IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 October 2022, accepted 1 November 2022, date of publication 7 November 2022, date of current version 11 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3220265

RESEARCH ARTICLE

An Investigation of Preprocessing Filters and Deep Learning Methods for Vessel Type Classification With Underwater Acoustic Data

LUCAS C. F. DOMINGOS[1,2], **PAULO E. SANTOS**[3], **PHILLIP S. M. SKELTON**[3], **(Member, IEEE), RUSSELL S. A. BRINKWORTH**[3], **AND KARL SAMMUT**[3], **(Senior Member, IEEE)**

[1]Department of Electrical and Electronics Engineering, Centro Universitário FEI, São Bernardo do Campo 09850-901, Brazil
[2]Department of Computer Vision, Instituto de Pesquisas Eldorado, Campinas 13083-898, Brazil
[3]Centre for Defence Engineering Research and Training, College of Science and Engineering, Flinders University, Tonsley, SA 5042, Australia

Corresponding author: Lucas C. F. Domingos (ldomingos@fei.edu.br)

ABSTRACT The illegal exploitation of protected marine environments has consistently threatened the biodiversity and economic development of coastal regions. Extensive monitoring in these – often remote – areas is challenging. Machine learning methods are useful in object detection and classification tasks and have the potential to underpin techniques for the development of robust monitoring systems to overcome this problem. However, development is hindered due to the limited number of publicly available labelled and curated datasets. Furthermore, there are relatively few open-source state-of-the-art methods to be used for evaluation. This paper presents an investigation of automated classification methods using underwater acoustic signals to infer the presence and type of vessels navigating in coastal regions. Various combinations of deep convolutional neural network architectures, and preprocessing filter layers, were evaluated using a new dataset based on a subset of the extensive open-source Ocean Networks Canada hydrophone data. Tests were conducted in which VGGNet and ResNet networks were applied to classify the input data. The data was preprocessed using either Constant Q Transform (CQT), Gammatone, Mel spectrogram, or a combination of these filters. With over 97% accuracy, using all three preprocessing representations simultaneously yielded the most reliable result. However, high accuracies of 94.95% were achieved using CQT as the preprocessing filter for a ResNet-based convolutional neural network, providing a trade-off between model complexity and accuracy; a result that is more than 10% higher than previously reported approaches. This more accurate classifier for underwater acoustics could be used as a reliable autonomous monitoring system in maritime environments.

INDEX TERMS Deep learning, hydrophones, marine environment, ship type, sound.

I. INTRODUCTION

Illegal fishing represents a serious problem for society in general, affecting not only the marine life through destructive trawling but also the local economy of coastal areas, which depends economically on this ecosystem for subsistence. Therefore, the detection and classification of illegal vessels situated in law-protected areas represent a poignant need for the surveillance and protection of the coastal ecosystem. Nowadays, there is a large number of applications that involve maritime classification tasks, such as the identification of underwater archaeological remains [1], the inspection of underwater structures for the offshore industry [2], [3], the surveillance of shorelines [4], the identification of vessels [5], as well as applications in environmental sciences, like counting and classifying the various marine species for biological research [6]. Also worth mentioning are studies relating the acoustic signals in the sea to environmental pollution, affecting not only the marine life [7], [8], [9], but also the human activities in port areas [10], [11], [12]. In this context, the

The associate editor coordinating the review of this manuscript and approving it for publication was Yougan Chen.

identification of vessels from acoustic data is selected as the domain of interest for this paper.

Some technologies, such as the Automatic Identification System (AIS), which contain Global Positioning System (GPS) data, and satellite images, can be applied in the surveillance of the marine environment. However, these approaches have limitations. For instance, the high costs and maintaining accurate instrument calibration are still challenges for satellite imagery [13]. GPS signals, on the other hand, can be masked or defrauded to limit the system capabilities or even to hide illegal activities. In contrast, the acoustic signals emitted by vessels captured using hydrophones provide a low-cost and fraud-resistant data source to be used in surveillance tasks, as acoustic signals can be difficult to omit or mask. Efforts in this area can also be the initial step towards determining how such signals can be analysed, becoming the gateway to understanding of how acoustic environmental pollution is affecting marine life.

As the classification of underwater acoustic signals gained importance, this task became unfeasible to be solved by traditional (time-frequency) methods, which were primarily conducted by humans operators, due to the complexity of the data. Time-frequency representations, such as those based on the Fourier transform or on temporal data segments [14], can be applied in different forms, such as a linear scale (e.g., short time Fourier transforms) or a logarithmic scale (e.g., Mel filter banks). Both strategies produce a two-dimensional time-frequency representation of the signal, which can be used to analyse the features of the sound. However, the underwater acoustic signal is a mixture of environmental, biological, and human-generated sounds. Therefore, it has a low signal-to-noise ratio (SNR) and a high degree of variability for the same source [15], [16], raising the difficulty of the recognition task. In order to cope with this issue, recent studies, mostly based on the application of Deep Learning (DL) methods [17], [18], [19], have shown promise in automatic data classification tasks.

Many DL methods for object detection and classification have been successfully developed in the past few years for computer vision applications [20]. The use-case presented in this paper allows for the application of these methods to other data domains, which could inherit from the solutions developed in the visual domain. In this context, as the time-frequency representations are two-dimensional representations of the acoustic signal, an opportunity arises to apply the DL strategies, originally developed for computer vision, to acoustic analysis. Numerous DL solutions for the acoustic domain are now based on Convolutional Neural Networks (CNNs) [21], [22]. Although they can be applied to raw audio, they are often applied to two-dimensional audio representations, such as spectrograms. The most recent studies have used VGGNet [23] or ResNet [24] models as base algorithms for this development, owing to the models' high accuracy in complex classification tasks. Despite the use of time-frequency representations as inputs to CNNs being an interesting approach, its future development depends on the representation used as input and the model which will receive it. Also, as the sound generated by sources in the underwater domain is dependent on the environment, it is of extreme importance that not only the type of two-dimensional representation matches the problem, but also its parameters must be optimised for the task.

This paper describes an investigation into the automated classification of four distinct classes of vessels in marine environments using DL. Single channel underwater acoustic signals obtained by research-grade hydrophones were used as input, and the impact of the application of distinct preprocessing methods on the DL classification task was explored. Classification results from two key state-of-the-art DL methods, VGGNet and ResNet, were compared. The impact of applying three distinct preprocessing filters, Mel Spectrogram, Constant Q Transform (CQT), and the Gammatone-like spectrograms (or just Gammatone), was also evaluated, as was the impact of a combination of these three filters into a three-channel representation. The complete pipeline was trained and tested on three scenarios characterised by the distance between the objects of interest and the hydrophone.

In order to better monitor and protect marine environments, there is a need for an autonomous monitoring system that can generate an alert whenever a particular class of vessel is detected in an area. Towards that end, the main contributions of this paper can be summarised as follows:

- Creation of an open-source pipeline for the classification of vessels from underwater acoustic signals using Machine Learning[1];
- Comparison of the Adam and Stochastic Gradient Descent (SGD) optimisers for spectrogram analysis;
- Evaluation of two different neural network architectures for acoustic classification (VGGNet and ResNet);
- Comparison of three different preprocessing filters (CQT, Mel Spectrogram, and Gammatone);
- Investigation into combining CQT, Mel Spectrogram, and Gammatone representations into a three-channel signal, generating a higher dimensional input signal to the network;
- Analysis of the relation between the distance of the object of interest to the hydrophone and the accuracy of classification methods;
- Presentation of a new open-source curated dataset containing underwater acoustic signals classified into different scenarios based on the distance from the vessel to the sensor.[23]

II. RELATED WORK

The task of underwater acoustic target classification is challenging due to the complex nature of the sound produced by vessels [25]. Usually, this sound is produced by the set of mechanical components in the vessel's propulsion system,

[1] https://github.com/lucascesarfd/underwater_snd
[2] https://github.com/lucascesarfd/onc_dataset
[3] http://ieee-dataport.org/9778

such as its engine, as well as by hydrodynamic interactions of the propeller. The former typically produces a broadband continuous spectrum, while the latter generates narrow band components whose spectrum consists of power at discrete frequencies [26]. As there are different types of vessels, in diverse states of upkeep, the sound produced by them is fundamentally distinct from one another, depending on the vessel's speed, the state of its mechanical parts, and the hydrodynamics of its design. Also, additional complexity exists due to the background sound produced by the region and the complexity of sound propagation in shallow waters, which causes multi-path reflections [27]. Environmental conditions, such as temperature, depth, salinity, pressure, and even precipitation, can directly influence how the signal travels from emitter to receiver [28]. In this context, some classical signal processing methods, such as Cepstral analysis [14], can improve the quality of the processed sound by reducing the effects of the reflections interference and scattering losses, but only if applied on signals for short ranges with a high SNR [29].

Early developments in the analysis and classification of underwater acoustic signals focused on time-frequency analyses, such as the use of Fourier transforms [14]. However, recent state-of-the-art methods are largely based on the application of deep-learning algorithms to solve similar tasks [17], [18], [19]. Advances in machine learning techniques mean CNNs are now being considered for underwater acoustic classification applications [25]. Consistent with this trend, the trade-off between the accuracy and model size of various CNN models for mine-like object detection from side-scan sonar images was investigated [30]. The comparative results reported suggest that deeper models (i.e., models with multiple layers) achieved less than 1% of accuracy improvement when compared with shallow models, at the cost of a 17x increase of computational requirements. This proved that smaller models can have a beneficial trade-off between processing time and accuracy. Similarly, the impact caused by distinct network topologies on the problem of underwater acoustic target classification is an important issue that has been recently considered [31]. A properly tuned model is capable of outperforming recent DL methods, such as a CNN-extreme learning machine [32], ResNet18 [24], and SqueezeNet [33]. The strong results presented in [30], [31] suggest that the search for the most suitable network topology, and the optimisation of its parameters, are essential tasks that should be considered in the development of any CNN-based classification system.

