Time and Cost gains enabled by Machine Learning for Environmental Impact Assessments

Mariana Semião[1], Bénédicte Hoofd[2], Erica Cruz[2], Diana Almeida[2], Susana Vieira[1], and Guilherme Vaz[2]

[1]Instituto Superior Técnico, Lisbon, Portugal
[2]blueOASIS, Ericeira, Portugal

Abstract: *An Environmental Impact Assessment (EIA) is a process aiming to assess a priori the impact that a large scale project would have on an ecosystem, to propose mitigation measures to minimise potential impact, and also to monitor the effectiveness of the measures during the execution of the project. The process often lasts up to 4 years. It is a mandatory step for the deployment of any new projects at sea such as the construction of a harbor terminal, the installation of an offshore wind farm or an aquaculture cage. One common threat that such projects pose is an increase of underwater noise. This topic is particularly important when assessing the impact on marine mammals and fish because they use underwater sound to interact for feeding, mating, or socializing. One of the methods to assess the impact of underwater noise on marine mammals is to estimate the presence/absence of animals in the specific region of the project using passive acoustic monitoring. This process is usually expensive: it requires the expensive installation of hydrophones, the manual recovery of the recorded data and the equipment with divers and ships, and finally the long manual analysis conducted by an expert to distinguish the different types of dolphins' vocalizations. This paper evaluates the time and cost gains that a Machine Learning algorithm can bring for the detection of acoustic sources of a specific site. A Deep Learning ensemble model is trained to detect dolphins in a coastal environment in Portugal, using a manually labelled dataset. The paper establishes the minimum requirements in terms of training dataset to allow for an automated, accurate, and fast analysis of the dolphins' behavior in the area. The requirements consider the size of the dataset, but also the class balance and data processing required for the analysis. It was found that for the analysis of the used dataset, the labelling efforts can be reduced by a factor of 15.*

Keywords: *Underwater Acoustics, Dolphin Detection, Deep Learning Ensemble Model, Feature Fusion, Environmental Impact Assessment*

1. INTRODUCTION

Underwater soundscapes result from a combination of anthropogenic and natural sources [1]. The contribution of each source will depend on the characteristics of both the sound and the environment that affect sound propagation [2]. Several marine species use sound as their main sense for relevant activities such as feeding, communication and breeding [3]. Due to the increasing of anthropogenic offshore activities, marine life is considerably impacted and this topic has been gaining attention from several stakeholders. Guidelines for underwater noise reduction from shipping, issued by International Maritime Organization [4], or the Marine Strategy Framework Directive, in Europe [5], reflect the need to manage underwater noise pollution. Additionally, Environmental Impact Assessment (EIA) studies, for new offshore projects, such as offshore renewables, oil&gas, and aquaculture, usually require promoters to assess the impact of underwater noise resulting from the project. These noise sources and their impact on marine environment can be monitored using passive acoustic monitoring systems, such as hydrophones. This is a long and costly task that can often impact their business model and cause delay in the projects' deployment, as an example, a three years cetacean monitoring can cost 370,000€. The regulations and recommendations force therefore the users to find new approaches to monitor their activity underwater.

Deep Learning (DL) applied to the field of underwater acoustics is a promising approach that can significantly decrease the time of the analysis and decrease its costs. For example, [6] and [7] proposed complex DL algorithms to analyse the spectogram of the sound data and they showed that Convolutional Neural Networks (CNN) are in fact a powerful way to conduct in depth analysis of recordings for the detection of specific sounds, like dolphin calls. Simpler models can however be used, such as the one presented by the authors in [8]. This work proposed to train a simple 3-layer CNN model using the spectogram of underwater recordings, in order to automate the identification of dolphins and ships. The model was trained and validated using two open source datasets: ShipsEar ([9]), with several types of vessels and background environment noise and DOSITS ([10]), with clean vocalization of dolphins recordings. The model showed very good results on the validation set based on those two datasets, but the model performed significantly worse when using a different dataset that was collected during the Robotic Experimentation and Prototyping using Maritime Uncrewed Systems (REPMUS), [11]. The authors showed that one critical factor in the training was the quality of the dataset and its relevance to a "real life" application. In this paper, the model architecture remained similar, but it the performance was improved by using REPMUS data as a training set.

