Transfer Learning for Distance Classification of Marine Vessels Using Underwater Sound

Decrop Wout, Deneudt Klaas, Parcerisas Clea, Schall Elena, and Debusschere Elisabeth

***Abstract*—Marine environments are increasingly affected by human activities, which generate underwater noise as a by-product. Acoustic data from these environments can offer valuable insights for tracking human activity and improving the monitoring of sensitive areas, such as marine protected areas (MPAs) and offshore wind farms. This study presents a convolutional neural network (CNN) trained to classify vessel distances from passive acoustic recordings. We constructed an open-source, diverse dataset by integrating 116 days of acoustic data from two stations in the Belgian part of the North Sea with automatic identification system data. The CNN was trained to classify acoustic clips into discrete distance bins, representing the proximity of the nearest vessel. Our results demonstrate that the model can effectively distinguish between distance categories using underwater sound alone, confirming the feasibility of passive acoustic monitoring for vessel activity. This technology provides an innovative approach to enhance MPA oversight and represents a first step in a promising pathway for conservation efforts.**

***Index Terms*—Automatic identification system (AIS), convolutional neural network (CNN), passive acoustics, underwater acoustics, vessel detection.**

I. Introduction

MARINE protected areas (MPAs) are distinct zones established for the protection of marine resources and ecosystem services. These areas play a crucial role in conserving biodiversity, safeguarding habitats, and maintaining ecosystem integrity, all while permitting sustainable resource use [1], [2]. In addition, MPAs provide vital buffers against the impacts of climate change [3]. The European Union has set an ambitious goal to designate 30% of its waters as protected areas by 2030, with 10% subject to stricter protection measures [4]. However, despite these objectives, MPAs often face challenges related to inadequate enforcement and insufficient national regulations, which can allow human activities to persist within their boundaries [5]. To ensure their effectiveness, robust monitoring systems are necessary to track human activities and enforce compliance with maritime regulations [6], [7], [8], [9], [10].

Anthropogenic activities, such as commercial shipping and fishing, pose significant threats to MPAs, both through physical disturbances and pollution. These activities contribute to underwater noise pollution, which disrupts marine ecosystems by affecting the behavior, communication, and stress levels of marine life [6], [7], [8], [11], [12]. Since 2008, underwater noise has been formally recognized as a pollutant by the Marine Strategy Framework Directive (MSFD) under Descriptor 11, emphasizing its global impact. This noise is a by-product of various human industrial activities, including the construction and operation of offshore wind farms (OWFs). With offshore wind capacity set to reach 60 GW by 2030 [13], noise from these projects is expected to contribute significantly to the overall underwater noise pollution.

Although noise is a critical concern, the presence and activity of ships within MPAs—including fishing activities—represent broader challenges that can undermine conservation efforts. Monitoring and managing these activities is essential for protecting MPAs effectively. Moreover, noise pollution is not confined to the MPA boundaries. Williams et al. [14] highlighted that noise-generating activities outside protected areas can still substantially raise noise levels within MPAs, affecting species that rely on acoustic cues. However, reducing anthropogenic noise from sources outside MPAs remains a significant challenge [15]. Given these threats, effective monitoring and management of vessel activity—both inside and near MPAs—are crucial for ensuring their ecological integrity.

To achieve this, various tools have been developed to monitor maritime traffic, such as the automatic identification system (AIS), the vessel monitoring system (VMS), and long-range identification and tracking. These systems, however, have technical and legal limitations [2]. Underwater sound, as an omnipresent and measurable phenomenon, has emerged as a complementary method for monitoring vessel activity and noise levels within MPAs. By utilizing acoustic data, it becomes possible to track ship movements and their associated noise footprints, offering a more holistic approach to understanding human activity in these critical zones. This study focuses on leveraging underwater sound as a tool to track vessel activity within MPAs, thereby supplementing existing monitoring systems and addressing gaps in current methodologies [16], [17], [18], [19].

Received 9 March 2025; revised 26 June 2025; accepted 26 July 2025. Date of publication 31 July 2025; date of current version 15 August 2025. This work was supported in part by the European Union Horizon Europe Programme through iMagine project under Grant 101058625, in part by the LifeWatch under Grant I001225N, in part by the European Union's Horizon Europe Research and Innovation Programme through iMagine project under Grant HORIZON-INFRA-2021-SERV-01, and in part by the Research Foundation - Flanders (FWO) under the framework of the Flemish contribution to LifeWatch, a landmark European Research Infrastructure on the European Strategy Forum on Research (ESFRI) roadmap. *(Corresponding author: Decrop Wout.)*

Decrop Wout, Deneudt Klaas, Parcerisas Clea, and Debusschere Elisabeth are with the VLIZ. Flanders Marine Institute (VLIZ), InnovOcean site, 8400 Oostende, Belgium (e-mail: wout.decrop@vliz.be; klaas.deneudt@vliz.be; clea.parcerisas@vliz.be; Elisabeth.Debusschere@vliz.be).

Schall Elena is with Alfred Wegener Institute (AWI), 27570 Bremerhaven, Germany (e-mail: elena.schall@awi.de).

Digital Object Identifier 10.1109/JSTARS.2025.3593779

VMS is regulated by the European Union (EC Council Regulation No. 1224/2009), whereas AIS follows international standards. The European satellite-based VMS aims to detect and identify fishing vessels through remote systems, whereas AIS was originally designed to prevent collisions at sea [20]. AIS operates as an automated tracking system used by vessel traffic services to identify maritime vessels by transmitting their unique identifiers, locations, and other relevant data. However, coverage limitations persist since AIS transmitters are mandatory only for specific classes of vessels [21]. For example, fishing boats smaller than 15 m and those on short voyages are exempt [2]. This leads to having over 30 million recreational vessels globally without such transmitters. For example, in shallow coastal and inner Danish waters, a significant proportion of vessels (83%) lack AIS capabilities [22]. Vessels without or with malfunctioning transmitters are known as *dark* vessels.

Besides vessel transmission techniques, optical remote sensing satellites can detect *dark* vessels (i.e., vessels operating without broadcasting AIS signals) by analyzing visual imagery [23], [24]. However, such methods face limitations, including dependence on sunlight, cloud cover, and resolution constraints. In addition, these methods require accurate calibration to ensure data precision [25]. Systems, such as SafeSeaNet, developed by the European Maritime Safety Agency, integrate satellite imagery with AIS data but offer a resolution of 500 m—insufficient for detailed vessel identification. Whereas complementary methods, such as drones or airborne vehicles, can monitor vessel activities, such as anchoring or fishing, they still pose challenges regarding safety, stability, and reliability [2], [26].

Recent advances leverage synthetic aperture radar (SAR) for vessel detection, particularly *dark* vessels [27], [28], [29], [30], [31], [32], [33]. Unlike optical satellites, SAR can detect vessels under all weather conditions and is unaffected by light or shadows, providing consistent detection capability [33]. Despite its advantages, SAR remains relatively expensive to implement on a large scale [34].

This research focuses on developing a novel approach for classifying vessel distances categories using underwater passive acoustics, without the need for costly satellite data or active transmitters, or as a complementary technique to existing methods. Although passive acoustic monitoring has proven to be a reliable tool for long-term vessel detection and surveillance [19], [35], [36], [37], thanks to the fact that all vessels produce sound and sound travels efficiently underwater [35], [38], our contribution lies in advancing distance classification specifically.

Traditional range estimation in acoustics often relies on matched-field processing (MFP) [39], [40], [41], [42], which has notable limitations, particularly in complex environments, such as shallow waters [43]. MFP requires detailed environmental data, and it struggles in noisy or dynamic environments, making it less reliable for accurate distance measurements [44]. Moreover, MFP techniques are computationally expensive, requiring extensive modeling and parameter tuning [45].

In contrast, machine learning techniques, such as convolutional neural networks (CNNs), offer a promising alternative by providing faster, more flexible, and more efficient solutions [19], [35], [36], [37], [46]. CNNs can process acoustic data directly from spectrograms, handling noisy and dynamic environments, such as shallow coastal waters, where traditional methods falter [47], [48], [49], [50], [51], [52]. Our approach leverages a CNN-based architecture that is not only capable of robust distance classification but also avoids the environmental constraints and high computational cost associated with MFP [44], [45].

Widely used acoustic classification models, such as VGGish, VGGNet, and ResNet, were originally developed for human-centric datasets (e.g., speech, music, and urban noise) and have been adapted with some success for vessel *type* recognition in underwater contexts [36], [53], [54], [55]. As noted by Ghani et al. [56], bioacoustic transfer learning still heavily relies on VGGish, despite it being outdated and surpassed on modern benchmarks. Its continued use stems more from inertia than suitability: VGGish requires substantial domain adaptation and struggles with overlapping biological, environmental, and anthropogenic signals.

The field is now shifting toward domain-specific solutions. For example, Li et al. [57] used contrastive language–audio pretraining (CLAP)-based embeddings to improve performance on large-scale underwater datasets, such as Oceanship. CLAP is a general-purpose model that learns shared representations of audio and text using large-scale contrastive learning [58], [59]. Several variants exist, including zero-shot and self-supervised versions, which allow flexible use across tasks without task-specific fine-tuning. BioLingual CLAP advances the CLAP framework by being trained specifically on bioacoustic data—including underwater recordings and full ecosystem soundscapes—producing embeddings that better capture the subtle, overlapping features found in marine environments. This makes it particularly effective for vessel detection in MPAs, where robust performance in noisy, natural settings is critical.

Building on this foundation, our innovation lies in using spectrograms not just for vessel classification, but for classifying vessel *distances* using a high-performance, fast, and scalable CLAP model. Unlike traditional methods that require environmental context, such as bathymetry, our approach operates solely on acoustic data and significantly accelerates the distance classification process, making it suitable for real-time applications. While we use the zero-shot version of BioLingual, our primary model is the *supervised* version of BioLingual CLAP. This choice is supported by findings from Kather et al. [60], who showed that even bioacoustic-specific models struggle with polyphonic PAM data unless supervised learning is used. Their results confirm that supervised BioLingual CLAP outperforms self-supervised and zero-shot variants when applied to complex, noisy marine recordings.

The dataset, collected over 116 days across all four seasons in the Belgian Part of the North Sea (BPNS), a highly dynamic, noisy, and shallow coastal region, ensures the model's robustness and generalizability to similar coastal environments throughout the year.

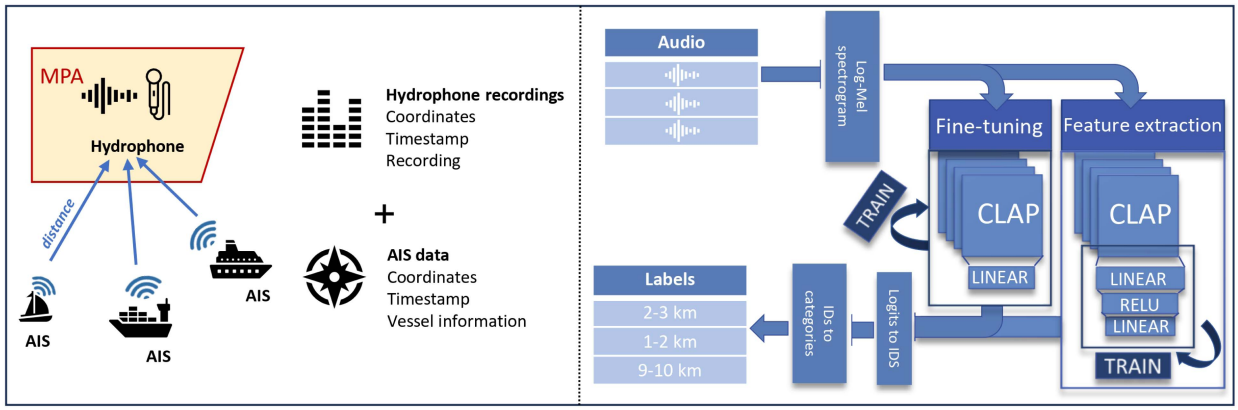


Fig. 1. Overview of the data integration and processing pipeline. On the left, the AIS coordinates are matched with hydrophone recordings to compute vessel proximity. On the right, the corresponding audio segments are converted into Log-mel spectrograms and processed via two approaches. 1) FE: The spectrograms are input into a pretrained CLAP model to extract high-level features, which are then used to train a three-layer neural network. 2) Fine-tuning: The spectrograms are used to train the pretrained CLAP model along with a linear classification layer.