Although CNNs can be applied directly to the audio signal [31], [34], [35], acoustic filters are frequently used as preprocessing layers to improve the quality of the resulting audio representations [25]. Therefore, not only should the CNN parameters be investigated, but also which filters and features best contribute to the development of effective DL methods for the underwater acoustic domain. To this end, recent work has investigated the effect of various preprocessing methods on the original audio signal, including magnitude Short-time Fourier transform (STFT) spectrum, complex-valued STFT spectrum, *Mel-log* spectrum, and Mel-frequency cepstral coefficients (MFCCs), as inputs to real-valued and complex-valued ResNet and DenseNet CNNs [36]. The results obtained using preprocessing filters were considerably better than the baseline approach where a CNN was directly applied to classify the raw audio signal. Similarly, *Mel-log* spectrograms, delta, and delta-delta features were also used as acoustic filters in a ship detection task using a CNN, where high accuracy in the detection and localisation of vessels was reported [37]. Other studies in the literature also successfully applied filters to the DL inputs, showing a consistent improvement in underwater audio classification tasks [38], [39], [40], [41]. This strongly suggests that, although CNNs are capable of learning distinct filters in their convolutional layers, there may be insufficient training time or data for the network to converge on the best solution. Therefore, superior results, in addition to smaller networks and reduced training time, are obtained with the use of appropriate preprocessing filters in the classification pipeline.

Analogous to the research described in the present paper, recent work has been driven by the advantage of using preprocessing filters to extract optimised features from the audio, also using stacks of multiple filters as inputs to the CNN models [42], [43]. The rationale behind this approach is to take advantage of the strengths of each method, feeding the network with different representations of the sound. For instance, a joint learning framework was developed to address the underwater acoustic target classification using MFCC, CQT, Gammatone, and Log-Mel feature extraction methods to feed a CNN-based architecture [42]. The comparison of the results obtained with individual approaches and their combination showed that superior outcomes could be achieved with the latter. Another relevant work used a fusion of the Mel-spectrogram, MFCC, chromatogram, spectral contrast, and Tonnetz filters, resulting in a one-dimensional representation, to improve the performance of a CNN model for the classification of underwater acoustic signals [43].

A summary of the classification methods cited in this section is shown in Table 1 which relates, for each method, the preprocessing applied (if any), the model architecture, the dataset used, the best reported accuracy, and the main contributions. A more complete up-to-date survey of this field can be found elsewhere [25].

There are a number of recent papers concentrating on the classification of underwater acoustic data. However, there is a pertinent need for a complete investigation into the application of DL algorithms for the task, an investigation that considers the optimisation of the DL model parameters, and the comparison between different preprocessing filters. Additionally, the impact of environmental variables on vessel classification is virtually non-existent in the related literature. These issues are taken into account in the research reported in this paper.

TABLE 1. An overview of the DL methods for classification of underwater acoustic signals.

Ref.	Preprocessing	Network	Dataset	Accuracy	Main contributions
[30]	Not Applicable	CNN	Mine-Like Objects images[b]	98.90%	The analysis of the trade-off between accuracy and model size of CNN models.
[31]	Not Applicable	UATC-DenseNet	Real-world passive sonar data[b]	98.85%	The analysis of the use of a number of convolutional blocks and layers, and different layer configurations and input features.
[34]	Not Applicable	Auditory Deep CNN	Ocean Network Canada signals[a]	81.96%	The use of a bank of multi-scale deep convolutional filters as a first processing stage, making possible the creation of an end-to-end NN.
[35]	Not Applicable	MSRDN	Ocean Network Canada signals[a]	83.15%	The development of a deep residual network using the soft-thresholding proposed in [44] and the convolution kernel proposed in [45].
[36]	STFT; Mel-log; MFCCs	ResNet and DenseNet	ShipsEar [27]	97.49%	The classification of synthetic mixed multitarget signals using CNNs.
[37]	Mel-log; delta; delta-delta	ResNext101	Shallow water Data[b]	85.00%	Direction-of-arrival prediction based on acoustic signals.
[38]	MFCC	Fully-Connected	Sonar dataset[b]	97.12%	The accuracy comparison of meta-heuristic algorithms and the use of a fully connected NN.
[39]	GFCC; MFCC	Fully-Connected	Six class dataset[b]	94.3%	A combination of MFCC and GFCC was used as feature extraction showing time performance improvement.
[40]	Cochlea model	Auditory Inspired CNN	Ocean Network Canada signals[a]	87.2%	The use of Gabor filter layers inspired on the a Cochlea model.
[41]	MFCC; CQT; GFCC; Mel-log; Cepstros; Wavelets	SCAE	DeepShip [41]	77.53%	The proposal of a new open source dataset and the comparison of various preprocessing strategies.
[42]	MFCC; CQT; Gammatone; Mel-log	CNN+DNN	Hydrophone datasets[b]	89.9%	The generation of a pipeline with the combination of the preprocessed features for acoustic target classification.
[43]	Mel-log; MFCC; chromatogram; spectral contrast; Tonnetz	CNN+LSTM	ShipsEar [27]	92.17%	The proposal of a hybrid neural network composed of CNN and LSTM, having a combination of the preprocessing strategies as input.

[a] Dataset composed of public data with nonpublic preprocessing techniques. [b] Dataset is proprietary and unavailable for reproduction.

III. DATASET

The data used in this work consisted of signals obtained from the Ocean Network Canada initiative,[4] captured during the deployment at the Strait of Georgia, Canada, from June 24 to November 3, 2017, representing typical pre-pandemic operations during the Summer and Autumn seasons. An icListen AF Hydrophone, located 147 meters below sea level, was used to obtain the acoustic signals. In addition, the positional information about the vessels was obtained using Automatic Identification System (AIS) data.

The first part of the annotation process focused on the translation and filtering of the AIS signals. These signals contained position, identification, speed, course, and other information about active maritime traffic. Some of the information contained in AIS data is not necessary for vessel classification tasks. Only messages related to *position report*, as well as *static and voyage related data*, were used. Duplicated messages, and messages that did not have positional arguments, were filtered out. The vessel's class was then inferred from the type of ship and cargo fields of the AIS messages, generating four categories: Tug, Passengership, Cargo, and Tanker. Using the positional coordinates, a geodesic distance calculation was performed to estimate the distance from the vessels in the area of interest to the hydrophone. As the update rate of AIS data is related to the vessel's size, cargo, velocity, etc., there are intervals where gaps appear in the AIS reports. It was deemed safe to linearly interpolate between these sparse data points to provide greater resolution of the vessel's distance to the hydrophone.

Different subsets of data were generated from the original data considering the distance from the vessel to the hydrophone picking up the vessel's sound. These subsets, or *scenarios*, were created considering *inclusion* and *exclusion* radii. The *inclusion radius* was defined as the radial distance when only one vessel was present at a specific moment, whereas the *exclusion radius* was the region in which there was no vessel within a fixed radial distance. To isolate a single vessel as much as possible, scenarios were generated as illustrated in Figure 1, where a vessel was within the inclusion radius while no other vessels were within the wider exclusion radius.

These scenarios facilitated the analysis of the classification accuracy concerning the distance between the object of interest and the sensor. As the problem of vessel classification using machine learning depends on the quality of the input data, it was expected that the sound emitted by distant sources would have a lower SNR and, thus, lower classification accuracy. The three scenarios considered in this work were created based on the available data: the first had an inclusion radius of 2 km and an exclusion radius of 3 km; the second had 3 km and 4 km as the inclusion and exclusion radii; and the third had radii of 4 km and 6 km. Table 2 summarises the

[4] https://www.oceannetworks.ca/

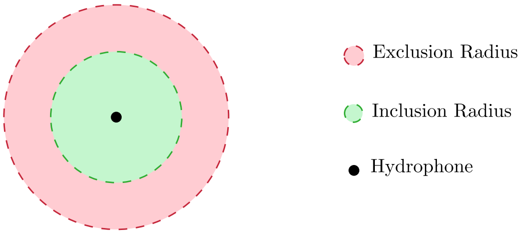

FIGURE 1. Diagram representing a scenario. Data was isolated where only a single vessel was within the inclusion zone while no other vessels were within the exclusion zone. This ensured a more reliable acoustic signature without interference from other vessels.

TABLE 2. The different scenarios considered for the dataset generation.

Scenario	Inclusion Radius (km)	Exclusion Radius (km)
First	2	4
Second	3	5
Third	4	6

scenario descriptions. The background class for each scenario was then generated based on the absence of vessels within the inclusion and exclusion radii combined. The final stage of the dataset formulation was the combination of every AIS instance, defined as the period that matched a specific scenario, with the acoustic data.

