Building ensemble of Resnet for dolphin whistle detection

Loris Nanni[1], Daniela Cuza[2], Sheryl Brahnam[3],

[1]Department of Information Engineering, University of Padua, Via Gradenigo 6, 35131 Padova, Italy; loris.nanni@unipd.it
daniela.cuza@studenti.unipd.it
[2]Department of Information Technology and Cybersecurity, Missouri State University, 901 S. National Street, Springfield, MO 65804, USA

Abstract: To effectively preserve marine environments and manage endangered species, it is necessary to employ efficient, precise, and scalable solutions for environmental monitoring. Ecoacoustics provides several benefits as it enables non-intrusive, prolonged sampling of environmental sounds, making it a promising tool for conducting biodiversity surveys. However, analyzing and interpreting acoustic data can be time-consuming and often demands substantial human supervision. This challenge can be addressed by harnessing contemporary methods for automated audio signal analysis, which have exhibited remarkable performance due to advancements in deep learning research. This paper introduces a research investigation into developing an automatic computerized system to detect dolphin whistles. The proposed method utilizes a fusion of various resnet50 networks integrated with data augmentation techniques. Through extensive experiments conducted on a publically available benchmark, our findings demonstrate that our ensemble yields significant performance enhancements across all evaluated metrics. The MATLAB/PyTorch source code is freely available at: https://github.com/LorisNanni/

Keywords— Convolutional Neural Network; dolphin whistle; ensemble; spectrogram classification

1. Introduction

Marine ecosystems play a critical role in maintaining the balance of our planet's ecosystem by supporting food security and contributing to climate regulation [1], making their preservation essential for the long-term sustainability of the earth's environment. Thus, there is a growing need to develop and test innovative monitoring systems to ensure the natural preservation of marine habitats. Modern technologies have already shown great potential in monitoring habitats and advancing our understanding of marine communities [2]. Acoustic methods are commonly used for underwater investigations because they detect and enable the classification of sensitive targets even in low visibility conditions. Passive acoustic technologies (PAM), such as underwater microphones, or hydrophones, are particularly attractive as they allow for non-invasive continuous monitoring of marine ecosystems without interfering with biological processes [3]. PAM has been shown to achieve various research and management goals by effectively detecting animal calls [4]. These objectives may include tracking and localizing animals [5, 6], species identification, identifying individuals [3, 7], analyzing distributions and behavior [8], and estimating animal density [9].

The bottlenose dolphin (Tursiops truncatus) is a highly intelligent marine mammal and a critical species for researchers studying marine ecosystems [10]. Like many other marine mammals, dolphins are acoustic specialists that rely on

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2023 doi:10.20944/preprints202306.0552.v1

sounds for communication, reproduction, foraging, and navigational purposes. The acoustic communication of dolphins employs a wide range of vocalizations, including clicks, burst-pulses, buzzes, and whistles [11]. Whistles, in particular, serve various social functions such as individual identification, group cohesion, and coordination of activities, such as feeding, resting, socializing, and navigation [12]. Understanding and accurately detecting dolphin vocalizations is essential for monitoring their populations and assessing their role within marine ecosystems.

Traditional bioacoustics tools and algorithms for detecting dolphins have relied on spectrogram analysis, manual signal processing, and statistical methods [13]. For example, the reference approach pursued in [14] applies three noise removal algorithms to the spectrogram of a sound sample. Then a connected region search is conducted to link together sections of the spectrogram that are above a predetermined threshold and close in time and frequency. A similar technique exploits a probabilistic Hough transform algorithm to detect ridges similar to thick line segments, which are then adjusted to the geometry of the potential whistles in the image via an active contour algorithm [15]. Other algorithmic methods aim to quantify the variation in complexity (randomness) occurring in the acoustic time series containing the vocalization, for example, by measuring signal entropy [16]. While these techniques have helped study dolphin vocalizations, they can be time-consuming and may not always provide accurate results due to the complexity and variability of the signals. Researchers have thus turned to machine learning methods to improve detection accuracy and efficiency.

Early machine learning studies in the field of dolphin detection applied traditional classifiers, such as Hidden Markov Models (HMM) [17] and Support Vector Machines (SVMs) [18]. For instance, in [19], a hidden Markov model was utilized for whistle classification; in [20], classification and regression tree analysis was employed along with discriminant function analysis for categorizing parameters extracted from whistles; in [21], a multilayer perceptron classifier was implemented for classifying short-time Fourier transforms (STFTs) and wavelet transform coefficient energies of whistles; lastly, in [15] a random forest algorithm and a support vector machine were combined to classify features derived from the duration, frequency, and cepstrum domain of whistles (see [22] for a review of the early literature).