II. Methodology

A. Conceptual Overview

This study integrates AIS data with acoustic recordings to create a strong-labeled dataset for deep learning classification, where each label corresponds to the distance between the nearest vessel and the hydrophone. By aligning AIS coordinates with time-stamped hydrophone data from the North Sea, we calculate these distances precisely. As illustrated in Fig. 1, the data integration process assigns each audio clip a distance-based label.

To prepare the data for distance classification, we segmented the audio into 10-s, nonoverlapping windows, categorizing each segment by its proximity to the nearest vessel. Categories were divided into 1-km bins: 0–1 km, 1–2 km, 2–3 km, 3–4 km, 4–5 km, 5–6 km, 6–7 km, 7–8 km, 8–9 km, 9–10 km, and 10+ km.

The model used was the contrastive language–audio model (CLAP-LAION) [61], which is built upon the CLAP model architecture [58], [59]. The pretrained CLAP-LAION model named BioLingual [62] was considered for transfer learning, as it was partly trained on underwater bioacoustics data. The BioLingual model was used following two different approaches.

1) *Feature extraction (FE):* High-level features were extracted from log-mel spectrograms of the audio data using the pretrained layers.
2) *Fine-tuning:* The pretrained weights were used for initialization, but the entire model was retrained.

In the FE approach, the extracted features were passed through three custom-made layers, whereas in the fine-tuning approach, the entire model was retrained and followed by just one linear layer, as shown in Fig. 1. Although fine-tuning is expected to yield slightly better performance due to the retraining of all layers, it is also significantly more computationally demanding. This tradeoff between performance and computational efficiency was anticipated, as fine-tuning requires full retraining of the model, whereas FE only requires adjustments to the final layers with pretrained weights. This is an important consideration when evaluating the two approaches.

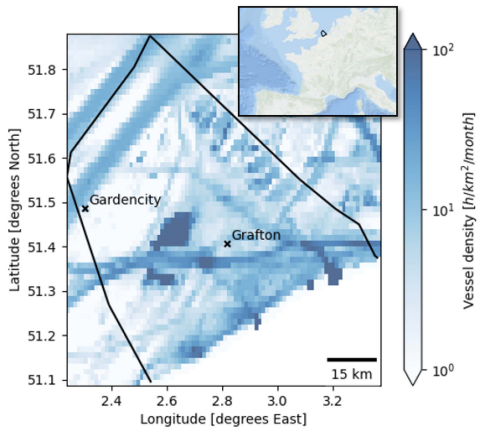


Fig. 2. Vessel density over in the Belgian part of the North Sea (BPNS) showing the locations of the hydrophone stations, GardenCity and Grafton. The figure highlights the proximity of these stations to major shipping lanes, which is crucial for accurately classifying vessel distances in our study.

B. Data Collection

The dataset comprises acoustic data from the Belgian part of the North Sea (BPNS) and AIS data. Acoustic data were collected from the LifeWatch Broadband Acoustic Network [63] using RESEA 320 RTSys recorders and Colmar GP1190M-LP hydrophones. These hydrophones, with a sensitivity of -180 dB/V re 1 μPa and a frequency range of -3 dB from 10 Hz to 170 kHz, were mounted 1 m above the seabed on steel mooring frames [64]. The hydrophone stations used for this study, GardenCity and Grafton, are positioned strategically near major shipping lanes, as depicted in Fig. 2. This strategic placement is ideal for the study's goal of classifying vessel distances, as it ensures high vessel activity, which improves the accuracy of the CNNs trained on the data. In addition, to ensure that the data are representative and usable throughout the entire year, recordings were collected from five different deployments spanning all seasons of 2022. The distribution of these recordings over time is shown in Fig. 5.

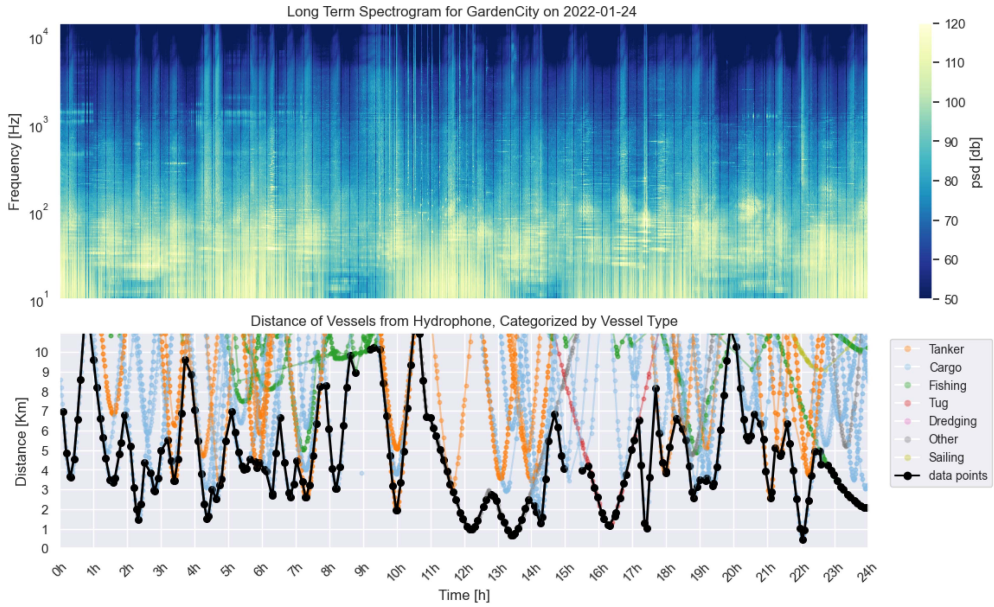


Fig. 3. Daily power spectrum and vessel proximity: The top section shows the power spectrum of an entire day's recording, whereas the bottom section illustrates the distance between nearby vessels and the hydrophone. High power spectrum values in higher frequencies appear to overlap with peaks in vessel proximity. Vessels are color-coded based on their type.

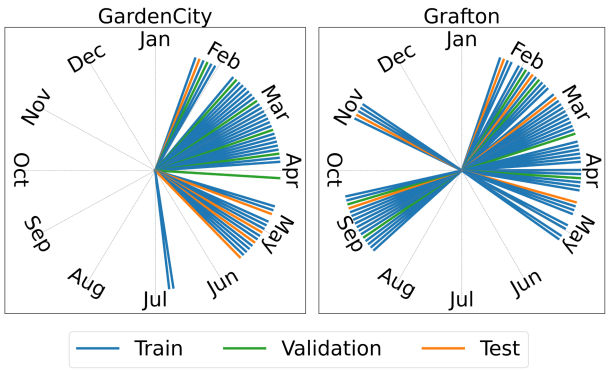


Fig. 4. Time distribution of acoustic data throughout 2022, color-coded by split type (training, validation, and testing). Data originate from two deployments in GardenCity and three in Grafton spread throughout the year.

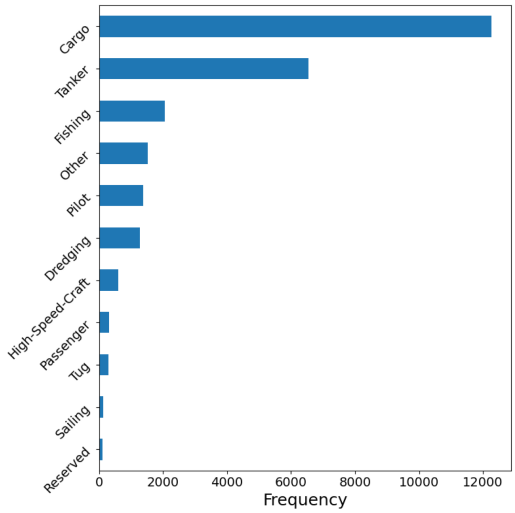


Fig. 5. Histogram showing the distribution of ship types in the dataset, where ship types with frequencies below 100 have been merged into the "other" category for better visualization. The smaller ship types, including "pleasure-craft" (75), "towing" (28), "law-enforcement" (25), "military" (25), "large-towing" (13), "S&R" (7), "antipollution-equipment" (3), and "diving" (1), have been grouped into "other." This grouping highlights the more frequent ship types, such as cargo, tanker, and fishing.

The BPNS is a very shallow area (maximum depth of 45 m [65]), which is a critical factor in determining the distance between a vessel and a hydrophone. The hydrophone stations, GardenCity and Grafton, are located at depths of 35 and 23 m, respectively. In such environments, sound propagation is influenced by several factors, including water depth, seabed composition, temperature gradients, and salinity, all of which can cause reflection, refraction, and scattering of sound waves [66], [67], [68]. These variables can lead to fluctuations in acoustic signal strength and travel time, making it more challenging to accurately determine distances.

The second part of the dataset comprises vessel coordinates data sourced from the AIS-Hub data network [69]. It is important to note that this hub also contains data about vessel speed, vessel category, and its current activity. The distance from the hydrophone to the vessel can be calculated at every timestamp based on the coordinates, resulting in semicontinuous annotated data.

C. Dataset Creation

The model's objective is to classify the distance category to the closest vessel. However, challenges arise due to vessels that are either *dark* or exhibit irregularities in their AIS transmissions.

To address these challenges, we developed a method to continuously annotate the recordings by determining the distance

to the nearest vessel using AIS data. Specifically, within each predefined time window—referred to as the *window frame*—we identified the vessel with the smallest recorded distance to the hydrophone. These minimum distances (*local minima*) represent the closest vessels within each time segment of the dataset.

The choice of window frame size is a critical parameter in this approach, as it defines the duration over which we search for the nearest vessel and assign its distance to the corresponding acoustic recording. A smaller window frame increases the temporal resolution of the dataset by providing more frequent distance annotations. However, this also raises the risk of missing the actual closest vessel if there are gaps in AIS data transmission within that short time period. Conversely, using a larger window frame reduces the likelihood of missing the closest vessel due to intermittent AIS updates but comes at the cost of decreased temporal resolution, as each distance annotation then represents a longer time span.

Window frame size is crucial to balance the tradeoff between data resolution and accuracy in vessel identification. A smaller window frame increases the dataset's temporal resolution by providing more data points but also raises the risk of missing the closest vessel if there are gaps in AIS data transmission. Conversely, a larger window frame reduces the risk of missing the closest vessel due to intermittent data but at the cost of decreased temporal resolution.

Through testing and filtering, a 6-min window frame was predominantly used across the dataset. However, only for Grafton data, a transition to a 5-min window frame was implemented during the latter half of 2022. As a result, within a 24-h period, the 6-min window produced 240 data points, whereas the 5-min window produced 288 data points. Despite these efforts, approximately 8% of the data had to be excluded due to irregularities in AIS data. An example of these inconsistencies is illustrated in Fig. 3, with additional examples available in the Appendix of the Supplementary Material. In this figure, certain segments of the data are excluded where the AIS data do not align with the power spectrum, indicating inconsistencies. For example, at 9 h, a new vessel briefly appears and then vanishes, but its trajectory is missing in the power spectrum, leading to its exclusion. Similarly, between 15 h and 15 h 30 min, a poor AIS sampling rate causes fluctuations in the vessel distance data, resulting in the omission of the time frames around these times (i.e., 8 h 50 min–9 h 10 min and 15 h–15 h 30 min).

Ultimately, 26 465 10-s audio segments were generated over 116 days, including 40 days with overlapping stations and 76 unique days. All acoustic recordings were converted to single channel, with a sampling rate of 48 kHz, and segmented into 10-s nonoverlapping windows.