This automatic annotation procedure could generate mislabelling in the dataset, therefore the results were further analysed and filtered to avoid this issue. A data cleaning process was performed, noting that the variation of the time domain amplitude of a vessel was greater than that of the background sound. First, a median filter (med()) was used to de-noise the original signal ($a(t)$, where t represents time). The resulting audio was subtracted from the original signal (Equation (1)) producing an audio signal ($g(t)$) free of DC offset. The standard deviation of $g(t)$ (represented as ϱ, as shown in Equation (2)) was used to generate a scalar value of the amplitude variation for each 1-second signal segment. The mean and standard deviation of the ϱ values were obtained from the vessel and background sounds, respectively ($\mu_{\varrho-\text{vessel}}, \sigma_{\varrho-\text{vessel}}$) and ($\mu_{\varrho-\text{back}}, \sigma_{\varrho-\text{back}}$). As expected, this analysis showed that the tagged vessel data delivered higher variation (ϱ) when compared with the background audio (i.e., $\mu_{\varrho-\text{vessel}} > \mu_{\varrho-\text{back}}$).

Individual segments tagged to contain a vessel, but with a ϱ value that was less than the overall standard deviation of the background increased by the mean ($\mu_{\varrho-\text{back}} + \sigma_{\varrho-\text{back}}$), were removed from the collection as they represented potentially mislabelled signals in the dataset.

$$g(t) = a(t) - \text{med}\,(a(t))\,. \tag{1}$$

$$\varrho = \sqrt{\frac{1}{N-1}\sum_{t=1}^{N}(g(t)-\overline{g})^2}\,. \tag{2}$$

In the equations above $\overline{g}$ represents the mean value of the signal with the median removed, and N is the number of audio recordings.

The final version of the dataset was composed of the three scenarios, summarised in Table 2, each one with audio files saved as raw, uncompressed, *WAV* files. Also, a Comma-Separated Value (CSV) file was generated with the annotation of the vessel type for each scenario. In this work, each audio file was divided into 1-second segments, which were used as inputs to the preprocessing filters. The complete data was divided into Training, Validation, and Test subsets, following the proportion of 85%, 10%, and 5%, respectively. As there was a class imbalance problem, only the Training subset was balanced using an oversampling strategy. An oversampling factor, Equation (3), was used to define the size of each class based on the class with the smallest length.

$$\text{factor} = \min\left(2, \frac{\text{L}}{\text{l}}\right). \tag{3}$$

In Equation (3), L represents the size in seconds of the class with the most data points, and l represents the size in seconds of the class with the fewest data points.

For each category, the audios were selected randomly to compose the dataset. If the size of the class did not reach the minimum size defined by factor (Equation (3)), the selection started again, gathering repeated audios until the desired length was achieved. However, uniqueness of each recording was enforced. Table 3 contains the duration of each subset for the dataset scenarios.

The next section introduces the concepts of each of the preprocessing filters used in this work, and the preprocessing pipeline is described in Section V-A.

IV. PREPROCESSING FILTERS

Two-dimensional representations of audio files can take the form of spectrograms, which represent the frequency distribution of the original signal over time. One of the possible ways to formulate such representations is using a window function applied along the length of the one-dimensional signal, dividing it into smaller (fixed) chunks. These chunks are then processed, generating the information about the frequencies in that period. Therefore, the horizontal axis of the resulting two-dimensional representation is highly dependent on the chosen initial time window. The vertical axis represents the frequency distribution of the sound and it is commonly represented either linearly, or logarithmically. For the problem of sound classification, the logarithm representation of the frequency is preferred over the linear representation, following the analogy with the human auditory system [46]. In this context, the present work focused on the application of three common methods for spectrogram generation based on non-linear frequency scales: Mel Spectrograms, Constant Q Transform (CQT), and Gammatone Spectrograms (as described below). These methods were used here to enhance features of the original signal, and their output served as input to the CNN models investigated in this work.

A. MEL SPECTROGRAMS

The Mel spectrogram is a representation of the short-term sound power spectrum. Mel's scale is empirically based

TABLE 3. The length, in seconds, of each dataset class for each of the three scenarios' subsets.

Scenario	Subset	Tug (s)	Passengership (s)	Cargo (s)	Tanker (s)	Background (s)	TOTAL (s)
First	Training	7302	7302	7302	7302	7302	**36510**
	Validation	902	74	1257	165	1362	**3760**
	Test	445	35	627	94	679	**1880**
Second	Training	13010	13010	13010	13010	13010	**65050**
	Validation	1253	1093	1919	180	2335	**6780**
	Test	689	551	931	95	1124	**3390**
Third	Training	10150	10150	10150	10150	10150	**50750**
	Validation	1249	647	1217	107	2010	**5230**
	Test	632	318	681	48	936	**2615**

on the way humans perceive sound [47]. The formulation of this scale consisted of submitting observers to different frequencies of sounds, while recording their perception and sensitivity to the stimulus. There are different mathematical formulations for the conversion between the frequency f in Hertz to m in Mels, such as that represented in Equation (4) [48].

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right). \tag{4}$$

Mel Spectrograms are commonly used in speech recognition analysis and music processing, where human perception is extremely relevant [49].

B. CQT

The Constant Q Transform (CQT) [50] uses a constant base scale (Q) to create a representation. This improves the resolution between frequencies of interest, while providing the means to solve the problem of fundamental frequency identification. In contrast with the classic Fourier transform, CQT is a bank of geometrically-spaced filters in which, for the k-th filter, the central frequencies are evaluated with Equation (5),

$$f_k = f_0 \times 2^{\frac{k}{b}} \tag{5}$$

where b represents the number of filters per octave. Thus, the relation between the distance of two adjacent filters is given by Equation (6),

$$\Delta_k = f_{k+1} - f_k = f_k(2^{\frac{1}{b}} - 1). \tag{6}$$

The quality factor Q (or constant Q) is defined as the ratio of frequency to resolution, as stated by Equation (7).

$$Q = \frac{f_k}{\Delta_k} = \left(2^{\frac{1}{b}} - 1\right)^{-1}. \tag{7}$$

The correct tuning of the quality factor Q can supply the needed information for the acoustic analysis, with resolution to distinguish adjacent musical notes, where a sound with harmonic frequency components will produce a constant pattern in the log frequency domain [50]. This representation also increases time resolution towards higher frequencies, resembling the human auditory system, while emphasising lower frequencies.

C. GAMMATONE SPECTROGRAMS

The gammatone filter was first defined as a filter bank capable of representing the shape of the impulse response function of the human auditory system [51]. A gammatone function can be obtained with the mathematical formulation shown in Equation (8),

$$g(t) = t^{n-1} \exp\left(-2\pi b_i t\right) \cos\left(2\pi f_i + \alpha_i\right) \tag{8}$$

where n is the filter order, i is the filter order number (ranging from 1 to the total number of filters), b is a bandwidth parameter, f is the filter centre frequency, and α is the phase of the impulse response.

The function defined on Equation (8) was used by [51] to summarise the *RevCor*, a representation of the correlation between a sound stimulus on the human ear and the response of a primary auditory fibre [52]. The first term of Equation (8) represents a gamma function, and the second term represents the tone of the stimulus. This representation has an amplitude characteristic that can be used to predict the human auditory response. It also has a minimum-phase characteristic, which is a preferred feature for an auditory filter bank [52]. Gammatone filter-banks facilitate the representation of the signal's time domain response, as gamma filters are broader on lower frequencies and narrow on higher ones, emphasising the lower spectrum.

The raw signal and its processed representations (CQT, Gammatone, and Mel spectrograms) are shown in Figure 2.

In this work, the three preprocessing methods described herein were used on the underwater acoustic signals to extract relevant features from the acoustic signal, generating the two-dimensional representations used as inputs to CNNs. Section V describes the successive stages of the implementation of this work.

V. DEVELOPMENT STEPS

This section describes the development steps performed to address the classification of underwater acoustic signals using preprocessing filters and CNNs, detailing the procedures and experimental setup.

A. PREPROCESSING FILTERS

Each entry in the original dataset was divided into one-second segments. Segments smaller than one second were padded

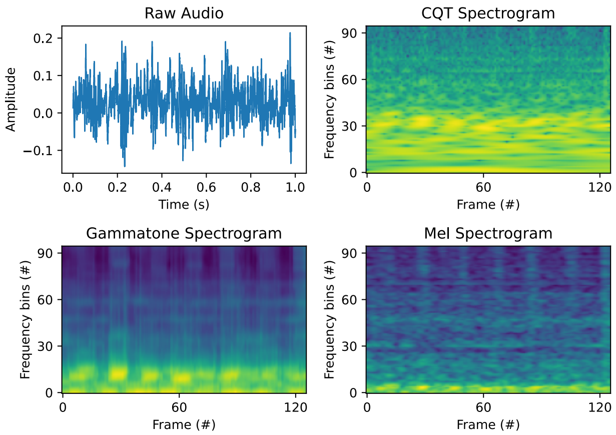

FIGURE 2. An example of audio used on this experiment. The image shows the raw audio signal in the time domain and its three preprocessed versions: the CQT, Gammatone, and Mel spectrograms.

TABLE 4. Description of the *Version 1* and *Version 2* sets of parameters used on the preprocessing generation.

