REDUCING PROCESSING REQUIREMENTS FOR RIGHT WHALE DETECTION FROM AUTONOMOUS SURFACE VEHICLES

W Vickers School of Computing Sciences, University of East Anglia, Norwich, UK
B Milner School of Computing Sciences, University of East Anglia, Norwich, UK
R Lee Gardline Environmental, Gardline Geosurvey Limited, Great Yarmouth, UK

1 INTRODUCTION

This work is concerned with investigating and comparing methods for detecting marine mammals from autonomous surface vehicles (ASVs) where processing power and communications are limited. Accurate detection of marine mammals is important for population monitoring and for mitigation as many species are endangered and protected by environmental laws. We consider the latter of these in the context of detecting North Atlantic right whales (*Eubalaena glacialis*) in the vicinity of potentially harmful subsea activities. Detecting their presence before they enter a mitigation zone both protects the animal and avoids the shutdown of costly offshore operations.

As the human population grows so does the demand for commercial shipping. With this comes increased ocean sound, much of which has recently been under scrutiny for impacting the wellbeing of marine mammals. Ship sounds such as propellers and engine noise are often the source of loud low frequency tones within the ocean. These have the potential to not only interfere with marine mammal communication but also effect their physiological stress levels resulting in possible fatalities[1]. Military sonar testing has also been hypothesised as the cause of mass cetacean fatalities in Greece 1996, with the post mortem report concluding that injuries were consistent with acoustic or impulsive signals causing cardiovascular collapse, which is often associated with extreme stress[1]. With a number of studies providing strong evidence to suggest physiological harm to marine mammals through anthropogenic noise it is logical to create techniques to help mitigate the future risk to mammals.

Detection has traditionally been made by human observers on-board ships, but more recently ASVs have been used[2]. Using an ASV limits the detection to an acoustic only signal, as opposed to visual with a human observer, however it provides a cheaper and more accessible alternative. ASVs typically employ passive acoustic monitoring (PAM) which processes acoustic signals from a hydrophone to determine if marine mammals are present. This presents a number of challenges that include performing audio analysis and detection with the limited processing power on an ASV whilst maximising detection accuracy.

A broad range of machine learning techniques have previously been applied to cetacean detection. For example, methods such as vector quantisation and dynamic time warping have been effective in detecting blue and fin whales from their frequency contours extracted from spectrograms[3]. Hidden Markov models (HMMs) have also been effective at recognising low frequency whale sounds using spectrogram features[4,5]. Further research utilised artificial neural networks (ANNs) for right whale detection, comparing its effectiveness to that of spectrogram correlation, with the ANN giving a better performance in samples with low signal-to-noise ratio (SNR)[6]. Neural networks were further used for classifying clicks of Blainville’s beaked whales, with a good performance recorded for correctly detecting beaked whale clicks[7]. The use of more advanced neural networks largely become popular in the 2010s with convolutional, deep and recurrent networks being widely used for speech, image and text classification[8,9,10] as their performance far outweighs previous techniques. Following on from the popularity of neural networks, a convolutional neural network (CNN) has been applied to right whale detection with spectrograms being used as input features. The CNN performed well achieving extremely high accuracies[11]. Since neural networks have proven to be highly accurate across many feature domains, this work investigates the use of a convolutional neural network (CNN) for right whale detection and considers both its accuracy and processing requirements. Further refinement of the CNN is then carried out in order to try and minimise total processing whilst maintaining a maximum accuracy.

The remainder of this paper is organised as follows. Section 2 describes the characteristics of right whales and their acoustic properties. Issues of detection from an ASV are highlighted in Section 3. Section 4 outlines the use of the CNN and how it has been developed for this application, with details on how features have been extracted and used. Finally, experiments and results are presented in Section 5.

2 BACKGROUND OF CETACEANS

Cetaceans are a large and diverse group of marine mammals. They are split into two suborders, odontocetes (toothed whales) and mysticetes (baleen whales). Odontocetes have teeth and feed on fish whilst mysticetes have a comb like structure (baleen) which helps them to feed on large amounts of crustaceans and zooplankton at once. Right whales are part of the mysticeti suborder and are known to move seasonally to feed and give birth[12].

Communication between whales is achieved primarily through sound. Large amounts of water make sight extremely difficult however sound propagation over hundreds of kilometres is very common. Most cetaceans can vocalise in several ways with whistles, clicks and burst pulses being the most common[13]. These methods of vocalisation have been predominantly recorded for use in the tasks of communication, feeding and navigation.

