723.

Ghani, B., Denton, T., Kahl, S., Klinck, H., 2023. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. Sci. Rep. 13 (1), 22876.

Gibb, R., Browning, E., Glover-Kapfer, P., Jones, K.E., 2019. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. Methods Ecol. Evol. 10 (2), 169–185.

Gibb, K.A., Eldridge, A., Sandom, C.J., Simpson, I.J., 2024. Towards interpretable learned representations for ecoacoustics using variational auto-encoding. Ecol. Inform. 80, 102449.

Gillespie, D., Mellinger, D., Gordon, J., Mclaren, D., Redmond, P., McHugh, R., Trinder, P., Deng, X., Thode, A., 2008. PAMGUARD: Semiautomated, open source software for real-time acoustic detection and localisation of cetaceans. J. Acoust. Soc. Am. 30 (5), 54–62.

Gong, Y., Lai, C.-I., Chung, Y.-A., Glass, J., 2022. Ssast: Self-supervised audio spectrogram transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36, pp. 10699–10709.

Guan, S., Brookens, T., 2023. An overview of research efforts to understand the effects of underwater sound on cetaceans. Water Biol. Secur. 2 (2), 100141. http://dx.doi.org/10.1016/j.watbs.2023.100141.

Gulli, A., Pal, S., 2017. Deep Learning with Keras. Packt Publishing Ltd.

Guo, P., Yang, H., Sano, A., 2023. Empirical study of mix-based data augmentation methods in physiological time series data. In: 2023 IEEE 11th International Conference on Healthcare Informatics. ICHI, IEEE, pp. 206–213.

Haider, U., Hanif, M., Kobayashi, H., Parajuli, L.K., Shimotoku, D., Rashid, A., Safeer, S., 2023. Bioacoustics signal classification using hybrid feature space with machine learning. In: 2023 15th International Conference on Computer and Automation Engineering. ICCAE, IEEE, pp. 376–380.

Han, S., Qubo, C., Meng, H., 2012. Parameter selection in SVM with RBF kernel function. In: World Automation Congress 2012. IEEE, pp. 1–4.

Havlik, M.-N., Predragovic, M., Duarte, C.M., 2022. State of play in marine soundscape assessments. Front. Mar. Sci. 9, 919418.

Hazen, E.L., Abrahms, B., Brodie, S., Carroll, G., Jacox, M.G., Savoca, M.S., Scales, K.L., Sydeman, W.J., Bograd, S.J., 2019. Marine top predators as climate and ecosystem sentinels. Front. Ecol. Environ. 17 (10), 565–574.

Henderson, E., Hildebrand, J., Smith, M., Falcone, E., 2012. The behavioral context of common dolphin (Delphinus sp.) vocalizations. Mar. Mam. Sci. 28 (3), 439–460.

Herr, H., Burkhardt-Holm, P., Heyer, K., Siebert, U., Selling, J., 2020. Injuries, malformations, and epidermal conditions in cetaceans of the strait of Gibraltar. Aquatic Mammals 46 (2).

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

Javier, R.F., Jaime, R., Pedro, P., Jesus, C., Enrique, S., 2023. Analysis of the underwater radiated noise generated by hull vibrations of the ships. Sensors 23 (2), 1035.

Juba, B., Le, H.S., 2019. Precision-recall versus accuracy and the role of large data sets. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33, pp. 4039–4048.

Kahl, S., Wood, C.M., Eibl, M., Klinck, H., 2021. BirdNET: A deep learning solution for avian diversity monitoring. Ecol. Inform. 61, 101236.

Kath, H., Serafini, P.P., Campos, I.B., Gouvêa, T.S., Sonntag, D., 2024. Leveraging transfer learning and active learning for data annotation in passive acoustic monitoring of wildlife. Ecol. Inform. 82, 102710.

Kershenbaum, A., Akçay, Ç., Babu-Saheer, L., Barnhill, A., Best, P., Cauzinille, J., Clink, D., Dassow, A., Dufourq, E., Growcott, J., et al., 2025. Automatic detection for bioacoustic research: a practical guide from and for biologists and computer scientists. Biol. Rev. 100 (2), 620–646.

Lapp, S., Rhinehart, T., Freeland-Haynes, L., Khilnani, J., Syunkova, A., Kitzes, J., 2023. OpenSoundscape: an open-source bioacoustics analysis package for Python. Methods Ecol. Evol. 14 (9), 2321–2328.

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

Liang, J., Nolasco, I., Ghani, B., Phan, H., Benetos, E., Stowell, D., 2024. Mind the domain gap: a systematic analysis on bioacoustic sound event detection. In: 2024 32nd European Signal Processing Conference. EUSIPCO, IEEE, pp. 1257–1261.