In addition, in [8] only the Mel spectrogram and pseudo-derivatives are used as features to identify dolphins. However, as presented in [12] using different features such as the Mel-frequency cepstral coefficients (MFCC), the Log-Mel (LM) and the Zero-crossing rate (ZCR), the results can be severely improved. Additionally, the enhancement of the Mel spectogram as proposed in [13] can also provide informative features that will be tested in the current work. Finally, model ensembling techniques, such as stacked generalization, can take advantage of the best qualities from multiple models to improve the global performance. Those different techniques were also tested here and their evaluation to the problem at hand is also presented.

The performance of a DL algorithm is not only related to its structure but also to the adequate selection of the hyperparameters. In [8], the dropout, learning rate and kernel regularizer were tuned, using Bayesian optimisation. However, a Hyperband Tuner can lead to better results and in a shorter time [14]. In fact, it runs random configurations for a fewer number of iterations per configuration, then, using previous results it selects candidates for longer runs, increasing

the number of iterations as it decreases the number of best configurations. In the current work, Hyperband optimisation was used as well, in order to optimise 6 hyperparameters.
In this paper, Sections 2 and 3 introduce the dataset used and how it is processed. In Section 4, the ensemble model is explained, presenting results and comparing the features extracted. In Section 5, the automated method is compared with manual labelling in terms of time gains. The papers then concludes with a proposition of future work to further increase the performance of the model and increase the gains.

2. DATASETS AND MANUAL LABELLING

The dataset used in this work corresponds to 35 `.wav` files of 10 minutes recorded during the REPMUS 2021 exercise. For each file, manual labelling was conducted by a marine biologist expert. The spectrogram of the file was visualised and listened to using the software *Audacity*. By going through each second of the recording, the biologist would clearly identify various dolphins sounds such as burst pulse sounds (BPS), gulps (GP), grunts (GR), whistles (WH) and squeaks (SQ), according to [15]. In all files used for the analysis, a distant vessel (VL) was always present. Any other identified sound that could not be classified was grouped into one label: unknown events (UV). A labeling `.txt` file was created for each recording. It contained the start time, end time and classification of each identified sound. An example of a typical 10 min spectrogram, a WH instance and SQ instance are presented in Figure 1.
After analysing the labels, it appeared that the dataset was highly imbalanced, with 97.21% more vessel instances than dolphin instances. To reduce the dataset imbalance, which often leads to a decrease of performance when used for DL applications, the classes WH and SQ were considered to be on dolphin class (DN) as they represent similar frequency noises, the classes BPS, GP and GR, which represented only 0.17% of the total instances where grouped with the VL class in the class background noise (BK). This class was then randomly undersampled to match the size of the class DN, leaving the classification as a binary dataset with only two classes: dolphin (DN) and background noise (BK).

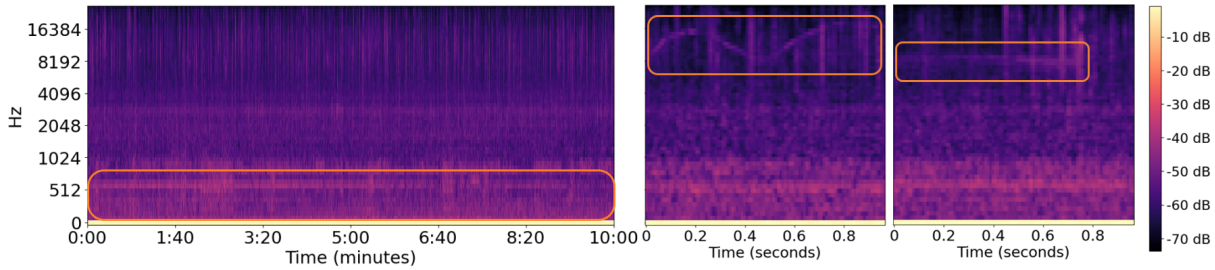

Figure 1: Left: typical spectrogram with a distant ship. Middle - example of whistle. Right - example of a squeak.