More recently, researchers have employed deep learning methods to detect whistle vocalizations. Deep neural networks have demonstrated great potential in sound detection generally [23] and underwater acoustic monitoring specifically [24]. The Convolutional Neural Network (CNN) is one of the best-known deep learners. Though commonly considered an image classifier, CNNs have been applied to whale vocalizations, significantly reducing the false-positive rates compared to traditional algorithms while at the same time enhancing call detection [25, 26]. In [27], the authors compared four traditional methods for detecting dolphin echolocation clicks with six CNN architectures, demonstrating the superiority of the CNNs. In [28], CNNs were shown to outperform human experts in dolphin call detection accuracy. CNNs have also been applied to automatically categorize dolphin whistles into distinct groups, as in [29], and to extract whistle contours either by leveraging peak tracking algorithms [30] or by training CNN-based models for semantic segmentation [31].

Several studies for dolphin whistle classification have used data augmentation on the training set to enhance the performance of CNNs by reducing overfitting and increasing the size and variability of the available datasets [29, 30, 32]. Dolphin vocalizations are complex and highly variable, as analyzed in [33]. Unsurprisingly, some traditional music data augmentation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2023 doi:10.20944/preprints202306.0552.v1

methods, such as pitch shifting, time stretching, and adding background noise, have proven effective at this classification task. When synthesizing dolphin calls, care should be taken to apply augmentations to the audio signal rather than to the spectrograms since altering the spectrogram could distort the time-frequency patterns of dolphin whistles, which would result in the semantic integrity of the labels being compromised [29, 34]. In [29], primitive shapes were interjected into the audio signal to generate realistic ambient sounds in negative samples, and classical computer vision methods were used to create synthetic time-frequency whistles, which replaced the training data. Generative Adversarial Networks (GANs) have also been employed to generate synthetic dolphin vocalizations [32]. The research underscores the efficacy of data augmentation and synthesis methods in enhancing both the precision and stability of dolphin whistle categorization models, especially in situations where the datasets are restricted or imbalanced.

The goal of this work is to continue exploring data augmentation techniques for the task of dolphin vocalization detection. Towards this end, we use the benchmark dolphin whistles dataset developed by Korkmaz et al. [28] but apply data augmentation on the original test set of spectrograms to enlarge it rather than on the training set. The training set contains all spectrograms obtained from audio files recorded between June 24th and June 30th, while the test set is composed of the spectrograms of audio files recorded between July 13th and July 15th, a three-day window. Aside from augmenting the test set, we extract a three-day window (June 24th - June 26th) from the training set as the validation set.

The proposed system outperforms previous state-of-the-art methods on the same dataset using the same testing protocol. We find our results interesting, especially since many misclassified audio samples are unclassifiable even by humans, so the classification result of our method is probably very close to maximum performance (AUC =1 is not obtainable). The main contribution of this study is the creation of a new baseline on this benchmark, along with a clear and repeatable criterion for testing various new developments in machine learning.

2. Materials and methods

2.1. Dataset

The dolphin whistle dataset developed by Korkmaz et al. [28] is a well-developed and relatively large set of patterns containing 108,317 spectrograms, of which 49,807 are tagged as noise and 58,510 as dolphin whistles. The test set contains 6,869 spectrograms. The data were collected with hydrophones during the summer of 2021 for 27 days from the dolphin's reef in Eilat, Israel. Following retrieval, a quality assurance (QA) process was conducted on the data to eliminate occasional disruptions and prolonged periods of noise. This QA procedure included the elimination of noise transients through wavelet denoising and identifying and removing cut-off events via thresholding and bias reduction.

2.1.1. Data Preprocessing and Tagging

The collected data were subjected to a bandpass filter in the range of 5–20 kHz to align with the majority of dolphins' whistle vocalizations. The data were then passed through a whitening filter designed to rectify the hydrophone's open circuit voltage response ripples and the sensitivity of the sound card. The recorded audio files, which consisted of two channels, were averaged before the creation of spectrograms to decrease noise. In addition, the preprocessing pipeline eliminated signal outliers based on their length, using the quartiles-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2023 doi:10.20944/preprints202306.0552.v1

based Tukey method [35], which led to the exclusion of signals that were longer than 0.78 s and shorter than 0.14 s.

The short-time fast Fourier transform of the signal was computed using MATLAB's spectrogram function from the digital signal processing toolbox to create the dolphin whistle spectrograms. SFFT was done with a Blackman function window with 2,048 points, periodic sampling, and a hop size achieved by multiplying the window length by 0.8. The subsequent spectrograms were computed by shifting the signal window by 0.4 s. These spectrogram images were finalized by applying a gray-scale colormap, converting the frequency to kHz and the power spectrum density to dB, and restricting the y-axis between 3 and 20 kHz to emphasize the most significant (dominant) frequency range [36].