Licciardi, A., Carbone, D., 2024. WhaleNet: a novel deep learning architecture for marine mammals vocalizations on watkins marine mammal sound database. IEEE Access.

Márquez-Rodríguez, A., Mohedano-Munoz, M.Á., Marín-Jiménez, M.J., Santamaría-García, E., Bastianelli, G., Jordano, P., Mendoza, I., 2025. A bird song detector for improving bird identification through deep learning: a case study from donana. Ecol. Inform. 103254.

McEwen, B., Bernard, C., Stowell, D., 2025. Stratified active learning for spatiotemporal generalisation in large-scale bioacoustic monitoring. BioRxiv 2025-2009.

McKenna, M.F., Ross, D., Wiggins, S.M., Hildebrand, J.A., 2012. Underwater radiated noise from modern commercial ships. J. Acoust. Soc. Am. 131 (1), 92–103.

Mcloughlin, M.P., Stewart, R., McElligott, A.G., 2019. Automated bioacoustics: methods in ecology and conservation and their potential for animal welfare monitoring. J. R. Soc. Interface 16 (155), 20190225.

Mellinger, D.K., Stafford, K.M., Moore, S.E., Dziak, R.P., Matsumoto, H., 2007. An overview of fixed passive acoustic observation methods for cetaceans. Oceanography 20 (4), 36–45.

Miksis-Olds, J.L., Nichols, S.M., 2016. Is low frequency ocean sound increasing globally? J. Acoust. Soc. Am. 139 (1), 501–511.

Mohamed, A., Lee, H.-y., Borgholt, L., Havtorn, J.D., Edin, J., Igel, C., Kirchhoff, K., Li, S.-W., Livescu, K., Maaløe, L., et al., 2022. Self-supervised speech representation learning: A review. IEEE J. Sel. Top. Signal Process. 16 (6), 1179–1210.

Mohebbi-Kalkhoran, H., 2022. Machine Learning Approaches for Classification of Myriad Underwater Acoustic Events Over Continental-Shelf Scale Regions with Passive Ocean Acoustic Waveguide Remote Sensing (Ph.D. thesis). Northeastern University.

Nadir, M., Adnan, S.M., Khan, M.U., et al., 2020. Marine mammals classification using acoustic binary patterns. Arch. Acoust. 45 (4), 721–731.

Navine, A.K., Denton, T., Weldy, M.J., Hart, P.J., 2024. All thresholds barred: direct estimation of call density in bioacoustic data. Front. Bird Sci. 3, 1380636.

Nshimiyimana, A., 2024. Acoustic data augmentation for small passive acoustic monitoring datasets. Multimedia Tools Appl. 83 (23), 63397–63415.

Nur Korkmaz, B., Diamant, R., Danino, G., Testolin, A., 2023. Automated detection of dolphin whistles with convolutional networks and transfer learning. Front. Artif. Intell. 6, 1099022.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Peng, C., Zhao, X., Liu, G., 2015. Noise in the sea and its impacts on marine organisms. Int. J. Environ. Res. Public Health 12 (10), 12304–12323.

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

Pons, M., De Stephanis, R., Verborgh, P., Genovart, M., 2022. Sharp decreases in survival probabilities in the long-finned pilot whales in strait of Gibraltar. Mar. Biol. 169 (4), 44.

Popper, A.N., Hawkins, A.D., 2018. The importance of particle motion to fishes and invertebrates. J. Acoust. Soc. Am. 143 (1), 470–488.

Ramos, R., 2019. Crossing the pillars of hercules: Understanding transoceanic migrations of seabirds throughout their breeding range. Ecol. Evol. 9 (8), 4760–4771.

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

Rojo-Nieto, E., Álvarez-Díaz, P., Morote, E., Burgos-Martín, M., Montoto-Martínez, T., Sáez-Jiménez, J., Toledano, F., 2011. Strandings of cetaceans and sea turtles in the Alboran sea and strait of Gibraltar: a long–term glimpse at the north coast (Spain) and the south coast (Morocco). Anim. Biodivers. Conserv. 34 (1), 151–163.

Ross, S.R.-J., O'Connell, D.P., Deichmann, J.L., Desjonquères, C., Gasc, A., Phillips, J.N., Sethi, S.S., Wood, C.M., Burivalova, Z., 2023. Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions. Funct. Ecol. 37 (4), 959–975.

Sayigh, L., Daher, M.A., Allen, J., Gordon, H., Joyce, K., Stuhlmann, C., Tyack, P., 2016. The watkins marine mammal sound database: an online, freely accessible resource. In: Proceedings of Meetings on Acoustics. Vol. 27, AIP Publishing.