3. AUTOMATED PRE-PROCESSING FOR DEEP LEARNING APPLICATION

The same pre-processing as in [8] and [16] is used to prepare the recording for its analysis with the CNN using the python library `librosa`. Using a sample rate of 52734, a Hann Window with hop of 1024 and length of 2048, each recording was transformed into a spectogram using

a Short-time Fourier Tansform (STFT). The values of the matrix were then squared to obtain the power spectral density. Next, the spectogram was converted into a Mel scale spectogram with 60 bands, and then split into windows corresponding to one second of the recording.
As presented in Section 1, different features were extracted after this step to test their effectiveness in improving the dolphin detection: the first and second pseudo-derivatives (Dev1, Dev2) followed by a scaling process were calculated as in [8] and [16]; the Log-Mel, which is the logarithm of the Mel scale spectogram (MS); the MFCC, which is the linear cosine transform of the Log-Mel spectrum; and finally the ZCR, that is defined in [12] as ”the number of zero crossings in the time domain within one second”. Given the shape of the output matrix from each of those pre-processing methods, the various features were stacked differently. The final matrix shape for each 1-second recording and each stacking process can be found in Table 1.

Features	**Model input shape**	**MCC**	**F1 Macro**	**F1 Weighted**
MS+Dev1+Dev2	(60, 50, 3)	**0.44932**	**0.41669**	**0.59995**
MS+MFCC+LM	(60, 50, 3)	0.29986	0.37493	0.49379
MS+LM+BW	(60, 50, 3)	0.35341	0.40915	0.50364
MS+LM+MFCC	(180, 50)	0.41209	0.41272	0.56950
MS+LM+MFCC+ZCR	(181, 50)	0.35178	0.37091	0.53518
MS+LM+ZCR	(121, 50)	0.32748	0.36613	0.52168
BW	(40,50)	**0.20127**	**0.59971**	**0.61755**

Table 1: F1-score and MCC for models with different features.

One of the challenges rising from the use of ”real life” data like REPMUS is that the spectogram obtained from the previous pre-processing is not as clear as the ones from datasets such as DOSITS, which makes the identification significantly more difficult. To overcome this issue, traditional computer vision techniques were used to enhance the spectogram image. As in [13], a contrast stretch is applied to the STFT spectogram to increase its dynamic grayscale range. This was done by subtracting 20 to where the value of the spectrogram was in the interval $[20, 70]$ or adding 80 when the value was in $[80, 120]$. This is followed by the binarization of the image, i.e. the transformation of a grayscale image into a black and white (BW) image, using a threshold of 100. This means that the grayscale values above or below 100 will be white or black, respectively. Different mathematical morphology techniques were tested as a final step of enhancement. Dilation and erosion consist in, respectively, expanding or shrinking white areas on the binary image (i.e. pixels with value of 1). Opening and closing result from the application of the previous two operators in a specific order: This modifies the image by eliminating noise in black or white, while maintaining the main information in the image. For the dataset in question, closing resulted in the more clear images. This method consists in simply dilating and then eroding, removing dark dots. These steps are presented in Figure 2. Because the dolphin whistle and squeaks correspond to high frequencies, and because the bottom of the binarized image is always just a white section, as shown in the most right image in Figure 2, the lower part of the binarized spectrogram was cut from the analysis.

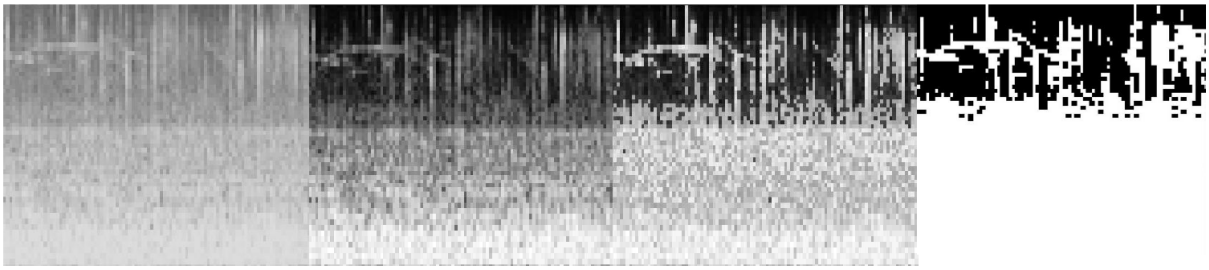

Figure 2: Left - original spectogram. Middle left - equalized histogram. Middle right - with stretched contrast. Right - binarized and closed.