Spectrograms were then manually labeled by a human expert in two steps: initial tagging and validation tagging. The first step involved precise annotation of 5 seconds spectrograms over ten days of data collection, which were used to train an initial version of a deep learning classifier. This classifier was then used to select new portions of recordings containing potential dolphin sounds, making tagging the remaining data in the validation phase more efficient. The validation phase only required the verification of positive samples detected by the preliminary deep learning classifier.

A human expert was tasked with identifying dolphin whistles as curving lines in the time-frequency domain and disregarding contour lines generated by shipping radiated noise. When the discrimination process was complex, the expert would directly listen to the recorded audio track to identify whistle-like sounds. The tagging resulted in a binary classification (whistle vs. noise) and a contour line marking the time-frequency characteristic of the identified whistle. This contour was used to assess the quality of the manual tagging by ensuring that the bandwidth of the identified whistle fell within the expected thresholds for a dolphin's whistle, specifically between 3 and 20 kHz. A second quality assessment was conducted by measuring the variance of the acoustic intensity of the identified whistle along the time-frequency contour, where the acoustic intensity of a valid whistle is expected to be stable.

2.1.2. Original training and test sets

As mentioned in the introduction, the training set [28] contained all the spectrograms obtained from audio files recorded between June 24th and June 30th, while the test set was composed of the spectrograms of audio files recorded between July 13th and July 15th, a three-day window. The rationale given by the authors for dividing the training and test sets in this manner was mainly to test the generalizability of models using completely disparate sets of recordings, as this would better assess detection accuracy amidst varying sea conditions.

As detailed in Section 2.3, we extract a validation set from the training set obtained from audio files recorded between June 24th and June 26th. We use the validation set for learning the weights of the weighted sum rule, and then the whole training set is fed into the networks for classifying the test set.

2.2. Baseline detection

PamGuard [14] is a widely used software designed to automatically recognize marine mammal vocalizations, it is a very interesting baseline method as it is widely used. The operational parameters of PamGuard were used as follows:

The "Sound Acquisition" module from the "Sound Processing" section was included to manage the data acquisition device and convey its data to other modules.

The "FFT (spectrogram) Engine" module from the "Sound Processing" section was incorporated to calculate spectrograms.

 doi:10.20944/preprints202306.0552.v1

The "Whistle and Moan Detector" module from the "Detectors" section was added for detecting dolphin whistles.

The "Binary Storage" module from the "Utilities" section was incorporated to preserve information from various modules.

A new spectrogram display was created by adding the "User Display" module from the "Displays" section.

Input spectrograms were devised utilizing the FFT analysis mentioned above, with identical parameters: FFT window length was assigned 2,048 points, and the hop size was set to the length multiplied by 0.8 using the Blackman window in the "FFT (spectrogram) Engine" module under software settings. The frequency range was determined between 3 and 20 kHz, and the "FFT (spectrogram) Engine Noise free FFT data" was chosen as the source of FFT data in the "Whistle and Moan Detector" module settings. During the creation of a new spectrogram display, the number of panels was assigned as 2 to visualize both channels. A detection by PamGuard was classified as a true positive if the signal window identified by the software overlapped with at least 5% of the ground truth signal interval. While this criterion may appear lenient, it allowed for the inclusion of many PamGuard detections that might have otherwise been disregarded.

2.3. Proposed approach

The approach proposed in this study is illustrated in Figure 1. Our method is based on the combination of ten ResNet50 networks. The data augmentation phase is applied only to the test set and not to the training set since it is already a large set of spectrograms. The data augmentation methods are selected using the validation set; moreover, by using the validation set, the weights of the weighted sum rule are fixed (see section 2.3.2). As illustrated in Figure 1, for each image of the test set, we classify three images: the original and two created by the data augmentation methods. The scores of these three images are combined by the weighted sum rule, where the weights are found using the validation set. The weighted sum rule is a machine learning approach that combines the predictions of multiple models, in which a factor weights the contribution of each model, here learned on the validation set. Altogether, we have ten ResNet50 networks (each obtained by simply reiterating training), which produce ten weighted sums. These ten scores (i.e., the output of the ten weighted sum rules, one for each network) are combined with the classic sum rule, obtaining the final score of the method

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2023 doi:10.20944/preprints202306.0552.v1