Parameter	Version 1			Version 2		
	Mel	Gamma	CQT	Mel	Gamma	CQT
Hop Length	512	512	512	256	256	256
Window	hann	hann	hann	hann	hann	hann
No. of bins	64	64	64	95	95	95
Min. Freq.	0	20	32.7	18	18	18
Max. Freq.	Auto	Auto	Auto	4186	4186	4186
Bins p. Oct.	–	–	12	–	–	12

with zeros. After that, the three proposed preprocessing methods, CQT, Gammatone, and Mel spectrogram, were applied to each audio file. Initially, to establish a baseline for the audio classification based on standard values found in the literature, the window chosen to generate the spectrograms had 1024 samples, with a hop length of 512, resulting in 64 frequency bins over 63-time intervals per data segment. This resulted in each method producing 64 × 63 element images. This set of parameters is referred to as *Version 1*.

A second set of parameters (*Version 2*) was obtained by means of an optimisation process. The majority of the power in the underwater acoustic signal was predominantly focused on the low-frequency band, below 3 kHz. To maintain a safe range above the maximum frequency, spectrograms were generated from 18 Hz (the minimum acceptable for the CQT representation for 1 second audios) to the frequency of 4186 Hz (C8 note and ≈ 1 kHz above the 3 kHz experimentally observed maximum value). Using a hop length of 256, which represented half of the value proposed on *Version 1*, the resulting representation (*Version 2*) had a size of 95 × 126. The values for the two versions of parameter sets are summarised in Table 4, where the values not related to a particular representation are marked with "–".

Inspired by the large variety of machine learning methods applied to three-channel images, such as colour images in, e.g., RGB (Red-Green-Blue) or HSV (Hue-Saturation-Value) colour spaces, the three preprocessing methods cited previously were combined into a single three-dimensional representation, which was then used as the input to the CNNs. This was motivated by providing the Neural Network with more complete representations, aiming to take advantage of the strengths of each of the preprocessing methods. Combining CQT, Gammatone, and Mel spectra resulted in data samples with dimensions of 64 × 63 × 3 for *Version 1* and 95 × 128 × 3 for *Version 2*. This representation is called *Complete*.

The next section presents the Deep Learning methods used to classify the representations obtained with the preprocessing filters described above.

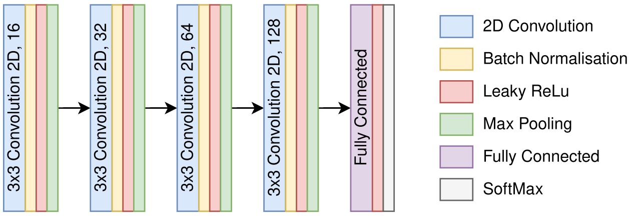

FIGURE 3. The VGG-based model architecture.

B. DEEP LEARNING MODEL DESIGN

As mentioned previously, this work used two distinct CNN models: VGGNet [23] and ResNet [24].

VGG-based methods can perceive granular spatial relations on images due their use of a 3 × 3 kernel size, the smallest possible size to capture the four cardinal directions (up, down, left, and right). This reduced kernel size also produces a good trade-off between classification accuracy and hyperparameter complexity. The implementation of this model in the present work contained two main modifications from the original VGGNet: 1) A Leaky ReLu was used as the activation function instead of a normal ReLu; and 2) A Batch Normalisation layer was added. Both changes aimed at reducing overfitting. The resulting model architecture is shown in Figure 3 and was composed of four feature extracting convolutional layers. The signal then passed through a Batch Normalisation and a Leaky ReLu activation layer associated with a Max Pooling layer, which resized the image by a factor of 2. Lastly, the classification weights were delivered by a fully connected layer.

The universal approximation theorem [53] states that a deep enough neural network is capable of approximating any complex function, although the vanishing gradient and the accuracy degradation problems become problematic as more layers are added to the models. ResNet addresses these issues by introducing the identity shortcut connection, bypassing one or more layers in a forward pass, defining Residual Blocks. These blocks facilitate the learning ability of the intermediary layers, reducing the vanishing gradient problem, and penalising the ones that could potentially degrade accuracy [24]. This work used ResNet18, with modifications

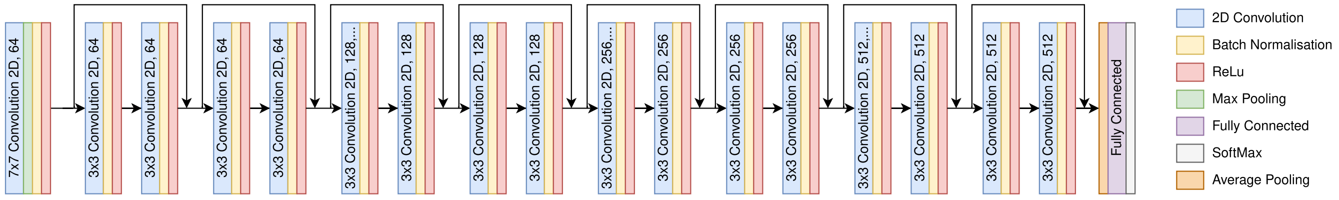

FIGURE 4. The ResNet18 model architecture.

to the input layer to match the preprocessed images. The ResNet18 architecture, shown in Figure 4, is composed of a convolutional layer followed by 8 Residual Blocks, each one formed by two other convolutional layers. As usual, the classification weights are generated by the final fully connected layers.

After the model definition, an optimiser had to be chosen, aiming to minimise the error in the training procedure. This work investigated the use of the Stochastic Gradient Descent (SGD) [54] and the Adam [55] optimisers. SGD is an iterative method that starts randomly and seeks the minimum value in the input function. It is the most common optimiser used in the literature. Adam, on the other hand, is an extension of SGD based on the combination of the Adaptive Gradient Algorithm (AdaGrad) and the Root Mean Square Propagation (RMSP). The use of SGD and Adam was compared in the tests executed in this work, where a learning rate of 0.001, decreasing exponentially, and a gamma value of 0.95, over 40 epochs, was used.

The resulting architecture was then composed of the preprocessed acoustic signals, produced by the four strategies described in Section V-A, applied to both CNN models (VGG-based and ResNet18). Each model was trained with batches of 8 images over 40 epochs using the *Categorical Cross-Entropy* loss function. The block diagram of the complete pipeline is shown in Figure 5. The preprocessing block is the representation of CQT, Gammatone, Mel, or the combination of all three preprocessing filters (Complete). The model block represents VGGNet or ResNet18.

VI. RESULTS

All training sessions were executed for the three chosen preprocessing filters in addition to the complete representation. Results are reported using micro-average accuracy, which measures the correct classifications of the classes combined. This provides a global overview of the model performance in realistic scenarios, i.e., with real observed class imbalances. Three additional metrics were used, providing complementary information: Precision, which represents the rate of correct positive predictions over the total positive predictions; Recall, which measures the rate of correct positive predictions over the real positive instances; and F1-score, which represents the weighted harmonic mean between precision and recall. These three additional metrics were evaluated using *macro-averaging*, which evaluates the classes separately before taking the average, aiming to obtain a balanced evaluation across the classes. The metrics were obtained using Equations (9), (10), (11), and (12), where K represents the number of classes in the dataset, TP stands for True Positive, TN for True Negative, FP for False Positive, and FN for False Negative.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{9}$$

$$\text{Precision} = \frac{1}{K} \times \sum_{k=1}^{K} \frac{TP_k}{TP_k + FP_k}. \tag{10}$$

$$\text{Recall} = \frac{1}{K} \times \sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k}. \tag{11}$$

$$\text{F1-score} = \frac{2}{K} \times \sum_{k=1}^{K} \frac{\text{Precision}_k \times \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}. \tag{12}$$

All work was performed using an Intel(R) Core(TM) i7-1065G7 machine, and implemented using PyTorch framework (version 1.11.0).

A. OPTIMISER SELECTION

SGD and Adam are two of the most common optimisers used in DL. However, their performances are domain dependent. This adds to the difficulty of selecting a standard approach for any classification problem. Therefore, the choice of a suitable optimiser is an essential step in the development of DL solutions. Table 5 presents the results of applying SGD and Adam to train a VGG-based classifier on the *first dataset scenario* described in Section III. This scenario provides the best SNR since the signals were collected at a short distance from the sensor. Also, the three preprocessing filters were applied using the *Version 1* parameters to maintain the same comparison basis.

The results represented in Table 5 show that SGD outperformed Adam for CQT, Mel, and the Complete representation, where the latter had the highest values (as shown in bold font in Table 5). Adam performed marginally better than SGD in the test where the Gammatone filter was used as the preprocessing method. In addition, the training session using Adam and the Mel spectrogram did not converge to a global minimum as the model predicted that almost everything belonged to the same class, as per the class-normalised confusion matrix shown in Figure 6.