4. MODEL ARCHITECTURE AND FEATURE EVALUATION

One main architecture, based on the work in [8] is used. It consists of a first convolutional layer with 24 neurons, followed by two convolutional layers with 48 neurons each. All three layers have kernel sizes of (5,5), strides of (1,1) and MaxPooling in between them. The first two layers also have padding. Following the convolutional layers, the output is flattened, and then followed by two dense layers with dropout. The final dense layer has 2 neurons, due to the 2 classes. The activation functions are ReLU (Rectified Linear Unit) on all layers except the final output layer which has an activation function of Softmax.
Each feature, as described in Section 3 is tested with that architecture. To do so, the processed data is split into 3 sets for training (70%), validation (15%), and testing (15%). Each set was then divided in two, to keep ”unseen data” to train the meta-learner, as later explained in this section. The model architecture is adapted to the shape of the input matrix as presented in Table 1. Each model is then trained for 50 epochs or stopped earlier if over-fitting is observed. Three metrics of comparison are used: - the Matthew's Correlation Coefficient (MCC), which is used to assess the quality of a problem with imbalanced data; - the F1-score (macro), which corresponds to the arithmetic average of every classes' F1-score; - the F1-score (weighted), which, similarly, corresponds to the weighted average of every classes' F1-score. F1-score Weighted is more relevant for imbalance data than Macro F1-score. It is important to note that the MCC varies between -1 and 1, consequently, an MCC of 0 corresponds to random predictions, meaning positive and negative MCC correspond to good and bad model performance, respectively. The comparison of all features through this metrics can be found in Table 1.

The best performing models are MS+Dev1+Dev2, hereafter referred to as *Model 1*, and BW model, referred as *Model 2*. To increase their performance, both models were optimized in terms of dropout and ADAM optimizer parameters (learning rate, β_1 and β_2, i.e. the exponential decay rate for the momentum and velocity term, respectively, and ϵ) using the Hyperband method. After optimization, *Model 1* achieved the following results: 0.71 and 0.74 F1-score for DN and BK, respectively; 0.72 Global Weighted F1-score and 0.4453 MCC. This represents a great improvement comparing to the validation on REPMUS done in [8], where dolphins were often wrongly detected and with low certainty. *Model 2* was tested after optimization, resulting in: 0.52 and 0.68 F1-score for DN and BK, respectively; 0.62 Global Weighted F1-score and 0.2020 MCC, as presented in Table 2.

As the previous two models work with different features from the same data, they can be stacked using a meta-learner, i.e., used simultaneously to predict more accurately the presence of dolphins, as presented in Figure 3. Stacked generalization (i.e. the ensemble technique used) uses a meta-learning algorithm that learns to aggregate the predictions of several other machine

learning models; it takes advantage of the best capabilities of each of these models, in order to outperform the ensembled models. Using data that was not used to train *Model 1* and *2*, the meta-learner was trained using the predictions of *Model 1* and *2* as inputs and the expected outputs. The model hyperparameters were optimized using Hyperband, in the same way as *Model 1* and *2*. After this optimization, the *Ensemble Model* yielded the following results: 0.70 and 0.81 F1-score for DN and BK, respectively; 0.76 Global Weighted F1-score and 0.5394 MCC, as presented in Table 2.

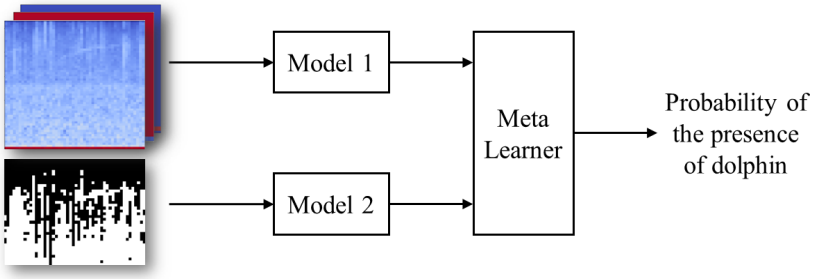

Figure 3: Ensemble model structure.