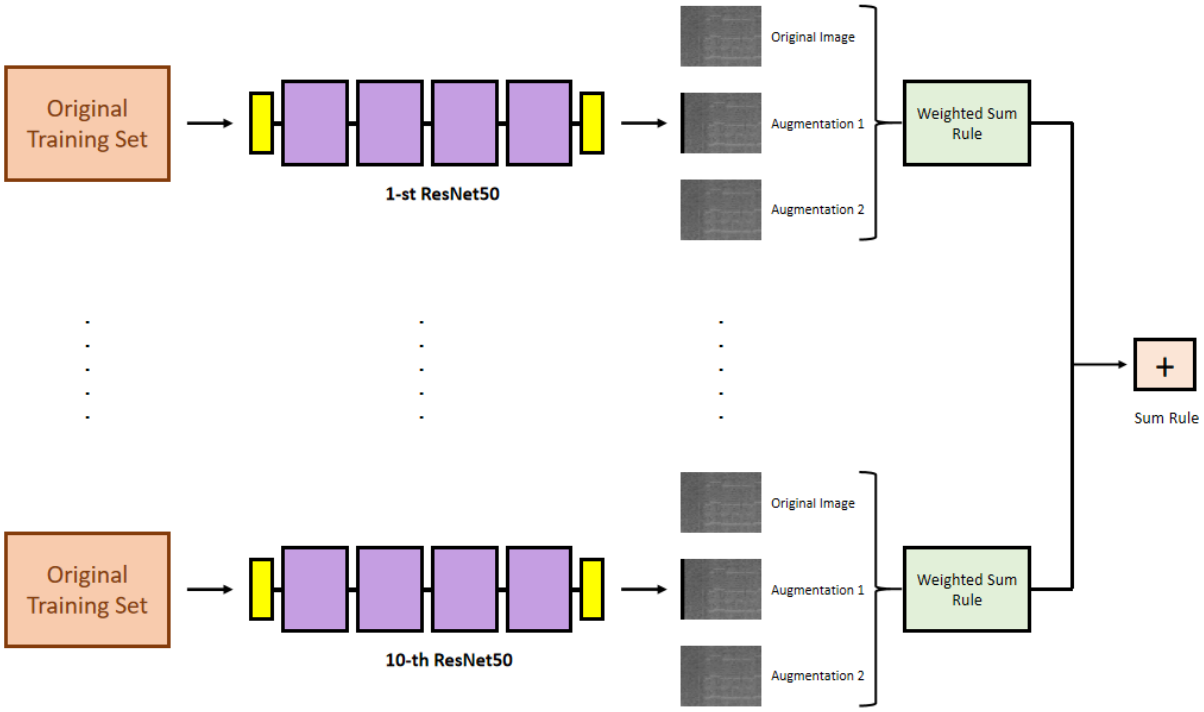


Figure 1. Proposed ensemble: For each image in the test set, we classify three images (the original and two augmented images) combined with the weighted sum rule.

2.3.1. ResNet50

ResNet50 is a convolutional neural network (CNN) architecture introduced by Microsoft Research in 2015 that belongs to a family of models called Residual Networks, or ResNets [37], which are widely used for various computer vision tasks, including image classification, object detection, and image segmentation. The key innovation of ResNet is the introduction of residual, or skip, connections for optimal gradient flow. ResNet enables the training of much deeper networks with improved performance by using skip connections. The name "ResNet50" signifies that this particular model has 50 layers.

The architecture of ResNet50 can be divided into several blocks. The input to the network is a 224×224 RGB image. The initial layer is a standard convolutional layer followed by a batch normalization layer and a ReLU activation function. This layer is followed by a max-pooling layer that reduces the spatial dimensions of the input. The main building blocks of ResNet50 are the residual blocks. Each residual block consists of a series of convolutional layers with batch normalization and ReLU activation. The output of these convolutional layers is added to the original input of the block through a skip connection. This addition operation allows the network to learn the residual information, i.e., the difference between the desired output and the input, which can be thought of as the "error" to be corrected.

ResNet50 contains several stacked residual blocks, with the number of blocks varying depending on the specific architecture. The model also includes bottleneck layers, which are 1×1 convolutional layers used to reduce the dimensionality of the feature maps, making the network more computationally efficient.

Towards the end of the network, a global average pooling layer spatially averages the feature maps, resulting in a fixed-length vector representation. This vector is fed into a fully connected layer with a softmax activation function, producing the final class probabilities.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2023 doi:10.20944/preprints202306.0552.v1

Overall, ResNet50 is a powerful and influential CNN architecture that has significantly advanced the field of computer vision. Its use of residual connections has paved the way for the development of even deeper and more accurate neural networks, and it continues to serve as a benchmark for many state-of-the-art models in the field.

2.3.2. Construction of the validation set

The original training and test sets in [28], as described in section 2.1.2, were used in this study. Unlike the original authors, however, we extracted a validation set from the training set using all the spectrograms related to the three-day recording period of June 24th to June 26th. The validation set was used to fix the parameters of the weights for combining the scores by sum rule of the different augmented spectrograms created for each test pattern. Our testing set was composed of the original and two augmented images. The data augmentation approaches are detailed in section 2.4.3.