Our focus is on right whales which, are one of the most endangered marine mammals[14] with a high possibly of extinction due to human activity within areas where they migrate, with as few as 350 individuals remaining. Right whale calls have been well documented[15] and this work focuses on their most commonly documented sound, a tonal up-sweep from approximately 60Hz to 250Hz typically lasting 1 second[16]. Tonal up-sweeps are believed to be used as contact calls and are produced by all ages of animal[17]. Examples of these tonal sounds are shown in Figure 1 which illustrates calls at different signal-to-noise ratios (SNRs) caused by marine noise. Calls, however, are not always consistent with one another and can often vary in duration, frequency range, by time of day, season and age of the animal[18]. Right whale vocalization patterns are also extremely variable with periods of silence regularly spanning many hours[19].

Calls can be difficult to hear, or visualise in spectrograms, as these low frequency bands are often congested with artificial sounds such as ship noise, drilling, piling, seismic exploration, or interference from other marine mammals[20]. These overlapping frequencies can cause large amounts of background noise in the signal making detection extremely difficult. Figure 1 shows three distinct levels of up-sweep visualisation. The top spectrograms show strong up-sweeps with little interference from background noise. The middle row shows strong calls amongst high levels of background noise. The bottom spectrograms show the most challenging scenario with weak calls embedded in large amounts of background noise, giving little indication of mammal presence.

Current methods of collecting cetacean data involve towing a hydrophone array from a ship and using trained observers to listen and watch the water for mammal activity. Visual surveys are often hindered by poor weather conditions (e.g. high seas, fog, presence of ice and darkness), uncertainty of species, and short surfacing intervals[21], causing this method to be extremely unpredictable for mammals that rarely surface. A combination of visual and acoustic monitoring from a ship will hypothetically yield the best result for detection however, ship time is expensive and often sporadically timed. Contrary to this, PAM only systems can record continuously without human interaction, giving a cheaper solution with a much higher likelihood of recording the animal of interest. Furthermore, being able to use an ASV with a PAM system will minimise local noises as no ship is needed for movement of the hydrophone.

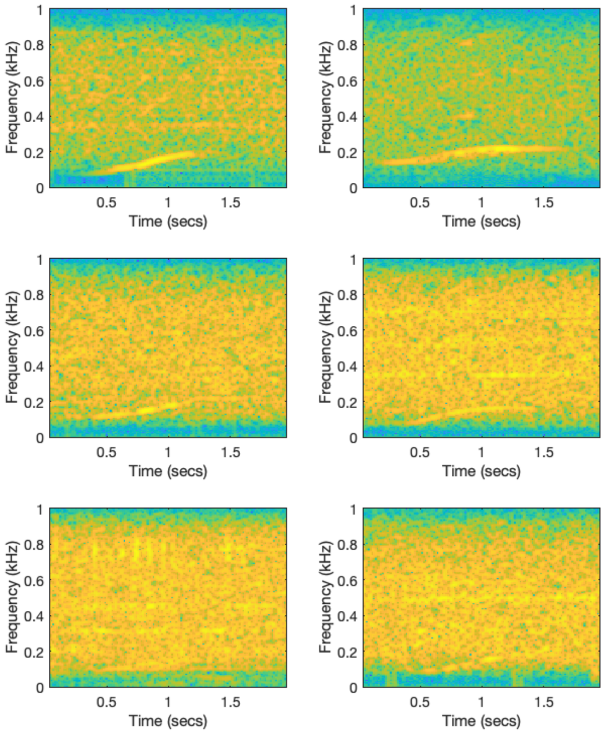

Figure 1: Example spectrograms showing up-sweep calls from right whales for: top row, high SNR; middle row, medium SNR; bottom row, low SNR.

3 LIMITATIONS OF AUTONOMOUS SURFACE VEHICLES

Deploying ASVs for marine mammal detection is much cheaper than employing human observers on-board ships and allows surveys to last several months[2]. For the task of mitigation monitoring a positive detection result needs to be communicated immediately so that mitigation measures can be put in place to protect the animal. This differs from, for example, population monitoring where data is stored on an ASV and then transferred and processed at a later time.