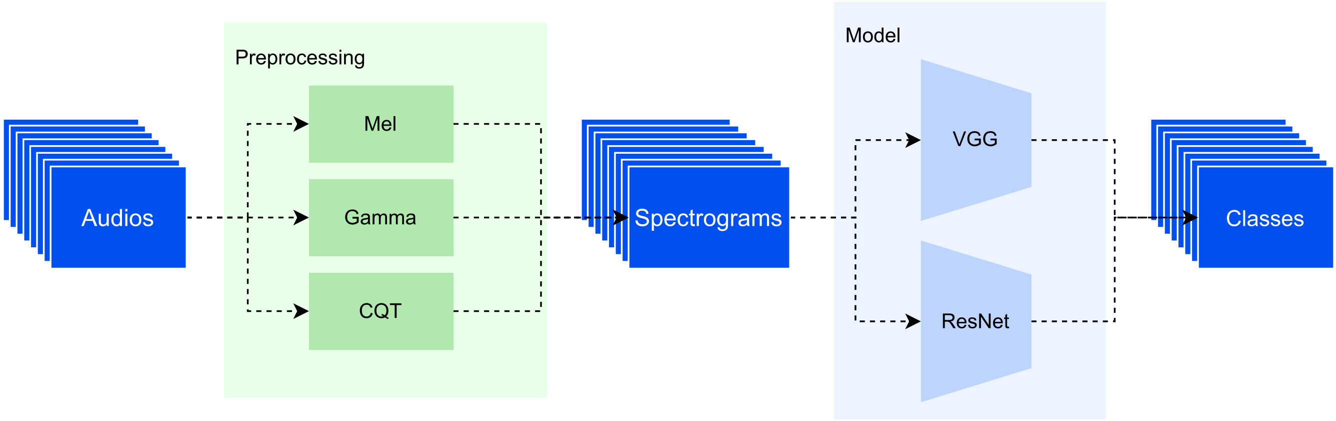

FIGURE 5. The block diagram of the complete solution. Preprocessing was performed either with CQT, Gammatone, or Mel spectrograms, or as a combination of all three. Either VGG or ResNet was used as the classification model, depending on the test condition.

TABLE 5. Comparison of SGD and Adam optimiser, showing accuracy, precision, recall, and F1-Score (F1). Using all three preprocessing methods (Complete) with the SGD optimiser produced the best results on all performance metrics.

	Method	Accuracy%	Precision%	Recall%	F1%
SGD	CQT	79.10	66.51	82.89	68.01
	Gamma	53.30	47.71	58.47	43.91
	Mel	54.47	51.23	61.74	44.40
	Complete	**84.46**	**71.48**	**85.56**	**75.42**
Adam	CQT	77.39	65.25	81.04	65.30
	Gamma	56.59	52.32	59.51	44.94
	Mel	5.26	21.00	20.16	2.23
	Complete	73.08	60.75	72.72	61.14

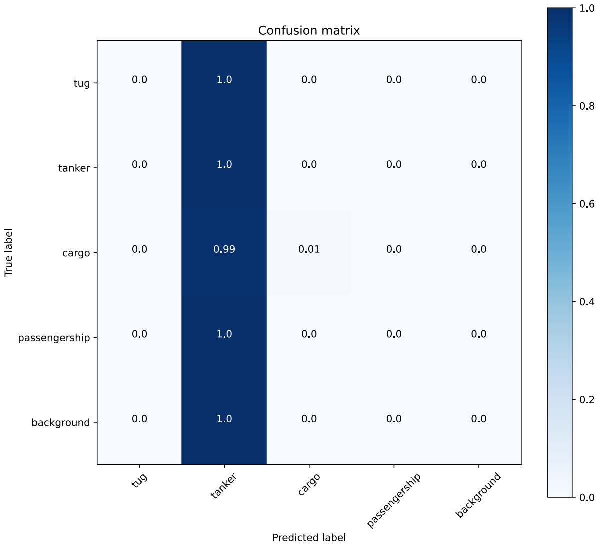

FIGURE 6. The confusion matrix for the execution of the Adam optimiser using the Mel spectrogram as the preprocessing filter, which produced erroneous predictions for the test dataset. In this case, all the input signals were considered as belonging to the *Tanker* class. For a confusion matrix showing the results for a more accurate classifier, see Figure 9.

TABLE 6. Comparison of *Version 1* and *Version 2* preprocessing representations, showing accuracy, precision, recall, and F1-Score (F1). The best performing condition was when the CQT method of preprocessing was used on the *Version 2* representation of the data. Note: VGG with SGD optimisation was used as the model classifier in this test.

	Method	Accuracy%	Precision%	Recall%	F1%
Vers. 1	CQT	79.10	66.51	82.89	68.01
	Gamma	53.29	47.71	58.47	43.91
	Mel	54.46	51.23	61.74	44.40
	Complete	84.46	71.48	85.56	75.42
Vers. 2	CQT	**86.32**	**73.97**	**88.34**	**77.91**
	Gamma	62.28	53.88	65.08	52.14
	Mel	41.86	53.85	45.41	27.92
	Complete	82.92	73.06	83.12	75.26

Comparing the best results for SGD and Adam, with CQT and Complete inputs, the higher accuracy was obtained with the SGD optimiser (84.46%), which is 7 percentage points larger accuracy than the best result obtained with Adam (77.39%). Also, the SGD approach was more stable during the training procedure.

B. PREPROCESSING OPTIMISATION

As mentioned in Section V-A, the *Version 1* parameters for the preprocessing filters were generated based on the information from the related literature, resulting in a 64 × 63 image. *Version 2* had images of dimensions 95 × 126, which were generated according to underwater acoustics features. An experiment was conducted to establish a comparison between these two representations, where the same baseline setup used in Section VI-A was applied: the VGG-based model trained on data from the first dataset scenario. As the results obtained in Section VI-A showed that the SGD optimiser produced better results, the experiments were only performed using this optimiser. The results of this test are summarised in Table 6.

Both Gammatone and Mel Spectrogram methods presented lower accuracy values when compared with CQT and Complete representations, as shown in Table 6. The worst CQT result, obtained with *Version 1*, was 17 percentage points better than the best result for Gammatone filter,

TABLE 7. Comparison of VGG-based and ResNet 18 models, showing accuracy, precision, recall, and F1-Score (F1). The ResNet model with all three preprocessing channels (Complete) yielded the best results in all analysis metrics.

	Method	Accuracy%	Precision%	Recall%	F1%
VGG	CQT	86.86	73.97	88.34	77.91
	Complete	82.92	73.06	83.12	75.26
ResNet	CQT	94.95	89.21	95.60	91.92
	Complete	**97.07**	**94.57**	**97.61**	**96.00**

obtained with *Version 2*. The Mel Spectrogram and Complete representations did not produce improvements with the *Version 2* parameters. However, even using the *Version 1* Mel Spectrogram results show an accuracy value that was 25 percentage points below that of CQT for the same parameters, representing a 31.15% drop in accuracy. Additionally, despite the accuracy drop of 1.54 percentage points between *Version 1* and *Version 2* for the Complete scenario, the precision improved 1.58 percentage points, showing a very close result for both versions of the Complete representation. On the other hand, when using the CQT method, *Version 2* had an accuracy improvement of 9.13% (7.22 percentage points) over *Version 1*. This improvement was likely due to the parameter optimisation process, which led to an increase in the temporal scale with the shorter hop length, and an improvement of the frequency representation with optimised frequency boundaries. These results suggest that the VGG-model, using *Version 2* of the preprocessing parameters, outperformed the results obtained with *Version 1*. Thus, *Version 2* was considered as the baseline setting in the remainder of this work.

C. MODEL EVALUATION

Following optimiser and preprocessing parameters selection, the next step in the development of the underwater acoustic signal classifier was the selection of the DL model. Training sessions were performed using both VGG-based and ResNet18 models. As the results obtained in the previous experiments suggested a better performance using the combination of SGD optimiser, with either CQT or the Complete representation (generated using *Version 2* parameters), this setup was selected for model evaluation. Table 7 shows the results obtained from these tests.

The results showed that the ResNet18 model outperformed VGG for both CQT and Complete preprocessing filters, presenting an improvement of 8.09 percentage points for the CQT, and 14.15 for the Complete, the latter being the best result obtained for this dataset scenario overall. The ResNet's capacity to have more intermediate layers proved to be suitable for the feature extraction stage, as it gave the model the ability to generalise the problem function better, thus resulting in higher classification performance. The loss obtained during the training stage (represented in Figure 7 and Figure 8) showed similar convergence behaviours for both models. Figure 8 also shows that the loss curves for CQT and Complete preprocessing methods (feeding a ResNet18 classifier) are almost identical. This contrasts with the curves shown in Figure 7 that represents a better performance for the CQT than Complete when applied to a VGG-based model.

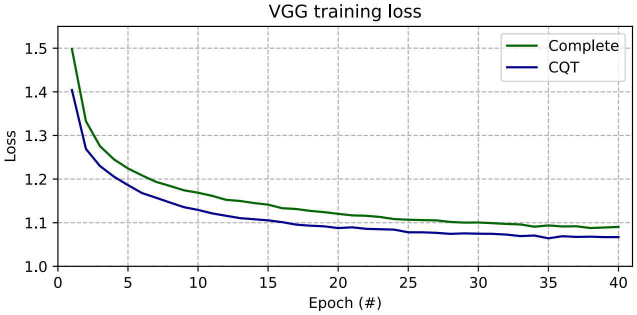

FIGURE 7. The loss obtained during the VGG-based model training stage, using CQT and Complete preprocessing filters.

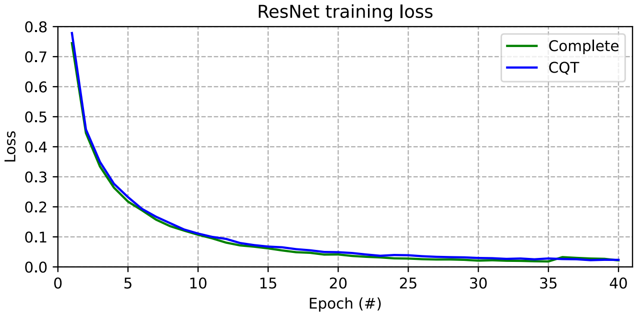

FIGURE 8. The loss obtained during the ResNet18 model training stage, using CQT and Complete preprocessing filters.