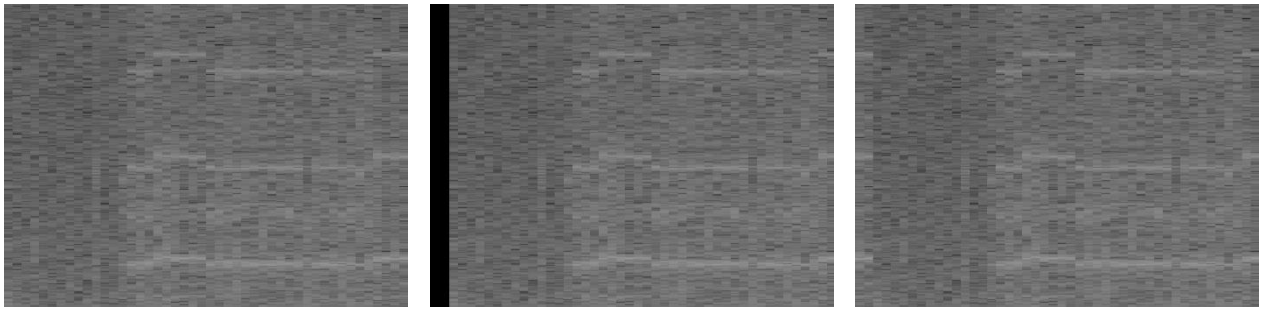
Using the validation set, we combined by weighted sum rule the following three spectrograms for each test pattern:

1) original pattern;
2) Random shift with black or wrap;
3) Symmetric alternating diagonal shift.

2.4.3. Construction of the test set

The following two data augmentation functions (see Figures 2/3) were used to generate the test set:

1. The Random shift with black or wrap (RS) augmentation function undertakes the task of randomly shifting the content of each image. The shift can be either to the left or right, determined by an equal probability of 50% for each direction. The shift's magnitude falls within a specified shift width. Upon performing the shift, an empty space is created within the image. To handle this void, the function uses one of two strategies, each of which is selected with an equal chance of 50%. The first strategy is to fill the space with a black strip, and the second is to wrap the cut piece from the original image around to the other side, effectively reusing the displaced part of the image. In our tests, we utilized a shift_width randomly selected between 1 and 90.
2. The symmetric alternating diagonal shift (SA) augmentation function applies diagonal shifts to distinct square regions within each image. Specifically, the content of a selected square region is moved diagonally in the direction of the top-left corner. The subsequent square region undergoes an opposite shift, with its content displaced diagonally towards the bottom-right corner. The size of the square regions is chosen randomly within the specified minimum and maximum size range.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2023 doi:10.20944/preprints202306.0552.v1

Figure 2. Illustration of the RS method. The first image showcases the original spectrogram. The second image presents the spectrogram after applying the random shift. The third image demonstrates the filled version of the spectrogram.

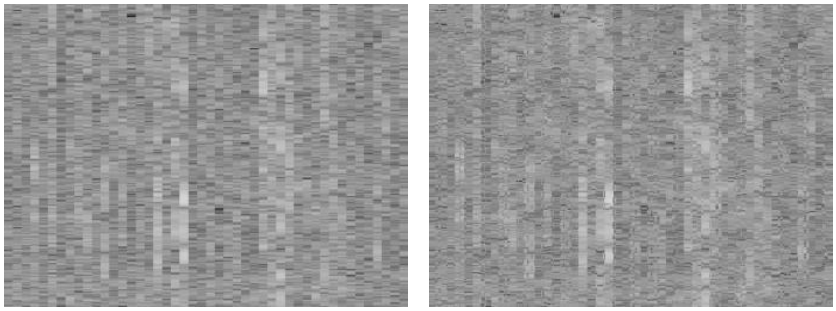


Figure 3. Illustration of the SA method. The first image showcases the original spectrogram. The second image presents the spectrogram after SA.

We tested many data augmentation methods, we reported, for sake of space, only the selected ones on the validation set.

3. Experimental Results

The protocol used in our experiments mirrored that proposed in [28]. However, we have used the validation set described in section 2.3.2 to learn which data augmentation methods to apply and the weights of the weighted sum rule. After choosing the weights with the validations set, we used the subdivision of the training and testing set described in [28]. We wish to stress that the validation set has been extracted from the training set, so there is no overfitting on the test set. We gauged the performance of the model on the distinct test set by calculating the same performance indicators used in [28]. The True Positive rate and the False Positive rate is used to ascertain Precision/Recall. These are used to generate the Receiver Operating Characteristic (ROC) curves and evaluate the corresponding Area Under the Curve (AUC):

$Precision = TP/TP + FP$;

$Recall = TP/TP + FN$;

$True\ Positive\ Rate = TP/TP + FN$;

$False\ Positive\ Rate = FP/FP + TN$;

where TP indicates True Positives, TN True Negatives, FP False Positives, and FN False Negatives.