The dataset was initially divided into training, validation, and testing sets, as shown in Fig. 5. To ensure data independence between the three sets, full days of one location were only included in one of the three sets. This means that for any given day and location, all the data from that day are entirely allocated to either the training, validation, or testing set, with no overlap across these sets. The distribution split is 79.4%, 10.6%, and 9.9% due to the uneven availability of data points across different days.

As shown in Fig. 5, the distribution of ship types indicates a highly imbalanced dataset, with the most frequent classes being "Cargo" and "Tanker," which together contribute the largest share of the data. These two ship types are structurally similar and likely produce similar acoustic signatures, which may explain their dominance in the dataset. The remaining categories, such as "Pleasure-Craft," "Towing," "Military," and others, account for a much smaller portion of the total. This imbalance could potentially skew model training if the smaller classes were not grouped together into the "Other" category.

In our analysis, we chose to focus on the distance classification to the nearest vessel as the primary variable for model classification, rather than the type of ship itself. This decision stems from the fact that distance is a more consistent and measurable feature across all ship types, regardless of class. Whereas ship type may be relevant in certain contexts, focusing on distance enables a more generalized approach to vessel detection and monitoring, reducing the model's susceptibility to class imbalances.

D. Architecture

The choice of classification over regression was based on both technical and ethical considerations. Technically, the pretrained BioLingual network [62] was specifically developed for classification tasks, with layers optimized to categorize data into discrete classes. This model is based on the CLAP-LAION architecture [61], implemented without feature fusion. As proposed by Robinson et al. [62], the architecture includes an audio encoder and projection layers, culminating in a final linear layer designed for general-purpose classification. It has also recently been used for ship type detection in large-scale underwater datasets [57].

Choosing classification rather than regression aligns with ethical and security considerations. Regression-based distance estimation could yield highly precise vessel localization, which may pose risks if misused for surveillance or unauthorized tracking. In contrast, using classification to estimate vessel *distance ranges* provides sufficient resolution for environmental monitoring and enforcement—such as in MPAs—while reducing the potential for misuse. This approach supports safe, open publication and broader accessibility in conservation-focused research.

For the fine-tuning approach, all layers of the pretrained model were retained and retrained on the downstream task. For the FE approach, all layers except the final classification layer were frozen, and a new model with two linear layers and one ReLU activation was trained from scratch using the extracted features as input.

In addition to fine-tuning and FE, we also evaluate a zero-shot approach using the unmodified BioLingual model to assess its generalization capabilities without any additional training. In this setup, vessel distance is divided into three descriptive ranges based on the actual distance in kilometers. Instances with a distance less than 3 km are labeled as *"vessel noise from nearby vessel,"* those between 3 and 7 km as *"vessel noise from distant vessel,"* and those greater than 7 km as *"vessel noise from very distant vessel."* These class labels are passed as text queries,

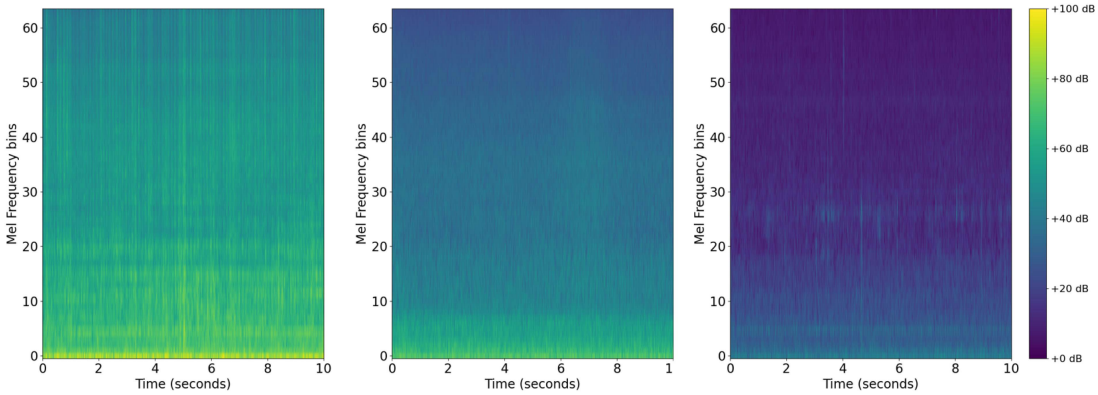


Fig. 6. Examples of log-mel spectrograms. The distances to the nearest vessels are 1112, 5057, and 8078 m from left to right. Log-mel spectrograms corresponding to closer vessels appear brighter than those for more distant vessels.

and each audio sample is classified according to the label with the highest cosine similarity in the shared audio–text embedding space. This setup enables classification without retraining, providing insight into the model's out-of-the-box performance for this task. The evaluation is performed on the test set from the GardenCity station, consisting of 1203 audio files collected during the first part of the year 2022.

E. Acoustic Preprocessing

Raw 10-s audio segments were converted to log-mel spectrograms using the preprocesser from the CLAP model, using a window size of 1024, a hop size of 320, and 64 mel bins as input to the encoder. Each 10-s audio file produced one log-mel spectrogram with a size of (1001, 64). Examples of the obtained log-mel spectrograms are shown in Fig. 6 together with the distance to the closest ship assigned to each 10-s segment.

A qualitative analysis was performed by mapping both the flattened log-mel spectrograms (1001 by 64) and the embeddings produced by the BioLingual model (1 by 512) into a 2-D space using UMAP [70]. This approach enabled a comparison between the input embeddings of the fine-tuning and the FE approach. By visualizing both sets of embeddings, we aimed to evaluate how effectively the BioLingual model captures relevant features, particularly in terms of distinguishing distances and potentially identifying specific activities.

F. Loss Function

In multiclass classification, the standard loss function is cross-entropy loss (CEL), which assumes that categories are independent of each other. However, in our application, the categories represent ordered distance ranges, and our goal is not merely to classify, but to estimate vessel distance as a continuous variable. Predicting a range of "1–2 km" when the true class is "2–3 km" should incur a smaller penalty than predicting a class that is much further away.

Although the model outputs a discrete class, we designed the loss function to reflect the underlying continuity of distance. This approach aligns more closely with regression, where errors further from the target incur larger penalties. Our custom loss function penalizes misclassifications proportionally to their distance from the true class—conceptually similar to mean-squared error (mse), but tailored to our domain through a distance-aware similarity transformation. In this way, the classification framework provides structure, while the loss function enforces regression-like behavior, enabling more accurate distance estimation.

Exploring transformation functions: The goal of the loss function was to impose exponentially larger penalties for greater differences between the classification and the true value. The difference can be expressed as the absolute value between their respective class indices. Each class represents a 1-km distance category—for instance, class 0 corresponds to 0–1 km, class 1 to 1–2 km, and so on—so the label y refers to the starting point of the corresponding distance interval.

Since the maximum possible difference is 10 km, we normalize this class difference to a range between 0 and 1 by dividing by 10. The resulting normalized distance difference (NDD) is defined in the following equation, where y_{true} and y_{pred} are the class indices of the true and predicted categories, respectively:

$$\text{NDD} = \frac{|y_{\text{true}} - y_{\text{pred}}|}{10}. \tag{1}$$

The distance similarity can then be expressed by subtracting the NDD from 1, as shown in (2). The distance similarity is denoted as x in the upcoming formulas

$$x = 1 - \text{NDD}. \tag{2}$$

To express the exponential relationship, we explored a formula based on the quadratic loss function (L2 loss), as shown in the L2 similarity transformation in (3). The penalties for the difference [from (2)] increase quadratically, as illustrated by the black graph in Fig. 7

$$\text{Transformation}_{\text{L2}} = x^2. \tag{3}$$

However, we also sought a more flexible formula that would allow for better adjustment to the relationships between categories. For this, we developed the custom similarity transformation (CST), as shown in (4). This function determines the similarity between classes using two adjustable parameters: a

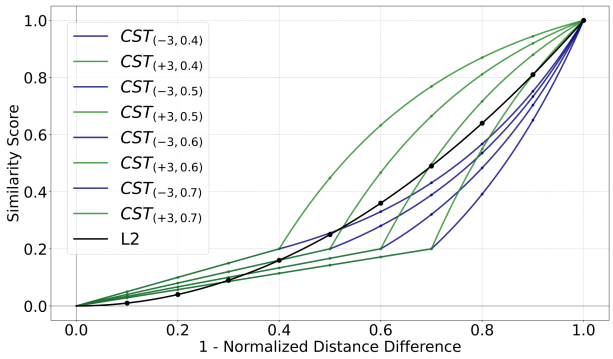


Fig. 7. Comparison of custom transformation functions. The L2 transformation is shown in black, as illustrated by the black graph in Fig. 7. The $\text{CST}_{(a,b)}$ value combinations are displayed in green for $a = +3$ and in blue for $a = -3$, using the custom transformation function given in (4). The x-axis represents the NDD from (2) (i.e., the distance difference between the predicted and actual vessel positions, scaled between 0 and 1). A smaller distance difference results in a higher similarity score, with a score of 1 indicating an exact match (no difference) and a score of 0 representing the maximum possible difference of 10 km. The y-axis shows the similarity score.

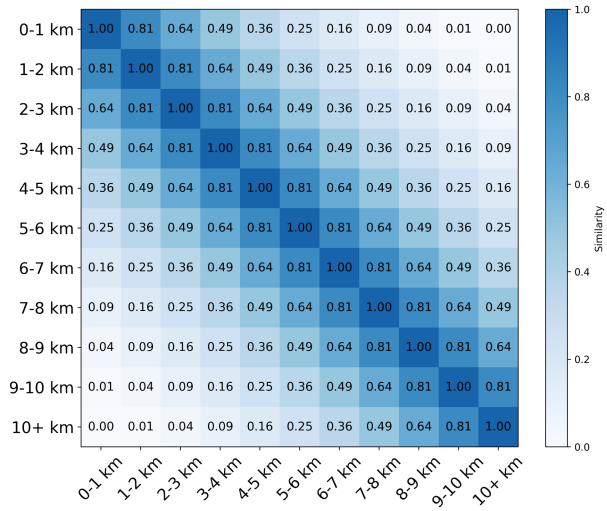


Fig. 8. Similarity matrix $\mathbf{S}$, computed using the L2 transformation function from (3).

and b. The parameter a controls the steepness of the transformation, whereas b determines the transition point where the transformation shifts from linear to exponential behavior. By adjusting these parameters, we fine-tuned the similarity between the classifications

$$\text{CST}_{a,b}(x) = \begin{cases} \frac{0.2}{b}x, & \text{if } x < b \\ 0.2 + \frac{(1-0.2)(1-e^{-a(x-b)})}{(1-e^{-a(1-b)})}, & \text{if } b \leq x \leq 1 \\ 1, & \text{if } x > 1. \end{cases} \tag{4}$$

Various versions of this CST are visualized in Fig. 7. The CST produces two distinct types of curves: higher values of a (green) result in a steep initial increase that gradually slows down, whereas lower (negative) a values (blue) lead to a more gradual initial rise that accelerates over time, resembling an exponential growth pattern. The parameter b determines the transition point where this behavior shifts.

For instance, when the distance difference is 1 km (i.e., $x = 0.9$), the L2 transformation assigns a similarity score of 81%, whereas the CST function—depending on the values of a and b—produces scores ranging between approximately 65% and 95%. However, for a distance difference of 5 km ($x = 0.5$), the L2 similarity drops to 25%, whereas the CST function yields values between roughly 15% and 45%.

The transformations have to be connected to the loss function. To do this, we transform the previously mentioned CEL.

Transforming CEL: For a single sample, CEL is defined as shown in the following equation:

$$\text{CEL} = -\log(p_y) \tag{5}$$

where p_y represents the predicted probability of the true class y.