Model	**Input Shape**	**MCC**	**F1 Macro**	**F1 Weighted**
Model 1 (MS+Dev1+Dev2)	(60, 50, 3)	0.44531	0.72266	0.72325
Model 2 (BW)	(40,50)	0.20200	0.60088	0.61733
Ensemble Model	[(60, 50, 3),(40,50)]	**0.53940**	**0.75480**	**0.75947**

Table 2: F1-score and MCC for optimized final models.

5. COMPARISON BETWEEN THE MANUAL AND AUTOMATED DETECTION

A sensitivity analysis was conducted to assess the dataset requirements to reach the performance achieved in Section 4. The influence of the amount of labelled data for a balanced dataset and the influence of the percentage of imbalance on the MCC and F1-score are presented in Figure 4. It can be seen that, with the current ensemble model, only 60 min of labelled recordings are necessary to reach an F1-score of 0.70 on DN class. This corresponds to 28 min of dolphin recordings and 32 min of background recordings. Considering 15 min of labor for the labelling of a 10 min recording, this corresponds to a labelling effort of only 90 min. However, more work is actually needed because dolphins sounds are relatively rare, so files with multiple dolphin noises are selected thanks to a quick screening of all the recordings before the start of the extensive and time consuming labelling process. This is particularly important given the high importance of having a balanced dataset, as shown in Figure 4 (right). Screening the recording is a rapid process: only half of the files are looked at, and the expert spends on average 1.5 min to screen a file of 10 min. Considering the current dataset with 560 files of 10 min, the screening process lasts 7h. After this initial screening, the expert must fully label 60 min of recordings. For recordings with many different sounds, an expert would take 20 min to label a 10 min file. This leads to a labelling time of 2 hours. Therefore, in total, the screening and the labelling effort are only 9h. After training the algorithm with those labeled files, all other files can then be automatically analysed in no-time instead of spending 140h to manually detect the dolphins. To use the same technique on another site, it is expected that a step of transfer learning will be

necessary, which means that the time for screening and manual label of a limited amount of files will still be required. This is nonetheless a significant improvement compared to a fully manual analysis.

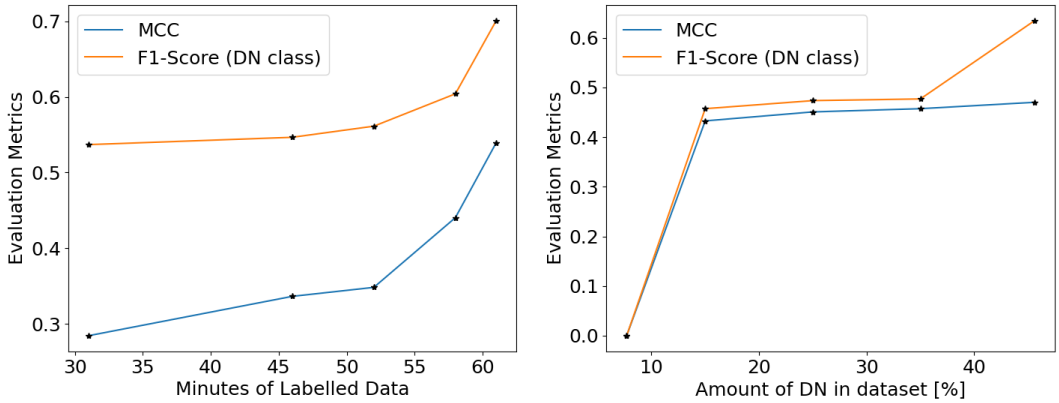

Figure 4: Sensitivity analysis: Left - dataset size, Right - dataset balance.

6. CONCLUSIONS

In this paper, a stacked ensemble model of two simple CNN was built, tested and optimised to detect dolphins from real life hydrophone recordings, reaching a Weighted F1-score of 0.7595 which corresponds to a significant improvement compared to the previous work using the same simple CNN. In addition, a sensitivity analysis was conducted to assess how much manual labelling is required to have enough data to accurately train the model. It was shown that a balanced dataset containing 30 min of recording for each class is enough to reach a Weighted F1-score of 0.76. The algorithm can then be used to automatically label with the same accuracy any recordings in the same environment. This represents a significant gain in time and costs for tasks such as EIA: just for the REPMUS exercice, the labelling time was divided by 15. The sensitivity analysis should however be continued with more data to find the optimal compromise between accuracy and effort in manual labelling. In addition, the technique of transfer learning should also be analysed to further decrease the dataset requirement to re-train the algorithm for the automatic detection of the same classes or different classes in a new site.