In Table 1, we compare a baseline ResNet50 with the proposed data augmented ResNet50 (named ResNet50_DA), increasing the size of the ensemble of ResNet50. ResNet50(x)_DA indicates combining by sum rule x ResNet50_DA networks.

	AUC
ResNet50(1)	0.960
ResNet50(1)_DA	0.964
ResNet50(5)_DA	0.972
ResNet50(10)_DA	**0.973**

Table 1. Area under the ROC curve.

We know that the performance increase recorded in Table 1 may not seem high compared to the baseline. However, we find our results interesting because many of the misclassified samples are unclassifiable by humans, so we are probably already very close to the maximum performance (AUC =1 not obtainable). Furthermore, our results create a new baseline on an available data set that can be repeated for testing other methods. Figure 4 reports the ROC

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2023 doi:10.20944/preprints202306.0552.v1

curve for ResNet50(1) vs ResNet50(10)_DA. That plot clearly shows that our proposed approach outperforms ResNet50(1).

In Table 2, we compare our proposed method with two other approaches using the same dataset with the same testing protocol, reporting a full set of performance indicators (accuracy, AUC, precision, and recall). Clearly, the proposed ensemble performs better than the methods reported in the literature, although with higher computational costs. We do not think this is a problem considering that the current computing power of GPUs and the developments expected in the coming years will reduce considerations of such costs. For example, using a NVIDIA 1080, we were able to classify a batch of 100 spectrograms in ~0.3 seconds (considering a stand-alone ResNet50). With a TitanRTX, we were able to classify a batch of 100 spectrograms in ~0.195 seconds (considering a stand-alone ResNet50).

	Accuracy	**AUC**	**Precision**	**Recall**
Pamguard [14]	66.4	---	75.5	19.5
[28]	92.3	0.960	90.5	89.6
ResNet50(10)_DA	**94.9**	**0.973**	**96.5**	**90.2**

Table 2. Comparison with the literature.

The ROC-curves obtained by ResNet50(10)_DA and ResNet50(1) are reported in Figure 4. It is very interesting to note that we obtain a True Positive rate of 0.9 with a False Positive rate of 0.02. Moreover, it is clear the ResNet50(10)_DA improves ResNet50(1).

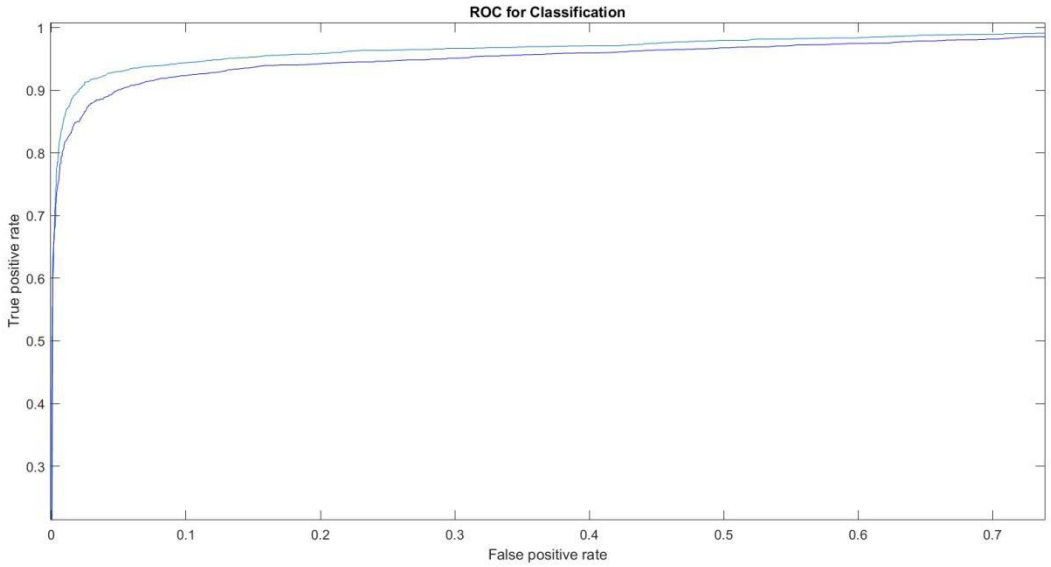


Figure 4. ROC-curve of the proposed system: light blue-> proposed ensemble; blue -> ResNet50(1).