The CEL can also be expressed in terms of the log softmax function. Given a vector of logits $\mathbf{z} \in \mathbb{R}^C$, the softmax function $\text{softmax}(\mathbf{z})$ converts these logits into predicted probabilities. The log softmax function, which is the logarithm of these probabilities, is defined as

$$\text{log\_softmax}_i = \log\left(\frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}\right). \tag{6}$$

Using log softmax, the CEL for a single sample can be rewritten as

$$\text{CEL} = -\text{log\_softmax}_y. \tag{7}$$

Custom Loss Function: To integrate the transformation functions into the CEL, we constructed a similarity matrix that captures the relative closeness of different categories. Let $\mathbf{S}$ be a similarity matrix of size $C \times C$, where C is the number of classes. Each entry s_{ij} represents the similarity between class i and class j, and $s_{*,n}$ denotes the similarity row corresponding to the true class n. The values of s_{ij} can be computed using either (3) or (4). The similarity matrix $\mathbf{S}$ based on L2 similarity is shown in Fig. 8.

To compute the custom loss, the similarity row $s_{*,n}$ for the true class n is selected and multiplied elementwise with the log softmax values from (6). The resulting values are then summed, as expressed in the following equation:

$$\text{loss}_{k,n} = -\sum_{i=1}^{C} \text{log\_softmax}_i \times s_{*,n} \tag{8}$$

where $\text{loss}_{k,n}$ is the loss for the sample k corresponding to true class n. The final loss is averaged over the batch, depending on the batch size.

The similarity matrix $\mathbf{S}$ ensures that the loss function accounts for both exact matches and the relative closeness of categories, leading to more nuanced classifications and improved performance.

TABLE I
HYPERPARAMETER RANGES FOR FINE-TUNING AND FE

	Parameter	Fine-Tuning	Feature Extraction
Hyperparameters	Learning Rate	[1e-3,1e-4,1e-5]	[1e-3,1e-4]
	Batch Size	[8,16]	[8,16,32]
	L2 Loss	[True, False]	[True, False]
	a Value (CST)	[-3, -2.9, ..., 2.9, 3]	[-3, -2.9, ..., 2.9, 3]
	b Value (CST)	[0.4, 0.5, 0.6, 0.7]	[0.4, 0.5, 0.6, 0.7]
Other	Number of Runs	10 000	32
	Max Epochs	100	100
	Early Stop	True	True
	Optimizer	Adam	Adam

G. Model Training

Two models were trained: one for FE and another for fine-tuning. Both were optimized using the loss function in (8), with a similarity matrix based on either the L2 transformation (3) or the CST (4). Training was conducted for up to 100 epochs, with early stopping triggered if the validation loss did not improve for four consecutive epochs.

For both models, training was performed using Ray Tune [71] to facilitate hyperparameter optimization. During training, Ray Tune explored a range of hyperparameters, including the CST parameters a and b, which were part of the search space. Due to computational constraints, the parameter search range for the fine-tuning approach was narrower and selected based on the best-performing configurations. Each training run consisted of at least three epochs, and the optimal configuration was determined based on the lowest real mse.

A summary of the fixed parameters and the hyperparameter ranges explored during tuning is provided in Table I. For fine-tuning, the search space was constrained based on the best-performing configurations to mitigate computational costs. The tested values for CST parameters a and b were selected carefully, excluding $a = 0$ to prevent an illogical outcome in the custom growth function (4). The loss function option determined whether the squared error loss (L2) or the CST was applied; CST was only used when L2 loss was disabled.

All operations, including both preprocessing and model training, were performed on a system running Linux 5.15.0-125-generic with an x86_64 CPU and an NVIDIA Tesla V100-PCIE-32 GB GPU (34.07 GB VRAM). The training utilized CUDA 12.1 and cuDNN 8.9.2. The experiments were conducted in a Python 3.10.13 environment using PyTorch 2.2.0.

For inference and potential real-time deployment, the model is designed to operate efficiently on standard CPUs without requiring GPU acceleration. Initial tests indicate that a single modern CPU core is sufficient for processing 10-s audio clips with low latency, enabling deployment on modest hardware setups, such as edge devices or dedicated monitoring stations.

H. Comparing Model Performance Across Environments

To evaluate the generalizability of our approach, we compare the performance of our trained model across multiple acoustic environments. In particular, we include an external benchmark using data from the Strait of Georgia, as introduced in [72]. This dataset is openly available on GitHub and includes underwater vessel recordings collected in a marine delta region with a depth of approximately 140 m.

We extracted 10-s WAV segments from this dataset in order to match the input format expected by our model. This allows us to assess how well a model trained on our primary dataset transfers to recordings from a geographically and acoustically distinct location. Such cross-dataset evaluation is crucial for understanding model robustness and identifying environment-specific performance differences.

In addition to this external comparison, we analyze model performance per recording station within our own dataset. This allows us to identify which stations—or by extension, which acoustic conditions or habitat types—are best handled by our approach. It also enables a fairer comparison across alternative models by quantifying their station-specific strengths and weaknesses.

Together, these comparisons provide a clearer view of the operational boundaries of our method and help guide future improvements in generalization across variable marine soundscapes.

III. RESULTS

A. Qualitative Analysis

The qualitative analysis through UMAP reveals the relationship between distance, station, and embedding model in the dataset. As shown in Fig. 9, the embeddings derived from both the log-mel spectrograms and the BioLingual model exhibit subtle clustering patterns when color-coded by distance and more distinct clusters when color-coded by station.

In the case of the spectrogram embeddings [see Fig. 9(a) and (b)], some trends related to distance are observed, but these trends are more strongly influenced by the station, as seen in Fig. 9(b). The environmental factors related to different stations seem to play a more prominent role in the organization of these embeddings. In contrast, the BioLingual embeddings [see Fig. 9(c) and (d)] exhibit a small dark-to-light gradient with respect to distance. This gradient is particularly pronounced, with a transition from a cluster of long-distance datapoints on the left to a concentrated short-distance cluster on the right.

B. Zero-Shot Classification

The zero-shot classification resulted in an overall accuracy of **0.280**. The confusion matrix in Fig. 10 shows that the model has a tendency to overpredict the majority class (*very distant*), with limited ability to distinguish between nearby and distant vessel sounds. Despite this, the result demonstrates that the model retains some, however limited, meaningful general-purpose acoustic representations.

This zero-shot performance serves as a lower bound baseline, illustrating the potential of leveraging audio–language models for maritime acoustic monitoring.

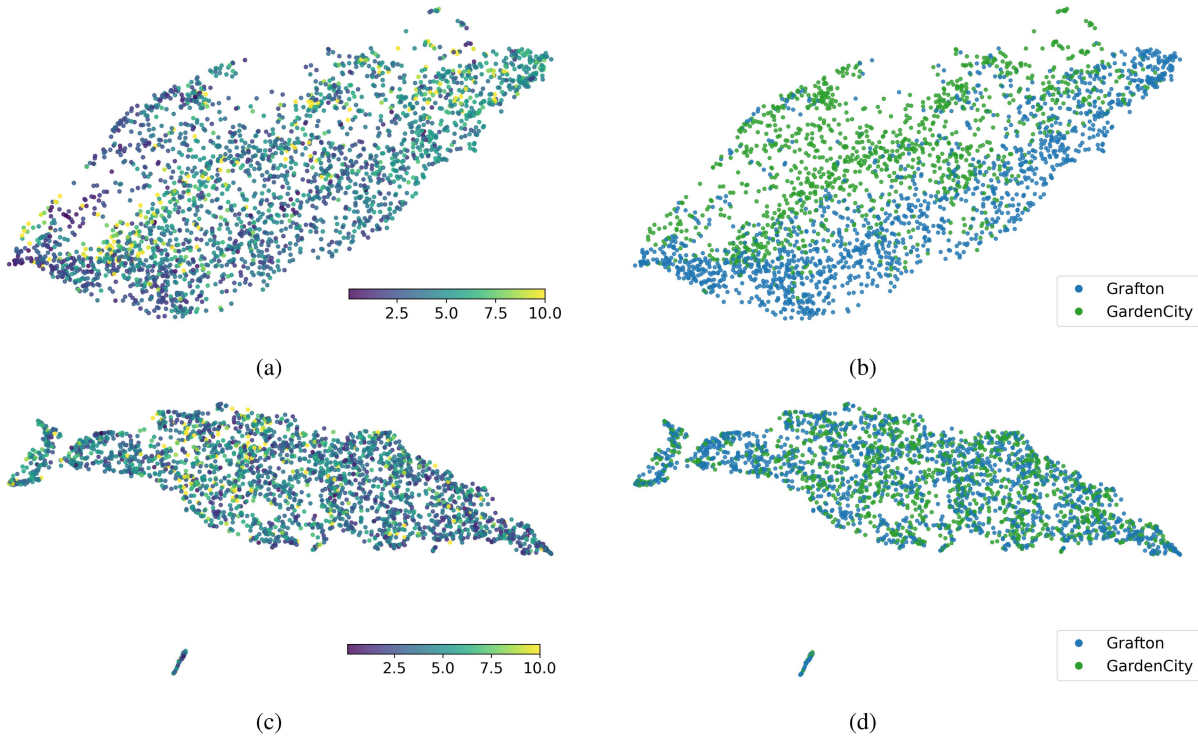


Fig. 9. UMAP 2-D visualizations of embeddings based on vessel distance and station characteristics, color-coded for clarity. (a) Flattened log-mel spectrogram embeddings color-coded by distance. (b) Flattened log-mel spectrogram embeddings color-coded by station. (c) BioLingual embeddings color-coded by distance. (d) BioLingual embeddings color-coded by station. These visualizations highlight how different embedding models represent distance and station characteristics. A total of 62 datapoints were excluded due to being shorter than 10 s, and the maximum distance was capped at 10 km. The visualizations are based on 2747 datapoints from the test set (after excluding the 62). UMAP was configured with 20 neighbors and a min dist value of 0.1 to generate the plots.

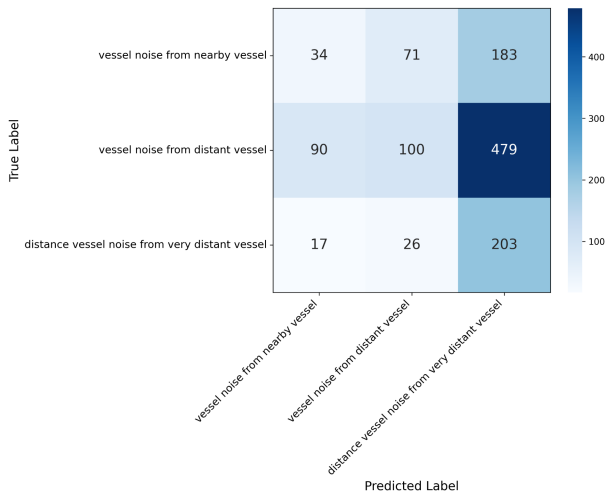


Fig. 10. Confusion matrix of the zero-shot classification results using the BioLingual model, evaluated on the GardenCity test set consisting of 1203 audio files from the first part of 2022. Rows represent true classes. Columns represent predicted classes.

TABLE II
SELECTED HYPERPARAMETER VALUES FOR BEST-PERFORMING MODELS AND OBTAINED RESULTS

	Parameter	Feature Extraction	Fine-Tuning
Hyperparameters	Learning Rate	1e-3	1e-4
	Batch Size	16	16
	a Value (Loss Function)	-	1.4
	b Value (Loss Function)	-	0.5
	L2 Loss	True	False
Results	MSE	2.9204	**2.519**
	RMSE	1.709	**1.587**
	Total Epochs	9	**7**
Time	Average Epoch Time	**35.8s**	1h52m
	Total Processing Time	**5.5m**	13h06m
	Average Pre-Processing Time	**1s**	1.5s
	Total Pre-Processing Time	**6h**	9h

The best results are highlighted in bold.

C. Models Performance

The selected hyperparameter values by Ray Tune are listed in Table II. The model evaluation is provided considering the best-performing values for each model. The fine-tuned model demonstrated slightly higher performance than the FE approach, but at the expense of computational efficiency. The full processing time (including training, validation, and testing) was approximately 143 times longer (13 h 06 min versus 5.5 m). In addition, the preprocessing phase (i.e., creating the embeddings) took 1.5 times longer (9 h versus 6 h).

This significant tradeoff between time and performance was expected, as fine-tuning requires retraining the entire model, whereas FE leverages pretrained layers with minimal adjustments.