Scuderi, A., Campana, I., Gregorietti, M., Moreno, E.M., García Sanabria, J., Arcangeli, A., 2024. Tying up loose ends together: Cetaceans, maritime traffic and spatial management tools in the strait of Gibraltar. Aquat. Conserv.: Mar. Freshw. Ecosyst. 34 (1), e4066.

Serra, O.M., Martins, F., Padovese, L.R., 2019. Automatic detection of estuarine dolphin whistles in spectrogram images. arXiv preprint arXiv:1909.04425.

Shiu, Y., Palmer, K., Roch, M.A., Fleishman, E., Liu, X., Nosal, E.-M., Helble, T., Cholewiak, D., Gillespie, D., Klinck, H., 2020. Deep neural networks for automated detection of marine mammal species. Sci. Rep. 10 (1), 607.

Stanley, J.A., Van Parijs, S.M., Hatch, L.T., 2017. Underwater sound from vessel traffic reduces the effective communication range in Atlantic cod and haddock. Sci. Rep. 7 (1), 14633.

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 Ecol. Evol. 10 (3), 368–380.

Tenan, S., Hernández, N., Fearnbach, H., de Stephanis, R., Verborgh, P., Oro, D., 2020. Impact of maritime traffic and whale-watching on apparent survival of bottlenose dolphins in the strait of Gibraltar. Aquat. Conserv.: Mar. Freshw. Ecosyst. 30 (5), 949–958.

Tokozume, Y., Ushiku, Y., Harada, T., 2017. Learning from between-class examples for deep sound recognition. arXiv 2017. arXiv preprint arXiv:1711.10282.

Tseng, S., Hodder, D.P., Otter, K.A., 2025. Setting BirdNET confidence thresholds: species-specific vs. universal approaches. J. Ornithol. 1–13.

Tuia, D., Kellenberger, B., Beery, S., Costelloe, B.R., Zuffi, S., Risse, B., Mathis, A., Mathis, M.W., Van Langevelde, F., Burghardt, T., et al., 2022. Perspectives in machine learning for wildlife conservation. Nat. Commun. 13 (1), 792.

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.

Van Merriënboer, B., Hamer, J., Dumoulin, V., Triantafillou, E., Denton, T., 2024. Birds, bats and beyond: Evaluating generalization in bioacoustics models. Front. Bird Sci. 3, 1369756.

Van Parijs, S.M., Clark, C.W., Sousa-Lima, R.S., Parks, S.E., Rankin, S., Risch, D., Van Opzeeland, I.C., 2009. Management and research applications of real-time and archival passive acoustic sensors over varying temporal and spatial scales. Mar. Ecol. Prog. Ser. 395, 21–36.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wang, Z.-T., Houser, D.S., 2023. Hearing research in cetaceans. http://dx.doi.org/10.3389/fmars.2023.1307317.

Wenkel, S., Alhazmi, K., Liiv, T., Alrshoud, S., Simon, M., 2021. Confidence score: The forgotten dimension of object detection performance evaluation. Sensors 21 (13), 4350.

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. Front. Mar. Sci. 9, 879145.

Williams, B., van Merriënboer, B., Dumoulin, V., Hamer, J., Fleishman, A.B., McKown, M., Munger, J., Rice, A.N., Lillis, A., White, C., et al., 2025. Using tropical reef, bird and unrelated sounds for superior transfer learning in marine bioacoustics. Philos. Trans. B 380 (1928), 20240280.

Williams, B., van Merriënboer, B., Dumoulin, V., Hamer, J., Triantafillou, E., Fleishman, A.B., McKown, M., Munger, J.E., Rice, A.N., Lillis, A., et al., 2024. Leveraging tropical reef, bird and unrelated sounds for superior transfer learning in marine bioacoustics. arXiv preprint arXiv:2404.16436.

Wisniewska, D.M., Johnson, M., Teilmann, J., Siebert, U., Galatius, A., Dietz, R., Madsen, P.T., 2018. High rates of vessel noise disrupt foraging in wild harbour porpoises (Phocoena phocoena). Proc. R. Soc. B: Biol. Sci. 285 (1872), 20172314.

Wood, C.M., Kahl, S., 2024. Guidelines for appropriate use of BirdNET scores and other detector outputs. J. Ornithol. 165 (3), 777–782.

WOPAM Project, 2025. World oceans passive acoustic monitoring day. URL: https://www.wo-pam.com/. (Accessed 25 April 2025).

Zhou, Z., Qu, Y., Zhu, B., Zhang, B., 2024. Detection of typical transient signals in water by XGBoost classifier based on shape statistical features: Application to the call of southern Right Whale. J. Mar. Sci. Eng. 12 (9), 1596.

Zimmer, W.M., 2011. Passive Acoustic Monitoring of Cetaceans. Cambridge University Press.