REFERENCES

[1] Merchant, N. D., et al.: ”Measuring acoustic habitats”, *Methods in Ecology and Evolution* **6**, 257–265 **(2015)**.

[2] Farcas, A., Thompson, P. M. and Merchant, N. D.: ”Underwater noise modelling for environmental impact assessment”, *Environmental Impact Assessment Review* **57**, 114–122 **(2016)**.

[3] Cruz, E., Lloyd, T., Lafeber, F. H., Bosschers, J., Vaz, G., and Djavidnia, S.: ”The SOUNDS project: towards effective mitigation of underwater noise from shipping in Europe”, *Proceedings of Meetings on Acoustics*, **(2022)**.

[4] IMO: ”Guidelines for the reduction of underwater noise from commercial shipping to address adverse imoacts on marine life”, **(2014)**.

[5] European Commission, Commission Decision (EU) 2017/848 of 17 May 2017: ”laying down criteria and methodological standards on good environmental status of marine waters and specifications and standardised methods for monitoring and assessment, and repealing Decision 2010/477/EU (Text with EEA relevance.)”, *Official Journal of the European Union* **125**, **(2017)**.

[6] Li, P., Wu, J., Wang, Y., Lan, Q. and Xiao, W.: ”STM: Spectrogram Transformer Model for Underwater Acoustic Target Recognition”, *Journal of Marine Science and Engineering* **10**, 1428 **(2022)**.

[7] Zou, H., Si, Y., Chen, C., Rajan, D. and Chng, E.: ”Speech Emotion Recognition with Co-Attention based Multi-level Acoustic Information”, *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7367-7371 **(2022)**.

[8] Dommergues, B., Cruz, E. and Vaz, G.: ”Optimization of underwater acoustic detection of marine mammals and ships using CNN”, *Proceedings of Meetings on Acoustics*, **(2022)**.

[9] Santos-Domínguez, D., Torres-Guijarro, S., Cardenal-López, A. and Pena-Guimenez, A.: ”ShipsEar: An underwater vessel noise database”, *Applied Acoustics* **113**, 64-69 **(2016)**.

[10] University of Rhode Island: *”Discovery of sounds in the seas”*.

[11] Dias, P. S., Costa, M., Pinto, J, Lima, K., Venancio, L., Aguiar, M. and Sousa, J. B.: ”Large Scale Unmanned Vehicles Oceanic Exercise REP(MUS)19 Field Report”, *2020 IEEE/OES Autonomous Underwater Vehicles Symposium (AUV)*, **(2020)**

[12] Liu, C., Hong, F., Feng, H., Zhai, Y. and Chen, Y.: ”Environmental Sound Classification Based on Stacked Concatenated DNN using Aggregated Features”, *Journal of Signal Processing Systems* **93**, 1287-1299 **(2021)**.

[13] Mankun, X., Xijian, P., Tianyun, L. and Mantian, X.:”A New Time-Frequency Spectrogram Analysis of FH Signals by Image Enhancement and Mathematical Morphology”, *Fourth International Conference on Image and Graphics (ICIG 2007)*, 610-615 **(2007)**.

[14] Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A. and Talwalkar, A.: ”Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization”, *Journal of Machine Learning Research* **18**, 1-52 **(2018)**.

[15] Luís, A. R., Alves,I. S., Sobreira, F. V, Couchinho, M. N. and dos Santos, M. E.: ”Brays and bits: information theory applied to acoustic communication sequences of bottlenose dolphins”, *Bioacoustics* **28**, 286-296 **(2018)**.

[16] Correia, A., Vicente, M., Sousa, J., Dommergues, B., Cruz, E. and Vaz, G.: ”Marine Acoustic Signature Recognition using Convolutional Neural Networks”, *SSRN Journal*, **(2022)**.