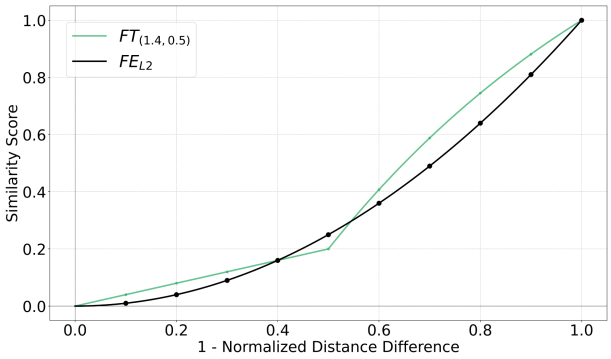


Fig. 11. Comparison of the best-performing loss functions. The quadratic loss function, which performed best for FE, is shown in black, whereas the most effective fine-tuning approach, $\text{CST}_{(a,b)}$ with $a = 1.4$ and $b = 0.5$, is displayed in green. The x-axis represents the distance similarity between the predicted and actual vessel positions. This similarity is computed by taking the absolute difference between the predicted and true distances and scaling it between 0 and 1. A smaller difference results in a higher similarity score, with a score of 1 indicating an exact match (no difference) and a score of 0 representing the maximum possible difference of 10 km. The y-axis shows the similarity score.

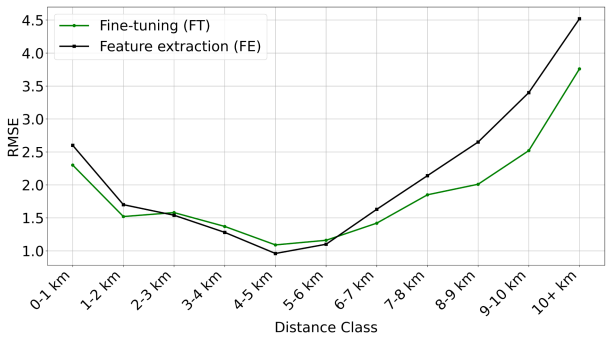


Fig. 12. RMSE per distance class for the fine-tuning (green) and FE (black) models.

The fine-tuned model achieved a root-mean-square error (RMSE) of 1.587, whereas the FE model reached 1.709. Regarding the selected loss functions, the quadratic loss function (L2) performed best for FE, whereas the fine-tuning approach benefited from a custom similarity function with $b = 0.5$ and $a = 1.4$. This function, shown in Fig. 12, is less strict on smaller errors, resulting in higher similarity scores.

We acknowledge the observation that comparing the models may raise concerns due to their different optimization characteristics. Fine-tuning typically involves updating a larger number of parameters and can require more training epochs and data to reach optimal performance. However, in our case, early stopping was guided by validation performance, and the fine-tuned model converged more quickly (in seven epochs) than the FE model (nine epochs). This fast convergence is likely due to the large amount of available training data, allowing the model to learn effectively within a few iterations. Therefore, the shorter training duration for the fine-tuned model does not reflect underfitting, but rather a more rapid convergence toward optimal performance.

The obtained confusion matrices for the best-performing fine-tuning and FE models, respectively, can be seen in Fig. 13.

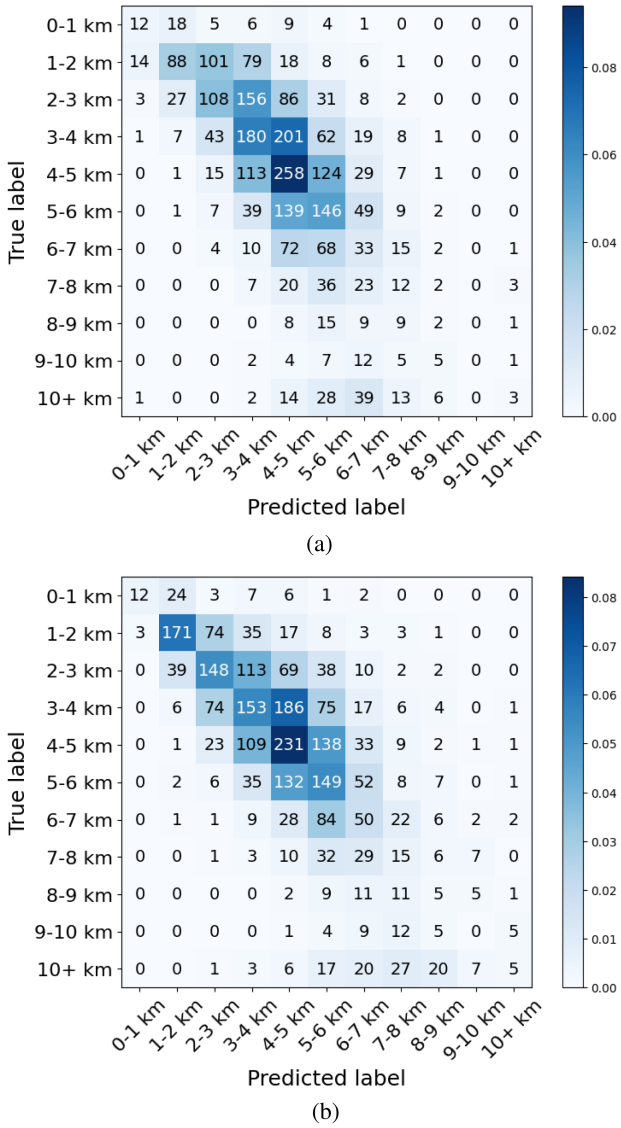


Fig. 13. Confusion matrices for the obtained models using the best-performing hyperparameters, evaluated in the independent test set for the approaches (a) FE and (b) fine-tuning. Values in the confusion matrix refer to the number of instances, whereas the coloring scale shows the percentage of instances in the total test set.

Most values are concentrated near the diagonal in both matrices, indicating that the models correctly predict the majority of classes. However, some misclassifications are evident. Underestimations—where vessels are classified as being closer than they actually are—are more common than overestimations. This is especially noticeable in the middle distance bins (e.g., 3–6 km), where vessel sounds may resemble closer range acoustic patterns.

Even though the FE approach results in a slightly lower average RMSE, it exhibits a tendency to overestimate vessel presence at short distances (e.g., 1–2 km and 2–3 km), while slightly underestimating presence at greater distances (e.g., 9–10 km and 10+ km). This behavior is reflected in Fig. 12, which shows the RMSE per distance class for both models. The fine-tuning approach, in contrast, performs better in the farthest classes,

especially the 10+ km bin, suggesting better generalization for weak and distant acoustic signals. On the other hand, the FE model may benefit from stronger general acoustic features but lacks specificity in the extreme classes.

Figs. 14 and 15 display the classifications of the best-performing model for an entire day's recording at two stations, comparing the fine-tuning and FE approaches. In both cases, the model predictions (blue lines) align more closely with the AIS ground truth (black lines) when using the fine-tuning approach. This outcome is expected, as the fine-tuned model consistently outperformed the FE model across evaluation metrics, reflecting its improved ability to learn and generalize from the training data. Specifically, it demonstrates greater sensitivity to nuanced acoustic patterns, allowing it to better detect distant vessel presence. This improved alignment is particularly evident in the outer distance bins (e.g., 9–10 km and 10+ km), supporting the earlier observations from the confusion matrices and RMSE plots.

Notably, in Fig. 15, both model's classifications diverge from the AIS data around 13 h. However, upon reviewing the audio and power spectrum, evidence of a vessel's presence emerges that is not captured by the AIS data. This indicates that the model successfully detected a *dark* vessel around 13 h, which the AIS failed to register. Such detections are particularly valuable, as they highlight the model's capacity to identify real-world vessel activity that eludes traditional AIS-based monitoring. These results demonstrate the potential of passive acoustic models to complement AIS data by filling observational gaps. While this is just one illustrative case, similar occurrences were noted at other time points and locations; additional examples are provided in the Supplementary Material.

D. Comparing Model Performance Across Environments

To assess the generalizability of our model across diverse acoustic settings, we evaluated its performance on the openly available Strait of Georgia dataset [72]. This environment differs substantially from our primary study region in both bathymetry (depths reaching 140 m) and vessel traffic characteristics. The dataset includes acoustic recordings collected within 2 km of known vessel positions, enabling direct input to our classifier using 10-s WAV segments.

Fig. 16 shows the confusion matrix of the model's predictions on this dataset. While the model was trained on shallow coastal environments (e.g., the BPNS), it still manages to classify a large proportion of samples within the correct or neighboring distance bins—primarily between 1–4 km—with rare predictions extending up to 7 km. Although overall performance is modest, the model does exhibit some capacity to transfer to this much deeper and acoustically distinct environment.

This result highlights both the limitations and potential of using a general acoustic distance classifier across domains. It suggests that while the model architecture can capture transferable features, fine-tuning is likely necessary to achieve high performance in each specific deployment context.

To further investigate environmental robustness, we also compare model performance across stations and deployment periods in our own dataset. This per-station analysis allows us to examine which acoustic conditions or habitat types are best supported by the model and whether certain deployments benefit more from fine-tuning. Table III summarizes RMSE results for both the FE and fine-tuning approaches, alongside key deployment metadata.

TABLE III
SUMMARY OF DEPLOYMENT RMSE FOR FE AND FINE TUNING MODELS

station	deployment period	Feature Extraction RMSE	Fine Tuning RMSE	proportion	depth
Grafton	01/2022–05/2022	1.239	1.175	0.350	23m
GardenCity	01/2022–05/2022	2.164	2.024	0.438	35m
Grafton	08/2022–10/2022	1.211	0.998	0.107	23m
Grafton	10/2022–11/2022	1.283	1.123	0.105	23m

Latitude and longitude for Grafton: approx. 51.406°N, 2.818°E; GardenCity: approx. 51.486°N, 2.305°E.

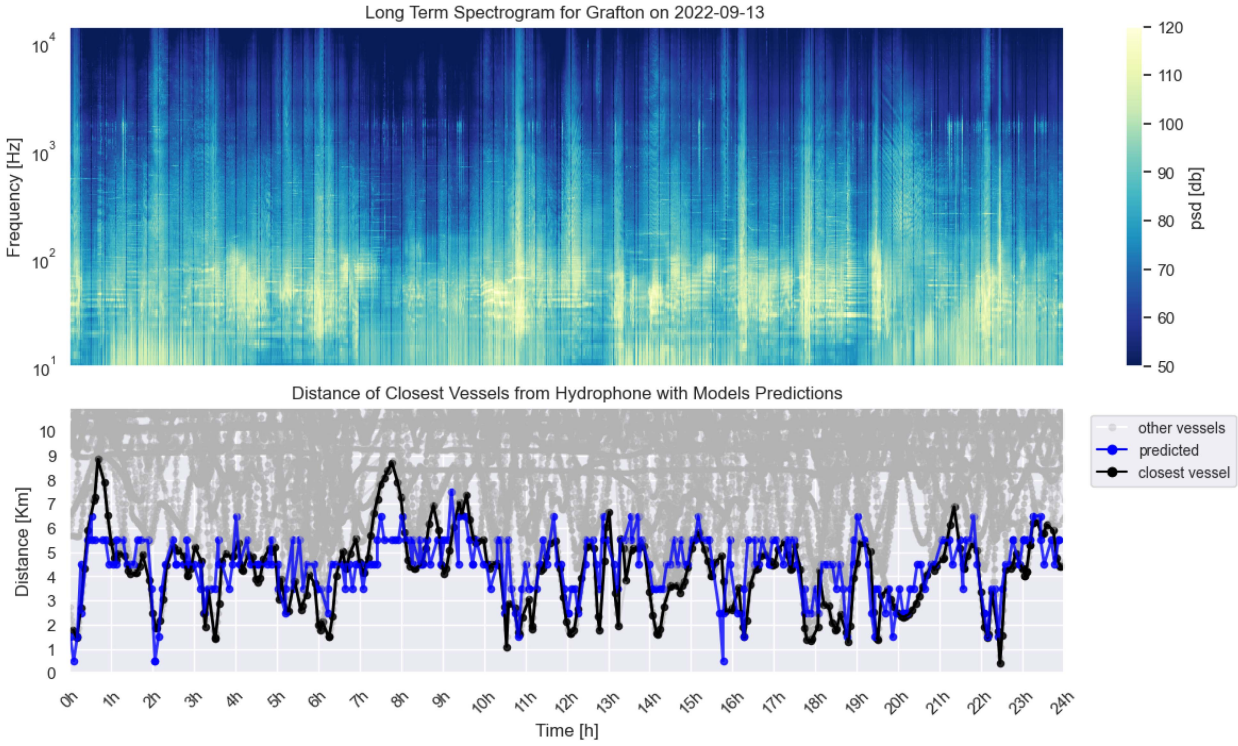
Fine-tuning consistently outperforms the FE model, showing lower RMSE values across all deployment periods and stations. This improvement is particularly notable at the Grafton station, where fine-tuning reduces RMSE by approximately 0.1–0.2 units. At GardenCity, which is located at a greater depth (35 m compared to 23 m at Grafton), both models experience higher errors. Although deeper waters are generally less acoustically complex than shallow environments, this result suggests that other site-specific factors—such as ambient noise levels or seafloor composition—may be affecting performance. Nevertheless, fine-tuning still provides better accuracy compared to FE.