Although the Complete representation, combined with the ResNet model, presented an improvement of 2.12 percentage points in accuracy, the CQT was able to obtain a similar value using only one-third of the input size and preprocessing, since Complete is a three-channel representation. This result suggests that a fair trade off between accuracy and model size is obtained when using the CQT as a single preprocessing method.

D. SCENARIOS VALIDATION

The final test executed in this work evaluated how the distance from the sensor to the target vessel influenced the classification results. Tests were conducted with the combination of methods that produced the best results, as reported in previous sections. The architecture, composed of the CQT preprocessing filter applied to the ResNet18 model, was used to compare the results obtained in training and testing on the three scenarios described in Section III. Also, a test using all of the data from the three dataset scenarios combined was performed, aiming to evaluate if the distance impacted the accuracy, or if the generalisation ability of the architecture

TABLE 8. Comparison of the final architecture (CQT preprocessing feeding into ResNet) on the proposed dataset scenarios, showing accuracy, precision, recall, and F1-Score (F1).

Scenario	Accuracy%	Precision%	Recall%	F1%
First	94.95	89.21	95.60	91.92
Second	94.45	94.57	93.89	94.17
Third	93.11	89.36	93.82	91.30
Combined	84.13	79.46	86.51	81.78

was capable of dealing with this variable. The results obtained from these tests are shown in Table 8.

As the different scenarios do not contain the same number of instances (or the same vessels), they are not directly comparable, making it difficult to precisely state which situation allowed the best outputs. However, these results suggest that there is no practical difference in accuracy between the individual scenarios. This means that range to target had minimal impact on system accuracy, at least up to the tested 6 km distance boundary. One explanation for this could be the depth and ocean temperature where the hydrophone was located, which provided the best context for underwater sound propagation [25], thereby not degrading the SNR sufficiently to invalidate the signal representation. However, combining all of the data from the three scenarios caused an accuracy drop of between 8.98 and 10.82 percentage points compared to the individual scenarios alone. The confusion matrix for the Combined scenario is shown in Figure 9. This suggests there is a negative influence in the data from the different scenarios that confused the model during training. In particular, results with the Combined scenario show a higher confusion rate between *background* and *tug* than when the scenarios were trained separately. This was probably due to the similar range of frequencies from these two classes, that may have been enhanced due to the combination of SNR from the various scenarios.

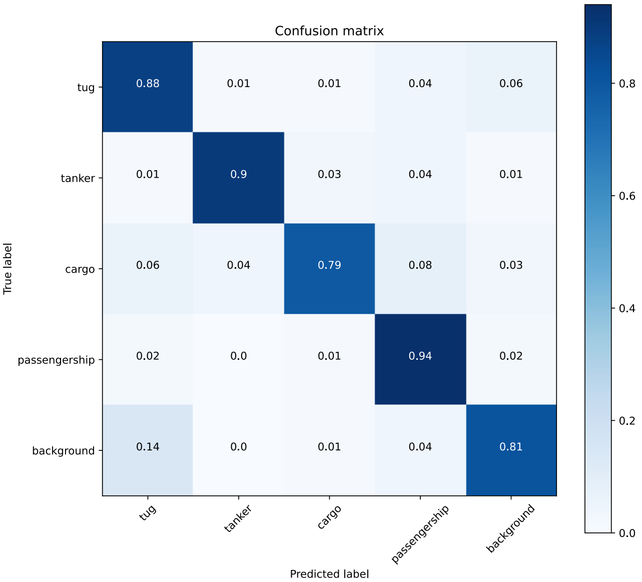

FIGURE 9. The confusion matrix for the execution of the Combined scenario using CQT as the preprocessing filter feeding into the ResNet18 model.

VII. DISCUSSION

This paper has reported experimental evaluations of the main aspects related to the development of a DL-based classifier for vessel types using underwater acoustic data.

The first tests reported focused on the selection of the most suitable elements to compose the classifier architecture, such as the optimiser and preprocessing methods. Initially, the two most commonly used optimisers, SGD and Adam, were tested and compared. The results reported in Section VI-A showed that Adam's performance was not satisfactory in this domain, owing to lower accuracy rates as a result of its inefficient treatment of local minima. In comparison, SGD produced higher accuracy and a more stable performance. This agrees with other studies (e.g., [56], [57]) that argue adaptive optimisation methods, like Adam, often generalise significantly worse than stochastic methods, such as SGD, since the strategy used by the former to escape saddle points causes difficulties in achieving flat global minima. In contrast, the momentum-based strategy of the latter provides a drift effect to escape saddle points without affecting the flat minima selection [57]. The tests performed herein seem to corroborate this hypothesis, as the best results obtained for SGD (using CQT) were 11.38 percentage points better than the performance (for the same preprocessing filters) obtained when using the Adam optimiser.

A second issue considered in this work was the selection of the most suitable representation of the signal to be used by the CNN. In our dataset, the CQT representation presented better performance, with an accuracy of around 86%, when compared with Gammatone and Mel spectrograms, with best accuracies of 62% and 54%, respectively, a minimum of 24 percentage points improvement. There are various possible explanations for this finding. One is that the acoustic signals generated by the vessels are predominantly composed of lower frequencies. However, the Gammatone filter does not emphasise low frequencies sufficiently, resulting in a lower classification performance. Mel spectrogram, on the other hand, does emphasise the lower frequencies, by mapping the frequency axis to the logarithmic Mel scale. However, it maintains the conversion from time-domain using fixed time windows, which negatively affects the temporal resolution. In contrast, CQT increases the time resolution towards higher frequencies while reducing the frequency resolution; this results in emphasising the lower-frequencies, which is akin to the human aural perception [50]. This feature makes the CQT spectrogram the most suitable representation for automated classifications of underwater acoustic data using CNN, owing to the nature of the convolutional layers.

The tests conducted with the Complete representation (all three preprocessing filters combined) aimed to obtain a preprocessing method that includes the advantages of each of the methods considered in this work. The results obtained showed that the classification accuracy obtained using this three-channel representation was marginally better than the best single filter (CQT) for ResNet, but not the VGG model. As the Complete representation combines multiple preprocessing methods, its generation and processing is more computationally expensive than applying each preprocessing method individually. Even in the cases where the Complete representation showed the best results (ResNet model), its performance was similar to that obtained using only the CQT spectrograms as input. This indicates that the Complete preprocessing method should only be used in situations requiring the highest possible accuracy or where computational cost is essentially irrelevant.

With respect to the DL model selection, the ResNet approach outperformed the VGG-based model for both CQT and Complete data representation methods by at least 8.09 percentage points in accuracy. This superior performance was likely due to the existence of residual blocks in the ResNet model, which reduce the probability of overfitting. It is worth mentioning that the ResNet model used had 17 convolutional layers, in contrast to four composing the VGGNet. Considering the relative complexity and accuracy results of both models tested in this work, we can conclude that, although ResNet18 produced the best classification results, VGG-based classifiers are still suitable models to be used in applications with limited computational resources.

The final test executed in this work evaluated the influence of the distance between the sensor to the targets with respect to the classification performance. Despite the fact that a minor degradation in accuracy was observed with respect to an increase in the distance to the sensor, the results obtained for the three scenarios showed similar figures. The tests using the combination of the data points from all three scenarios presented the worst results compared with the performance values obtained for each of its constituent scenarios. This was probably due to the fact that, although a minor variation in the SNR (resulted from the distance between target and sensor) did not affect the results obtained in each of the individual scenarios, this difference was large enough to increase the complexity of the audio patterns contained in the combined dataset, thus hindering the capacity of a simple classifier to find a suitable generalisation that represented accurately the distinct classes. It should be noted that the largest (by far) point of confusion in the Combined dataset was the classifier detecting the presence of a tug when there was no vessel present. While additional data could help overcome this error, another option for future investigation is a two-stage detect and classify process. A computationally simple detection algorithm could be used to determine the possible presence of a vessel and then the classifier used to determine what sort of vessel it is. This has the potential to both improve the error rate when no vessels are present, while reducing the computational resources as the classification network would not have to run continuously. It is likely that a detection algorithm of sufficient sensitivity will detect a vessel before sufficient structure is present for a classification algorithm to accurately classify said vessel.

TABLE 9. Comparison of the ResNet classification results obtained in the present work with the results reported in [41] obtained on the DeepShip dataset, showing accuracy (Acc.), precision (Prec.), recall (Rec.), and F1-Score (F1).

Model	Dataset	Acc.%	Prec.%	Rec.%	F1%
ResNet	1st Scenario	94.95	89.21	95.60	91.92
	2nd Scenario	94.45	94.57	93.89	94.17
	3rd Scenario	93.11	89.36	93.82	91.30
	Combined Scenarios	84.13	79.46	86.51	81.78
SCAE	DeepShip	77.53	78	77	77
Residual		76.98	77	77	77
CNN		76.35	76	76	76
Inception		76.16	76	76	76
DNN [41]		73.11	73	73	73
SVM	DeepShip	72.24	72	72	72
RF		69.71	70	70	69
KNN		62.71	64	63	63
Naive Bayes [41]		53.97	57	53	52

Considering related work developed with data from the ONC initiative, accuracy values of the order of 80% were previously reported using raw audio data, where time-frequency filter dependency was not considered in DL pipelines [34], [35]. An accuracy value of around 87% was reported as a result of the application of a bio-inspired cochlea model preprocessing filter to a CNN-based classification [40]. In addition, a comparison of various deep learning methods was conducted in [41] using an analogous set of ONC raw data that was used in the present paper. However, it was reported that 77.53% was the highest accuracy obtained in that work. That work was developed on a dataset called DeepShip, whose recordings were divided into 613 files, which varies from about 6 seconds to 1530 seconds. Only the identification of a single vessel within a range of 2 km from the hydrophone was used to generate the data, and the background noise recordings were added from a distinct source. Table 9 shows a summary of the best results obtained in the present work (first block), against the results reported in [41].

