CENTRO UNIVERSITÁRIO FEI

LUCAS CESAR FERREIRA DOMINGOS

MACHINE LEARNING METHODS FOR VESSEL TYPE CLASSIFICATION WITH UNDERWATER ACOUSTIC DATA

São Bernardo do Campo

2022

LUCAS CESAR FERREIRA DOMINGOS

MACHINE LEARNING METHODS FOR VESSEL TYPE CLASSIFICATION WITH UNDERWATER ACOUSTIC DATA

Master's Thesis, presented to the Centro Universitário FEI to obtain the title of Master in Electrical Engineering. Oriented by Prof. Dr. Paulo Eduardo Santos.

São Bernardo do Campo

2022

Cesar Ferreira Domingos, Lucas.
Machine learning methods for vessel type classification with underwater acoustic data / Lucas Cesar Ferreira Domingos. São Bernardo do Campo, 2022.
89 f. : il.

Dissertação - Centro Universitário FEI.
Orientador: Prof. Dr. Paulo Eduardo Santos.

1. Aprendizado Profundo. 2. Acústica. 3. Hidrofones. 4. Subaquática. 5. Embarcações. I. Eduardo Santos, Paulo, orient. II. Título.

Elaborada pelo sistema de geração automática de ficha catalográfica da FEI com os dados fornecidos pelo(a) autor(a).

centro universitário FEI

APRESENTAÇÃO DE DISSERTAÇÃO
ATA DA BANCA EXAMINADORA

Programa de Pós-Graduação Stricto Sensu em Engenharia Elétrica

Mestrado

PGE-10

Aluno(a): Lucas Cesar Ferreira Domingos **Matrícula:** 120306-6

Título do Trabalho: Machine learning methods for vessel type classification with underwater acoustic data

Área de Concentração: Inteligência Artificial Aplicada à Automação e Robótica

Orientador(a): Prof. Dr. Paulo Eduardo Santos

Data da realização da defesa: 20/10/2022

ORIGINAL ASSINADA

Avaliação da Banca Examinadora:

A defesa de mestrado do candidato Lucas Domingos iniciou-se às 9:00am (horário de Brasília) do dia 20/10/2022 e contou com a presença de todos os membros da banca avaliadora: Prof. Carlos E. Thomaz, Prof. Flávio Buiochi e eu (orientador) Prof. Paulo Santos, convidados e familiares também estavam presentes. A banca ocorreu por meio de vídeo conferência, sem ter havido alguma ocorrência que mereça destaque. O aluno apresentou o seu trabalho em 45min, respeitando o prazo previsto, e respondeu aos questionamentos dos membros da banca de maneira satisfatória. A banca considerou o trabalho com nível suficiente para justificar sua aprovação por unanimidade.

A Banca Julgadora acima-assinada atribuiu ao aluno o seguinte resultado:

APROVADO ☒ REPROVADO ☐

MEMBROS DA BANCA EXAMINADORA
Prof. Dr. Paulo Eduardo Santos
Prof. Dr. Carlos Eduardo Thomaz
Prof. Dr. Flávio Buiochi

Aprovação do Coordenador do Programa de Pós-graduação

Prof. Dr. Carlos Eduardo Thomaz

To my parents, Fernando and Amelia, who always supported me.

ACKNOWLEDGMENT

Thanks to Phillip Skelton, Russel Brinkworth and Karl Sammut for their useful comments, spending time correcting and suggesting changes to this present work.

Special thanks to the Instituto de Pesquisas Eldorado, for providing financial assistance and time to finish the requirements for this work.

“To the person who pleases him, God gives wisdom, knowledge and happiness, but to the sinner he gives the task of gathering and storing up wealth to hand it over to the one who pleases God. This too is meaningless, a chasing after the wind.”

Ecclesiastes 2:26

RESUMO

A identificação de embarcações em ambientes de tráfego controlado pode ser benéfica para manutenção da biodiversidade e proteção dos ambientes costeiros de regiões protegidas, gerando contribuições para a comunidade local e para o ecossistema. Nesse âmbito, vê-se latente a necessidade de melhores técnicas de identificação e classificação de embarcações, proporcionando mecanismos para melhora destes sistemas. Sinais sonoros subaquáticos são mais difíceis de serem mascarados ou omitidos durante a navegação de uma embarcação quando comparados com outras fontes de dados, proporcionando uma fonte confiável e resistente a fraudes para sistemas de classificação, porém, estes sofrem interferências das condições do meio em que se encontram. Neste trabalho, uma metodologia foi proposta para realizar a classificação de sinais sonoros subaquáticos provenientes de embarcações utilizando técnicas de aprendizado de máquina, considerando também as variáveis ambientais, como a distância entre os hidrofones e as embarcações. Uma comparação relativa à performance das redes neurais convolucionais mais comuns foi realizada utilizando a arquitetura da VGG e da ResNet 18. Também foram realizadas comparações entre os três filtros de pré-processamentos comumente presentes na literatura, os espectrogramas Mel, os filtros Gamma, e a transformada de constante Q, proporcionando um estudo sobre o impacto de tais variáveis na classificação final. Devido a escassez de conjuntos de dados anotados para estudo deste problema, um conjunto de dados anotados foi proposto utilizando como base os sinais sonoros da iniciativa Ocean Canada Network. Os resultados obtidos atingiram a acurácia de 94.95% no conjunto de dados proposto usando CQT como filtro de pré-processamento para uma rede neural convolucional baseada na ResNet. Os códigos fontes para reprodução dos testes, assim como para obtenção do dataset, estão disponibilizados de maneira gratuita e pública para fins acadêmicos.
Palavras-chave: Aprendizado Profundo. Acústica. Hidrofones. Subaquática. Embarcações.

ABSTRACT

Vessel identification in a controlled traffic environment can be beneficial for biodiversity maintenance and coastal environment surveillance in protected regions, generating contributions to the local community and the ecosystem. In this context, there is a latent need for better techniques for identifying and classifying vessels, providing mechanisms to improve these systems. Underwater sound signals are more challenging to be masked or omitted, during the navigation of a vessel, when compared to other data sources, providing a reliable and fraud-resistant source for classification systems, however, they suffer interference from the conditions of the environment in which they are used. In this work, a methodology was proposed to perform the underwater acoustic classification, using signals produced by vessels, using machine learning techniques, and also considering environmental variables, such as the distance between the hydrophones and the target vessels. A comparison regarding the performance of the most common convolutional neural networks was performed using the VGG and ResNet 18 architectures. Comparisons were also made between the three preprocessing filters commonly present in the literature, the Mel spectrograms, the Gamma filters, and the constant Q transform, providing a study on the impact of such variables in the final classification. Due to the scarcity of annotated datasets to study this problem, an annotated dataset was proposed based on the sound signals of the Ocean Canada Network initiative. The results obtained reached the accuracy of 94.95% on the proposed dataset using CQT as the preprocessing filter for a ResNet-based convolutional neural network. The source codes for reproducing the tests, as well as for obtaining the dataset, are freely and publicly available for academic purposes.
Keywords: Deep Learning. Sound. Marine Environment. Ship Type. Hydrophones

LIST OF FIGURES

Figure 1 – Examples of visual representation for the three preprocessing filters: Mel Spectrogram, Gammatonegram and CQT. 52
Figure 2 – The baseline VGG-based model architecture. 52
Figure 3 – Location of the Hydrophone (in red), at Strait of Georgia, Canada. 53
Figure 4 – The icListen AF Hydrophone (BORNHÖFT, 2022) on the left side of the figure and the Sea-Bird SBE 16plus SEACAT Recorder (ELECTRONICS, 2022) on the right side of the figure. 54
Figure 5 – Diagram representing a scenario. 54
Figure 6 – Examples of audios from the five categories present in the dataset. 58
Figure 7 – An example of audio used on this experiment. The image shows the raw audio signal, on time domain, and its three preprocessed versions using the *Version 2* parameters. 59
Figure 8 – The VGG-based model architecture. 59
Figure 9 – The ResNet 18 model architecture. 60
Figure 10 – The block diagram of the complete solution. 60
Figure 11 – Confusion matrix with baseline results using the CQT preprocessing filter. 63
Figure 12 – The confusion matrix of the execution for Adam optimiser using the Mel spectrogram as preprocessing filter. 65
Figure 13 – The confusion matrix of the execution for the Combined scenario the CQT as preprocessing filter and ResNet 18 model. 69

LIST OF TABLES

Table 1 – An overview of Machine Learning (ML) methods for classification from frequency and time-frequency analysis. 36

Table 2 – An overview of ML methods for classification using Biologically-Inspired filtering algorithms. 39

Table 3 – An overview of methods for classification using convolutional filters alone. 43

Table 4 – A summary of some datasets available. 46

Table 5 – Summary of the *Version 1* set of parameters, used on the baseline preprocessing generation. 52

Table 6 – The different scenarios considered for the dataset generation. 55

Table 7 – The length, in seconds, of each dataset class for each of the three scenarios' subsets. 57

Table 8 – The detailing of the *Version 1* and *Version 2* sets of parameters used in the preprocessing generation. 58

Table 9 – Comparison of evaluated methods showing accuracy, precision, recall, and F1-Score (F1) 62

Table 10 – Comparison of SGD and ADAM optimiser, showing accuracy, precision, recall, F1-Score (F1), and total training time. 64

Table 11 – Comparison of *Version 1* and *Version 2* preprocessing representations, showing accuracy, precision, recall, F1-Score (F1), and total training time. 66

Table 12 – Comparison of VGG-based and ResNet 18 models, showing accuracy, precision, recall, F1-Score (F1), and total training time. 67

Table 13 – Comparison of the final architecture on the proposed dataset scenarios, showing accuracy, precision, recall, F1-Score (F1), and total training time. 68

ACRONYMS

ACT	Anisotropic Chirplet Transforms
AdaGrad	Adaptive Gradient Algorithm
Adam	Adaptive moment estimation
ADCNN	Auditory Perception-inspired Deep Convolutional Neural Network
AIS	Automatic Identification System
ANN	Artificial Neural Network
AP	Average Precision
AUC	Area Under the Curve
AUSCC	Area Under the Sample Complexity Curve
BI	Bookmarker Informedness
CNN	Convolutional Neural Network
CQT	Constant Q Transform
CSDM	Cross-Spectral Density Matrices
CSV	Comma-Separated Value
CTD	Conductivity Temperature and Depth
DCT	Discrete Cosine Transform
DL	Deep Learning
DWT	Discrete Wavelet Transform
EFP	Efficient Feature Pyramid
ELM	Extreme Learning Machine
EM	Electromagnetic
FEI	Centro Universitário FEI
FN	False Negative
FP	False Positive
FT	Fourier transforms
GFCC	Gammatone Frequency Cepstral Coefficients
GPS	Global Positioning System
HSV	Hue Saturation Value
IEEE	Institute of Electrical and Electronics Engineers
ILSVRC	ImageNet Large Scale Visual Recognition Challenge

KNN	K-Nearest Neighbors algorithm
LVQ	Learning Vector Quantisation
mAP	Mean Average Precision
MCC	Mathews Correlation Coefficient
MDPI	Multidisciplinary Digital Publishing Institute
MDT	Marine Debris Turntable
MFCC	Mel Frequency Cepstrum Coefficients
ML	Machine Learning
MSGD	Mini-batch Gradient Descent Method
MSRDN	Multiscale Residual Deep Neural Network
NN	Neural Network
ONC	Ocean Networks Canada
PCA	Principal component analysis
RA	Random Accuracy
ReLu	Rectified Linear Units
revcor	Reverse Correlation
RMSP	Root Mean Square Propagation
ROC	Receiver Operating Characteristics
SCAE	Separable Convolution Autoencoder
SGD	Stochastic Gradient Descent
SNR	Signal-to-Noise Ratio
SOM	Self-Organised Mapping
STFT	Short-Time Fourier transform
SVM	Support-Vector Machine
TFFNet	Time-Frequency Feature Network
TN	True Negative
TP	True Positive
UATC	Underwater Acoustic Target Classification
VG	Vanishing Gradient
VGG	Visual Geometry Group
VPCN	Vehicle Propeller Cavitation Noise

CONTENTS

1 INTRODUCTION 16
1.1 Objective 18
1.2 Thesis Structure 18
2 UNDERWATER ACOUSTICS AND DEEP LEARNING BACKGROUND 20
2.1 Fundamentals of underwater acoustics 20
2.2 Main concepts in classic acoustic processing methods 23
2.2.1 Mel Spectrograms 25
2.2.2 Gammatone Spectrograms 26
2.2.3 Constant Q Transform 26
2.3 An introduction to Deep Neural Networks 28
2.3.1 AlexNet 30
2.3.2 VGG 31
2.3.3 ResNet 31
2.4 Evaluation metrics for Machine Learning detection and classification 32
3 LITERATURE REVIEW ON UNDERWATER ACOUSTIC CLASSIFICATION 35
3.1 Classification using frequency and time-frequency analysis 35
3.2 Classification using biologically-inspired feature extraction filters 38
3.3 Classification using only CNNs 42
3.4 Available datasets 45
4 A VESSEL-TYPE CLASSIFICATION METHOD 48
4.1 Gap statement 48
4.2 Motivation 48
4.3 Development steps for underwater acoustic classification 50
4.3.1 Baseline development for vessel-type classification with existing datasets . 50
4.3.2 Generation of a dataset of underwater acoustic signals 53
4.3.3 A method for underwater classification with the proposed dataset 57
5 EXPERIMENTS AND RESULTS 61
5.1 Environment setup 61
5.2 Results of the vessel-type classification with Machine Learning 61
5.2.1 Baseline development on DeepShip dataset 61

5.2.2 **Deep Learning method using the proposed dataset** 63
6 **DISCUSSION** 70
7 **CONCLUDING REMARKS** 74
REFERENCES 76
INDEX 86

1 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 (ORENGO; GARCIA-MOLSOSA, 2019), the inspection of underwater structures for the offshore industry (XU et al., 2016; CHEN et al., 2019), the surveillance of shorelines (DÄSTNER et al., 2018), the identification of vessels (CHOI; CHOO; LEE, 2019), as well as applications in environmental sciences, like counting and classifying the marine life behaviour for biological research (TERAYAMA et al., 2019). Also worth mentioning are studies relating the acoustic signals in the sea to environmental pollution, affecting not only the marine life (ERBE et al., 2019; MERCHANT et al., 2014; ROSSI et al., 2016), but also the human activities in port areas (NASTASI et al., 2020; MCKENNA et al., 2012; BOCANEGRA et al., 2022). In this context, the present thesis selected the identification and classification of vessels from underwater acoustic data as the domain of interest.

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 cost and the accurate instrument calibration are still challenges for satellite imagery (DUBOVIK et al., 2021). 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 vessel-emitted acoustic signals are harder to omit and its impact can be perceived in many underwater environments (NEENAN et al., 2016; FERRAO, 2018; JÉZÉQUEL; BONNEL; CHAUVAUD, 2021). The effort in this area can also be the initial step toward understanding how those 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 high complexity of the data. Time-frequency representations,

such as those based on the Fourier transform or on temporal data segments (CHILDERS; SKINNER; KEMERAIT, 1977), 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 (BOASHASH, B., 2003; GHOSH et al., 1996), raising the difficulty of the recognition task. In order to cope with this issue, recent studies are mostly based on the application of Deep Learning (DL) methods (YANG, Haesang et al., 2020b,a,c), which became promising methods for 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 (CHAI et al., 2021). The use case presented in this thesis 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 developed for computer vision for acoustic analysis. Numerous DL solutions for the acoustic domain are now based on Convolutional Neural Network (CNN) (RUMELHART; HINTON; WILLIAMS, 1986; LECUN et al., 1990). 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 Visual Geometry Group (VGG)Net (SIMONYAN; ZISSERMAN, 2015) or ResNet (HE et al., 2016) 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 CNN being an interesting approach, its growth depends on the representation used and the model which will receive it. Also, as the sound generated by the targets 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 solution.

The present work describes an investigation of the automated classification of single-channel underwater acoustic signals of four distinct classes of vessels in marine environments, using DL applied on time-frequency representations of the sounds. It also presents a study of different preprocessing filters and an analysis of the results with varying distances from the ships to the sensors.

1.1 OBJECTIVE

The objective of this research is to develop a methodology for performing vessel classification using single-channel underwater acoustic signals while considering environmental variables, such as the distance between the sensor and the vessel, and comparing the accuracy of the various types of existing preprocessing. The main contributions of this work can be summarised as follows:

a) The creation of an open-source pipeline for the classification of vessels from underwater acoustic signals using Machine Learning[1];
b) The comparison of the Adam and Stochastic Gradient Descent (SGD) optimisers for spectrogram analysis;
c) The comparison of the three different preprocessing filters (CQT, Mel Spectrogram, and Gammatone);
d) A combination of the CQT, Mel Spectrogram, and Gammatone representations into a three-channel input signal, generating a more complete input signal to the network;
e) The analysis of the relation between the distance of the object of interest to the hydrophone and the accuracy of the various classification methods evaluated in this work;
f) The proposal 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[2,3].

1.2 THESIS STRUCTURE

The background knowledge needed to understand the theory behind the problem was presented in Chapter 2, followed by an extensive review of the literature, including three main different types of approaches and a summary of the available datasets (see Chapter 3). The gap statement and motivations behind this work were addressed at Chapter 4, together with the description of the solution stages. The detailed description of the preliminary experiments, as well as the results obtained, were reported in Chapter 5. The analysis and discussion of the results were handled in Chapter 6.

[1] `https://github.com/lucascesarfd/underwater_snd`
[2] `https://github.com/lucascesarfd/onc_dataset`
[3] `http://ieee-dataport.org/documents/vtuad-vessel-type-underwater-acoustic-data`

The sections that constitute the Chapter 2 and Chapter 3 have been published in (DOMINGOS et al., 2022b). Summarised versions of the Chapters 4 and 5 were submitted for submission in (DOMINGOS et al., 2022a).

2 UNDERWATER ACOUSTICS AND DEEP LEARNING BACKGROUND

This chapter presents a summary of underwater acoustics, classic signal processing methods and general deep learning algorithms. These constitute the background knowledge needed to understand and analyse the current techniques for the autonomous classification of underwater acoustic data in the maritime domain.

2.1 FUNDAMENTALS OF UNDERWATER ACOUSTICS

The acoustic signal produced by a vessel moving in the sea is mainly composed of a broadband component, which has a continuous spectrum, generated by the propeller and its hydrodynamic interactions. It is also comprised of a narrow band component, whose spectrum consists of line components at discrete frequencies, owing to the propulsion system and other mechanical parts (VAHIDPOUR; RASTEGARNIA; KHALILI, 2015). The automatic classification of this type of signal is a challenging task as the acoustic signal is also dependent on the vessel's speed, the age and state of the propulsion system, the highly variable background noise and the diversity of sound propagation mechanisms in the ocean. The latter aspect is also a source of complexity in active sonar applications. As identified in Section 3 below, the characteristics of underwater sound propagation, however, is largely ignored in most literature concerning machine learning-based underwater acoustic signal classifiers, even though it is essential for interpreting the accuracy of the classification results. To provide an appropriate context for this issue, this section presents a summary of underwater acoustics, mainly based on (KUPERMAN; ROUX, 2007; ABRAHAM, 2019; URICK; AGENCY, 1979).

Sound propagation in the ocean is dependent on the properties of the water column, such as temperature, salinity and pressure, and effects related to the ocean floor, such as scattering and reverberation. For a temperature T (measured in Celsius), a depth below the surface z (in meters) and salinity S (parts per thousand), the sound speed c in the ocean can be represented by the following function (ABRAHAM, 2019)[1]:

$$c = 1449.2 + 4.6T + 0.055T^2 + 1.39(S - 35) + 0.016z. \tag{1}$$

[1]N.B., Equation 1 is one possible way of defining the dependency of the speed of sound with oceanographic variables, other formulations are presented in (LEROY; ROBINSON; GOLDSMITH, 2008).

Sound rays respect Snell's law of refraction, which expresses that the rate between the cosine of the ray angle $\theta(z)$, with respect to the horizontal plane, at local sound speed $c(z)$ and depth z, is constant:

$$\frac{cos\theta(z)}{c(z)} = constant. \tag{2}$$

Therefore, a negative sound-speed gradient (e.g. in a thermocline) causes sound rays to bend downwards; the opposite happens with a positive gradient (e.g. in a deep ocean where changes in pressure are more significant than temperature changes). In other words, sound bends toward regions of low sound speed (KUPERMAN; ROUX, 2007). This implies that distinct sound profiles should be considered for warmer and colder geographic areas, periods of the day or seasons of the year. Ocean volume also causes attenuation of the sound signal, which is directly proportional to the squared acoustic frequency.