IV. Discussion

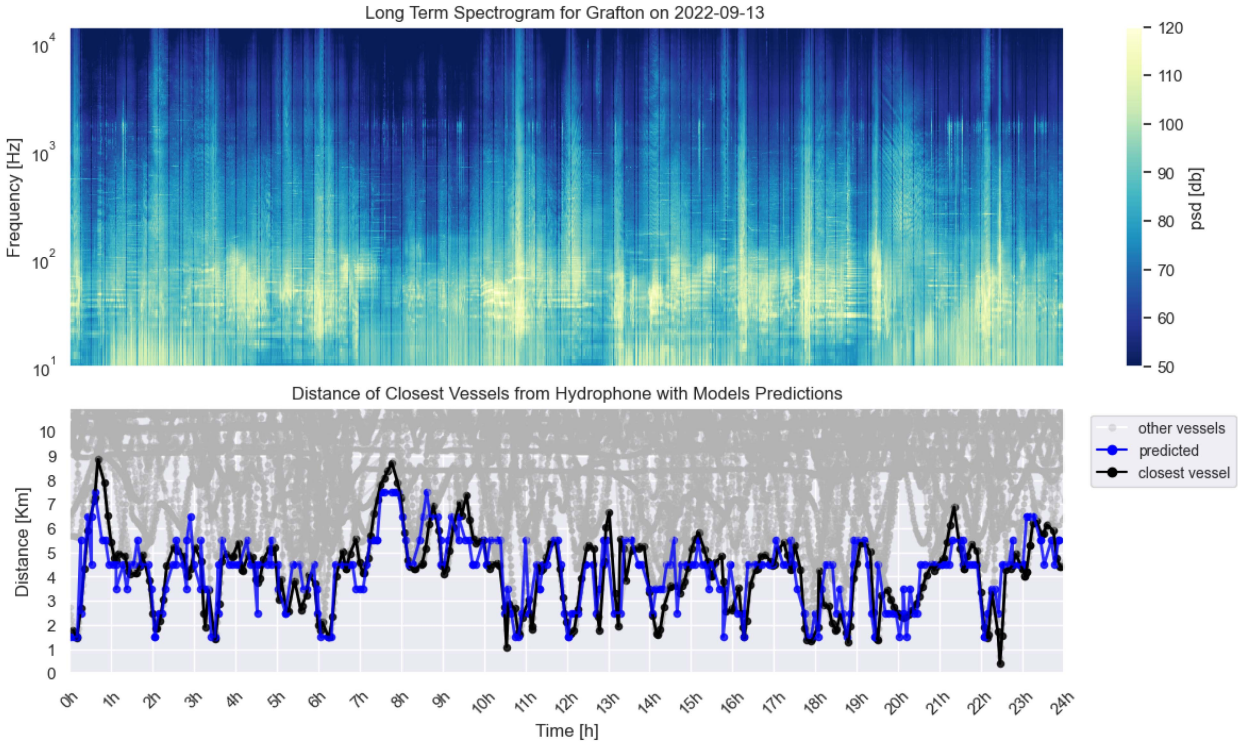
Both models developed in this study show significant promise for monitoring human activities in MPAs. They offer a valuable tool for enhancing our ability to monitor maritime traffic, particularly in regions where traditional methods, such as AIS, are limited or unavailable. By classifying vessel distances based on sound production, these models can help detect *dark* vessels—those that do not transmit AIS signals—making them a useful complement to existing tracking systems. However, to effectively identify specific vessels, integration with AIS or other identification systems remains essential.

Due to the absence of established baselines for vessel distance estimation from passive acoustic data, we compared model performance to a zero-shot classification approach using the BioLingual model. This comparison provides an initial reference point in a field lacking prior benchmarks. Consequently, the models developed here can serve as foundational baselines for future research aiming to improve vessel distance classification from underwater acoustic signals.

This study demonstrates that transfer learning from bioacoustics, specifically underwater sound, is a powerful method for speeding up the training of models aimed at classifying vessel distances. Whereas the fine-tuned model performed slightly better than the pretrained model used as a feature extractor, it can be concluded that the embeddings from the pretrained model are less informative for capturing all the necessary information to classify vessel distance. Nevertheless, the FE model, although performing somewhat lower, still provides a valuable tool for

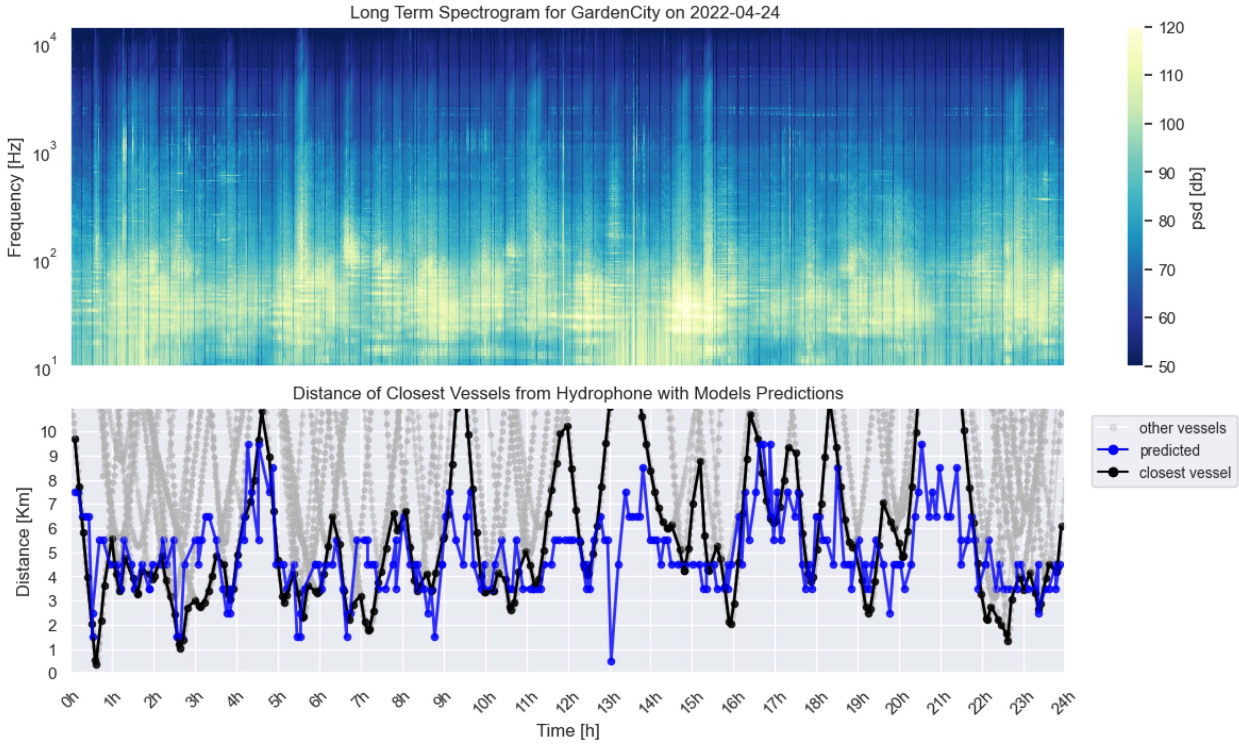


(a)

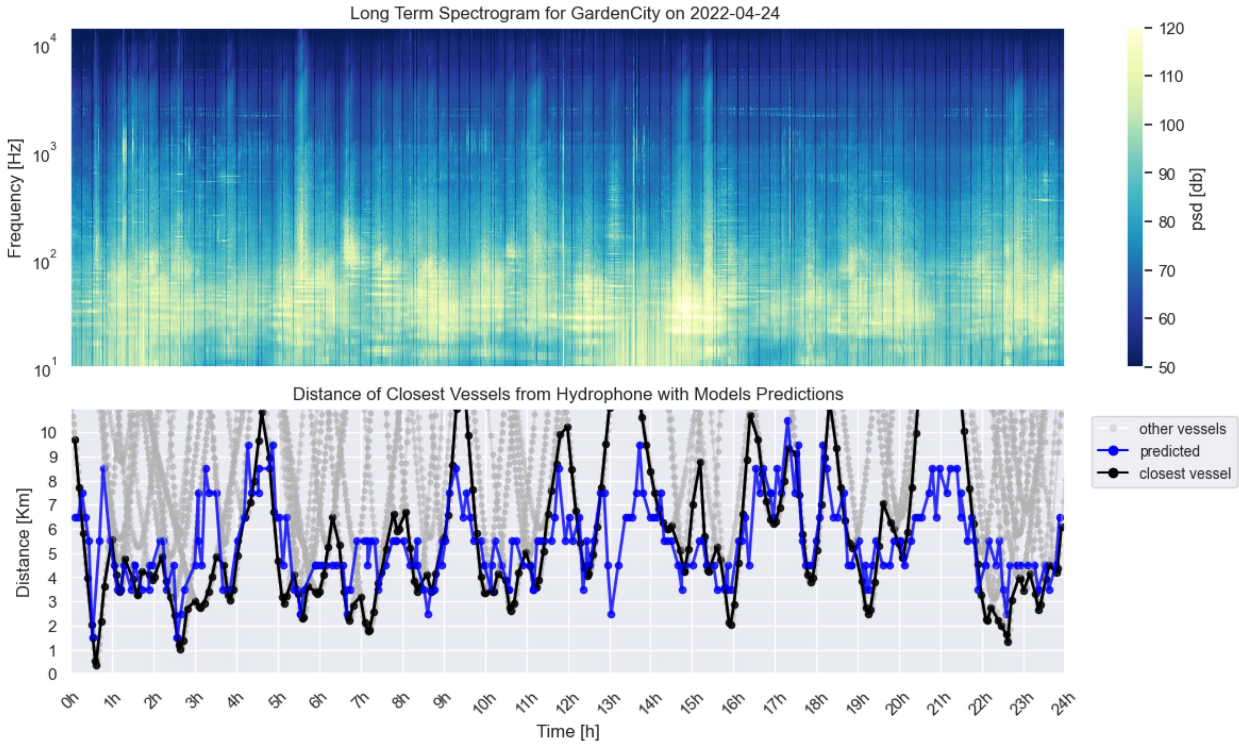


(b)

Fig. 14. Comparison of the power spectrum in relation to vessel proximity from Grafton on 13 September 2022. The top section of each figure displays the power spectrum for the entire day's recording, while the bottom section shows the distances between nearby vessels and the hydrophone. AIS data are shown in black, model classifications in blue, and other AIS vessels in gray. Each classification represents a distance category of 1 km width (e.g., 0–1 km, 1–2 km), and the corresponding dot is positioned at the center of the predicted interval (e.g., 0.5 km for the 0–1 km category). (a) Results from the FE model. (b) Fine-tuning model.