Although it is virtually impossible to faithfully reproduce the results reported in [41] (as the code used to generate them is not publicly available) it is possible to observe that the results reported in that paper are similar to those obtained in the present work for the Combined dataset (84%). This already places the baseline results obtained in the present work within the state-of-the-art of the field. However, the research reported here achieved superior results when distinct scenarios were considered with respect to the distance to the sensor, obtaining an accuracy of 95% or 97% depending on the preprocessing filtering method used (CQT or Complete, respectively), as shown in Table 9.

VIII. CONCLUSION

This paper presented a machine learning approach for vessel type classification using underwater acoustic data. CQT, Gammatone, and Mel Spectrogram filters were used to preprocess the acoustic data, aiming to extract relevant features of the signals. Striving to achieve a better representation of the signal, the combination of the three outputs into three-channel data was also investigated. The results showed that the CQT and the Combined approaches achieved the highest accuracy results for the dataset used in this paper. This study also compared the SGD and Adam optimiser performance applied to the vessel type classification problem, showing that the SGD optimiser is more stable and presents a better generalisation than Adam. Concerning the deep learning model, results showed that ResNet18 yielded the highest evaluation metrics when compared to the VGGNet model.

A new dataset was defined, using the raw data from Ocean Networks Canada, which include the underwater acoustic signals annotated with the related vessel types. Three distinct scenarios were defined with respect to the distance between the target vessel to the hydrophone used to capture the signal. These three scenarios were compared using the proposed pipeline, achieving a maximum accuracy of 94.95% when CQT was used to preprocess the data fed into a ResNet classification model. Higher accuracy (97%) was achieved if all three preprocessing methods (CQT, Gammatone, and Mel Spectrogram) were used simultaneously. However, the increase in computational costs may not be worth the slight accuracy improvement. Furthermore, a test was conducted combining the three distance scenarios into a single dataset, resulting in a classification accuracy of 84.13% for the CQT preprocessed data.

A complete pipeline for the classification of underwater acoustic signals was proposed in this paper, whose source code and data are publicly available. However, some issues were not addressed and will be considered in future work. Despite the promising results obtained in classification using the isolated distance scenarios, the results for the combined approach have a large scope for improvement, since the accuracy across distinct scenarios had a variation of 10 percentage points. Future work will also consider the application of other machine learning models, combined with novel biologically inspired filters, to the task of classification of vessels from acoustic data, in order to investigate better trade offs between accuracy and model size in image classification tasks for the classification of acoustic data.

REFERENCES

[1] H. A. Orengo and A. Garcia-Molsosa, "A brave new world for archaeological survey: Automated machine learning-based potsherd detection using high-resolution drone imagery," *J. Archaeological Sci.*, vol. 112, Dec. 2019, Art. no. 105013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0305440319301001

[2] C. Xu, J. Chen, D. Yan, and J. Ji, "Review of underwater cable shape detection," *J. Atmos. Ocean. Technol.*, vol. 33, no. 3, pp. 597–606, Mar. 2016. [Online]. Available: https://journals.ametsoc.org/view/journals/atot/33/3/jtech-d-15-0112_1.xml

[3] W. Chen, F. Hou, and D. Chen, "Development of tactile imaging for underwater structural damage detection," *Sensors*, vol. 19, no. 18, p. 3925, Sep. 2019, doi: 10.3390/s19183925.

[4] K. Dastner, B. V. H. Z. Roseneckh-Kohler, F. Opitz, M. Rottmaier, and E. Schmid, "Machine learning techniques for enhancing maritime surveillance based on GMTI radar and AIS," in *Proc. 19th Int. Radar Symp. (IRS)*, Jun. 2018, pp. 1–10.

[5] J. Choi, Y. Choo, and K. Lee, "Acoustic classification of surface and underwater vessels in the ocean using supervised machine learning," *Sensors*, vol. 19, no. 16, p. 3492, Aug. 2019.

[6] K. Terayama, K. Shin, K. Mizuno, and K. Tsuda, "Integration of sonar and optical camera images using deep neural network for fish monitoring," *Aquacultural Eng.*, vol. 86, Aug. 2019, Art. no. 102000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0144860919300640

[7] C. Erbe, S. A. Marley, R. P. Schoeman, J. N. Smith, L. E. Trigg, and C. B. Embling, "The effects of ship noise on marine mammals—A review," *Frontiers Mar. Sci.*, vol. 6, Oct. 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmars.2019.00606

[8] N. D. Merchant, E. Pirotta, T. R. Barton, and P. M. Thompson, "Monitoring ship noise to assess the impact of coastal developments on marine mammals," *Mar. Pollut. Bull.*, vol. 78, nos. 1–2, pp. 85–95, Jan. 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0025326X13006802

[9] E. Rossi, G. Licitra, A. Iacoponi, and D. Taburni, "Assessing the underwater ship noise levels in the north tyrrhenian sea," *Adv. Exp. Med. Biol.*, vol. 875, pp. 943–999, Sep. 2016.

[10] M. Nastasi, L. Fredianelli, M. Bernardini, L. Teti, F. Fidecaro, and G. Licitra, "Parameters affecting noise emitted by ships moving in port areas," *Sustainability*, vol. 12, no. 20, p. 8742, Oct. 2020. [Online]. Available: https://www.mdpi.com/2071-1050/12/20/8742

[11] M. F. McKenna, D. Ross, S. M. Wiggins, and J. A. Hildebrand, "Underwater radiated noise from modern commercial ships," *Acoust. Soc. Amer.*, vol. 131, no. 1, pp. 92–103, Jan. 2012, doi: 10.1121/1.3664100.

[12] J. A. Bocanegra, D. Borelli, T. Gaggero, E. Rizzuto, and C. Schenone, "A novel approach to port noise characterization using an acoustic camera," *Sci. Total Environ.*, vol. 808, Feb. 2022, Art. no. 151903.

[13] O. Dubovik, G. L. Schuster, F. Xu, Y. Hu, H. Bösch, J. Landgraf, and Z. Li, "Grand challenges in satellite remote sensing," *Frontiers Remote Sens.*, vol. 2, Feb. 2021, Art. no. 619818.

[14] D. G. Childers, D. P. Skinner, and R. C. Kemerait, "The cepstrum: A guide to processing," *Proc. IEEE*, vol. 65, no. 10, pp. 1428–1443, Oct. 1977.

[15] B. Boashash, *Time Frequency Signal Analysis and Processing: A Comprehensive Reference*. Amsterdam, The Netherlands: Elsevier, 2003. [Online]. Available: https://books.google.com.au/books?id=AtcSlAEACAAJ

[16] J. Ghosh, K. Turner, S. Beck, and L. Deuser, "Integration of neural classifiers for passive sonar signals," in *Multidimensional Systems Signal Processing Algorithms and Application Techniques* (Control and Dynamic Systems), vol. 77, C. T. Leondes, Ed. San Diego, CA, USA: Academic, 1996, pp. 301–338. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0090526796800338

[17] H. Yang, K. Lee, Y. Choo, and K. Kim, "Underwater acoustic research trends with machine learning: General background," *J. Ocean Eng. Technol.*, vol. 34, no. 2, pp. 147–154, Apr. 2020.

[18] H. Yang, S.-H. Byun, K. Lee, Y. Choo, and K. Kim, "Underwater acoustic research trends with machine learning: Active SONAR applications," *J. Ocean Eng. Technol.*, vol. 34, no. 4, pp. 277–284, Aug. 2020.

[19] H. Yang, K. Lee, Y. Choo, and K. Kim, "Underwater acoustic research trends with machine learning: Passive SONAR applications," *J. Ocean Eng. Technol.*, vol. 34, no. 3, pp. 227–236, Aug. 2020.

[20] J. Chai, H. Zeng, A. Li, and E. W. T. Ngai, "Deep learning in computer vision: A critical review of emerging techniques and application scenarios," *Mach. Learn. Appl.*, vol. 6, Dec. 2021, Art. no. 100134. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666827021000670

[21] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, Oct. 1986.

[22] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, ''Handwritten digit recognition with a back-propagation network,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 1990, pp. 396–404.

[23] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' 2014, *arXiv:1409.1556*.

[24] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[25] L. C. F. Domingos, P. E. Santos, P. S. M. Skelton, R. S. A. Brinkworth, and K. Sammut, ''A survey of underwater acoustic data classification methods using deep learning for shoreline surveillance,'' *Sensors*, vol. 22, no. 6, p. 2181, Mar. 2022.

[26] V. Vahidpour, A. Rastegarnia, and A. Khalili, ''An automated approach to passive sonar classification using binary image features,'' *J. Mar. Sci. Appl.*, vol. 14, no. 3, pp. 327–333, Sep. 2015.

[27] D. Santos-Domínguez, S. Torres-Guijarro, A. Cardenal-López, and A. Pena-Gimenez, ''ShipsEar: An underwater vessel noise database,'' *Appl. Acoust.*, vol. 113, pp. 64–69, Dec. 2016.