In more general terms, the influences of oceanographic properties on the sound propagation paths can be classified into three primary classes: *very short range*, *deep water*, and *shallow water propagation* (KUPERMAN; ROUX, 2007). Very short range propagation includes the direct and the surface reflected paths, where the former refers to sound waves that travel without interacting with the sea surface or bottom; the latter refers to sound waves that are reflected by the air/sea interface. The interference of these two paths creates a Lloyd mirror pattern effect (MCMAHON et al., 2012), which makes object detection and classification in shallow waters, such as in a harbour or port, a challenging task (FERGUSON et al., 2017). Deep water (or long-range) propagation paths can mostly be characterised by Snell's law of refraction, apart from the *bottom bounce* effect. In this effect, sound rays are reflected by the ocean floor in a process dependent on the signal frequency and sensitive to seabed characteristics. In shallow waters, where the depths are up to a few hundred meters from the surface to the ocean bottom, the effects of surface reflection, bottom bounce and distinct temperature regimes in distinct seasons have to be taken into account. During the summer period, following Snell's law, sound rays bend more toward the bottom than during the winter months. This implies that the bottom bounce effect is more prominent during the hotter periods of the year and, therefore, the sound propagation in shallow waters has higher losses in the summer than during the winter. The rough winter surface conditions have also to be considered, as they generate considerable scattering losses at high frequencies (achieving 0.1 db/km at 600 Hz and 1 db/km at 6000 Hz) since more energy is needed to keep the mechanical vibrations in these situations (URICK; AGENCY, 1979).

Scattering, due to rough boundaries or small obstacles, is another process that causes loss in the acoustic signal. In contrast to reflection, scattering happens with wavelengths at the order of the obstacles, causing parts of the acoustic field to be randomised (WEILL, 2014). Scattering due to rough surfaces cause a frequency-dependent attenuation of the acoustic field, whereas volume scattering decreases with depth and presents variations at distinct periods of the day (KUPERMAN; ROUX, 2007), usually due to near-surface bubbles, bubble clouds or biological obstacles, such as air-filled swim bladders and zooplankton.

Classical signal processing methods, such as cepstral analysis (CHILDERS; SKINNER; KEMERAIT, 1977), can attenuate some of the effects of reflection interference and scattering losses, but only at short ranges with a high SNR (FERGUSON et al., 2017). Nevertheless, these effects represent some of the main challenges for sonar systems and they should be considered in the evaluation of any underwater acoustic classifier since misclassifications are likely to result if the confounding effects of temperature, depth and boundary conditions are not taken into account by the system.

Sound Navigation and Ranging (SONAR) systems

In general terms, a sonar system consists of a sensor, or an array of sensors, converting acoustic pressure underwater to electrical voltage (ABRAHAM, 2019). The most common example is the hydrophones, which are essentially transducers or *underwater microphones*.

Passive sonar systems detect acoustic signals emitted by the objects of interest, such as ship's propellers, seismic/volcanic or biological signals, for instance. The application of these systems assumes that all the information necessary for the detection, classification and tracking of underwater objects is available in the signal emitted by the objects themselves. Active sonar systems, on the other hand, have sound projectors that are transducers, converting electrical voltages to acoustic pressures. These projectors emit acoustic pulses, called *transmit waveform*, whose reflections are measured by the hydrophones. In these systems, the information necessary for achieving the inferential objectives reside in the reflection and scattering that the original signal suffers upon interacting with objects in the environment. Any sonar application, involving machine learning processing or not, should take into account the limitations of the sensing apparatus. Usually, this is measured by accounting for each part of the system, including the system's components, the effects of the underwater environment, and the characteristics of sound or scattering from the target. This accountability is summarised in sonar equations 3 and

4 (ABRAHAM, 2019), resulting in the SNR of the system. For a source level (SL), a propagation loss (PL), a noise power level in the processing band (NL) and an array gain (AG) (all measured in decibels), the basic *passive* sonar equation is given by:

$$SNR_{Passive} = SL - PL - NL + AG. \tag{3}$$

The basic *active* sonar equation (Equation 4) also takes into account the target strength (TS), and the propagation losses from the sound projector (PL_a) and from the object of interest (PL_b), considering SL as the source level of the sound projector:

$$SNR_{Active} = SL - PL_a + TS - PL_b - NL + AG. \tag{4}$$

Equation 3 and Equation 4 incorporate the main underwater sonar variables including temperature, depth, salinity, air/sea and bottom/sea interface, rough boundaries and biological obstacles. Although, most existing research on the application of machine learning methods for the classification of underwater acoustic data ignore these effects and do not evaluate the classification tasks concerning the performance of the sonar system in distinct oceanographic situations. Ignoring these variables limits the use of ML methods in applications that demand a high level of trust, such as in defence systems or the inspection of underwater built structures, such as deep-sea mining facilities. In these applications, classic acoustic processing strategies, which do not involve machine learning algorithms, are still the first-choice methods because these effects can be explicitly accounted for.

Many DL algorithms use classic acoustic signal processing strategies to preprocess the sonar data, aiming to extract the most relevant features of the original sound to feed the networks. The subsequent section will describe some of the most common preprocessing methods found in the literature.

2.2 MAIN CONCEPTS IN CLASSIC ACOUSTIC PROCESSING METHODS

Acoustic signals are produced by mechanical vibrations propagating in all directions. They can be captured by transducers and represented as one-dimensional signals that oscillate in amplitude through time. Besides the many applications that can be developed using the temporal information of audio signals, some characteristics of this kind of data are best obtained in the frequency domain. In the particular case of underwater acoustic signal processing, the frequency

domain is usually more informative (ABRAHAM, 2019). To this end, Fourier transforms (FT) are used to obtain the frequency content of a time-domain signal (BOASHASH, Boualem, 2016). The result of an FT on an auto-correlation function is called *power spectrum* that, informally, represents how much of the original signal is at a particular frequency.

Features related to amplitude and frequency can be directly analysed from the power spectrum, but the information on how the frequency varies with time is lost in this representation. To fill this gap, two-dimensional time-frequency features can be extracted by passing a windowing function through the time signal and extracting the FT for each window of the original data. This procedure is called the Short-Time Fourier transform (STFT), and the unified resultant output is called the *spectrogram* of the signal. Spectrograms are typically plotted with time on the x-axis, frequency on the y-axis, and magnitude on the z-axis. They can be thought of as a series of time segmented FT put in chronological order.

The analysis of the harmonic behaviour of the spectrographic data is also an important feature for sound analysis in general (TSAI; LIN, 2010; LEE et al., 2009; GOPALAN, 2005), and can be obtained with the extraction of their cepstrum representation. The cepstrum is the result of the inverse FT of the signal spectrum logarithm and is used to obtain the periodic structures in spectra (OPPENHEIM; SCHAFER, 2004). The cepstrum is also commonly understood as the power spectrum of the logarithm of the power spectrum (ABRAHAM, 2019). As the term cepstrum is a reversion of the first syllable of the spectrum, operations on cepstra are also known as quefrency analysis, a semi-reversed spelling of frequency (BOGERT; HEALY; TUKEY, 1963).

Due to the fact that FT is composed of a sum of sinusoidal functions, this representation is not well localised in time and space, since sine waves are functions with an infinite duration. To improve the representation of signals that have abrupt changes in time and space domains, the wavelet transform was introduced. A wavelet is a wave whose oscillation has a finite duration, which is defined in time and has a zero mean (DONALD et al., 2009). Wavelets can be represented in many different formats, such as morlet, mexican hat, biorthogonal and others. Using the concepts of scaling, representing the frequency and duration of the wavelet, and shifting, representing the time positioning of the wave, this representation can capture both, high and low-frequency features. The output of a wavelet transform is a matrix whose coefficients are functions of the scale and time information. Wavelet decomposition proved to be suitable for analysing signals that contain information at different frequencies and time (TZANETAKIS;

ESSL; COOK, 2001; ENDELT; COUR-HARBO, 2004; CHUAN; VASANA; ASAITHAMBI, 2012).

The approaches cited above were developed to improve the representation of time series and add a frequency component to the resulting signals. However, they are generalised to any one-dimensional signals, not only acoustic ones. Looking to the problem of sound classification, the logarithm representation of the frequency is preferred over the linear, following the analogy with the human auditory system (OXENHAM, 2018). This behaviour demands the development of approaches exclusively dedicated to the acoustic domain. In this context, the present work focused on the application of three typical methods for spectrogram generation based on non-linear frequency scales: The Mel Spectrograms, the Gammatone Spectrograms, and the Constant Q transform.

2.2.1 Mel Spectrograms

The Mel Spectrogram is a representation of the short-term sound power spectrum. Mel's scale (STEVENS; VOLKMANN; NEWMAN, 1937) is empirically based on the way humans perceive sound, on a logarithmic scale. The work conducted by (STEVENS; VOLKMANN; NEWMAN) consisted in submitting observers to different frequencies of sounds and recording their perception and sensitivity to the stimulus. Thus, the resulting scale differs from the musical and frequency scales, as neither is subjective.

There are different mathematical formulations for the conversion between the f in heartz to m in mels, however one of the most popular was first defined in (O'SHAUGHNESSY, 1987), as shown in Equation 5.

$$m = 2595 \log_{10}(1 + \frac{f}{700}) \tag{5}$$

A typical practice in the literature is to represent the frequency variations using the cepstrum of the Mel scale. The coefficients extracted from that cepstrum are called the Mel Frequency Cepstrum Coefficients (MFCC). The general process to extract the MFCC from the signal is described in detail in (LOGAN, 2000). Informally, MFCC is obtained by first sampling the original signal and then extracting the amplitude spectrum of each sample. After this process, the signal amplitude is converted to a logarithmic scale and the Mel scale. Finally, a Discrete Cosine Transform (DCT) is taken to obtain the final form of the MFCC.

One common criticism of the Mel Scale is that the initial experiment was harmed by methodological bias (GREENWOOD, 1997), as the five observers (two authors and three others) were not instructed to listen to the notes on ascending and descending orders. However, this scale remains commonly used in analysis involving speech recognition and music processing, where human perception is especially relevant, due to its characteristic of emphasising audible sound on the proper scale. The following section, on the other hand, will address another audio processing method, which was mathematically based on the human perception of sounds.

2.2.2 Gammatone Spectrograms

The Gammatone filter was first defined as a suitable filter bank capable of representing the shape of the impulse response function of the human auditory system (JOHANNESMA, 1972). A gammatone function can be obtained with the mathematical formulation presented on Equation 6.

$$g(t) = t^{n-1} \exp\left(-2\pi b_i t\right) \cos\left(2\pi f_i + \alpha_i\right) \tag{6}$$

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 6 was used by (JOHANNESMA) to summarise the Reverse Correlation (revcor), a representation of the correlation between a sound stimulus on the human ear and the response of a primary auditory fibre (PATTERSON et al., 1987). The first element of the Equation 6 represents a gamma function, and the cosine element represents the tone of the stimulus. This representation has an amplitude characteristic that can predict human auditory masking data well, and has the minimum-phase characteristic, which (PATTERSON et al., 1987) argues is a preferred feature for an auditory filter bank.

Gammatone filter-banks facilitate the representation of the human response to the acoustic signals, as gamma filters are broader on lower frequencies and narrow on higher ones, emphasising the lower spectrum. The section below describes another approach to sound representation, not based on human perception but musical scale properties.

2.2.3 Constant Q Transform

The harmonic frequency components of sounds are geometrically spaced, following the same distance proportion independently of the fundamental frequency. As FT represents a bank of

sinusoidal filters equally spaced in frequency, this property makes the linear representation of frequencies, such as the STFT spectrograms, inefficient in properly representing acoustic signals. Taking the *C* and *D* musical notes as an example: on the lower octaves the *C0* (16.35 Hz) and *D0* (18.35 Hz) are only 2 Hz apart from each other, however, in the higher octaves, *C8* (4186 Hz) and *D8* (4699 Hz) have a linear difference of 513 Hz. This points out that the linear spacing cannot emphasise both octaves on the same representation using fixed frequency distances.

The solution proposed on (BROWN, 1991), the Constant Q Transform (CQT), uses a constant Q to create a representation, improving the visualisation resolution while providing the means to solve the problem of fundamental frequency identification, aiming to represent the human perception of sounds. The CQT, in contrast with the FT, is a bank of geometrically spaced filters, where, for the k-th filter, central frequencies can be evaluated with the following equation:

$$f_k = f_0 \times 2^{\frac{k}{b}} \tag{7}$$

Where b represents the number of filters per octave. From the Equation 7 is possible to derive the relation between the distance of two adjacent filters, as shown in Equation 8.

$$\Delta_k = f_{k+1} - f_k = f_k(2^{\frac{1}{b}} - 1) \tag{8}$$

The quality factor Q (or constant Q) is then defined as the ratio of frequency to resolution, as stated by the Equation 9.

$$Q = \frac{f_k}{\Delta_k} = (2^{\frac{1}{b}} - 1)^{-1} \tag{9}$$

The properly tunning of the quality factor Q can supply the needed information for 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 (BROWN, 1991). This representation also increases time resolution towards higher frequencies, resembling the auditory system and emphasising lower frequencies, which is a preferred characteristic for underwater acoustic problems where the sounds to be analysed are mostly located on the lower frequency bands (see Section 2.1).

When applied correctly, time-frequency analysis can yield useful insights into data. It does this by extracting and emphasising important signal characteristics such as how frequency components (spectrogram) or harmonics of the frequency components (cepstrogram) change over time. However, these approaches are nothing more than changing the way the data is represented

and presented. They do not perform any sort of classification or categorisation of the information within the data. Traditionally, those higher-level pattern matching tasks were performed by humans or basic linear or statistical models. However, over the last few years machine learning in general, and neural networks in particular, have become the dominant way to classify patterns in noisy data. The following section will cover the basic understanding of deep neural networks.

2.3 AN INTRODUCTION TO DEEP NEURAL NETWORKS

There has been significant recent developments in DL methods, mostly pushed by object detection and classification in images (LIU; XU; WANG, 2019; HOSSAIN et al., 2019) and visual question answering (WU et al., 2017). Spectrographic and cepstrographic representations of acoustic data are analogous to visual images, representing a signal across two dimensions (x, y spatial dimensions for images and time, frequency for spectrograms). Therefore, to introduce applications of DL in the classification of sonar data, a brief introduction to the application of these methods in image classification is in order.

Informally, a CNN (RUMELHART; HINTON; WILLIAMS, 1986; LECUN et al., 1990) extends the concept of Artificial Neural Network (ANN) adding a set of layers that function as *feature filters* (the convolutional layers) that, by enhancing specific aspects of the input, learn the features to be classified. More specifically, a CNN is usually structured in many stages. Initially, the input data is processed by a sequence of convolutional layers with activation functions, and pooling layers. The convolutional layers compute a sum of products, named *convolution*, with their input and a set of kernel weights, called *receptive fields*. This operation is applied at every spatial location of each element of the input data using several spatial increments called *strides*. A bias value is added to the result of this operation at each location, and the final value obtained is passed through an activation function. Activation layers controls what information should be reinforced or attenuated to the next neurons based on a simple mathematical operation. Initially, most computationally expensive functions were used in DL, such as the Sigmoid (mapping the values to ranges from 0 to 1) or Hyperbolic Tangent (producing a range from -1 to 1). Nowadays, the Rectified Linear Units (ReLu) offered a significant improvement in training time, as it is a simple $\max\{0,x\}$ function which does not need expensive calculations to be evaluated. Improvements to the ReLu function are being used to minimise the impact of the zeroed side of the function, which harm the gradient descendent for those neurons. The Leaky ReLu is one of the most commons, as it attributes a small constant to the zeroed side, defined as $\max\{0.1x,x\}$.

The results of the activation layer generates the inputs to the next layer as an array called *feature map*. Feature maps represent the extracted features from the original input (e.g. edges, points or blobs, considering images). The Pooling layers are responsible for merging similar features by a sub-sampling process, which can be based on the mean values of each feature map (representing an Average Pooling) or based on the maximum values of the feature maps (producing a Max Pooling).

Usually, the final stage of a CNN is a traditional fully-connected neural network, that learns class probabilities for the features picked out by the convolutional layers. The learning process of CNN occurs as an extension of back-propagation in fully-connected neural networks, which propagates the gradient of an objective function through all the Neural Network (NN) parameters (weights). This procedure is computed by gradient descent, for which various optimisation methods exist, such as Stochastic Gradient Descent (SGD) and Adaptive moment estimation (Adam), amongst others (SUN et al., 2020). SGD, the most common optimiser used in the literature, is an iterative method that starts randomly and seeks the minimum value in the input function. Adam, on the other hand, is an extension of the SGD based on the combination of the Adaptive Gradient Algorithm (AdaGrad) and the Root Mean Square Propagation (RMSP), claiming to be *"straightforward to implement, is computationally efficient, has little memory requirements, is invariant to diagonal rescaling of the gradients, and is well suited for problems that are large in terms of data and/or parameters"* (KINGMA, Diederik P; BA, 2014). Recently, an important stage on DL architectures became popular: the normalisation layer. Normalisation aims to transforms the set of available data to a similar scale, aiming to stabilise the gradient descent step and improve the convergence. Some of the most common approaches are the Batch Normalisation (which uses the batch information to re-scale the data) and the Layer Normalisation (which uses the layer/channel information).

Another issue that any work in DL should consider is overfitting, which occurs when the model fits exactly with its training data but does not generalise well to data outside the training set (BROWNLEE, 2018). There are numerous strategies to tackle this issue, such as *early stopping*, which consists in stop the training before the model incorporates noise; *expand the training set*, as more data implies a more accurate model; *feature selection*, which identifies the most relevant features to be learned, ignoring redundant ones; *ensemble methods*, that aggregate the output of a set of classifiers, selecting the best output by a voting process; and, finally, *regularisation*, that, in general, limits the amount of variance in the model by penalising input parameters with large coefficients. Regularisation methods have received great attention from

recent research (TIAN; ZHANG, 2022) since they are solely related to the algorithms, and not to data quality or classifier competitions. One particular regularisation method worth mentioning in this thesis is *Dropout*, which ignores randomly chosen neurons, with a certain probability, during the training phase, so that a reduced network is obtained as a result. This strategy reduces the possibility of co-dependency among neurons, which is one factor that causes overfitting.

Most DL models for object recognition that exist today are variations of the basic CNN structure briefly explained above, with a distinct number, type and distribution of convolutional and pooling layers, or with some extra constraints on the way the data is processed. There are three most prominent CNN architectures that have a considerable influence on the image classification area: AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012); VGG network (SIMONYAN; ZISSERMAN, 2014); and ResNet (HE et al., 2016).

2.3.1 AlexNet

The AlexNet model (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) was proposed to solve the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (RUSSAKOVSKY et al., 2015), a competition for the automatic object detection and classification based on the ImageNet (DENG et al., 2009) database. The model achieved an accuracy of more than 10% better than the other state-of-the-art proposals in the competition. The original AlexNet architecture was composed of eight layers with learnable parameters, comprising five convolutional layers (with kernel sizes of 11x11, 5x5, 3x3, 3x3, 3x3), with layers 1, 2 and 5 each followed by a Max Polling layer, and finally three successive fully-connected layers.

One of the major contributions of this model was the introduction of the ReLu function as the activation layer. Before AlexNet, the most commonly used activation layers were based on *sigmoid* functions and *tanh* functions. ReLu offered a significant improvement in training time and also minimised the Vanishing Gradient (VG) problem (HOCHREITER et al., 2001). As the DL networks go deeper (increasing the number of layers), there is a tendency for the value of gradients in the backpropagation algorithm to decrease. Due to the limits of *sigmoid* (0 to 1) and *tanh* (-1 to 1), the values of the derivatives vary between 0 - 0.25 and 0 - 1, respectively, which causes the weight updates to be smaller as the error is propagated to the inputs, impacting the ability of the first layers to learn. The ReLu function does not have the same limitation, since it has gradient 1 when its input is greater than zero, and 0 otherwise, and so the backpropagation value does not diminish as it moves through the layers of the network. This finding paved the

way for other DL applications, which became to be vastly applied to image computing. The next section will describe one of the improvements which used the AlexNet model findings as a starting point.