(a)



(b)

Fig. 15. Comparison of the power spectrum in relation to vessel proximity from GardenCity on 24 April 2022. The top section of each figure displays the power spectrum for the entire day's recording, while the bottom section shows the distances between nearby vessels and the hydrophone. AIS data are shown in black, model classifications in blue, and other AIS vessels in gray. Each classification represents a distance category of 1 km width (e.g., 0–1 km, 1–2 km), and the corresponding dot is positioned at the center of the predicted interval (e.g., 0.5 km for the 0–1 km category). (a) Results from the FE model. (b) Fine-tuning model.

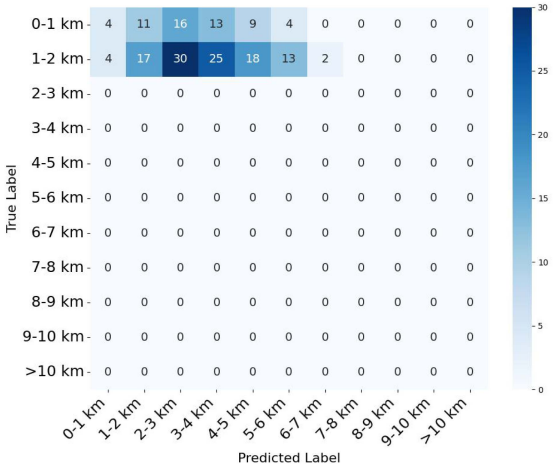


Fig. 16. Confusion matrix showing classification performance of the fine-tuned model on the Strait of Georgia dataset [72].

classifying underwater sounds. This suggests that the BioLingual model's features are suitable for distinguishing underwater sounds and could be beneficial for researchers looking to retrain the model on their own data.

Furthermore, the performance evaluation across different stations and deployment periods reveals that environmental factors, such as depth and site-specific acoustic properties, significantly influence model accuracy. Interestingly, the Grafton station—located in shallower waters—consistently yielded lower RMSEs compared to the deeper GardenCity station. This is counterintuitive, as shallow waters are typically more acoustically complex due to increased reverberation and multipath effects. These results suggest that additional, unaccounted-for variables may be influencing performance, highlighting the need for models to incorporate environmental metadata. The ability of the fine-tuned model to adapt to different sites underscores the benefit of transfer learning in accommodating diverse marine conditions. We also suggest potential strategies for improvement, such as incorporating environmental metadata and training on larger, more varied datasets.

When applying the model to the publicly available DeepShip dataset from the Strait of Georgia in Canada—an area with substantially different bathymetric and oceanographic conditions—performance dropped notably, with most predictions clustering between 1 and 4 km, despite all vessels being located within 2 km of the hydrophone. This suggests that while the model is capable of generalizing to some extent, it does not transfer seamlessly across environments with differing sound propagation characteristics.

Together, these findings emphasize that environment-specific retraining or fine-tuning may be necessary to achieve optimal performance in diverse marine regions. The ability of the fine-tuned model to adapt to new sites nonetheless demonstrates the utility of transfer learning, especially when combined with larger, ecologically varied datasets and contextual environmental features.

Future improvements could focus on enhancing the model's ability to classify specific vessel activities, such as fishing, anchoring, or potentially illegal operations. The model's ability to cluster certain activities, as shown in the embeddings on the bottom of Fig. 9(c) and (d), indicates that there is potential for expanding its functionality to provide a more comprehensive understanding of maritime traffic. However, the limited availability of accurately annotated activity data has prevented the model from being trained on these activities to date.

It is noteworthy that whereas some activity clustering is observed, inconsistencies in the AIS database—such as instances where vessels are inaccurately labeled as "underway"—present challenges for training. Similarly, some underestimations in the model's classifications, where vessel distances appear shorter than recorded AIS values, may indicate potential errors in the AIS data. This raises the possibility that the model is capturing real distances more accurately than AIS in certain cases. However, further investigation is required to validate this hypothesis and determine whether the model is indeed correcting AIS inaccuracies or if these deviations stem from other factors.

In addition, it is challenging to determine exactly how many of these underestimations represent detections of *dark* vessels. Since these vessels do not transmit AIS signals, they may be incorrectly labeled as nearby vessels rather than being explicitly identified. However, it is reasonable to conclude that the model's performance is likely higher than what is suggested by the RMSE alone, as the AIS data themselves contain inaccuracies that could affect the evaluation metrics. Despite these limitations, there is an evident opportunity for improvement, and with the acquisition of more precise annotated data, the model could advance in its ability to capture and interpret various vessel activities more effectively.

The development of these models can play a crucial role in safeguarding MPAs and other zones of interest, such as OWFs, restoration sites, and cable areas, where monitoring vessel activities is essential for environmental protection and regulatory enforcement. Importantly, the model's inference can be performed efficiently on standard CPU hardware, making it suitable for deployment in real-time monitoring scenarios without requiring high-end GPUs. This facilitates implementation on modest or edge computing devices in remote marine environments, enabling continuous, automated oversight with minimal infrastructure demands.

The development of these models can play a crucial role in safeguarding MPAs and other zones of interest, such as OWFs, restoration sites, and cable areas, where monitoring vessel activities is essential for environmental protection and regulatory enforcement. Importantly, the model's inference can be performed efficiently on standard CPU hardware, making it suitable for deployment in real-time monitoring scenarios without requiring high-end GPUs. This facilitates implementation on modest or edge computing devices in remote marine environments, enabling continuous, automated oversight with minimal infrastructure demands.

The proposed ship detection model aligns with key policy objectives outlined in the EU Biodiversity Strategy for 2030, which aims to protect 30% of marine areas, including 10% under strict protection [4]. Ensuring effective management of these areas remains a major challenge, as many MPAs, including Natura 2000 sites, suffer from illegal or unregulated activities. By detecting vessel presence regardless of AIS status through

passive acoustic monitoring, the model provides objective data on human activity and potential noncompliance within these zones. It can serve as a complementary decision-support tool alongside existing spatial monitoring and enforcement measures, helping authorities identify areas of recurring pressure and adapt protection strategies accordingly.

In addition, the model contributes to the MSFD, specifically Descriptor 11, by supplying valuable data on the contribution of ship activity to underwater noise levels, supporting broader assessments of good environmental status in European seas. With further developments in edge computing and satellite connectivity, a scalable surveillance network could be established to monitor MPA activity in near real time, enabling responsive enforcement actions by local authorities even in remote or high-risk marine areas.

The need for efficient monitoring systems is especially relevant in multiuse zones, such as the Vlaamse Banken, a section of the BPNS. Future iterations of the model could distinguish between different types of vessel activities in such areas, providing actionable data to enforce regulations more effectively. This highlights the wide applicability of the model, which could be tailored for various marine environments worldwide, including areas with different regulatory frameworks and conservation priorities.

V. Conclusion

In conclusion, this study presents a valuable approach to monitoring vessel activity in MPAs using underwater acoustic data. The presented models currently provide reliable vessel distance classification, which is a crucial first step in enhancing maritime surveillance. It forms the basis for further developments, including vessel activity, which is needed for more targeted enforcement and conservation efforts.

The future work of this research includes its adaptability to other marine areas worldwide, particularly where traditional monitoring methods are insufficient or cost-prohibitive. By expanding the model to include activity detection and integrating it with complementary technologies, such as SAR and drones, the overall effectiveness of marine monitoring could be greatly improved.

Further research should focus on testing the model in diverse marine environments and developing its capacity to classify vessel activities. These advancements will ensure that the model contributes meaningfully to global marine conservation strategies and strengthens the management of MPAs in both strictly protected and multiuse zones.

Acknowledgment

The authors would like to thank the project managers and all partners involved for their efforts in fostering the creation of open-access image repositories for AI-based image analysis services. Finally, the authors are also grateful to the VLIZ researchers and crew of the *RV Simon Stevin* for their practical support during the monthly sampling campaigns and to the Flemish Ministry of Mobility and Public Works (DAB VLOOT) for operating the vessel and facilitating the surveys. Special thanks go to ChatGPT for its invaluable assistance in providing constructive feedback, refining sections of this article, and aiding in the development of the ideas presented in this work.

References

[1] E. T. Game et al., "Pelagic protected areas: The missing dimension in ocean conservation," *Trends ecol. Evol.*, vol. 24, no. 7, pp. 360–369, 2009.
[2] O. Delfour-Samama and C. Leboeuf, "Review of potential legal frameworks for effective implementation and enforcement of MPAs in the high seas," *ICES J. Mar. Sci.*, vol. 71, pp. 1031–1039, Aug. 2014.
[3] A. M. E. Arefin, "Proposal of a marine protected area surveillance system against illegal vessels using image sensing and image processing," *Acta Ecologica Sinica*, vol. 38, no. 2, pp. 111–116, 2018.
[4] E. Commission and D.-G. f. Environment, "EU biodiversity strategy for 2030–Bringing nature back into our lives," Publications Office of the European Union, 2021.
[5] J. Reker et al. *Marine Protected Areas in Europe's Seas. An Overview and Perspectives for the Future*. Copenhagen, Denmark: Eur. Environ. Agency, 2015, pp. 1–35.
[6] B. Halpern et al., "Recent pace of change in human impact on the world's ocean," *Sci. Rep.*, vol. 9, 2019, Art. no. 11609.
[7] J. Tournadre, "Anthropogenic pressure on the open ocean: The growth of ship traffic revealed by altimeter data analysis," *Geophysical Res. Lett.*, vol. 41, pp. 7924–7932, 2014.
[8] C. Duarte et al., "The soundscape of the anthropocene ocean," *Science*, vol. 371, 2021, Art. no. eaba4658.
[9] B. Causey, "Enforcement in marine protected areas," in *Marine Protected Areas: Principles and Techniques for Management*. Dordrecht, The Netherlands: Springer, 1995, pp. 119–148.
[10] P. Guidettia et al., "Italian marine reserve effectiveness: Does enforcement matter?," *Biol. Conservation*, vol. 141, pp. 699–709, 2008.
[11] J. A. Hildebrand, "Anthropogenic and natural sources of ambient noise in the ocean," *Mar. Ecol. Prog. Ser.*, vol. 395, pp. 5–20, 2009.
[12] R. Williams et al., "Impacts of anthropogenic noise on marine life: Publication patterns, new discoveries, and future directions in research and management," *Ocean Coastal Manage.*, vol. 115, pp. 17–24, 2015.
[13] E. Commission, "EU strategy on offshore renewable energy (com/2020/741)," 2020. Accessed: Jan. 26, 2025.
[14] R. Williams, C. Erbe, E. Ashe, and C. W. Clark, "Quiet (ER) marine protected areas," *Mar. Pollut. Bull.*, vol. 100, no. 1, pp. 154–161, 2015.
[15] L. T. Hatch and K. M. Fristrup, "No barrier at the boundaries: Implementing regional frameworks for noise management in protected natural areas," *Mar. Ecol. Prog. Ser.*, vol. 395, pp. 223–244, 2009.
[16] K. W. Chung, A. Sutin, A. Sedunov, and M. Bruno, "Demon acoustic ship signature measurements in an urban harbor," *Adv. Acoust. Vib.*, vol. 2011, 2011, Art. no. 952798.
[17] A. Sutin et al., "Stevens passive acoustic system for underwater surveillance," in *Proc. 2010 Int. WaterSide Secur. Conf.*, 2010, pp. 1–6.
[18] L. Fillinger et al., "Towards a passive acoustic underwater system for protecting harbours against intruders," in *Proc. 2010 Int. WaterSide Secur. Conf.*, 2010, pp. 1–7.
[19] E. L. Ferguson, R. Ramakrishnan, S. B. Williams, and C. Jin, "Deep learning approach to passive monitoring of the underwater acoustic environment," *J. Acoust. Soc. Amer.*, vol. 140, pp. 3351–3351, 2016.
[20] J. L. Shepperson, N. T. Hintzen, C. L. Szostek, E. Bell, L. G. Murray, and M. J. Kaiser, "A comparison of VMS and AIS data: The effect of data coverage and vessel position recording frequency on estimates of fishing footprints," *ICES J. Mar. Sci.*, vol. 75, no. 3, pp. 988–998, 2018.
[21] M. Robards et al., "Conservation science and policy applications of the marine vessel automatic identification system (AIS)—A review," *Bull. Mar. Sci.*, vol. 92, no. 1, pp. 75–103, 2016.
[22] L. Hermannsen, L. Mikkelsen, J. Tougaard, K. Beedholm, M. Johnson, and P. T. Madsen, "Recreational vessels without automatic identification system (AIS) dominate anthropogenic noise contributions to a shallow water soundscape," *Sci. Rep.*, vol. 9, Oct. 2019, Art. no. 15477.
[23] L. Yao, Y. Liu, and Y. He, "A novel ship-tracking method for GF-4 satellite sequential images," *Sensors*, vol. 18, 2018, Art. no. 2007.
[24] S. Voinov, F. Heymann, R. Bill, and E. Schwarz, "Multiclass vessel detection from high resolution optical satellite images based on deep neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 166–169.
[25] O. Dubovik et al., "Grand challenges in satellite remote sensing,", Frontiers in Remote Sensing, vol. 2, p. 619818, 2021.