Finally, we report the confusion matrix obtained by our proposed ensemble with the previous baseline on the same dataset. Even this test shows the reliability of the proposed method reduces the number of false noise and false whistle classifications with respect to the previous baselines.

	Here		**[28]**		**Pamguard [14]**	
	Noise	**Whistle**	**Noise**	**Whistle**	**Noise**	**Whistle**
Noise	4124	88	3963	249	4044	168
Whistle	260	2397	277	2380	2139	518

 doi:10.20944/preprints202306.0552.v1

Table 3. Confusion matrices.

4. Conclusion

The surge in human activities in marine environments has led to an influx of boats and ships that emit powerful acoustic signals, often impacting areas larger than 20 square kilometers. The underwater noise from larger vessels can surpass 100 PSI, disturbing marine mammals' hearing, navigation, and foraging abilities, particularly for coastal dolphins [38, 39]. Therefore, the monitoring and preservation of marine ecosystems and wildlife becomes paramount.

However, conventional monitoring technologies depend on detection methods that are less than ideal, thereby hindering our capacity to carry out extensive, long-term surveys. While automatic detection methods could significantly enhance our survey capabilities, their performance is typically subpar amidst high background noise levels. In this paper, we illustrated how deep learning techniques involving data augmentation can identify dolphin whistles with remarkable accuracy, positioning them as a promising candidate for standardizing the automatic processing of underwater acoustic signals.

Despite the need for additional research to confirm the efficacy of such techniques across various marine environments and animal species, we are confident that deep learning will pave the way for developing and deploying economically feasible monitoring platforms. We hope our new baseline will further the comparison of future deep learning techniques in this area.

Finally, we should stress the main cons of using this data set as a benchmark: the training and test set are from the same region (Dolphin's Reef in Eilat, Israel), and samples were collected using the same acoustic recorder.

Acknowledgments: We want to acknowledge the support that NVIDIA provided us through the GPU Grant Program. We used a donated TitanX GPU to train deep networks used in this work.

References

1. Halpern, B., et al., *Recent pace of change in human impact on the world's ocean. Sci. Rep. 9, 11609*. 2019.
2. Danovaro, R., et al., *Implementing and innovating marine monitoring approaches for assessing marine environmental status.* Frontiers in Marine Science, 2016. **3**: p. 213.
3. Gibb, R., et al., *Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring.* Methods in Ecology and Evolution, 2019. **10**(2): p. 169-185.
4. Desjonquères, C., T. Gifford, and S. Linke, *Passive acoustic monitoring as a potential tool to survey animal and ecosystem processes in freshwater environments.* Freshwater Biology, 2020. **65**(1): p. 7-19.
5. Macaulay, J., et al., *Passive acoustic tracking of the three-dimensional movements and acoustic behaviour of toothed whales in close proximity to static nets.* Methods in Ecology and Evolution, 2022. **13**(6): p. 1250-1264.
6. Wijers, M., et al., *CARACAL: a versatile passive acoustic monitoring tool for wildlife research and conservation.* Bioacoustics, 2021. **30**(1): p. 41-57.
7. Ross, S.R.J., et al., *Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions.* Functional Ecology, 2023.
8. Kowarski, K., *Humpback whale singing behaviour in the western north atlantic: from methods for analysing passive acoustic monitoring data to understanding humpback whale song ontogeny*. 2020.
9. Arranz, P., et al., *Comparison of visual and passive acoustic estimates of beaked whale density off El Hierro, Canary Islands.* The Journal of the Acoustical Society of America, 2023. **153**(4): p. 2469-2469.
10. Lusseau, D., *The emergent properties of a dolphin social network.* Proc Biol Sci, 2003. **270 Suppl 2**(Suppl 2): p. S186-8.
11. Lehnhoff, L., et al., *Behavioural Responses of Common Dolphins Delphinus delphis to a Bio-Inspired Acoustic Device for Limiting Fishery By-Catch.* Sustainability, 2022. **14**(20): p. 13186.