2.3.2 VGG

Improving the idea behind the AlexNet construction, the VGG (SIMONYAN; ZISSERMAN, 2014) was proposed to cope with the need to reduce the number of parameters necessary to obtain each feature map. The idea behind this architecture is that the features learned with larger kernel sizes can be reproduced using a combination of smaller kernels, such as the 3x3 reducing the number of parameters to be computed. A reduced set of parameters represent a faster execution and an improved robustness to overfitting. VGG-based methods can perceive granular spatial relations on images due their kernel size, the smallest possible size to capture the four different directions (up, down, left, right). This reduced kernel size also produces a good trade-off between classification accuracy and hyperparameter complexity.

The VGG model is composed of blocks with the combination of convolutional layers attached to ReLu, followed by Max Pooling layers, as the feature extraction stage. The classification stage is composed of fully-connected layers and a softmax activation function. The most common architectures are the VGG-16 and the VGG-19, where 16 and 19 represents the number of convolutional layers. The universal approximation theorem (CYBENKO, 1989) states that a deep enough neural network is capable of approximating any complex function, therefore, besides the smaller receptive fields, the VGG-based methods were becoming deeper as more convolutional layers were added to them. However, the vanishing gradient and the accuracy degradation problems become problematic, as more layers are added to the models. The ResNet model, described below, addressed this issue.

2.3.3 ResNet

The use of ReLu activation functions greatly improved the robustness of deep learning models concerning the VG problem, however, DL models were still challenging to train, and the accuracy of deeper models failed to perform better than shallower networks. ResNet (HE et al., 2016) addressed this problem by introducing an *identity shortcut connection* between the convolutional layers, bypassing one or more layers in a forward pass, defining the residual blocks. These blocks facilitate the learning ability of the intermediary layers, reducing the VG problem, and penalising

the ones that could potentially degrade the accuracy (HE et al., 2016). Residual blocks are based on using a copy of the input data as base elements for the output values, being composed of two sequential weight layers and one shortcut (identity layer) which connects the input directly to the output of the block. Sometimes, the weight layers, which can be convolutional, will not have the same dimensions, so a different type of shortcut is used, such as a convolutional with size reduction or even padding applied on the input feature.

The residual blocks aim to solve the *degradation problem*. This states that the accuracy function increases with network depth only up to a certain number of layers, after which, adding more layers to the network causes the accuracy to decrease and the loss to increase relative to the shallower network. This is due to the complexity present in deeper models, which impacts the learning ability of the back-propagation algorithm. Identity shortcuts affect the updating of weights in the intermediate layers (those that are in parallel with the identity), generating adaptability of the learning process only in cases where accuracy is improved. Thus, only the relevant layers (in terms of accuracy improvement) are considered in the final forward-pass of the trained network, and those layers that cause the degradation generally converge to zero weights.

The trade-off between model size and performance must be taken into account when developing a DL solution, aiming to optimise the computational resources while finding an optimal prediction result. Therefore, the analysis of the results obtained during the experiments is an indispensable aspect of DL pipelines, as they will conduct the development. The following section will cover the most commonly used metrics applied to the classification systems evaluation.

2.4 EVALUATION METRICS FOR MACHINE LEARNING DETECTION AND CLASSIFICATION

Results of ML classification procedures in general can be summarised by a confusion matrix, which is a two-dimensional matrix representing in one of its dimensions, usually its *rows*, the correct instances of each class, whereas the other dimension, usually its *columns*, represents instances of the predicted class. They can also be normalised to show the predictions in a 0 to 1 range. In a binary classification task, with a distribution of P real positive instances and N real negative instances, the entries in a confusion matrix represent the number of True Positive (TP), which are the correct predictions of the positive class, shown in its main diagonal; True Negative (TN), which are the correct predictions of the negative class; False Positive (FP), representing the wrong predictions of the positive class; and, False Negative (FN), which are wrong predictions

of the negative class. The most usual metrics for evaluating a classification task are based on calculating rates from subsets of these values, such as *accuracy* (Equation 10), that measures the overall portion of correct classifications; *precision* (Equation 11) that is the fraction of *correct* positive instances identified by the classifier over all instances identified as positive for each class; *recall*, or *sensitivity* (Equation 12), that measures the fraction of correct positive instances identified by the classifier over all positive instances for each class; among others (POWERS, 2008; SOKOLOVA; LAPALME, 2009).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{11}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{12}$$

The *F1 score* (Equation 13) provides the weighted harmonic mean between precision and recall. Also important for evaluation purposes is the concept of *support*, which is the number of instances of each class in the test dataset. An imbalanced support has to be corrected, e.g. by sampling or rebalancing, in order to avoid bias in these metrics.

$$\text{F1-score} = \frac{Precision \times Recall}{Precision + Recall} \tag{13}$$

The metrics cited above can be extended to multiclass classification tasks by *micro-averaging* or *macro-averaging*. The latter treats all classes equally by calculating the metrics for each class independently before taking the average, whereas the former averages over the combined contributions of all classes (SOKOLOVA; LAPALME, 2009). The area under the precision-recall curve for each class gives another important metric called Average Precision (AP), whereas the Mean Average Precision (mAP) score is obtained by taking the mean of AP over all classes.

Sokolova and Lapalme (2009) present an analysis of 24 evaluation metrics for ML classification tasks, showing the existence of invariance properties across the metrics. These properties allowed the definition of a taxonomy matching the appropriate metrics with qualities of the datasets, such as representativeness of class distribution, reliability of class labels, and the unimodality or multimodality of the classes. Besides that, current literature mostly assumes accuracy or mAP as the main evaluation metric.

The theory addressed in this chapter provides the needed background for the basic understanding and application of DL for underwater acoustics. The next chapter will cover the available DL solutions in the literature, providing a categorisation according to the most common approaches.

3 LITERATURE REVIEW ON UNDERWATER ACOUSTIC CLASSIFICATION

This chapter contains a review of the methods used for classification of underwater acoustic signals. These methods are separated in three main classes: methods which uses data from frequency and time-frequency analysis; methods that apply biologically-inspired feature-extraction filters as pre-processing; and methods that use convolutional Neural Networks alone. Tables summarising these methods are presented below, relating the datasets used, the highest test result reported and a short description of the main contribution of each work. Then, a review of the available datasets will be presented.

3.1 CLASSIFICATION USING FREQUENCY AND TIME-FREQUENCY ANALYSIS

Frequency and time-frequency analysis has been used to pre-process acoustic signals in order to enhance the most distinctive data features, facilitating the classification process. Table 1 presents a summary of important methods related to this idea, as described below.

Conventional algorithmic approaches to vessel classification and ranging from hydrophone data perform well at close range (<180 m) but their performance falls off quickly at longer ranges. These methods are also inaccurate in the presence of noise (FERGUSON; LO; THURAISINGHAM, 2005). Using cepstrographic inputs to a CNN can overcome the shortcomings of the conventional approaches. The combination of cepstral filtering and CNNs greatly extends detection ranges and robustness to noise compared to an algorithmic method, but has slightly reduced accuracy at short range due to expressive bias (FERGUSON et al., 2017). Analogously, the performance of a standard CNN in the classification of surface Vehicle Propeller Cavitation Noise (VPCN) in shallow water has shown to be improved by analysing the amplitude variation of signal to detect the fundamental frequencies of VPCN (BACH; VU; NGUYEN, 2021). Results described in (BACH; VU; NGUYEN, 2021) suggest that this method outperforms traditional classification methods. However, the fundamental frequencies of the target object need to be known beforehand, and it is assumed that they are independent of environmental conditions, which may not be true in the general case.

Ship classification from acoustic data is considered a binary classification problem in (CHOI; CHOO; LEE, 2019) aiming at identifying surface from underwater vessels. As the audio signals are received on a N-element vertical line array, each element of the original array is considered to generate two initial representations: the phone-space and the mode-space Cross-Spectral Density Matrices (CSDM) (pCSDM and mCSDM, respectively). A CSDM is a

time-frequency representation used to compare different signals, similar to the covariance matrix, representing the distribution of power across the frequency spectrum over the time. The absolute values of each CSDM element were used as the source for generating the two other matrices. Altogether, the generated matrices were used as input to several distinct machine learning models: Random Forest, Support-Vector Machine (SVM), a Feed-forward NN and a CNN. All algorithms presented excellent performance, with an error rate below 5%. The training procedure employed in (CHOI; CHOO; LEE, 2019) used only simulated data considering a source with high SNR (2dB to 5dB); whereas the validation task was performed on a pair of simulated datasets: the first had a SNR of 2dB to 5dB; and the second had a SNR of -1dB to 0dB. The results confirm that the proposed machine learning methods were robust with respect to SNR. However, it is unclear

Table 1 – An overview of ML methods for classification from frequency and time-frequency analysis. In the 'Metric' column, ACCY stands for accuracy, AP stands for Average Precision, and mAP stands for Mean Average Precision.

Reference	Preprocessing	Network	Dataset	Metric	Main contributions
(FERGUSON et al., 2017)	Cepstrum and Cepstrogram	CNN	Recorded boat radiated noise[a]	99.78% AP	The cepstral representation of inputs for a NN and an improved distance estimation from acoustic noisy sources.
(CHOI; CHOO; LEE, 2019)	pCSDM and mCSDM	CNN, RF, SVM and FNN	Simulated acoustic data[a]	>95% ACCY	Comparison between CNN and FNN.
(MIAO et al., 2021)	ACT	TFFNet	Whale FM	92.1% mAP	The combination of ACT with the EFP to generate a high resolution time-frequency representation.
(KIM et al., 2021)	DWT	CNN	Underwater acoustic signals[a]	100% ACCY	The wavelet transform to obtain noise robustness and data augmentation as data source. The results converged with an 8-fold reduction in the number of epochs.
(CINELLI; CHAVES; LIMA, 2018)	Spectrogram, delta and delta-delta frequencies	Fully-Connected and CNN	Brazilian Marine dataset[a]	88.1% AP	A comparative study of the accuracy of different ML architectures for various input layers.
(VAHIDPOUR; RASTEGARNIA; KHALILI, 2015)	Image Histograms	Fully-Connected	Five class acoustic dataset[a]	95.13% ACCY[b]	The generation of a short-time Fourier transform-based binary image as input, improving noise robustness.
(BACH; VU; NGUYEN, 2021)	Signal demodulation	CNN	ShipsEar	90% ACCY	An algorithm for detecting the fundamental frequencies of a signal according to the amplitude variation, improving the performance of a CNN.

[a] Dataset is proprietary and unavailable for reproduction.
[b] Results obtained at 10dB of SNR.

whether this result would generalise to real sonar data, under various environmental conditions. Also, results suggest that phone-spaced data inputs are more suitable to ML methods than mode space, as the former are independent from environmental information.

In the need to improve the time frequency representation of underwater acoustic signals, a method for sonar classification based on Anisotropic Chirplet Transforms (ACT) and Deep Learning was proposed (MIAO et al., 2021). Chirps are signals whose frequency increases or decreases with time. Similar to the relation of Waves to Wavelet (cf. Section 2.2), a chirplet is the time limited representation of a chirp. They were first proposed to identify small icebergs fragments using sonar data (MANN; HAYKIN, 1991), as those signals propagate on a radial chirp waveform. ACT was used in (MIAO et al., 2021) as a pre-processing stage for underwater acoustic communication dataset and for identifying whale sounds, then it was fed into a deep CNN, called Time-Frequency Feature Network (TFFNet). Also, an Efficient Feature Pyramid (EFP) technique was defined by aggregating the context information of the features maps at different scales, widening the network instead of increasing its depth. This caused more features to be learned from the input data. The combination of ACT with the EFP method was compared against Random Forest and SVM, achieving higher accuracy with less resources (memory and time). This result was attributed to the fact that ACT generated a high resolution time-frequency representation, providing features that would not be achieved using other known transforms, such as STFT.

Discrete Wavelet Transform (DWT) was used in the classification of underwater acoustic signals with noise robustness in a method composed of four stages (KIM et al., 2021): white noise elimination based on the DWT; an imaging stage, where the spectrogram of the discrete wavelet coefficients was obtained; a data augmentation method; and the final classification stage. The data resulting from this four-tier process was, then, used as input to a standard CNN, composed of four convolutional layers, a dropout and a fully-connected layer. The use of DWT was shown to be effective at enhancing noise robustness in the system, achieving 99.7% classification accuracy with a SNR of 0dB, greatly outperforming other 6 different CNN architectures. The reason behind the improvement on noisy environments was attributed to the application of a threshold, compromised between the Garrote and hyperbola functions, and also to the DWT coefficients on the noise elimination stage, impacting directly on the performance of the wavelet denoising. However, the source of the original data used as input is unclear.

When using spectrograms of hydrophone data a basic four convolutional layer CNN has been shown to outperform a fully-connected NN with either 0 or 512-neuron hidden

layers (CINELLI; CHAVES; LIMA, 2018). Furthermore, improved results were observed when both spectrograms and delta frequency images (i.e. the first difference of signal features, an approximation of the first derivative) were used as inputs over the use of spectrograms alone or in combination with delta-delta frequency images (i.e. the second difference of signal features). However, due to possible overfitting of the system resulting from low databases size, high inter-class similarity and a large number of model parameters, this finding may have limitations. This problem could be avoided using data augmentation techniques, but the lack of background noise information provided limited the application of such techniques in this context.

A fully-connected neural network was applied to recognise ships in hydrophone data (VAHIDPOUR; RASTEGARNIA; KHALILI, 2015). In order to do this, the original signal was treated by traditional image processing methods, mainly a STFT, to generate a binary image, based on the frequency spectrum of signal segmentation. This process was shown to improve the analysis of short-time transient sound behaviours, generating a better representation of the signal to be inputted to the binary image generation. An experiment with three different SNR (without noise; 5dB and 10dB) compared the accuracy of the proposed model with a fractal-based method, showing that NN have a better noise robustness for all analysed scenarios, and also indicating that, for the five class classification problem, the feature extraction method proposed achieved a high performance recognition with just 11 features. Although the innovative feature extraction method presented better results, the characteristics and origin of the database used are not discussed, therefore the results cannot be properly evaluated or repeated.

Research has found that the classification of acoustic signals is improved when the data is presented in a time-frequency representation, such as a spectrogram or cepstrogram. Moreover, pre-filtering in the time-frequency domain further improves classification as does the inclusion of convolutional layers within the classifier model.

3.2 CLASSIFICATION USING BIOLOGICALLY-INSPIRED FEATURE EXTRACTION FILTERS

Using as analogy the methods by which biological systems filter audio signals, a class of classifiers has evolved that uses bio-inspired filters as a pre-processing step for detection and classification (Table 2).

Ship-type classification from hydrophone data using CNNs was investigated in (SHEN et al., 2020; SHEN; YANG; LI, 2019; SHEN et al., 2018), where an algorithm was proposed

that used a cochlea model to decompose the acoustic signal into target features. This was implemented as a convolutional layer with auditory filters that resemble the Gabor Neural Network Le et al. (2020), which uses Gabor filters to assist the detection of mine-like objects. The original (non-annotated) data used in this study is freely available at Ocean Networks Canada (ONC)[1]. The results showed that the proposed "auditory-inspired CNN" resulted in a

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

Table 2 – An overview of ML methods for classification using Biologically-Inspired filtering algorithms. In the 'Metric' column, ACCY stands for accuracy, and AP stands for Average Precision.

Reference	Preprocessing	Network	Dataset	Metric	Main contributions
(SHEN et al., 2020; SHEN; YANG; LI, 2019; SHEN et al., 2018)	Cochlea model	Auditory Inspired CNN	Ocean Network Canada signals[a]	87.2% ACCY	The use of Gabor filter layers inspired on the a Cochlea model.
(IRFAN et al., 2021)	Cepstros, Mel spectrograms, MFCC, CQT, GFCC and Wavelets	SCAE	DeepShip	77.53% ACCY	The proposal of a new open source dataset and the comparison of various preprocessings.
(WANG, D. et al., 2018)	Mel-filter bank	Inception, Xception, VGG and Densenet	Whale FM	84.40% ACCY	The accuracy comparison of traditional methods and CNNs for the same scenarios.
(KHISHE; MOHAMMADI, 2019)	MFCC	Fully-Connected	Sonar dataset[b]	97.12% ACCY	The accuracy comparison of meta-heuristic algorithms and the use of fully connected NN.
(WANG, X. et al., 2019)	GFCC and MFCC	Fully-Connected	Six class dataset[b]	94.3% ACCY	A combination of MFCC and GFCC was used as feature extraction showing time performance improvement.
(SUN; WANG, 2022)	STFT, Mel-log and MFCCs	ResNet and DenseNet	ShipsEar	97.49%	The classification of synthetic mixed multitarget signals using CNNs.
(WANG, D. et al., 2019)	Mel-log, delta, and delta-delta	ResNext101	Shallow water Data[b]	85.00%	Direction-of-arrival prediction based on acoustic signals.
(CHEN, L. et al., 2022)	MFCC, CQT, Gammatone, and Log-Mel	CNN+DNN	Hydrophone datasets[b]	89.9%	The generation of a pipeline with the combination of the preprocessed features for acoustic target classification.
(HAN et al., 2022)	Mel-log, MFCC, chromatogram, spectral contrast, and Tonnetz	CNN+LSTM	ShipsEar	92.17%	The proposal of a joint neural network composed of CNN and LSTM, applied on a combination of the preprocessing strategies.

[a] Dataset uses public data with non-public preprocessing techniques.
[b] Dataset is proprietary and unavailable for reproduction.

high accuracy (87.2%), outperforming a standard CNN (83.2%), a stack of autoencoders (CAO et al., 2016) (81.4%) and various traditional object recognition algorithms, such as a combination of wavelet transform with Principal component analysis (PCA) (SHI; XU, 2013) (74.6%) and a SVM (YANG, H. et al., 2016) (68.2%). However, Shen et al. (2020) do not report the exact devices that were used, the precise period of data collection, or the labelled data used. Therefore, even though the original raw data is freely available, it is impossible to reproduce the results reported in that work.

The work proposed in Irfan et al. (2021) fills the gaps opened by the work of Shen et al. (2020), involving the creation of an open-source reference dataset. The final set, called DeepShip, is dedicated to classifying vessel types using underwater acoustic signals, and consists of approximately 47 hours of recordings from 265 different vessels belonging to four classes: Cargo, Passenger Ship, Tanker and Tug. For classification, the authors proposed a Separable Convolution Autoencoder (SCAE). Using six different signal preprocessing techniques (Cepstros, Mel spectrograms, MFCC, CQT, Gammatone Frequency Cepstral Coefficients (GFCC) and Wavelets), this model proved to have a better accuracy in the classification task, achieving 77.53% for CQT filter, when compared to other traditional models, such as CNN (76.35%), SVM (72.24%) and the K-Nearest Neighbors algorithm (KNN) (62.71%).

The classification of large-scale acoustic signals was evaluated in (WANG, D. et al., 2018) with a comparison among four popular CNN models (VGGnet, Inception, Xception, and DenseNet) and a classical baseline approach. The main aim of this comparison was to evaluate the improvement of using CNNs instead of classical acoustic classification methods. *Mel-filter bank features* were used to extract features from the signals. As the Mel scale emphasises lower frequencies over the higher ones (approximating the human auditory perceptions), using Mel-filter banks resulted in more lower frequency filters than filters on higher frequencies. This work showed that the logarithmic encoding of frequencies produced results 10 percentage points higher than linear encoding, which is further evidence that the use of preset pre-processing stages can result in improvements over random and generic starting points for a NN.

As CNNs can be understood as a combination of feature extraction layers (convolutions) and a classifier (the final densely connected layers), a few other recent contributions investigate the use of a single fully-connected neural network to classify passive sonar data, receiving the extracted features from bio-inspired methods as inputs. For instance, a fully-connected neural network was used in (KHISHE; MOHAMMADI, 2019) to classify data from two hydrophones picking out signals from three distinct propellers running on on a closed-water circuit cavitation

tunnel. The raw data from the hydrophones were pre-processed into a feature vector formed by the discrete cosine transforms of the MFCC. The classifier used meta-heuristic algorithms for optimising the weights and biases of the NN, aiming to avoid local minima, while accelerating learning convergence. The results presented were only compared with other meta-heuristic algorithms training the same neural network, but not against state-of-the-art deep learning methods. The data used is not available to reassess these results, making reproduction, or even comparison with the performance of modern classification systems, impossible. Nevertheless, the use of meta-heuristics (usually inspired by insect behaviours) for optimising the NN training opens an avenue for investigating the application of these methods to finding the optimal structure of a NN, or for fitting the hyperparameters of more complex architectures (KHISHE; MOSAVI; MORIDI, 2016).