[26] J. Wang, K. Zhou, W. Xing, H. Li, and Z. Yang, "Applications, evolutions, and challenges of drones in maritime transport," *J. Mar. Sci. Eng.*, vol. 11, no. 11, 2023, Art. no. 2056.
[27] G. Margarit and A. Tabasco, "Ship classification in single-pol SAR images based on fuzzy logic," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 8, pp. 3129–3138, Aug. 2011.
[28] R. G. Meyer, W. Kleynhans, and C. P. Schwegmann, "Small ships don't shine: Classification of ocean vessels from low resolution, large swath area SAR acquisitions," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 975–978.
[29] N. Ødegaard, A. O. Knapskog, C. Cochin, and J.-C. Louvigne, "Classification of ships using real and simulated data in a convolutional neural network," in *Proc. IEEE Radar Conf.*, 2016, pp. 1–6.
[30] C. Bentes, A. Frost, D. Velotto, and B. Tings, "Ship-iceberg discrimination with convolutional neural networks in high resolution SAR images," in *Proc. EUSAR 2016: 11th Eur. Conf. Synthetic Aperture Radar*, 2016, pp. 1–4.
[31] C. Bentes, D. Velotto, and S. Lehner, "Target classification in oceanographic SAR images with deep neural networks: Architecture and initial results," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 3703–3706.
[32] M. Kang, X. Leng, Z. Lin, and K. Ji, "A modified faster R-CNN based on CFAR algorithm for SAR ship detection," in *Proc. 2017 Int. Workshop Remote Sens. Intell. Process.*, 2017, pp. 1–4.
[33] Z. Zhang, L. Zhang, J. Wu, and W. Guo, "Optical and synthetic aperture radar image fusion for ship detection and recognition: Current state, challenges, and future prospects," *IEEE Geosci. Remote Sens. Mag.*, vol. 12, no. 4, pp. 132–168, Dec. 2024.
[34] L. Cinelli, G. Chaves, and M. Lima, "Vessel classification through convolutional neural networks using passive sonar spectrogram images," Proceedings of the Simp'osio Brasileiro deTelecomunicac¸ ˜oes e Processamento de Sinais (SBrT 2018),Armac¸ ˜ao de Buzios, Brazil, pp. 21–25, May. 2018.
[35] B. M. Howe, J. Miksis-Olds, E. Rehm, H. Sagen, P. F. Worcester, and G. Haralabus, "Observing the oceans acoustically," *Front. Mar. Sci.*, vol. 6, 2019, Art. no. 426.
[36] J. Choi, Y. Choo, and K. Lee, "Acoustic classification of surface and underwater vessels in the ocean using supervised machine learning," *Sensors*, vol. 19, 2019, Art. no. 3492.
[37] R. Zaheer, I. Ahmad, D. Habibi, K. Y. Islam, and Q. V. Phung, "A survey on artificial intelligence-based acoustic source identification," *IEEE Access*, vol. 11, pp. 60078–60108, 2023.
[38] A. Rahman and F. Stillinger, "Propagation of sound in water. A molecular-dynamics study," *Phys. Rev. A*, vol. 10, pp. 368–378, 1974.
[39] H. P. Bucker, "Use of calculated sound fields and matched-field detection to locate sound sources in shallow water," *J. Acoust. Soc. Amer.*, vol. 59, no. 2, pp. 368–373, 1976.
[40] E. K. Westwood, "Broadband matched-field source localization," *J. Acoust. Soc. Amer.*, vol. 91, no. 5, pp. 2777–2789, 1992.
[41] A. B. Baggeroer, W. A. Kuperman, and P. N. Mikhalevsky, "An overview of matched field methods in ocean acoustics," *IEEE J. Ocean. Eng.*, vol. 18, no. 4, pp. 401–424, Oct. 1993.
[42] Z.-H. Michalopoulou and M. B. Porter, "Matched-field processing for broad-band source localization," *IEEE J. Ocean. Eng.*, vol. 21, no. 4, pp. 384–392, Oct. 1996.
[43] D. Gingras, "Robust broadband matched-field processing: Performance in shallow water," *IEEE J. Ocean. Eng.*, vol. 18, no. 3, pp. 253–264, Jul. 1993.
[44] C. Debever and W. Kuperman, "Exploring the limits of matched-field processing," *J. Acoust. Soc. Amer.*, vol. 128, pp. 2431–2431, 2010.
[45] M. Wazenski, "Reduction in computational requirements with matched field processing," The Journal of the Acoustical Societyof America, vol. 104, no. 3 Supplement, pp. 1782–1782, 1998.
[46] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *IEEE Access*, vol. 8, pp. 42200–42216, 2020.
[47] H. Niu, E. Ozanich, and P. Gerstoft, "Ship localization in Santa Barbara channel using machine learning classifiers," *J. Acoust. Soc. Amer.*, vol. 142, no. 5, pp. EL455–EL460, 2017.
[48] H. Niu, P. Gerstoft, and E. Ozanich, "Source localization in an ocean waveguide using supervised machine learning," *J. Acoust. Soc. Amer.*, vol. 142, pp. 1176–1188, Jan. 2017.
[49] Y. Wang and H. Peng, "Underwater acoustic source localization using generalized regression neural network," *J. Acoust. Soc. Amer.*, vol. 143, no. 4, pp. 2321–2331, 2018.
[50] Z. Huang, J. Xu, Z. Gong, H. Wang, and Y. Yan, "Source localization using deep neural networks in a shallow water environment," *J. Acoust. Soc. Amer.*, vol. 143, no. 5, pp. 2922–2932, 2018.
[51] R. Chen and H. Schmidt, "Model-based convolutional neural network approach to underwater source-range estimation," *J. Acoust. Soc. Amer.*, vol. 149, pp. 405–420, Jan. 2021.
[52] J. Yangzhou, Z. Ma, and X. Huang, "A deep neural network approach to acoustic source localization in a shallow water tank experiment," *J. Acoust. Soc. Amer.*, vol. 146, no. 6, pp. 4802–4811, 2019.
[53] D. Stowell, "Computational bioacoustics with deep learning: A review and roadmap," *PeerJ*, vol. 10, 2022, Art. no. e13152.
[54] Z. Bai, Y. Zhu, X. Li, H. Wang, X. Yang, and W. Kong, "Underwater acoustic signal recognition with small transfer learning-based training samples," in *Proc. OES China Ocean Acoust.*, 2024, pp. 1–5.
[55] L. C. F. Domingos, P. E. Santos, P. S. M. Skelton, R. S. A. Brinkworth, and K. Sammut, "An investigation of preprocessing filters and deep learning methods for vessel type classification with underwater acoustic data," *IEEE Access*, vol. 10, pp. 117582–117596, 2022.
[56] B. Ghani, T. Denton, S. Kahl, and H. Klinck, "Global birdsong embeddings enable superior transfer learning for bioacoustic classification," *Sci. Rep.*, vol. 13, no. 1, 2023, Art. no. 22876.
[57] Z. Li et al., "Oceanship: A large-scale dataset for underwater audio target recognition," in *Advanced Intelligent Computing Technology and Applications*, D.-S. Huang, C. Zhang, and W. Chen, Eds., Singapore: Springer, 2024, pp. 475–486.
[58] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP learning audio concepts from natural language supervision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.
[59] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," in ICASSP2024-2024 IEEE International Conference on Acoustics, Speechand Signal Processing (ICASSP), pp. 336–340, IEEE, 2024.
[60] V. S. Kather, B. Ghani, and D. Stowell, "Clustering and novel class recognition: Evaluating bioacoustic deep learning feature extractors", presented at the Forum acusticum Euronosie 2025 : 11th Convention of the European Acoustics Association, Málaga, 2025
[61] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.
[62] D. Robinson, A. Robinson, and L. Akrapongpisak, "Transferable models for bioacoustics with human language supervision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 1316–1320.
[63] C. Parcerisas, D. Botteldooren, P. Devos, and E. Debusschere, "Broadband acoustic network dataset," 2021.
[64] A. Calonge et al., "Multi-purpose seabed moorings: The missing link for sustained long-term observations in dynamic shallow seas," Manuscript submitted for publication, 2024.
[65] S. Degraer et al. *De Macrobenthosatlas Van Het Belgisch Deel Van de Noordzee*. Brussel: Federaal Wetenschapsbeleid, 2006.
[66] R. J. Urick, *Principles of Underwater Sound*, 2nd ed. New York, NY, USA: McGraw-Hill, 1975.
[67] F. B. Jensen, W. A. Kuperman, M. B. Porter, H. Schmidt, and A. Tolstoy, *Computational Ocean Acoustics*, vol. 2011. Berlin, Germany: Springer, 2011.
[68] J. H. Simpson and J. Sharples, *Introduction to the Physical and Biological Oceanography of Shelf Seas*. New York, NY, USA: Cambridge Univ. Press, 2012.
[69] AISHub, "AIS data exchange," 2018. Accessed: Aug. 2, 2018. [Online]. Available: http://www.aishub.net
[70] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2020
[71] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica, "Tune: A research platform for distributed model selection and training," 2018
[72] M. Irfan, Z. Jiangbin, S. Ali, M. Iqbal, Z. Masood, and U. Hamid, "DeepShip: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification," *Expert Syst. Appl.*, vol. 183, 2021, Art. no. 115270.

Decrop Wout received the B.S. and M.S. degrees bioscience engineering from Ghent University, Ghent, Belgium, in 2020 and 2022, respectively.

His research focuses on machine learning related to marine science.

Deneudt Klaas received the B.S. degree in biology/biological sciences in 1998 and the M.S. degree in zoology/animal biology in 2000, both from Ghent University, Ghent, Belgium.

His research interests include marine biodiversity science, ocean technology, data science, and scientific project management.

Parcerisas Clea received the B.S. degree in Industrial Engineering in 2016 and the M.S. degree in Industrial Engineering in 2018, both from the Polytechnic University of Catalonia, Barcelona, Spain, and the Ph.D. degree from Ghent University, Ghent, Belgium, in 2024.

Her research focuses on developing open-source tools for marine soundscape analysis.

Schall Elena received the B.S. degree in biology from the University of Bremen, Bremen, Germany, in 2014, and the M.S. degree in marine biology and biological oceanography from the University of Groningen, Groningen, the Netherlands, in 2016, with a focus on bioacoustics, biodiversity, ecology, and sustainability, and the Ph.D. degree from the University of Oldenburg, Oldenburg, Germany, in 2021.

Her research focuses on marine bioacoustics.

Debusschere Elisabeth received the B.S. degree in 2009, the M.S. degree in Marine Sciences in 2011, and the Ph.D. degree in Marine Sciences in 2016, all from Ghent University, Ghent, Belgium. The doctoral research focused on the effects of high-intensity impulsive sound on young European sea bass (Dicentrarchus labrax), with particular attention to pile driving during offshore wind farm construction.

Her research interests include bioacoustics and specialized offshore experiments/studies using passive and active acoustics.