12. Papale, E., et al., *The Social Role of Vocal Complexity in Striped Dolphins.* Frontiers in Marine Science, 2020. **7**.
13. Oswald, J.N., J. Barlow, and T.F. Norris, *Acoustic identification of nine delphinid species in the eastern tropical Pacific Ocean.* Marine Mammal Science, 2003. **19**(1): p. 20-037.
14. Gillespie, D., et al., *Automatic detection and classification of odontocete whistles.* The Journal of the Acoustical Society of America, 2013. **134**(3): p. 2427-2437.
15. Serra, O., F. Martins, and L.R. Padovese, *Active contour-based detection of estuarine dolphin whistles in spectrogram images.* Ecological Informatics, 2020. **55**: p. 101036.
16. Siddagangaiah, S., et al., *Automatic detection of dolphin whistles and clicks based on entropy approach.* Ecological Indicators, 2020. **117**: p. 106559.
17. Parada, P.P. and A. Cardenal-López, *Using Gaussian mixture models to detect and classify dolphin whistles and pulses.* J. Acoust. Soc. Am, 2014. **135**(6): p. 3371-3380.
18. Jarvis, S., et al. *Automated classification of beaked whales and other small odontocetes in the tongue of the ocean, bahamas*. In *OCEANS 2006*. 2006. IEEE.
19. Ferrer-i-Cancho, R. and B. McCowan, *A law of word meaning in dolphin whistle types.* Entropy, 2009. **11**(4): p. 688-701.
20. Oswald, J.N., et al., *A tool for real-time acoustic species identification of delphinid whistles.* The Journal of the Acoustical Society of America, 2007. **122**(1): p. 587-595.
21. Mouy, X., M. Bahoura, and Y. Simard, *Automatic recognition of fin and blue whale calls for real-time monitoring in the St. Lawrence.* The Journal of the Acoustical Society of America, 2009. **126**(6): p. 2918-2928.
22. Usman, A.M., O.O. Ogundile, and D.J. Versfeld, *Review of automatic detection and classification techniques for cetacean vocalization.* Ieee Access, 2020. **8**: p. 105181-105206.
23. Abayomi-Alli, O.O., et al., *Data Augmentation and Deep Learning Methods in Sound Classification: A Systematic Review.* Electronics, 2022. **11**(22): p. 3795.
24. Testolin, A. and R. Diamant, *Combining denoising autoencoders and dynamic programming for acoustic detection and tracking of underwater moving targets.* Sensors, 2020. **20**(10): p. 2945.
25. Jiang, J.-j., et al., *Whistle detection and classification for whales based on convolutional neural networks.* Applied Acoustics, 2019. **150**: p. 169-178.
26. Zhong, M., et al., *Beluga whale acoustic signal classification using deep learning neural network models.* The Journal of the Acoustical Society of America, 2020. **147**(3): p. 1834-1841.
27. Buchanan, C., et al. *Deep convolutional neural networks for detecting dolphin echolocation clicks*. in *2021 36th International Conference on image and vision computing New Zealand (IVCNZ)*. 2021. IEEE.
28. Nur Korkmaz, B., et al., *Automated detection of dolphin whistles with convolutional networks and transfer learning.* Frontiers in Artificial Intelligence, 2023. **6**: p. 1099022.
29. Li, L., et al., *Automated classification of Tursiops aduncus whistles based on a depth-wise separable convolutional neural network and data augmentation.* The Journal of the Acoustical Society of America, 2021. **150**(5): p. 3861-3873.
30. Li, P., et al. *Learning deep models from synthetic data for extracting dolphin whistle contours*. in *2020 International Joint Conference on Neural Networks (IJCNN)*. 2020. IEEE.
31. Jin, C., et al., *Semantic segmentation-based whistle extraction of Indo-Pacific Bottlenose Dolphin residing at the coast of Jeju island.* Ecological Indicators, 2022. **137**: p. 108792.
32. Zhang, L., et al., *Dolphin vocal sound generation via deep WaveGAN.* Journal of Electronic Science and Technology, 2022. **20**(3): p. 100171.
33. Kershenbaum, A., L.S. Sayigh, and V.M. Janik, *The encoding of individual identity in dolphin signature whistles: How much information is needed?* PloS one, 2013. **8**(10): p. e77671.
34. Padovese, B., et al., *Data augmentation for the classification of North Atlantic right whales upcalls.* The Journal of the Acoustical Society of America, 2021. **149**(4): p. 2520-2530.
35. Tukey, J.W., *Comparing individual means in the analysis of variance.* Biometrics, 1949: p. 99-114.
36. Jones, B., et al., *Sounds produced by bottlenose dolphins (Tursiops): A review of the defining characteristics and acoustic criteria of the dolphin vocal repertoire.* Bioacoustics, 2020. **29**(4): p. 399-440.

 doi:10.20944/preprints202306.0552.v1

37. He, K., et al., *Deep Residual Learning for Image Recognition.* 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: p. 770-778.
38. Ketten, D.R., *Underwater ears and the physiology of impacts: comparative liability for hearing loss in sea turtles, birds, and mammals.* Bioacoustics, 2008. **17**(1-3): p. 312-315.
39. Erbe, C., et al., *The Effects of Ship Noise on Marine Mammals—A Review.* Frontiers in Marine Science, 2019. **6**.