MFCC can be used as input to fully-connected neural networks and directed to the problem of underwater acoustic recognition of ship types, marine mammals and underwater background noise with weak targets (WANG, X. et al., 2019). The method proposed in (WANG, X. et al., 2019) used a combination of GFCC and mode decomposition on the MFCC to extract features from the signal, that were then input to the neural network. Results obtained with this method suggest that by not using a CNN to learn the features saves computing time, as no redundant features are generated. However, comparisons were only executed against variations of the author's own algorithm, and not with respect to state-of-the-art CNNs.

The work proposed in Sun and Wang (2022) has investigated the comparison of various pre-processing methods on the original audio signal, including magnitude STFT spectrum, complex-valued STFT spectrum, *Mel-log* spectrum, and MFCCs, as inputs to real-valued and complex-valued ResNet and DenseNet CNNs. 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, Log-mel 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 (WANG, D. et al., 2019).

Recent work has been driven by the advantage of using preprocessing filters to extract optimised features from the audio, using also stacks of multiple filters as inputs of the CNN models (CHEN, L. et al., 2022; HAN et al., 2022). 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 (CHEN, L. et al., 2022). The comparison of the results obtained with individual approaches and their combination showed that superior results could be reached with the latter. Another relevant work uses a fusion of the Mel-spectrogram, MFCC, chromatogram, spectral contrast, and tonnetz filters, resulting in a one-dimensional representation, to evaluate the performance of a CNN model for the classification of underwater acoustic signals (HAN et al., 2022).

Much of the work presented in this section was developed following a *problem solving* strategy, whose focus is on error minimisation under specific conditions, when the real effort should instead be employed on a *problem understanding* approach, where the background knowledge about the interrelation between environment conditions, SNR, and machine learning tools dictates the future development of suitable combinations of filters with DL models. Nevertheless, in general, the use of bio-inspired pre-processing steps in ML classification procedure has shown to reduce the complexity of the ML training procedure, accelerating the learning process while avoiding local maxima.

3.3 CLASSIFICATION USING ONLY CNNS

It is common to have a data preprocessing stage before applying acoustic signals into a DL model, aiming to extract the relevant features of the signal and improve its representation. Generally, in a time series classification problem, the strategy is to first transform the data and represent it as a spectrogram or cepstrogram, then obtain two-dimensional features by applying distance-based or feature-based methods (XING; PEI; KEOGH, 2010). However, considering that the DL models as universal functions approximators, the optimised representation of the signal can be also learned from its original time-domain data. Therefore, the use of unprocessed audio signals to feed convolutional kernels became also an option to tackle classification in the acoustic domain, as summarised in Table 3 and discussed below.

An end-to-end NN called Auditory Perception-inspired Deep Convolutional Neural Network (ADCNN) (YANG, Honghui et al., 2019) was defined for underwater acoustic target recognition. It used a bank of multi-scale deep convolutional filters as the first stage of the model to decompose the raw time domain signal into an set of distinct frequency domain signals. Each convolutional filter was followed by a max-pooling layer and several fully-connected layers. After this, a fusion stage was defined in order to unify those components into a more informative representation, where the main distinguishing features were prominent. The data flow then goes

through a decision layer, which produces the classification result. This process can be understood as an adaptive strategy where the sub-networks learn, directly from the input data, how to extract the intrinsic characteristics of the acoustic signal. The ADCNN model was able to execute each stage of the data-processing and classification pipeline, starting from noise decomposition, passing through feature extraction and finishing with the classification output.

Processing the original raw data, Hu et al. (2018) proposed a method for identification of underwater noise data based on CNN and Extreme Learning Machine (ELM) (HUANG; ZHU; SIEW, 2006), which, typically, provides good generalisation performance when compared with the densely connected NN for classification. Instead of using a classical filter as input features generator, the work proposed the extraction of advanced signal features using multi-layer convolutions. Additionally, by removing the final fully connected layer of the CNN and attaching a ELM as the classifier, the model was able to achieve 93.04% recognition rate on a dataset of civil ships, representing an improvement of, at least, 6.79% when compared with other approaches that used classical features as inputs and fully connected layers as classifiers. The final analysis showed that the features generated by the one-dimensional convolutional layers acted as band-pass filters, similarly to the ADCNN model described previously.

Similarly, a CNN model, called Underwater Acoustic Target Classification (UATC) DenseNet (DOAN; HUYNH-THE; KIM, 2020), was designed to classify underwater acoustic

Table 3 – An overview of methods for classification using convolutional filters alone. In the 'Metric' column, ACCY stands for accuracy.

Reference	Network	Dataset	Metric	Main contributions
(YANG, Honghui et al., 2019)	Auditory Deep CNN	Ocean Network Canada signals[a]	81.96% ACCY	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.
(HU et al., 2018)	CNN + ELM	Civil Ship[b]	93.04% ACCY	The substitution of the final fully connected layer of a CNN for an ELM classifier, improving the generalisation of the model.
(DOAN; HUYNH-THE; KIM, 2020)	UATC-DenseNet	Real-world passive sonar data[b]	98.85% ACCY	The analysis of the use of a number of convolutional blocks and layers, and different layer configurations and input features.
(TIAN et al., 2021)	MSRDN	Ocean Network Canada signals[a]	83.15% ACCY	The development of a deep residual network using the soft-thresholding proposed in (ZHAO et al., 2020) and the convolution kernel proposed in (HANNUN et al., 2019).

[a] Dataset uses public data with non-public preprocessing techniques.
[b] Dataset is proprietary and unavailable for reproduction.

targets from raw audio signals. This model is composed of a deep CNN with dense architecture, that was trained to recognise 12 classes of signals. The dataset used was labelled by a sonar expert, containing 11 different classes of sounds and one class corresponding to the ambient sound. The impact on the final model accuracy was investigated by tuning the number of convolutional filters, the depth (number of layers) of the neural network model, the different layer configuration and the different input features. The results showed that a particular model architecture (with 6 convolutional filters with a kernel size of 1x15) performed better when compared to various other models. The final model used the original sound data as input, without pre-processing layers or filters, and outperformed other classical methods, such as CNN-ELM (HU et al., 2018), ResNet18 (HE et al., 2016) and SqueezeNet (IANDOLA et al., 2016), by approximately 0.9%-4.6% at 0-dB SNR. The authors claimed that this performance improvement was due to the fact that the UATC-DenseNet optimises the use of features represented in multiple layers by an appropriate use of skip-connections (an argument that resembles the justification for the success of ResNet, cf. Section 2.3). Moreover, this work pointed to an important future direction that this type of research should take, which is the identification of topologies and parameters that are more appropriate for this kind of classification problem.

A Multiscale Residual Deep Neural Network (MSRDN) (inspired by (ZHAO et al., 2020)) was developed in (TIAN et al., 2021) to classify passive underwater acoustic signals from modelling the original signal waveform directly. The aim of using unprocessed original signal in this work was to avoid the loss of the waveform fine structure when converting it to time-frequency domain and to reduce the dependency of the classification task on the window size of FT and the hop length of the FT window. Tests using ONC[2] public dataset have shown that MSRDN outperformed other DL methods whose inputs were based on time-frequency representations.

The gain in the automatic classification of raw audio signal relies on the possibility of extracting information without imposing any *a priori* hypothesis on the nature of the object to be identified and the medium in which it is immersed. This idea was explored two decades ago, and led to an early US patent (CHAKRABORTY et al., 2002). In this work, a system that combines an unsupervised Self-Organised Mapping (SOM) network (KOHONEN, 1982) with Learning Vector Quantisation (LVQ) (SATO; YAMADA, 1996) was proposed to provide a high performance in classifying the roughness of the ocean floor from raw sonar data. This line of research, implemented in modern state-of-the-art hardware, which facilitates embedded parallel

[2] `https://oceannetworks.ca`

processing, could lead to deeper, more instances of LVQ networks and finer LVQs, possibly leading to more detailed representations of the seabed.

Virtually every efficient convolutional neural network in use today was first developed for image processing tasks. Therefore, most applications of CNNs to acoustic signals are based on adapting the methods initially developed for images. The development of dedicated acoustic feature-extraction strategies (using convolutional layers alone) to process audio signals is still an issue to be explored in this field.

The review of the DL applications presented on the Sections 3.1, 3.2, and 3.3, pointed out to some gaps in the available literature. There is insufficient concern about the environmental issues related to acoustic propagation, as the majority of the architectures presented here were based on image classifiers adaptation to the acoustic domain. Besides this approach presenting promising results, there is a need to consider specific acoustic processing issues. One of the steps taken in this direction was the application of time-frequency and biologically-inspired filters, which are gaining the field attention. However, these approaches are applied to a problem-solving strategy, commonly without proper application or tuning of parameters. The present work aims to solve some of the gaps presented above by stabilising a method for the processing filter parameters optimisation and the distance between target and sensor consideration as an important environmental factor in the conduction of the tests.

A significant portion of the works described in the previous sections was developed in proprietary datasets, unavailable for reproduction. The subsequent section will summarise some available datasets that can be used for underwater acoustic classification solutions development.

3.4 AVAILABLE DATASETS

CNNs have proved to be an important tool for automatic data classification. However, they are *data hungry*, as a large amount of labelled data points is needed to properly train and validate the models. This is a critical issue in the classification of underwater acoustic data, since most datasets are not publicly available, owing to the financial and technical complexity in obtaining such data and also to their potential defence-sensitive information. Therefore, much work in this area is conducted using synthetic data only (DENOS et al., 2017), or on a limited set of real data augmented with semi-synthetic examples for training (HUO; WU; LI, 2020; BERG; HJELMERVIK, 2018; LE et al., 2020). There are, however, a few datasets commonly used in the literature, that are summarised in Table 4.

One common criticism of the synthetic or augmented datasets usage in machine learning is that these might be created to *fit* the model rather than the other way around, thus generating a biased process. Therefore, there is a need for a complete dataset that comprises the information needed for this research area. In this way, the present work proposes a dataset for the vessel-type classification using underwater acoustic signals, which will be publicly available, containing the

Table 4 – A summary of some datasets available.

Dataset	**Description**
Ocean Network Canada (HEESEMANN et al., 2014)	This (non-annotated) dataset is curated by the University of Victoria, Canada and contains a continuous monitoring of the west and east coasts of Canada and the Arctic. It is available from `https://oceannetworks.ca`
DeepShip: An Underwater Acoustic Benchmark Dataset (IRFAN et al., 2021)	DeepShip is a benchmark dataset for underwater ship classification which consists of 47 hours and 4 min of recordings of 265 different ships belonging to four different classes (background sound was not available). This dataset is available for download at `https://github.com/irfankamboh/DeepShip`
ShipsEar: An underwater vessel noise database (SANTOS-DOMÍNGUEZ et al., 2016)	ShipsEar is a database containing underwater recordings of ship and boat sounds, which has 90 recordings of 11 different vessel types. It also has some useful information about the recordings, such as channel depth, wind, distance, location, to cite a few. This dataset is available for download at `http://atlanttic.uvigo.es/underwaternoise/`
Five-element acoustic dataset (PERRINE et al., 2009)	The main purpose of this dataset is to facilitate research on Doppler correction techniques for underwater acoustic transmissions. The dataset is composed of 360 communication packets with duration of 0.5 sec generated by a transceiver and captured by 5 hydrophones at nine different positions, and is available for download at `http://users.ece.utexas.edu/~bevans/projects/underwater/datasets/`
Fish classification with Dual-Frequency Identification Sonar (DIDSON) (MCCANN et al., 2018)	Fishery acoustic observation data was collected using Dual-Frequency Identification Sonar (DIDSON) with the purpose of classifying fish species. From 100 hours of data, 524 clips were extracted with eight species labelled. The dataset is available for download at `https://osf.io/sxek6/`
Passive sonar spectrogram images derived from time-series (LAMPERT; O'KEEFE, 2013)	The main purpose of this dataset is to facilitate solutions for the problem of detecting tracks in a spectrogram. It contains 4142 spectrograms generated from synthetic and also real small-boat data. The dataset is available for download at `https://sites.google.com/site/tomalampert/data-sets?authuser=0`
MDT dataset (VALDENEGRO-TORO; PRECIADO-GRIJALVA; WEHBE, 2021)	This dataset was obtained from a forward-looking sonar (ARIS Explorer 300) placed in a water tank in which a rotating turntable was used to allow various poses for the objects observed. The MDT dataset contains 2471 images with 12 classes of object, including bottle, pipe, platform and propeller and it is available from `https://github.com/mvaldenegro/marine-debris-fls-datasets`

acoustic data and the physical environment variables that can impact the signal representation. The next chapter will describe the development of the proposed dataset and the pipeline used in the proposal of a method to approach the underwater acoustic classification problem using DL.

4 A VESSEL-TYPE CLASSIFICATION METHOD

Underwater acoustic classification is a challenging task, as the sound propagation in the water has a complex behaviour and demands the consideration of various environmental features. As DL gained attention showing the potential to solve problems unfeasible by humans alone, their application in this area is increasing and presenting promising results. Besides that, this advance is still in the early stages and needs more attention on tools, methodologies, and datasets development.

This chapter will present the gap found in the literature that this work aims to fulfil and the motivation behind this objective. Also, the development steps conducted in this way will be described, focusing on three main stages: the initial approach involving the usage of a variation of the most common DL architecture found in the literature, the VGG, applied to one of the existing datasets, aiming to establish a baseline to be used as the base on the subsequent developments; the formulation and proposal of a new annotated dataset, which seeks to fill the gaps stated on the Chapter 3; and the development of a complete method, starting from the baseline, to the development of an underwater acoustic classification approach, considering the distance from target to sensor as a significant factor, and optimising the preprocessing of the audio signals.

4.1 GAP STATEMENT

There is an absence of works in the literature that use environmental variables as elements to be considered in the classification of vessels using underwater sound signals. On the other hand, a visible increase in works involving the application of different preprocessing could be noticed. The objective of this work is to establish a methodology to perform the classification of vessels using single-channel underwater acoustic signals, considering environmental factors, such as the distance from the sensor to the vessel, and comparing, in terms of accuracy, the main types of existing preprocessing.

4.2 MOTIVATION

In recent years, computer vision has pushed forward the field of DL to solve tasks such as object classification and detection. The area of sound and acoustic processing benefited from such advances, adapting these techniques to solve similar problems applied to the one-dimensional sound signal. Despite that, there are intrinsic characteristics of the acoustic signals that were not

considered, and these need to be better explored and understood. As the audio signal is naturally different from the visual stimulus, the development of dedicated acoustic Machine Learning models can have a beneficial impact on the classification of such signals.

In the underwater domain, the task of classification has increased in complexity, mainly because of the behaviour of the sound propagation profile, which is influenced by the height of the water column and the variation of physical characteristics, caused by the seasonal changes of the medium and external influences of the environment. The combination of physical characteristics with the audio signal on the classification task can pave the way for understanding the correlation between them. This approach has not yet been properly explored in the current scientific scenario, and the impact of this on classification accuracy needs to be understood to improve such methods.

The use of bio-inspired algorithms and time-frequency representations to preprocess the acoustic signal claims to contribute to the feature extraction stage of the deep learning approaches, facilitating the convergence and improving the classification accuracy. Nevertheless, this statement cannot be validated without performance comparisons between the distinct approaches presented in the literature. To achieve valid results, this task needs to be performed using the same database as comparison and applying reliable metrics, which consider not only the prediction rates but the number of class occurrences as well.

When classifying moving targets, it is vital to understand how the variation of the distance from the target to the sensor is influencing the accuracy of the classification, as the attenuation of the original sound by the medium can directly impact the received signal power. As the quality of Machine Learning models is heavily based on the data used for the training and validation steps, there is a need for an appropriate annotated dataset composed of the audio signals and the distance information. This is another important question that remains unexplored, as there is no publicly annotated data available in the literature to use as a starting point for that kind of research.

Machine Learning has historically proven to be an adequate strategy to solving pattern recognition problems, and, given the background presented on the Section 3, there is a long way to go to consider all the implications involved in classifying underwater acoustic signals. Given the unexplored points mentioned on this section, a end-to-end approach that classifies the acoustic data and has well-defined specifications, regarding the positioning of sensors on the underwater domain, is only achievable with the understanding of, at least, three main topics: the variation of the distance from sensor, the comparison of different acoustic data processing

approaches, and the comparison of the available ML architectures. The proposed pipeline, that aims to fulfil the topics stated above, is presented below.

4.3 DEVELOPMENT STEPS FOR UNDERWATER ACOUSTIC CLASSIFICATION

The implementation of this work focuses on three main stages:

a) The development of a baseline machine learning strategy to classify vessel-type underwater acoustic data, using an available dataset present on the literature;
b) The development of a new dataset, containing the acoustic signals, the environmental features, and the information of distance between target and sensor structured into different scenarios;
c) The study of the impact of the new variables, contained at the proposed dataset, on the final accuracy, comparing two distinct models with three different preprocessing filters.

The next sections will cover each one of the stages addressed in this work to solve the stated problems.

4.3.1 Baseline development for vessel-type classification with existing datasets

Most works cited in the literature used proprietary datasets unavailable for reproduction or have private machine learning source codes. In this work, the objective was not re-implement the current models, as they can introduce unexpected behaviour, but propose a new method to address the problem of underwater acoustic classification. With that in mind, this development stage focused on a baseline development to use as a starting point for the subsequent steps, using a publicly available dataset.

This stage proposed to evaluate the performance of three different preprocessing methods used as inputs to a VGG-based CNN. The chosen dataset was the DeepShip (IRFAN et al., 2021) with the addition of background noise recordings for a baseline. The DeepShip dataset is composed of data from four commercial ship classes: oil tanker, tug, passenger ship and cargo ship. Those recordings are divided into 613 files, which vary from about 6 seconds to 1530 seconds. Only the identification of a single vessel within a range of 2 kilometres from the hydrophone was used to generate the data.

Besides all the files contained in the dataset, the original set does not have information about background sound. The data included on the original set, instead, was downloaded from

the internet from various websites. In this work, those signals were replaced by ones generated using the same source as the original vessel data: the ONC initiative. The background sound used was generated considering the absence of vessels within the range of 2 kilometres from the hydrophone, using the AIS data as reference for vessel identification, and was captured during the deployment from 24 June 2017 to 03 November 2017. To maintain the same length as the other four classes of the original DeepShip dataset, only the first 12 hours of recording were considered. Therefore, the final dataset is composed of five classes (Cargo, Passenger Ship, Tanker, Tug and Background Sound) and has a total length of, approximately, 59 hours. The final dataset was preprocessed to have a sample rate of 32000 and 1-second segments, padding with zeros when the size was smaller than needed.

This baseline development focused on the application of three preprocessing filters commonly present in the literature and also applied in the (IRFAN et al., 2021) work. The hyperparameters chosen to generate such representations were based on typical values found in the other works involving acoustic preprocessing. The first processing method was the Mel Spectrogram, which chosen window had 1024 samples, with a hop length of 512, resulting in 64 frequency bins over 63 time intervals *per* data segment. The second method was the CQT, chosen mainly for its characteristic of emphasising lower frequencies, which are predominant in vessel sounds. The parameters used to generate such representation were a hop length of 512 samples (the same period as the Mel spectrogram), producing signals with a size of 64x63, with frequencies starting at 32.7 Hz (a C1 on a musical scale) using 12 bins per octave. Third, a Gammatone Spectrogram was used for comparison, mainly due to its characteristic of mimicking the human auditory system. To match the same resultant spectrogram shape obtained with the other preprocessing filters, the parameters chosen were windows of 1024 samples with a hop length of 512 and 64 frequency bins, starting from 20Hz. The three preprocessing methods used the Hann function as the windowing method, as it has a low aliasing effect. This set of parameters (see summary at Table 5) is referenced in this work as *Version 1* parameters. Also, sample images representing an audio signal processed by the three filters cited above were shown at Figure 1.