[28] D. A. Abraham, *Underwater Acoustic Signal Processing Modeling, Detection, and Estimation*. Cham, Switzerland: Springer, 2019.

[29] E. L. Ferguson, R. Ramakrishnan, S. B. Williams, and C. T. Jin, ''Convolutional neural networks for passive monitoring of a shallow water environment using a single sensor,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2657–2661.

[30] D. Gebhardt, K. Parikh, I. Dzieciuch, M. Walton, and N. A. V. Hoang, ''Hunting for naval mines with deep neural networks,'' in *Proc. OCEANS*, 2017, pp. 1–5.

[31] V.-S. Doan, T. Huynh-The, and D.-S. Kim, ''Underwater acoustic target classification based on dense convolutional neural network,'' *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[32] G. Hu, K. Wang, Y. Peng, M. Qiu, J. Shi, and L. Liu, ''Deep learning methods for underwater target feature extraction and recognition,'' *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–10, Sep. 2018.

[33] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, ''SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size,'' 2016, *arXiv:1602.07360*.

[34] H. Yang, J. Li, S. Shen, and G. Xu, ''A deep convolutional neural network inspired by auditory perception for underwater acoustic target recognition,'' *Sensors*, vol. 19, no. 5, p. 1104, Mar. 2019.

[35] S. Tian, D. Chen, H. Wang, and J. Liu, ''Deep convolution stack for waveform in underwater acoustic target recognition,'' *Sci. Rep.*, vol. 11, no. 1, p. 9614, Dec. 2021, doi: 10.1038/s41598-021-88799-z.

[36] Q. Sun and K. Wang, ''Underwater single-channel acoustic signal multitarget recognition using convolutional neural networks,'' *J. Acoust. Soc. Amer.*, vol. 151, no. 3, pp. 2245–2254, Mar. 2022.

[37] D. Wang, L. Zhang, C.-C. Bao, M. Shu-qing, and Y. Wang, ''Passive ship localization in a shallow water using pre-trained deep learning networks,'' in *Proc. 23rd Int. Congr. Acoust., Integrating, 4th EAA Euroregio*, 2019, pp. 9–13.

[38] M. Khishe and H. Mohammadi, ''Passive sonar target classification using multi-layer perceptron trained by salp swarm algorithm,'' *Ocean Eng.*, vol. 181, pp. 98–108, Jun. 2019.

[39] X. Wang, A. Liu, Y. Zhang, and F. Xue, ''Underwater acoustic target recognition: A combination of multi-dimensional fusion features and modified deep neural network,'' *Remote Sens.*, vol. 11, no. 16, p. 1888, Aug. 2019, doi: 10.3390/rs11161888.

[40] S. Shen, H. Yang, X. Yao, J. Li, G. Xu, and M. Sheng, ''Ship type classification by convolutional neural networks with auditory-like mechanisms,'' *Sensors*, vol. 20, no. 1, p. 253, Jan. 2020.

[41] 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, Nov. 2021, Art. no. 115270.

[42] L. Chen, F. Liu, D. Li, T. Shen, and D. Zhao, ''Underwater acoustic target classification with joint learning framework and data augmentation,'' in *Proc. 5th Int. Conf. Artif. Intell. Big Data (ICAIBD)*, May 2022, pp. 23–28.

[43] X. C. Han, C. Ren, L. Wang, and Y. Bai, ''Underwater acoustic target recognition method based on a joint neural network,'' *PLoS ONE*, vol. 17, no. 4, Apr. 2022, Art. no. e0266425.

[44] M. Zhao, S. Zhong, X. Fu, B. Tang, and M. Pecht, ''Deep residual shrinkage networks for fault diagnosis,'' *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4681–4690, Jul. 2020.

[45] Y. H. Awni, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, ''Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network,'' *Nature Med.*, vol. 25, pp. 65–69, Jan. 2019.

[46] A. J. Oxenham, ''How we hear: The perception and neural coding of sound,'' *Annu. Rev. Psychol.*, vol. 69, no. 1, pp. 27–50, Jan. 2018.

[47] S. S. Stevens, J. Volkmann, and E. B. Newman, ''A Scale for the measurement of the psychological magnitude pitch,'' *J. Acoust. Soc. Amer.*, vol. 8, no. 3, pp. 185–190, Jan. 1937.

[48] D. O'Shaughnessy, *Speech Communications: Human and Machine*. Reading, MA, USA: Addison-Wesley Publishing Company, 1987.

[49] S. K. Pandey, H. S. Shekhawat, and S. R. M. Prasanna, ''Deep learning techniques for speech emotion recognition: A review,'' in *Proc. 29th Int. Conf. Radioelektronika (RADIOELEKTRONIKA)*, Apr. 2019, pp. 1–6.

[50] J. C. Brown, ''Calculation of a constant Q spectral transform,'' *J. Acoust. Soc. Amer.*, vol. 89, no. 1, pp. 425–434, 1991.

[51] P. Johannesma, ''The pre-response stimulus ensemble of neurons in the cochlear nucleus,'' in *Proc. Symp. Hearing Theory*. Eindhoven, The Netherlands: IPO, 1972.

[52] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, ''An efficient auditory filterbank based on the gammatone function,'' presented at the Meeting IOC Speech Group Auditory Modelling. Malvern, U.K.: RSRE, Institute of Acoustics on Auditory Modelling, vol. 2, Dec. 1987.

[53] G. Cybenko, ''Approximation by superpositions of a sigmoidal function,'' *Math. Control Signals Syst.*, vol. 2, no. 4, pp. 303–314, Dec. 1989, doi: 10.1007/BF02551274.

[54] H. Robbins and S. Monro, ''A stochastic approximation method,'' *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, 1985, doi: 10.1214/aoms/1177729586.

[55] D. P. Kingma and J. Ba, ''Adam: A method for stochastic optimization,'' 2014, *arXiv:1412.6980*.

[56] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, ''The marginal value of adaptive gradient methods in machine learning,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 120–190.

[57] Z. Xie, X. Wang, H. Zhang, I. Sato, and M. Sugiyama, ''Adaptive inertia: Disentangling the effects of adaptive learning rate and momentum,'' in *Proc. 39th Int. Conf. Mach. Learn.* (Proceedings of Machine Learning Research), vol. 162, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds. Baltimore, MD, USA: PMLR, Jul. 2022, pp. 24430–24459. [Online]. Available: https://proceedings.mlr.press/v162/xie22d.html

LUCAS C. F. DOMINGOS was born in São Caetano do Sul, São Paulo, Brazil, in 1996. He received the bachelor's degree in electrical engineering from the Centro Universitário FEI, São Bernardo do Campo, Brazil, in 2018, where he is currently pursuing the master's degree in artificial intelligence applied to robotics and automation. His research interests include software development for embedded devices, embedded applications for machine learning, and deep learning applied for acoustic and visual problems. Since 2021, he labors in computer vision solutions development, applying machine learning and classic image processing methods, at the Instituto de Pesquisas Eldorado, Campinas, Brazil.

PAULO E. SANTOS received the Ph.D. degree in artificial intelligence from Imperial College London, U.K., in 2003, working on the development of spatial reasoning systems for mobile robots. From 2003 to 2005, he was a Research Assistant at the School of Computing, University of Leeds, U.K. He was part of a leading research group in AI and robotics in Sao Paulo, Brazil (2005–2019), conducting a number of research projects of industrial interest. Since 2019, he has been an Associate Professor at the College of Science and Engineering, Flinders University, Adelaide, SA, Australia, and also a Full Member of the CNRS International Research Laboratory CROSSING and a Full Member of the Centre for Defence Engineering Research and Training, College of Science and Engineering, Flinders University.

PHILLIP S. M. SKELTON (Member, IEEE) received the Ph.D. degree in biologically inspired vision systems for robotics from the University of South Australia, in 2020. He joined Flinders University, as a Postdoctoral Research Associate, in 2020, where he is currently a Full Member of the Centre for Defence Engineering Research and Training, College of Science and Engineering. He works across all aspects of maritime autonomous systems and oversees a variety of projects. His current research interests include developing adaptive biologically inspired signal processing algorithms for multi-modal autonomous perception tasks in the underwater domain, and the complex task of tuning these algorithms using evolutionary computation techniques.

RUSSELL S. A. BRINKWORTH received the Ph.D. degree in neuroscience from the University of Adelaide, in 2004. He joined the Insect Vision Laboratory, as a Postdoctoral Researcher then as an ARC Research Fellow. He moved to become a Lecturer in mechanical engineering, in 2010. He moved to the University of South Australia as the Program Director for Autonomous Systems, in 2011. In late 2019, he joined the College of Science and Engineering, Flinders University, as an Associate Professor of autonomous systems. He is currently a Full Member of the Centre for Defence Engineering Research and Training.

KARL SAMMUT (Senior Member, IEEE) received the Ph.D. degree from the University of Nottingham, U.K., in 1992. From 1992 to 1995, he was employed as a Postdoctoral Fellow with the Politecnico di Milano, Italy, and Loughborough University, U.K. He commenced his appointment at Flinders University, in 1995, where he is currently a Professor with the College of Science and Engineering. He works as the Co-Director of the Centre for Defence Engineering Research and Training, College of Science and Engineering, Flinders University, and the Theme Leader of the Maritime Autonomy Group. His research interests include navigation, optimal guidance and control systems, and mission planning systems for autonomous marine surface and underwater vehicles.