The CNN used in this preliminary development is based on VGGNet with a reduced number of convolution layers and modifications to match the input signal. Four feature extraction convolutional layers were used, with the kernel size of 3x3 and a ReLu activation layer. To downscale the image by a factor of 2, emphasising the most activated presence of a feature, a Max Pooling layer was used. Additionally, a Fully Connected layer was used to generate the classification weights, which finally passed through a Softmax function to generate the

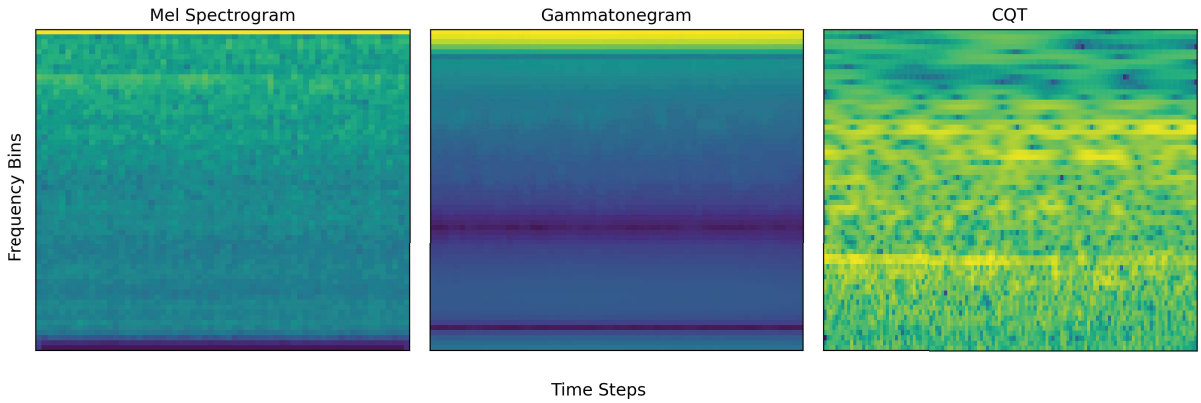

Figure 1 – Examples of visual representation for the three preprocessing filters: Mel Spectrogram, Gammatonegram and CQT.

Parameter	Mel	Gamma	CQT
Hop Length	512	512	512
Window	hann	hann	hann
N° of bins	64	64	64
Min. Freq.	0	20	32.7
Max. Freq.	Auto	Auto	Auto
Bins p. Oct.	–	–	12
Out. Length	64x63	64x63	64x63

Table 5 – Summary of the *Version 1* set of parameters, used on the baseline preprocessing generation.

classification logits. The model was trained with batches of 25 images and a learning rate of 0.001, during 40 epochs with the ADAM optimiser. The loss function used was the *Categorical Cross-Entropy* loss. The architecture is shown in Figure 2.

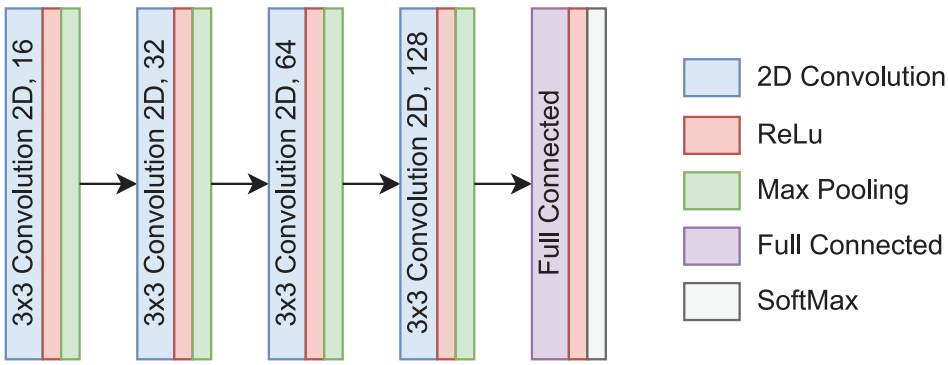

Figure 2 – The baseline VGG-based model architecture.

The results obtained on this stage will be described at Chapter 5.

Besides the development of the baseline focused no the DeepShip dataset, this source of audio signals did not contained the environmental variables needed to address the classification

problem. To include the environmental variables and provides a new public dataset, the subsequent stage of the work focused on the generation of an annotated underwater acoustic dataset.

4.3.2 Generation of a dataset of underwater acoustic signals

The data used in this work consists of signals obtained from the ONC initiative[1], captured during the deployment at the Strait of Georgia, Canada (see Figure 3), from June 24 to November 3, 2017. 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 the data from the AIS. The Conductivity Temperature and Depth (CTD) recorder used to capture the environmental signals was the Sea-Bird SBE 16plus SEACAT Recorder, a high accuracy recorder designed for fixed-side deployments. Figure 4 contains a picture of each one of the devices.

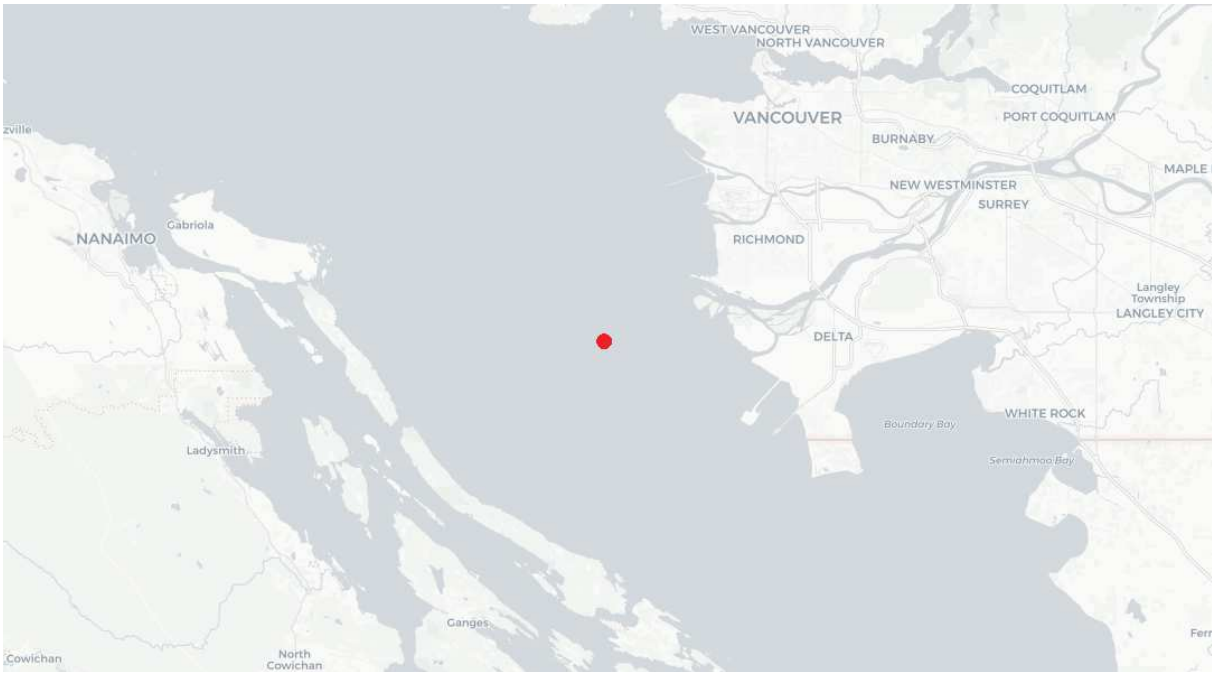

Figure 3 – Location of the Hydrophone (in red), at Strait of Georgia, Canada.

The first part of the annotation process focused on the translation and filtering of the AIS signals. These signals contain position, identification, speed, course, and other information about active ships in maritime traffic. Some of the information contained in AIS is not necessary for vessel classification tasks, only messages related to *position report*, *static and voyage related data*, and *static data report* were used. Duplicated messages, and messages that did not have positional arguments, were filtered out and ignored in this work. The vessel's class was then inferred from the type of ship and cargo fields of the AIS messages, generating four categories:

[1] `https://www.oceannetworks.ca/`

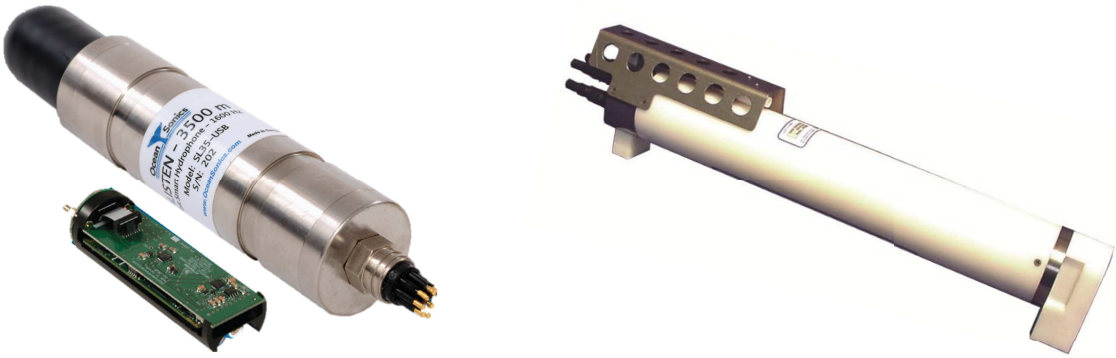

Figure 4 – The icListen AF Hydrophone (BORNHÖFT, 2022) on the left side of the figure and the Sea-Bird SBE 16plus SEACAT Recorder (ELECTRONICS, 2022) on the right side of the figure.

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* is defined as the radial distance where only one vessel is present at a specific moment, whereas the *exclusion radius* is the region in which there is no vessel within a fixed radial distance. To isolate a single vessel as much as possible, scenarios were generated as illustrated in Figure 5, where a vessel is within the inclusion radius while no other vessels are within the wider exclusion radius.

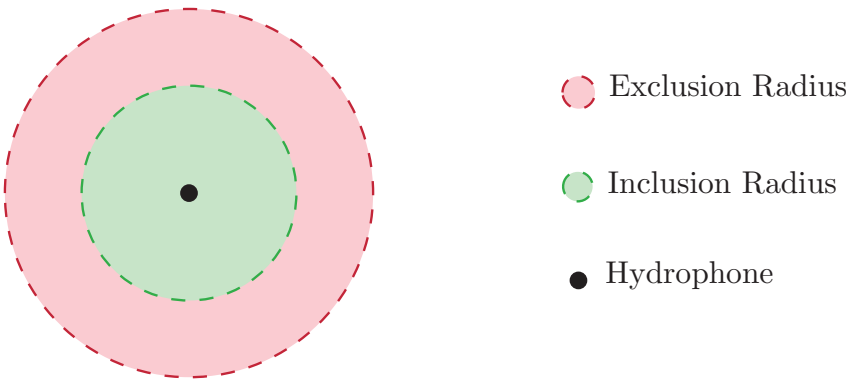

Figure 5 – Diagram representing a scenario.

These scenarios facilitate 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 scenario has an inclusion radius of 2km and an exclusion radius of 3km; the second is defined within the interval of 3km and 4km between inclusion and exclusion radius; the third scenario is defined within the interval between 4km and 6km. Table 6 summarises the scenarios descriptions.

The application of the three scenarios combined into a more extensive dataset can also be explored, as each audio segment was placed into a single scenario, thus, there is no repeated data in the dataset. The three scenarios combination produces a dataset that ignores the distance from target to sensor, serving as a baseline for comparing the results obtained from the separated versions.

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

Table 6 – The different scenarios considered for the dataset generation.

The next stage of the dataset formulation was the combination of every AIS instance, defined as the time period that matches a specific scenario, with the acoustic data and with the CTD data. The environmental information considered as relevant to this work, obtained from the CTD recorder, is composed of five different signals: temperature, measured in Celsius; conductivity, measured in siemens per metre; pressure, in decibar; salinity, measured in psu; and sound speed, measured in meters per second. For every different instance, the average of the measures was considered.

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 ($\mathrm{med}()$) was used to de-noise the original signal ($a(t)$, where t represents time). The resulting audio was subtracted from the original signal (Equation (14)) producing an audio signal ($g(t)$) free of DC offset. The standard deviation of $g(t)$ (represented as ϱ, as shown in Equation (15)) 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.

$$g(t) = a(t) - \text{med}\left(a(t)\right). \tag{14}$$

$$\varrho = \sqrt{\frac{1}{N-1}\sum_{t=1}^{N}\left(g(t) - \overline{g}\right)^2}. \tag{15}$$

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 6, 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 (16), was used to define the size of each class based on the class with the smallest length.

$$\text{factor} = \min\left(2, \frac{\text{L}}{\text{l}}\right). \tag{16}$$

In Equation (16), 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 (16)), the selection started again, gathering repeated audios until the desired length was achieved. However, uniqueness of each recording was enforced. Table 7 contains the duration of each subset for the dataset scenarios.

Once the source of data was obtained, annotated, cleaned, and organised, the logical sequence was the evaluation of the performance of the DL model using this new dataset. The

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

Table 7 – The length, in seconds, of each dataset class for each of the three scenarios' subsets.

following section will describe the tests and ML training sections performed with those acoustic signals.

4.3.3 A method for underwater classification with the proposed dataset

This stage of the development focused on the UATC on the dataset generated at Section 4.3.2 using preprocessing filters and CNNs, detailing the procedures and experimental setup. First, each entry in the dataset was divided into one-second segments, padding with zeros the values with smaller sizes. After that, the three proposed preprocessing methods, CQT, Gammatone, and Mel Spectrogram, were applied to each audio file. Initially, for comparison with the same approach used on Section 4.3.1, the same preprocessing parameters were used (*Version 1*), resulting in 64 frequency bins over 63-time intervals *per* data segment.

A second set of parameters, referenced as *Version 2*, was obtained by means of an optimisation process. Analysing the frequency spectrum of the vessel's signal, it was possible to notice that the underwater acoustic signal produced was mostly focused on the low-frequency band, which in our dataset was mostly below 3kHz, as exemplified in Figure 6. To maintain a considerable safe range above the maximum frequency, spectrograms were generated from 18Hz (the minimum acceptable for the CQT representation for 1s audios) to the frequency of 4186Hz (C8 note and $\approx$ 1kHz above the 3kHz experimental value). Using a hop length of 256, which represents half of the value proposed on *Version 1*, the resulting representation (*Version 2*) had a length of 95x126. The values for the two versions of parameter sets are compared and summarised in Table 8, where the values not related to a particular representation were marked with "–".

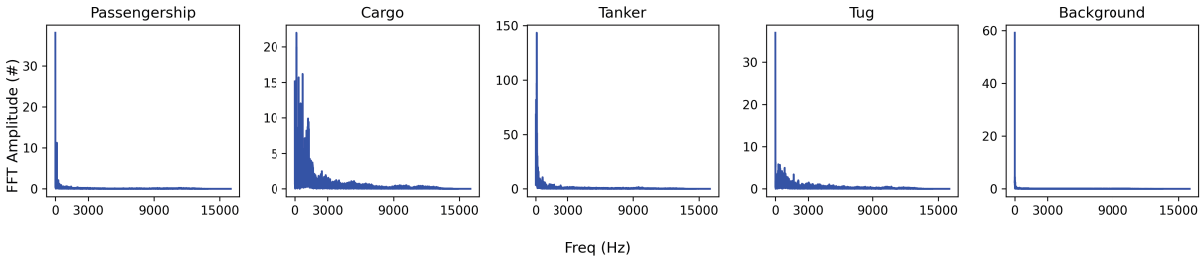

Figure 6 – Examples of audios from the five categories present in the dataset.

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
N° 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

Table 8 – The detailing of the *Version 1* and *Version 2* sets of parameters used in the preprocessing generation.

The raw audio signal and its processed representations (CQT, Gammatone and Mel spectrograms), obtained with *Version 2* parameters, are exemplified in Figure 7.

Inspired by the great number of machine learning methods applied to three-channel images, such as colour images in RGB or Hue Saturation Value (HSV) colour spaces, the three preprocessing methods cited above were combined in a three-dimensional representation, that was used as input to the CNNs tested in this work. This was motivated by providing the Neural Network with more complete representations, aiming to take advantage of the strengths of each of the assumed preprocessing methods. Combining Mel Spectrogram, CQT, and Gammatone resulted in data samples with dimensions of 64x63x3 for *Version 1* and 95x128x3 for *Version 2* parameters. In this work, this representation is referred as *Complete*.

This stage of the implementation used two distinct CNN models as base approach: the VGGNet (SIMONYAN; ZISSERMAN, 2015) and the ResNet (HE et al., 2016). The implementation of the former model suffered two main modifications from the baseline approach (see Section 4.3.1): a Leaky ReLu was used as activation function, instead of a ReLu function, and a Batch Normalisation layer was added, both changes aiming to reduce overfitting, as the new dataset was divided in smaller scenarios. The resulting model architecture is shown in Figure 8 and was composed of four convolutional layers, which were used to extract the features, a Batch Normalisation and a Leaky ReLu activation layers associated with a Max Pooling layer,

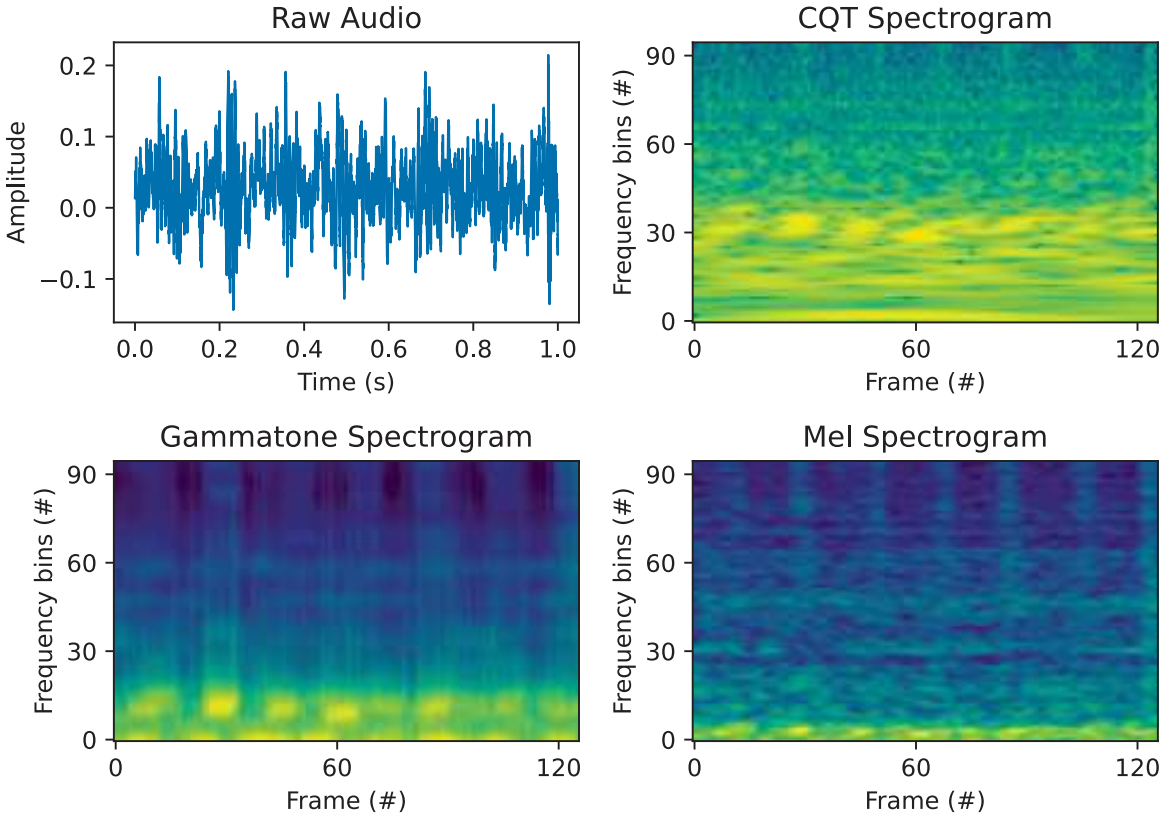

Figure 7 – An example of audio used on this experiment. The image shows the raw audio signal, on time domain, and its three preprocessed versions using the *Version 2* parameters.

which resizes the image by a factor of 2. Lastly, the classification weights are delivered by a fully connected layer.

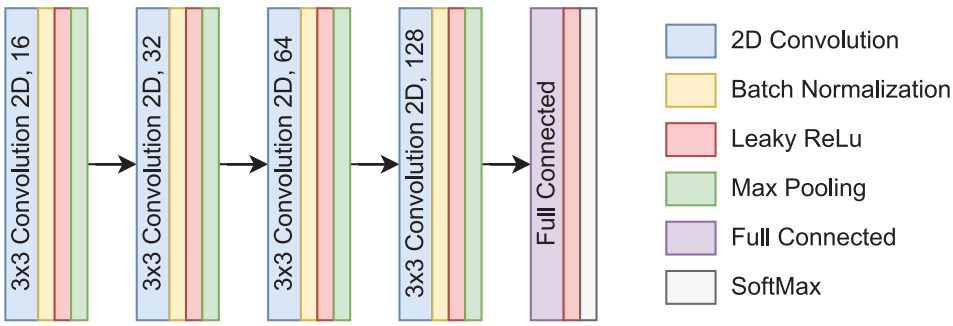

Figure 8 – The VGG-based model architecture.

This stage of the work also used the ResNet18 model, with modifications on the input layer to match the preprocessed images. The ResNet18 architecture, shown in Figure 9, 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. As there is a lack of open source codes or pre-trained models in this domain (see Section 3), both VGG and ResNet18 models were trained without the application of transfer learning techniques.

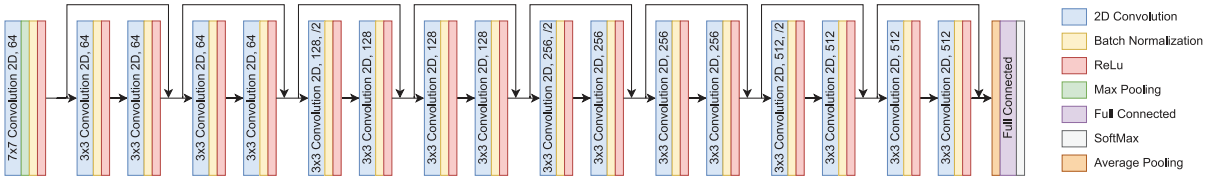

Figure 9 – The ResNet 18 model architecture.

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 SGD (ROBBINS; MONRO, 1951) and the Adam (KINGMA, Diederik P.; BA, 2015) optimisers. Their use was compared in the tests executed in this work, using a learning rate of 0.001, decreasing exponentially, with a gamma value of 0.95, over 40 epochs.

The resulting architecture was then composed of the preprocessed acoustic signals, produced by the four strategies described in this section, applied to both VGG-based and ResNet18 models. 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 10, where the preprocessing block is the representation of Mel Spectrogram, Gammatone Spectrogram, CQT, or the combination of the three preprocessing filters, and the model block represents VGGNet or ResNet18.

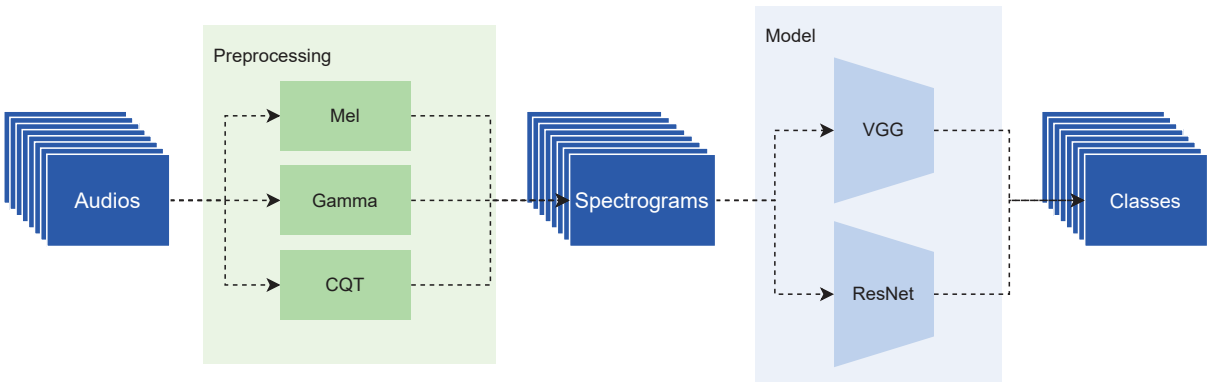

Figure 10 – The block diagram of the complete solution.

5 EXPERIMENTS AND RESULTS

After the definition of the method used in this work, training sessions were executed for each approach described in Section 4.3. The Section 5.1 describes the metrics used to instrument the experiment, and than the results are reported on Section 5.2.

5.1 ENVIRONMENT SETUP

All the tests executed in this work were executed on a CPU using an Intel(R) Core(TM) i7-1065G7 machine with Ubuntu 20.04.4 LTS operating system, and were implemented using PyTorch framework (version 1.11.0).

Results were reported using Micro-averaged Accuracy, providing a global overview of the model performance in the real-world scenario. Three additional metrics were used, providing complementary information: Precision, Recall, and F1-score. Those three latter metrics were evaluated using *macro-averaging*, aiming to obtain a balanced evaluation across the classes.

5.2 RESULTS OF THE VESSEL-TYPE CLASSIFICATION WITH MACHINE LEARNING

The experiments were instrumented with metrics evaluation after every training epoch, ensuring the convergence and correctness of the execution. Aiming to evaluate the performance of the models using data not applied in the training stage, the results described in the following sections were reported on the test subset of the respective datasets, and were obtained after the training execution was completed.

5.2.1 Baseline development on DeepShip dataset

As this stage focused on the model development, the DeepShip dataset (IRFAN et al., 2021) was used with addition of background noise recordings for a baseline. The original dataset was randomised and divided into train, validation and test subsets, which represent 70%, 20%, and 10% of the total data respectively. Training sessions were executed for the three preprocessing filters using *Version 1* parameters (see Section 4.3.1 as reference). The results are shown in Table 9.

The CQT approach outperformed the other two methods with an accuracy of 83.92% compared to 68.86% and 64.81% for the Mel Spectrogram and Gammatone Spectrogram, respectively. To discriminate the results comparing the success rate for each existing class in the

	Method	Accuracy %	Precision %	Recall	F1 %
CNN	CQT	**83.92**	**84.04**	**83.75**	**83.81**
	Gammatone Spec.	64.81	63.69	64.09	63.19
	Mel Spec.	68.86	67.80	68.38	67.80
SCAE	CQT	77.53	78	77	77
	Gammatone Spec.	73.59	74	73	73
	Mel Spec.	70.18	70	70	69

Table 9 – Comparison of evaluated methods showing accuracy, precision, recall, and F1-Score (F1)

set, a confusion matrix was also plotted in Figure 11. It shows that the *background* class has the best results, with 98% of true positives, and *cargo* has the lowest, with 73%. The reason for the reduced accuracy with respect to cargo ship class is likely due to the variability within the class. Cargo ships can transport various types of cargo, goods or materials, which can influence weight, motor power, number of propellers, size and other physical characteristics. All of these can heavily influence on how the transmission medium will be disturbed, and consequently, how the sound will be perceived.

Table 9 also shows a comparison between the results obtained for the proposed CNN with the SCAE proposed in (IRFAN et al., 2021), using the same preprocessing layers. The metrics reported for SCAE were extracted from original paper (IRFAN et al., 2021). The results obtained in the present work outperformed the best approach found in the target study, the proposed SCAE model, by a margin of 6 percentage points. Although there are many possible reasons which influence this improvement, the fact that the background sound used on both training sections were obtained from different sources must be taken into account. While the initial model implemented on this work used the background sound from ONC sources, the same origin of the vessel sounds, (IRFAN et al., 2021) used various websites as source for their data, possibly introducing confounding factors. Also, for the CQT, the windowing was obtained with different parameters, which can directly influence the temporal relation captured by the model. Knowing that VGG-based methods have a strong capability to perceive spatial relations on images, the different parameters for windowing can strongly impact the results. Another important difference between the two works is the size of the resultant spectrogram. While the CNN model used directly the two-dimensional representation, with size of 64x63, the spectrogram used by SCAE was re-scaled to match the desired input size (from 40x43 to 48x48) generating additional errors from interpolation.

For both Gammatone and Mel Spectrograms, Table 9 shows that the results using SCAE model were better than the proposed method. However, the results obtained with Mel spectrogram were only 1.2 percentage point away from each other. These results can be heavily influenced by the different preprocessing parameters used to feed the DL models, as the representations obtained on (IRFAN et al., 2021) were produced on audios with 250 ms of duration, generating spectrograms of size 40x43 and 64x43 for Mel and Gammatone, respectively. These results pointed to CQT as preferred filter to the underwater classification, and reinforced the importance of tuning the parameters of the feature extraction layers, guiding the experiments of the DL method with the proposed dataset, described below.

5.2.2 Deep Learning method using the proposed dataset

The experiments performed in this stage focused on the baseline model application on the proposed dataset and the adjustments and optimisations performed to improve the metrics.

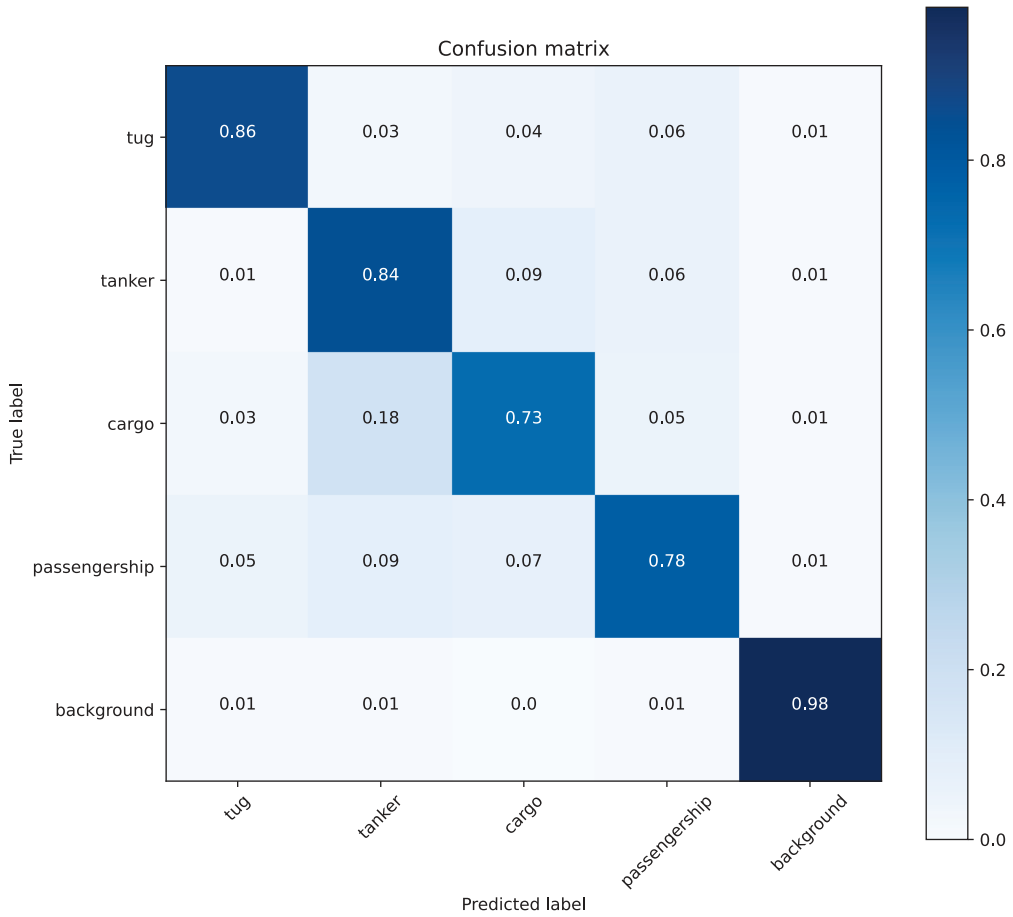

Figure 11 – Confusion matrix with baseline results using the CQT preprocessing filter.

The training sessions reported in the following sections were executed for the three chosen preprocessing filters using the proposed dataset and the Complete representation.

Optimiser selection

SGD and Adam are two of the most common optimisers used in DL. However, their performances are domain dependent, which 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 10 presents the results of applying Adam and SGD to train a VGG-based classifier on the the *first dataset scenario* described in Section 4.3.2, as 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.

	Method	**Accuracy%**	**Precision%**	**Recall%**	**F1%**	**Time (h)**
SGD	CQT	79.10	66.51	82.89	68.01	4.1
	Gamma	53.30	47.71	58.47	43.91	4.0
	Mel	54.47	51.23	61.74	44.40	**3.9**
	Complete	**84.46**	**71.48**	**85.56**	**75.42**	4.5
ADAM	CQT	77.39	65.25	81.04	65.30	4.8
	Gamma	56.59	52.32	59.51	44.94	4.2
	Mel	5.26	21.00	20.16	2.23	4.5
	Complete	73.08	60.75	72.72	61.14	5.2

Table 10 – Comparison of SGD and ADAM optimiser, showing accuracy, precision, recall, F1-Score (F1), and total training time.

The results represented in Table 10 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 10). Adam only performed better than SGD in the test where the Gammatone filter was used as preprocessing method, and yet the metrics obtained are very close to the SGD approach. In addition, the training session using Adam and Mel spectrogram did not converge to a global minimum, and the model mostly predicted the same class, as per the class-normalised confusion matrix shown in Figure 12.

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

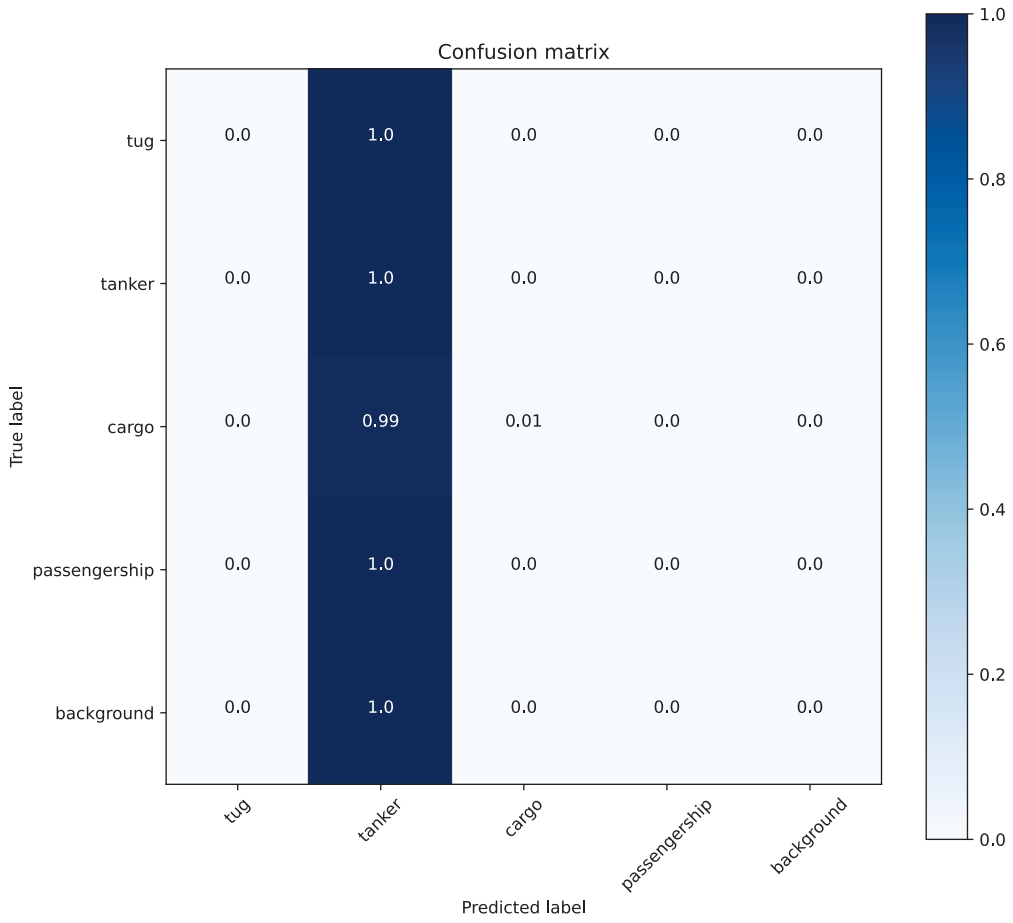

Figure 12 – The confusion matrix of the execution for Adam optimiser using the Mel spectrogram as preprocessing filter.

Preprocessing optimisation

As mentioned in Section 4.3.3, the *Version 1* parameters for the preprocessing filters were generated based on the information from the related literature, resulting in a 64x63 image. *Version 2* had images of dimensions 95x126, 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 5.2.2 was applied: the VGG-based model trained on data from the first dataset scenario. As the results obtained in Section 5.2.2 showed that the SGD optimiser was more suited to this problem, the experiments were only performed using this optimiser. The results of this test were summarised in Table 11.

Both Mel Spectrogram and Gammatone methods presented lower accuracy values when compared with CQT and Complete representations, as shown in Table 11 values show. The worst CQT result, obtained with the *Version 1*, is 17 percentage points better than the best result for Gammatone filter, obtained with the *Version 2*. The Mel Spectrogram and Complete

	Method	Accuracy%	Precision%	Recall%	F1%	Time (h)
Vers. 1	CQT	79.10	66.51	82.89	68.01	4.1
	Gamma	53.29	47.71	58.47	43.91	4.0
	Mel	54.46	51.23	61.74	44.40	**3.9**
	Complete	84.46	71.48	85.56	75.42	4.5
Vers. 2	CQT	**86.32**	**73.97**	**88.34**	**77.91**	10.5
	Gamma	62.28	53.88	65.08	52.14	9.4
	Mel	41.86	53.85	45.41	27.92	9.6
	Complete	82.92	73.06	83.12	75.26	21.9

Table 11 – Comparison of *Version 1* and *Version 2* preprocessing representations, showing accuracy, precision, recall, F1-Score (F1), and total training time.

representations did not present improvements with the *Version 2* parameters. However, even using the *Version 1* Mel spectrogram results, this preprocessing method accuracy was 25 percentage points below that of CQT for the same parameters, which represents a 31.15% accuracy drop. Additionally, despite the accuracy drop of 1.54 percentage points for the Complete scenario, the precision obtained in this case improved 1.58 percentage points, showing a very close result for both versions of this representation. On the other hand, comparing the metrics obtained with *Versions 1 and 2* for the CQT method, it is possible to notice that *Version 2* had a significant improvement in accuracy, close to 9.13% (7.22 percentage points). This improvement was probably due to the parameter optimisation process, that led to an increase in the temporal scale with the shortest 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.

Model evaluation

After the investigation of the preprocessing parameters and the optimiser selection, the next step in the development of the underwater acoustic signal classifier proposed in this work was the selection of the DL model. The VGG and ResNet architectures were chosen in this work based on their ubiquitous presence in the related literature and their potential to solve complex problems from distinct domains. Training sessions were performed using the VGG-based model and ResNet18. As the results obtained in the previous experiments suggested a better performance using the SGD optimiser combined with CQT or the Complete representation generated using

Version 2 parameters, these methods were chosen in the model comparison experiments described here. Table 12 shows the results obtained in these tests.

	Method	Accuracy%	Precision%	Recall%	F1%	Time (h)
VGG	CQT	86.86	73.97	88.34	77.91	**10.5**
	Complete	82.92	73.06	83.12	75.26	21.9
ResNet	CQT	94.95	89.21	95.60	91.92	53.5
	Complete	**97.07**	**94.57**	**97.61**	**96.00**	77.1

Table 12 – Comparison of VGG-based and ResNet 18 models, showing accuracy, precision, recall, F1-Score (F1), and total training time.

The results showed that the ResNet18 model outperformed the VGG-based for both CQT and Complete preprocessing filters, presenting an improvement of 8.09 percentage points for the CQT, and 14.15 for the Complete, being the latter 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 gives the model the ability to generalise the problem function better, thus resulting in higher classification performance. 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 one-third of the input size, as the 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 preprocessing approach alone.

In this context, the final set of tests (described below) evaluates the impact of the distance from the object of interest and the sensor, considering the three different dataset scenarios (described in Section 4.3.2) using the CQT as preprocessing filter combined with the ResNet18 model.

Scenarios validation

The final test executed in this work intended to evaluate how the distance from the sensor to the target vessel influenced the classification results. Based on the tests with the best results reported above, 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 4.3.2. Also, an experiment using all the data from the three dataset scenarios combined was performed, aiming to evaluate if the distance is impacting the accuracy, or if the generalisation ability of the architecture is capable of dealing with this variable. The results obtained in this test are shown in Table 13.

Scenario	Accuracy%	Precision%	Recall%	F1%	Time (h)
First	94.95	89.21	95.60	91.92	53.5
Second	94.45	94.57	93.89	94.17	101.3
Third	93.11	89.36	93.82	91.30	74.1
Combined	84.13	79.46	86.51	81.78	197.8

Table 13 – Comparison of the final architecture on the proposed dataset scenarios, showing accuracy, precision, recall, F1-Score (F1), and total training time.

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 result has higher metrics. However, these results suggest that there is no significant difference in accuracy between the different scenarios. This could be explained by the depth and the ocean temperature where the hydrophone was located, which provided the best context for underwater sound propagation (DOMINGOS et al., 2022b), not degrading the SNR sufficiently to invalidate the signal representation. However, it was possible to notice that combining all the data in the training session caused an accuracy drop of, at least, 8.98 percentage points. The confusion matrix for the Combined scenario is shown on Figure 13. This suggests that there is a negative influence in the data from the different scenarios that confused the model during the training stage, as discussed in more details in the next section.

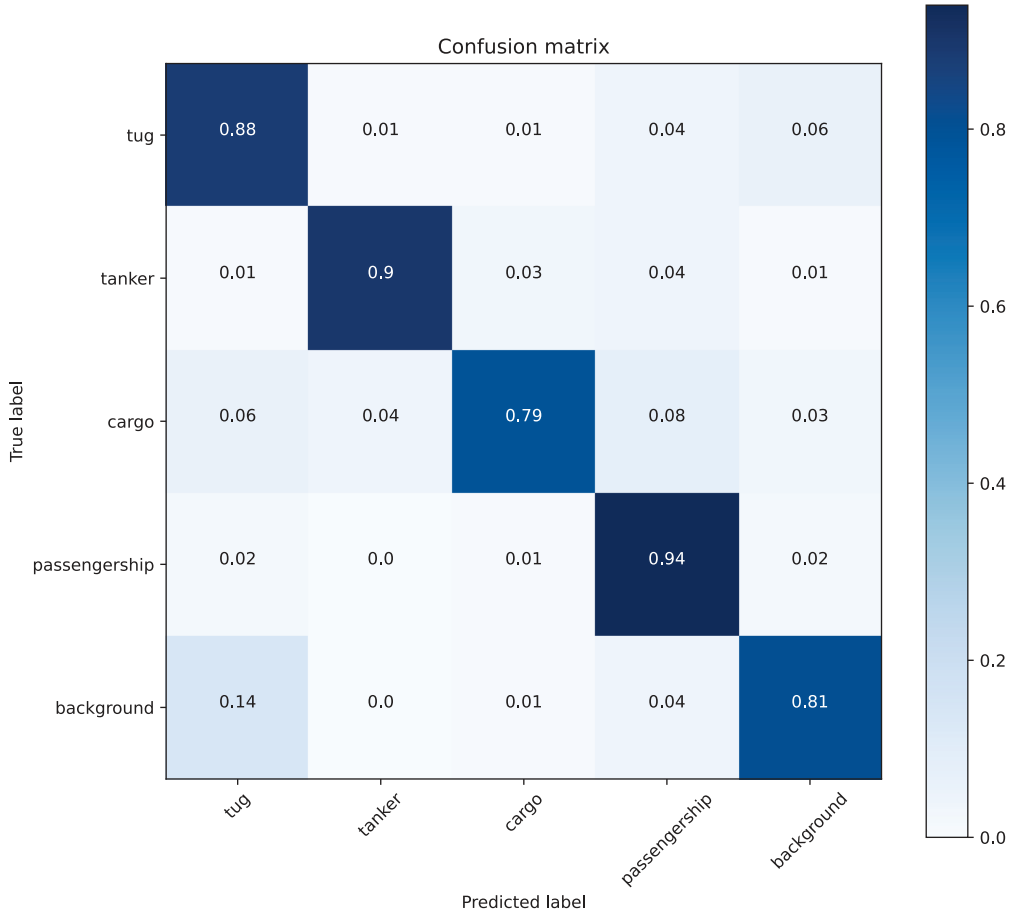

Figure 13 – The confusion matrix of the execution for the Combined scenario the CQT as preprocessing filter and ResNet 18 model.

6 DISCUSSION

The work reported in this thesis shows 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 focused on the baseline establishment for the classification of underwater acoustic signals (see Section 4.3.1). The results, produced by a proposed VGG-based architecture, were obtained on the DeepShip (IRFAN et al., 2021) dataset and compared with the SCAE model. They showed that the experiment with the CQT processed inputs obtained the best results for both approaches, with the highest metrics obtained with the proposed VGG-based method. DL architectures are function approximators whose performance is strongly dependent on the quality of the signals used as input, for instance, if they have a high signal-to-noise ratio or appropriately represent the real-world scenario. The values obtained in this experiment showed that the proper preprocessing methods parameter choice has to be the focus of the modelling. With better input representations, the models can be simplified and still achieve similarly, or even better, classification results. In addition, this result was also influenced by the different sets of background sounds, which, in this work, were gathered from the same source of the original signals. This also showed that the environmental conditions have to be considered in the solution, as the sound propagates different depending on the environment.

After the baseline formulation, the following tests were performed on the new proposed dataset. The first experiment 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, Adam and SGD, were tested and compared. The results reported in Section 5.2.2 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, whereas SGD presented higher accuracy and a more stable performance. This agrees with other studies (e.g. (WILSON et al., 2017; XIE et al., 2022)) that argue that 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 (XIE et al., 2022). The tests performed in the present work 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.

Another issue considered in this work was the selection of the most suitable representations of the signal to be used by the CNN. In our dataset, the CQT representation presented significantly better performance, with an accuracy of around 86%, when compared with Gammatone and Mel spectrograms, with the accuracy of 62% and 54%, a minimum of 17 percentage points improvement. There are various possible explanations for this finding. One of them is that the acoustic signals generated by the vessels are predominantly composed of lower frequencies; however, the Gammatone filter does not emphasise those signals properly, resulting in a lower classification performance. Mel spectrogram, on the other hand, emphasises the lower frequencies by mapping the frequency axis to the Mel scale, but maintaining the conversion from time-domain using fixed time windows, which affects negatively 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(BROWN, 1991). 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 CQT filter produced the best classification results, with an accuracy of around 86%, when compared with Gammatone and Mel spectrograms, with the accuracy of 62% and 54%. This is due to some characteristics of this filter that are desirable in the preprocessing of underwater acoustic signals, such as the exponential frequency resolution of constant-Q transform, which enhances the lower-spectral frequency resolution. Besides, the CQT representation produced a more pronounced 2D pattern than the other spectrograms, as can be observed in Figure 7. 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 aimed to obtain a preprocessing method that includes the advantages of each of the preprocessing methods considered in this work. The results obtained showed that the classification accuracy obtained using this three-channel representation was better than the other in some tests, but it did not achieve the best performance overall. This was probably due to confounding effects that a more complex data representation brings to a DL classifier. Moreover, as the Complete representation combines the other preprocessing methods, its generation can be time expensive. Even in the cases where the Complete representation showed the best results, its performance was similar to that obtained using only the CQT spectrograms as input.

With respect to the DL model selection, the ResNet approach overcame 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 probably 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 worse results compared with the performance values obtained for each of its constituting 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.

Considering related work developed with data from the ONC initiative, accuracy values of the order of 80% were reported using raw audio data, where time-frequency filter dependency was not considered in DL pipelines (YANG, Honghui et al., 2019; TIAN et al., 2021). An accuracy value of around 87% was reported as a result of the application a bio-inspired cochlea model preprocessing filter to a CNN classification (SHEN et al., 2020). Although it is virtually impossible to reproduce these results, as the annotated data and code used to generate them are not publicly available, the results reported are similar to those obtained in the present work for the combined dataset (84%), which shows that our general approach achieved state-of-the-art results. However, the research reported in the present thesis achieved superior results when distinct scenarios were considered with respect to the distance to the sensor, obtaining accuracy at the order of 94%.

A previous experiment using various preprocessing filters for vessel-type classification, applied on a publicly available annotated version of the ONC underwater acoustic signals (IRFAN et al., 2021), also obtained the best classification results using CQT as preprocessing

filters, although achieving the best accuracy values of around 77%, when using an auto-encoder architecture.

7 CONCLUDING REMARKS

This thesis presented a machine learning approach for vessel type classification using underwater acoustic data. CQT, Mel Spectrogram, and Gammatone 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 proposed. The results showed that the CQT and the Combined approaches achieved the highest accuracy results for the dataset used in this work. This study also compared the Adam and SGD 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 presented the highest evaluation metrics when compared to the VGGNet model.

A new dataset was defined, using the raw data from ONC, 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 accuracy of 94.95% using CQT as the preprocessing filter for a ResNet-based convolutional neural network, providing a trade-off between model complexity and accuracy. Furthermore, a test was also executed using the whole dataset, combining the three scenarios into a single dataset, resulting in a classification accuracy of 84.13% with the same preprocessing and model.

A complete pipeline for the classification of underwater acoustic signals was proposed in this thesis, whose source code and data will be made publicly available. However, some issues were not addressed, which will be considered in future work. For instance, despite the promising results obtained in classification using the isolated dataset 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. Some strategies can be taken to address this issue, such as the formulation of a joint DL approach composed of an initial model to automatically detect the distance between the target vessel to the hydrophone (probably based on the SNR) and a second stage to properly classify the data.

Future work will also consider the application of other machine learning models to the task of classification of vessels from acoustic data, such as the MobileNet (SANDLER et al., 2018) and the SqueezeNet (TAN; LE, 2019), as these models have a good trade-off between accuracy

and model size in image classification tasks but are yet to be explored for the classification of acoustic data.

The present work also reported an extensive literature review containing classification methods that use convolutional Neural Networks alone (see Section 3.3). However, these strategies were not applied in the current solution, representing another task in future works. This approach can benefit the preprocessing optimisation through automatically learning a CNN based processing method, providing representations focused on the application of DL models. To achieve this result, Auto Encoders (HINTON; SALAKHUTDINOV, 2006), Variational Auto Encoders (KINGMA; WELLING, 2013), and U-Net (RONNEBERGER; FISCHER; BROX, 2015) models can be explored.

REFERENCES

ABRAHAM, Douglas A. **Underwater Acoustic Signal Processing Modeling, Detection, and Estimation**. Switzerland AG: Springer Nature, 2019.

BACH, Nhat Hoang; VU, Le Ha; NGUYEN, Van Duc. Classification of Surface Vehicle Propeller Cavitation Noise Using Spectrogram Processing in Combination with Convolution Neural Network. **Sensors**, v. 21, n. 10, 2021. ISSN 1424-8220. DOI: 10.3390/s21103353. Available from: https://www.mdpi.com/1424-8220/21/10/3353.

BERG, H.; HJELMERVIK, K. T. Classification of anti-submarine warfare sonar targets using a deep neural network. In: OCEANS 2018 MTS/IEEE Charleston. [S.l.: s.n.], 2018. P. 1–5. DOI: 10.1109/oceans.2018.8604847.

BOASHASH, B. **Time Frequency Signal Analysis and Processing: A Comprehensive Reference**. [S.l.]: Elsevier, 2003. ISBN 9780080443355. Available from: https://books.google.com.au/books?id=AtcSlAEACAAJ.

BOASHASH, Boualem (Ed.). **Time-Frequency Signal Analysis and Processing**. Second Edition. Oxford: Academic Press, 2016.

BOCANEGRA, J. A. et al. A novel approach to port noise characterization using an acoustic camera. **Sci Total Environ**, v. 808, p. 151903, 2022. ISSN 0048-9697. DOI: 10.1016/j.scitotenv.2021.151903.

BOGERT, B. P.; HEALY, Mjr; TUKEY, J. W. The quefrency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. **proceedings of the symposium on time**, 1963.

BORNHÖFT. **icListen - The new standard in broadband digital marine acoustics**. [S.l.: s.n.], 2022. [Online; accessed September 01, 2022]. Available from: https://www.xn--bornhft-e1a.de/wp-content/uploads/2017/01/Ocean-Sonics-icListen-HF-AF-and-LF-Brochure.pdf.

BROWN, Judith C. Calculation of a constant Q spectral transform. **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 89, n. 1, p. 425–434, 1991.

BROWNLEE, J. **Better Deep Learning: Train Faster, Reduce Overfitting, and Make Better Predictions**. [S.l.]: Machine Learning Mastery, 2018. Available from: https://books.google.com.au/books?id=T1-nDwAAQBAJ.

CAO, Xu et al. Deep learning-based recognition of underwater target. In: 2016 IEEE International Conference on Digital Signal Processing (DSP). [S.l.]: Ieee, 2016. P. 89–93. DOI: 10.1109/icdsp.2016.7868522.

CHAI, Junyi et al. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. **Machine Learning with Applications**, v. 6, p. 100134, 2021. ISSN 2666-8270. DOI: https://doi.org/10.1016/j.mlwa.2021.100134. Available from: https://www.sciencedirect.com/science/article/pii/S2666827021000670.

CHAKRABORTY, Bishwajit et al. **System for classifying seafloor roughness,US6763303B2**. [S.l.: s.n.], 2002.

CHEN, Liang et al. Underwater Acoustic Target Classification with Joint Learning Framework and Data Augmentation. In: IEEE. 2022 5th International Conference on Artificial Intelligence and Big Data (ICAIBD). [S.l.: s.n.], 2022. P. 23–28.

CHEN et al. Development of Tactile Imaging for Underwater Structural Damage Detection. **Sensors**, Mdpi Ag, v. 19, n. 18, p. 3925, Sept. 2019. ISSN 1424-8220. DOI: 10.3390/s19183925. Available from: http://dx.doi.org/10.3390/s19183925.

CHILDERS, D.G.; SKINNER, D.P.; KEMERAIT, R.C. The cepstrum: A guide to processing. **Proceedings of the IEEE**, v. 65, n. 10, p. 1428–1443, 1977. DOI: 10.1109/proc.1977.10747.

CHOI, Jongkwon; CHOO, Youngmin; LEE, Keunhwa. Acoustic classification of surface and underwater vessels in the ocean using supervised machine learning. **Sensors**, Multidisciplinary Digital Publishing Institute, v. 19, n. 16, p. 3492, 2019.

CHUAN, Ching-Hua; VASANA, Susan; ASAITHAMBI, Asai. Using wavelets and gaussian mixture models for audio classification. In: IEEE. 2012 IEEE International Symposium on Multimedia. [S.l.: s.n.], 2012. P. 421–426.

CINELLI, Lucas; CHAVES, Gabriel; LIMA, Markus. Vessel Classification through Convolutional Neural Networks using Passive Sonar Spectrogram Images. In: SIMPÓSIO Brasileiro de Telecomunicações e Processamento de Sinais (SBrT 2018). [S.l.: s.n.], 2018. DOI: 10.14209/sbrt.2018.340.

CYBENKO, G. Approximation by superpositions of a sigmoidal function. **Mathematics of Control, Signals, and Systems (MCSS)**, Springer London, v. 2, n. 4, p. 303–314, Dec. 1989. ISSN 0932-4194. DOI: 10.1007/bf02551274. Available from: http://dx.doi.org/10.1007/BF02551274.

DÄSTNER, Kaeye et al. Machine Learning Techniques for Enhancing Maritime Surveillance Based on GMTI Radar and AIS. In: IEEE. 2018 19th International Radar Symposium (IRS). [S.l.]: Ieee, 2018. P. 1–10. DOI: 10.23919/irs.2018.8447961.

DENG, J. et al. ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09. [S.l.: s.n.], 2009.

DENOS, Killian et al. Deep learning applied to underwater mine warfare. In: OCEANS 2017-Aberdeen. [S.l.]: Ieee, 2017. P. 1–7. DOI: 10.1109/oceanse.2017.8084910.

DOAN, Van-Sang; HUYNH-THE, Thien; KIM, Dong-Seong. Underwater acoustic target classification based on dense convolutional neural network. **IEEE Geoscience and Remote Sensing Letters**, Ieee, 2020.

DOMINGOS, Lucas CF et al. A machine learning method for vessel type classification with underwater acoustic data (under submission). **Access**, IEEE, 2022.

DOMINGOS, Lucas CF et al. A survey of underwater acoustic data classification methods using deep learning for shoreline surveillance. **Sensors**, mdpi, v. 22, n. 6, p. 2181, 2022.

DONALD, D.A. et al. **Feature Selection in the Wavelet Domain: Adaptive Wavelets**. [S.l.]: Elsevier, 2009. v. 3, p. 647–679. DOI: 10.1016/b978-044452701-1.00033-8. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84856083257&doi=10.1016%2fB978-044452701-1.00033-8&partnerID=40&md5=a887b3edef34b60b6f72e6724ac74505.

DUBOVIK, O. et al. Grand Challenges in Satellite Remote Sensing. **Front. Remote Sens.**, 2021.

ELECTRONICS, Seabird. **User Manual: 16plus IM V2**. [S.l.: s.n.], 2022. [Online; accessed September 01, 2022]. Available from: https://www.seabird.com/asset-get.download.jsa?id=59797800901.

ENDELT, Line Ortoft; COUR-HARBO, Anders la. Wavelets for sparse representation of music. In: IEEE. PROCEEDINGS of the Fourth International Conference onWeb Delivering of Music, 2004. EDELMUSIC 2004. [S.l.: s.n.], 2004. P. 10–14.

ERBE, Christine et al. The Effects of Ship Noise on Marine Mammals—A Review. **Frontiers in Marine Science**, v. 6, 2019. ISSN 2296-7745. DOI: 10.3389/fmars.2019.00606. Available from: https://www.frontiersin.org/article/10.3389/fmars.2019.00606.

FERGUSON, Brian G; LO, Kam W; THURAISINGHAM, Ranjit A. Sensor position estimation and source ranging in a shallow water environment. **IEEE Journal of Oceanic Engineering**, Ieee, v. 30, n. 2, p. 327–337, 2005.

FERGUSON, E. L. et al. Convolutional neural networks for passive monitoring of a shallow water environment using a single sensor. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.: s.n.], 2017. P. 2657–2661. DOI: 10.1109/icassp.2017.7952638.

FERRAO, Ryan R. **Underwater Masked Carrier Acoustic Communication: Modeling and Analysis**. [S.l.], 2018.

GHOSH, Joydeep et al. Integration Of Neural Classifiers For Passive Sonar Signals. In: LEONDES, Cornelius T. (Ed.). **Multidimensional Systems Signal Processing Algorithms and Application Techniques**. [S.l.]: Academic Press, 1996. v. 77. (Control and Dynamic Systems). P. 301–338. DOI: https://doi.org/10.1016/S0090-5267(96)80033-8. Available from: https://www.sciencedirect.com/science/article/pii/S0090526796800338.

GOPALAN, Kaliappan. Robust watermarking of music signals by cepstrum modification. In: IEEE. 2005 IEEE International Symposium on Circuits and Systems. [S.l.: s.n.], 2005. P. 4413–4416.

GREENWOOD, Donald D. The Mel Scale's disqualifying bias and a consistency of pitch-difference equisections in 1956 with equal cochlear distances and equal frequency ratios. **Hearing research**, Elsevier, v. 103, n. 1-2, p. 199–224, 1997.

HAN, Xing Cheng et al. Underwater acoustic target recognition method based on a joint neural network. **PloS one**, Public Library of Science San Francisco, CA USA, v. 17, n. 4, e0266425, 2022.

HANNUN, A. Y. et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. **Nat Med**, v. 25, n. 1, p. 65–69, 2019. ISSN 1078-8956 (Print) 1078-8956. DOI: 10.1038/s41591-018-0268-3.

HE, Kaiming et al. Deep residual learning for image recognition. In: PROCEEDINGS of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2016. P. 770–778.

HEESEMANN, Martin et al. Ocean Networks Canada: From Geohazards Research Laboratories to Smart Ocean Systems. **Oceanography**, June 2014. Available from: https://doi.org/10.5670/oceanog.2014.50.

HINTON, Geoffrey E; SALAKHUTDINOV, Ruslan R. Reducing the dimensionality of data with neural networks. **science**, American Association for the Advancement of Science, v. 313, n. 5786, p. 504–507, 2006.

HOCHREITER, S. et al. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In: KREMER, S. C.; KOLEN, J. F. (Eds.). **A Field Guide to Dynamical Recurrent Neural Networks**. [S.l.]: IEEE Press, 2001.

HOSSAIN, MD Zakir et al. A comprehensive survey of deep learning for image captioning. **ACM Computing Surveys (CsUR)**, ACM New York, NY, USA, v. 51, n. 6, p. 1–36, 2019. DOI: 10.1145/3295748.

HU, Gang et al. Deep learning methods for underwater target feature extraction and recognition. **Computational intelligence and neuroscience**, Hindawi, v. 2018, 2018.

HUANG, Guang-Bin; ZHU, Qin-Yu; SIEW, Chee-Kheong. Extreme learning machine: theory and applications. **Neurocomputing**, Elsevier, v. 70, n. 1-3, p. 489–501, 2006.

HUO, Guanying; WU, Ziyin; LI, Jiabiao. Underwater object classification in sidescan sonar images using deep transfer learning and semisynthetic training data. **IEEE access**, Ieee, v. 8, p. 47407–47418, 2020. DOI: 10.1109/access.2020.2978880.

IANDOLA, Forrest N et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. **arXiv preprint arXiv:1602.07360**, 2016.

IRFAN, Muhammad et al. DeepShip: An Underwater Acoustic Benchmark Dataset and a Separable Convolution Based Autoencoder for Classification. **Expert Systems with Applications**, Elsevier, p. 115270, 2021.

JÉZÉQUEL, Youenn; BONNEL, Julien; CHAUVAUD, Laurent. Potential for acoustic masking due to shipping noise in the European lobster (Homarus gammarus). **Marine Pollution Bulletin**, Elsevier, v. 173, p. 112934, 2021.

JOHANNESMA, PLM. The pre-response stimulus ensemble of neurons in the cochlear nucleus. In: IPO. SYMPOSIUM on Hearing Theory, 1972. [S.l.: s.n.], 1972.

KHISHE, M.; MOSAVI, M.R.; MORIDI, A. Classification of sonar target using hybrid particle swarm and gravitational search. **Mar. Technol.**, v. 3, n. 1, p. 1–13, 2016.

KHISHE, Mohammad; MOHAMMADI, Hassan. Passive sonar target classification using multi-layer perceptron trained by salp swarm algorithm. **Ocean Engineering**, Elsevier, v. 181, p. 98–108, 2019. DOI: 10.1016/j.oceaneng.2019.04.013.

KIM, Kyong-Il et al. A method for underwater acoustic signal classification using convolutional neural network combined with discrete wavelet transform. **International Journal of Wavelets, Multiresolution and Information Processing**, World Scientific, p. 2050092, 2021.

KINGMA, Diederik P; BA, Jimmy. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014.

KINGMA, Diederik P; WELLING, Max. Auto-encoding variational bayes. **arXiv preprint arXiv:1312.6114**, 2013.

KINGMA, Diederik P.; BA, Jimmy. Adam: A Method for Stochastic Optimization. In: ICLR (Poster). [S.l.: s.n.], 2015. Available from: http://arxiv.org/abs/1412.6980.

KOHONEN, Teuvo. Self-organized formation of topologically correct feature maps. **Biological Cybernetics**, v. 43, n. 1, p. 59–69, Jan. 1982. ISSN 0340-1200. DOI: 10.1007/bf00337288. Available from: http://dx.doi.org/10.1007/BF00337288.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. Imagenet classification with deep convolutional neural networks. **Advances in neural information processing systems**, v. 25, p. 1097–1105, 2012.

KUPERMAN, William; ROUX, Philippe. Underwater Acoustics. In: SPRINGER Handbook of Acoustics. New York, NY: Springer New York, 2007. P. 149–204. ISBN 978-0-387-30425-0. DOI: 10.1007/978-0-387-30425-0_5. Available from: https://doi.org/10.1007/978-0-387-30425-0%5C%5F5.

LAMPERT, Thomas A; O'KEEFE, Simon EM. On the detection of tracks in spectrogram images. **Pattern recognition**, Elsevier, v. 46, n. 5, p. 1396–1408, 2013.

LE, Hoang Thanh et al. Deep gabor neural network for automatic detection of mine-like objects in sonar imagery. **IEEE Access**, Ieee, v. 8, p. 94126–94139, 2020.

LECUN, Yann et al. Handwritten digit recognition with a back-propagation network. In: ADVANCES in neural information processing systems. [S.l.: s.n.], 1990. P. 396–404.

LEE, Chang-Hsing et al. Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features. **IEEE Transactions on Multimedia**, Ieee, v. 11, n. 4, p. 670–682, 2009.

LEROY, Claude C.; ROBINSON, Stephen P.; GOLDSMITH, Mike J. A new equation for the accurate calculation of sound speed in all oceans. **The Journal of the Acoustical Society of America**, v. 124, n. 5, p. 2774–2782, 2008. DOI: 10.1121/1.2988296. eprint: https://doi.org/10.1121/1.2988296. Available from: https://doi.org/10.1121/1.2988296.

LIU, Xiaoxiao; XU, Qingyang; WANG, Ning. A survey on deep neural network-based image captioning. **The Visual Computer**, Springer, v. 35, n. 3, p. 445–470, 2019. DOI: 10.1007/s00371-018-1566-y.

LOGAN, Beth. Mel frequency cepstral coefficients for music modeling. In: CITESEER. IN International Symposium on Music Information Retrieval. [S.l.: s.n.], 2000.

MANN, Steve; HAYKIN, Simon. The chirplet transform: A generalization of Gabor's logon transform. In: CITESEER. VISION interface. [S.l.: s.n.], 1991. v. 91, p. 205–212.

MCCANN, Erin et al. An underwater observation dataset for fish classification and fishery assessment. **Scientific data**, Nature Publishing Group, v. 5, n. 1, p. 1–8, 2018.

MCKENNA, Megan F. et al. Underwater radiated noise from modern commercial ships. **The Journal of the Acoustical Society of America**, v. 131, n. 1, p. 92–103, 2012. DOI: 10.1121/1.3664100. eprint: https://doi.org/10.1121/1.3664100. Available from: https://doi.org/10.1121/1.3664100.

MCMAHON, K. G. et al. Horizontal Lloyd mirror patterns from straight and curved nonlinear internal waves. **The Journal of the Acoustical Society of America**, v. 131, n. 2, p. 1689–1700, 2012. DOI: 10.1121/1.3666004. eprint: https://doi.org/10.1121/1.3666004. Available from: https://doi.org/10.1121/1.3666004.

MERCHANT, Nathan D. et al. Monitoring ship noise to assess the impact of coastal developments on marine mammals. **Marine Pollution Bulletin**, v. 78, n. 1, p. 85–95, 2014. ISSN 0025-326x. DOI: https://doi.org/10.1016/j.marpolbul.2013.10.058. Available from: https://www.sciencedirect.com/science/article/pii/S0025326X13006802.

MIAO, Yongchun et al. Underwater Acoustic Signal Classification Based on Sparse Time–Frequency Representation and Deep Learning. **IEEE Journal of Oceanic Engineering**, Ieee, 2021.

NASTASI, Marco et al. Parameters Affecting Noise Emitted by Ships Moving in Port Areas. **Sustainability**, v. 12, n. 20, 2020. ISSN 2071-1050. DOI: 10.3390/su12208742. Available from: https://www.mdpi.com/2071-1050/12/20/8742.

NEENAN, Sarah TV et al. Does masking matter? Shipping noise and fish vocalizations. In: THE effects of noise on aquatic life II. [S.l.]: Springer, 2016. P. 747–753.

O'SHAUGHNESSY, Douglas. **Speech communications: Human and machine**. [S.l.]: Addison-Wesley Publishing Company, 1987. P. 150. ISBN 9780201165203.

OPPENHEIM, A. V.; SCHAFER, R. W. From frequency to quefrency: a history of the cepstrum. **IEEE Signal Processing Magazine**, v. 21, n. 5, p. 95–106, 2004. DOI: 10.1109/msp.2004.1328092.

ORENGO, H.A.; GARCIA-MOLSOSA, A. A brave new world for archaeological survey: Automated machine learning-based potsherd detection using high-resolution drone imagery. **Journal of Archaeological Science**, v. 112, p. 105013, 2019. ISSN 0305-4403. DOI: https://doi.org/10.1016/j.jas.2019.105013. Available from: http://www.sciencedirect.com/science/article/pii/S0305440319301001.

OXENHAM, Andrew. How We Hear: The Perception and Neural Coding of Sound. **Annual Review of Psychology**, v. 69, Jan. 2018. DOI: 10.1146/annurev-psych-122216-011635.

PATTERSON, Roy D et al. An efficient auditory filterbank based on the gammatone function. In: 7. A meeting of the IOC Speech Group on Auditory Modelling at RSRE. [S.l.: s.n.], 1987. v. 2.

PERRINE, Kenneth A et al. **University of Texas Applied Research Laboratory Nov. 2009 Five-Element Acoustic Underwater Dataset**. [S.l.], 2009.

POWERS, David. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. **Mach. Learn. Technol.**, v. 2, Jan. 2008.

ROBBINS, Herbert; MONRO, Sutton. A Stochastic Approximation Method. **The Annals of Mathematical Statistics**, Institute of Mathematical Statistics, v. 22, n. 3, p. 400–407, 1951. DOI: 10.1214/aoms/1177729586. Available from: https://doi.org/10.1214/aoms/1177729586.

RONNEBERGER, Olaf; FISCHER, Philipp; BROX, Thomas. U-net: Convolutional networks for biomedical image segmentation. In: SPRINGER. INTERNATIONAL Conference on Medical image computing and computer-assisted intervention. [S.l.: s.n.], 2015. P. 234–241.

ROSSI, E. et al. Assessing the Underwater Ship Noise Levels in the North Tyrrhenian Sea. **Adv Exp Med Biol**, v. 875, p. 943–9, 2016. ISSN 0065-2598 (Print) 0065-2598. DOI: 10.1007/978-1-4939-2981-8_116.

RUMELHART, David E.; HINTON, Geoffrey E.; WILLIAMS, Ronald J. Learning representations by back-propagating errors. **nature**, Nature Publishing Group, v. 323, n. 6088, p. 533–536, 1986. DOI: 10.1038/323533a0.

RUSSAKOVSKY, Olga et al. Imagenet large scale visual recognition challenge. **International journal of computer vision**, Springer, v. 115, n. 3, p. 211–252, 2015. DOI: 10.1007/s11263-015-0816-y.

SANDLER, Mark et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In: PROCEEDINGS of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2018. P. 4510–4520.

SANTOS-DOMÍNGUEZ, David et al. ShipsEar: An underwater vessel noise database. **Applied Acoustics**, Elsevier, v. 113, p. 64–69, 2016.

SATO, Atsushi; YAMADA, Keiji. Generalized Learning Vector Quantization. In: TOURETZKY, D.; MOZER, M. C.; HASSELMO, M. (Eds.). **Advances in Neural Information Processing Systems**. [S.l.]: MIT Press, 1996. v. 8. Available from: https://proceedings.neurips.cc/paper/1995/file/9c3b1830513cc3b8fc4b76635d32e692-Paper.pdf.

SHEN, S.; YANG, H.; LI, J. Improved Auditory Inspired Convolutional Neural Networks for Ship Type Classification. In: OCEANS 2019 - Marseille. [S.l.: s.n.], 2019. P. 1–4. DOI: 10.1109/oceanse.2019.8867312.

SHEN, Sheng et al. Auditory Inspired Convolutional Neural Networks for Ship Type Classification with Raw Hydrophone Data. **Entropy**, 2018.

SHEN, Sheng et al. Ship type classification by convolutional neural networks with auditory-like mechanisms. **Sensors**, Multidisciplinary Digital Publishing Institute, v. 20, n. 1, p. 253, 2020. DOI: 10.3390/s20010253.

SHI, Min; XU, Xi. Underwater target recognition based on wavelet packet entropy and probabilistic neural network. In: 2013 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC 2013). [S.l.]: Ieee, 2013. P. 1–3.

SIMONYAN, Karen; ZISSERMAN, Andrew. Very Deep Convolutional Networks for Large-Scale Image Recognition. In: BENGIO, Yoshua; LECUN, Yann (Eds.). **3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings**. [S.l.: s.n.], 2015. Available from: http://arxiv.org/abs/1409.1556.

SIMONYAN, Karen; ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, 2014.

SOKOLOVA, Marina; LAPALME, Guy. A systematic analysis of performance measures for classification tasks. **Information Processing & Management**, v. 45, n. 4, p. 427–437, 2009. ISSN 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2009.03.002. Available from: https://www.sciencedirect.com/science/article/pii/S0306457309000259.

STEVENS, Stanley Smith; VOLKMANN, John; NEWMAN, Edwin Broomell. A scale for the measurement of the psychological magnitude pitch. **The journal of the acoustical society of america**, Acoustical Society of America, v. 8, n. 3, p. 185–190, 1937.

SUN, Qinggang; WANG, Kejun. Underwater single-channel acoustic signal multitarget recognition using convolutional neural networks. **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 151, n. 3, p. 2245–2254, 2022.

SUN, Shiliang et al. A Survey of Optimization Methods From a Machine Learning Perspective. **IEEE Transactions on Cybernetics**, v. 50, n. 8, p. 3668–3681, 2020. DOI: 10.1109/tcyb.2019.2950779.

TAN, Mingxing; LE, Quoc. Efficientnet: Rethinking model scaling for convolutional neural networks. In: PMLR. INTERNATIONAL conference on machine learning. [S.l.: s.n.], 2019. P. 6105–6114.

TERAYAMA, Kei et al. Integration of sonar and optical camera images using deep neural network for fish monitoring. **Aquacultural Engineering**, v. 86, p. 102000, 2019. ISSN 0144-8609. DOI: https://doi.org/10.1016/j.aquaeng.2019.102000. Available from: http://www.sciencedirect.com/science/article/pii/S0144860919300640.

TIAN, Shengzhao et al. Deep convolution stack for waveform in underwater acoustic target recognition. **Scientific Reports**, v. 11, n. 1, p. 9614, 2021. ISSN 2045-2322. DOI: 10.1038/s41598-021-88799-z. Available from: https://doi.org/10.1038/s41598-021-88799-z.

TIAN, Yingjie; ZHANG, Yuqi. A comprehensive survey on regularization strategies in machine learning. **Information Fusion**, v. 80, p. 146–166, 2022. ISSN 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2021.11.005. Available from: https://www.sciencedirect.com/science/article/pii/S156625352100230X.

TSAI, Wei-Ho; LIN, Hao-Ping. Background music removal based on cepstrum transformation for popular singer identification. **IEEE transactions on audio, speech, and language processing**, Ieee, v. 19, n. 5, p. 1196–1205, 2010.

TZANETAKIS, George; ESSL, Georg; COOK, Perry. Audio analysis using the discrete wavelet transform. In: CITESEER. PROC. conf. in acoustics and music theory applications. [S.l.: s.n.], 2001. v. 66.

URICK, R.J.; AGENCY, United States. Defense Advanced Research Projects. **Sound Propagation in the Sea**. [S.l.]: Darpa, 1979. Available from: https://books.google.com.au/books?id=KgpRAAAAMAAJ.

VAHIDPOUR, Vahid; RASTEGARNIA, Amir; KHALILI, Azam. An automated approach to passive sonar classification using binary image features. **Journal of Marine Science and Application**, v. 14, July 2015. DOI: 10.1007/s11804-015-1312-z.

VALDENEGRO-TORO, Matias; PRECIADO-GRIJALVA, Alan; WEHBE, Bilal. Pre-trained Models for Sonar Images. **CoRR**, abs/2108.01111, 2021. arXiv: 2108.01111. Available from: https://arxiv.org/abs/2108.01111.

WANG, Dezhi et al. Large-scale whale call classification using deep convolutional neural network architectures. In: IEEE. 2018 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC). [S.l.: s.n.], 2018. P. 1–5.

WANG, Dezhi et al. Passive Ship Localization in a Shallow Water Using Pre-trained Deep Learning Networks. In: PROCEEDINGS of the 23rd International Congress on Acoustics : integrating 4th EAA Euroregio 2019. [S.l.: s.n.], 2019. P. 9–13.

WANG, Xingmei et al. Underwater Acoustic Target Recognition: A Combination of Multi-Dimensional Fusion Features and Modified Deep Neural Network. **Remote Sensing**, Mdpi Ag, v. 11, n. 16, p. 1888, Aug. 2019. ISSN 2072-4292. DOI: 10.3390/rs11161888. Available from: http://dx.doi.org/10.3390/rs11161888.

WEILL, Alain. Acoustic Waves, Scattering. In: ENCYCLOPEDIA of Remote Sensing. New York, NY: Springer New York, 2014. P. 13–16. ISBN 978-0-387-36699-9. DOI:

10.1007/978-0-387-36699-9_3. Available from:
https://doi.org/10.1007/978-0-387-36699-9%5C%5F3.

WILSON, Ashia C et al. The marginal value of adaptive gradient methods in machine learning. **Advances in neural information processing systems**, v. 30, 2017.

WU, Qi et al. Visual question answering: A survey of methods and datasets. **Computer Vision and Image Understanding**, v. 163, p. 21–40, 2017. Publisher: Elsevier. DOI: 10.1016/j.cviu.2017.05.001.

XIE, Zeke et al. Adaptive Inertia: Disentangling the Effects of Adaptive Learning Rate and Momentum. In: CHAUDHURI, Kamalika et al. (Eds.). **Proceedings of the 39th International Conference on Machine Learning**. [S.l.]: Pmlr, July 2022. v. 162. (Proceedings of Machine Learning Research), p. 24430–24459. Available from:
https://proceedings.mlr.press/v162/xie22d.html.

XING, Zhengzheng; PEI, Jian; KEOGH, Eamonn. A brief survey on sequence classification. **ACM Sigkdd Explorations Newsletter**, ACM New York, NY, USA, v. 12, n. 1, p. 40–48, 2010.

XU, Chunying et al. Review of Underwater Cable Shape Detection. **Journal of Atmospheric and Oceanic Technology**, American Meteorological Society, Boston MA, USA, v. 33, n. 3, p. 597–606, 2016. DOI: 10.1175/jtech-d-15-0112.1. Available from:
https://journals.ametsoc.org/view/journals/atot/33/3/jtech-d-15-0112%5C%5F1.xml.

YANG, H. et al. Underwater acoustic target recognition using SVM ensemble via weighted sample and feature selection. In: IEEE. 2016 13th International Bhurban Conference on Applied Sciences and Technology (IBCAST). [S.l.: s.n.], 2016. P. 522–527. DOI: 10.1109/ibcast.2016.7429928.

YANG, Haesang et al. Underwater Acoustic Research Trends with Machine Learning: Active Sonar Applications. **Journal of Ocean Engineering and Technology**, v. 34, n. 4, p. 277–284, 2020. Publisher: The Korean Society of Ocean Engineers. DOI: 10.26748/ksoe.2020.015.

YANG, Haesang et al. Underwater Acoustic Research Trends with Machine Learning: General Background. **Journal of Ocean Engineering and Technology**, v. 34, n. 2, p. 147–154, 2020. Publisher: The Korean Society of Ocean Engineers. DOI: 10.26748/ksoe.2020.015.

YANG, Haesang et al. Underwater Acoustic Research Trends with Machine Learning: Passive Sonar Applications. **Journal of Ocean Engineering and Technology**, v. 34, n. 3, p. 227–236, 2020. Publisher: The Korean Society of Ocean Engineers. DOI: 10.26748/ksoe.2020.015.

YANG, Honghui et al. A deep convolutional neural network inspired by auditory perception for underwater acoustic target recognition. **Sensors**, Multidisciplinary Digital Publishing Institute, v. 19, n. 5, p. 1104, 2019.

ZHAO, Minghang et al. Deep Residual Shrinkage Networks for Fault Diagnosis. **IEEE Transactions on Industrial Informatics**, v. 16, n. 7, p. 4681–4690, 2020. DOI: 10.1109/tii.2019.2943898.

INDEX

A

ACT, 37

AdaGrad, 29

Adam, 29, 60, 64, 70

ADCNN, 42, 43

AIS, 16, 51, 53–55

ANN, 28

AP, 33

C

CNN, 17, 28–30, 35–41, 43–45, 50, 51, 57, 58, 62, 71, 72, 75

CQT, 27, 40, 51, 57, 58, 60–67, 70–72, 74

CSDM, 35, 36

CTD, 53, 55

D

DCT, 25

DL, 17, 23, 28–32, 34, 42, 44, 45, 47, 48, 56, 63, 64, 66, 70, 72, 74, 75

DWT, 37

E

EFP, 37

ELM, 43, 44

F

FN, 32

FP, 32

FT, 24, 26, 27, 44

G

GFCC, 40, 41

GPS, 16

H

HSV, 58

I

ILSVRC, 30

K

KNN, 40

L

LVQ, 44, 45

M

mAP, 33

MDT, 46

MFCC, 25, 40, 41

ML, 23, 32, 33, 36, 37, 39, 42, 50, 57

MSRDN, 44

N

NN, 29, 36–38, 40–43

O

ONC, 39, 44, 51, 53, 62, 72, 74

P

PCA, 40

R

ReLu, 28, 30, 31, 51, 58

revcor, 26

RMSP, 29

S

SCAE, 40, 62, 63, 70

SGD, 29, 60, 64–66, 70, 74

SNR, 17, 22, 23, 36–38, 42, 44, 55, 64, 68, 72, 74

SOM, 44

STFT, 24, 27, 37, 38

SVM, 36, 37, 40

T

TFFNet, 37

TN, 32

TP, 32

U

UATC, 43, 44, 57

V

VG, 30, 31

VGG, 17, 30, 31, 40, 48, 50, 51, 58–60, 62, 64–67, 70, 72, 74

VPCN, 35