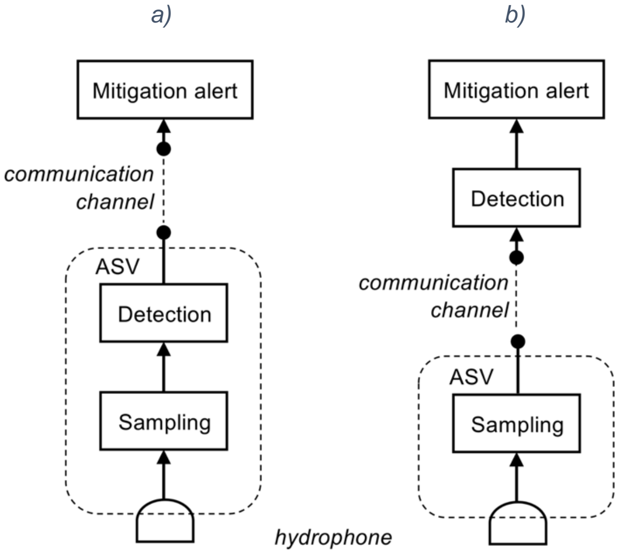

Figure 2: Alternative ASV architectures, showing a) 'thick' ASV with on-board detection and b) 'thin' with detection done remotely.

Two potential ASV architectures can be considered for mitigation monitoring and can be termed ‘thick’ and ‘thin’. The ‘thick’ ASV samples the acoustic data from the hydrophone and inputs this into an on-board detection algorithm with positive detections transmitted for mitigation alert. The ‘thin’ ASV performs only the sampling on-board and transmits the data remotely for detection processing and mitigation alerts. Providing communication beyond a few miles, where a wireless modem could be employed, requires a satellite link. For the ‘thin’ ASV, the communication costs are generally prohibitive as a permanent satellite link is necessary. Furthermore, transmission would likely exceed the 2.4kbps limit for the Iridium network and thereby require a connection to the Inmarsat network which is substantially more expensive and has much higher power consumption (100 W, as opposed to 2.5 W). Based on these limitations of the ‘thin’ ASV architecture, we consider only the ‘thick’ ASV and explore how processing requirements can be minimised. To reduce false alarms (and the potentially large associated costs) with the ‘thick’ ASV architecture, the segment of audio associated with a detection can be transmitted for a human to check, with the frequency of occurrence of this unlikely to be prohibitive.

4 CNN-BASED DETECTION

Convolutional neural networks have been hugely successful in the field of image classification[22]. CNNs work by sliding gradient filters over an input feature, computing the dot product for each filter region along with a weight value (generated randomly initially). This process is repeated for multiple filters, producing a feature map of values that the CNN has learnt. An activation function is then applied to each of the filters output to normalise the data. If a value is high it determines that the feature is likely to be present. Often pooling is then carried out to reduce the dimensionality of the data and to minimise computation of small transformations, whilst aiming to retain the most important details. Class probabilities are then returned often after multiple iterations of convolutional and pooling layers. Gradients of the error in respect to the initial weights are then utilised by a gradient descent function to adjust the filter weights. This can be carried out multiple times in order to minimise the output error and produce a more accurate system[23].

4.1 FEATURE EXTRACTION

The CNN-based detection is implemented by first extracting a time-frequency spectral feature from the audio signal and then inputting this into a CNN to predict the presence of a whale. The time-frequency feature, $\boldsymbol{X}$, is created using a sliding window that transforms short-duration frames of audio into log power spectral vectors. Specifically, an N-point frame of time-domain samples is extracted from the audio, Hamming windowed and a Fourier transform computed. The upper $N/2$ frequency points are discarded, and the remaining points logged. Analysis windows are advanced by S samples to compute each new spectral vector. For an audio recording comprising of T samples, a total of $\left\lceil\frac{T-N+1}{S}\right\rceil$ spectral vectors are computed. This gives the total number of time-frequency points, D, as

Equation 1

$$D = \left\lceil\frac{T-N+1}{S}\right\rceil \times \frac{N}{2}$$

Within each time-frequency matrix, normalisation is applied so all elements, $x(t,f)$, are in the range 0 to 1.

Time-frequency feature vectors were systematically created in an attempt to establish which temporal-spectral resolution gave the best detection accuracy. By varying the temporal and spectral resolutions independently, processing complexity could be assessed against detection accuracy. This would ideally result in a system which can obtain a high accuracy, whilst greatly reducing processing requirements. An explanation of the experiments that alter temporal and spectral resolutions can be found in Section 5.

4.2 CNN DEVELOPMENT

A number of CNN architectures were considered with highest accuracy found using three convolutional layers. Each of these has a max pooling layer followed by a final dense layer. The size of the input varies according to the time and frequency resolution of the feature extraction and this is investigated in section 5. In all convolutional layers, 3×3 filters are applied with zero-padding at the edges, with 32, 64 and 128 in each layer, respectively, with a ReLU activation function following a dropout of 0.5. Again, other filter sizes and numbers of filters in each layer were tested, with the highest accuracy attained using this configuration. The final dense layer uses a sigmoid activation function to give a probability of whale detection.

Preliminary tests were carried out to evaluate a suitable number of epochs for use in the final system. 40 – 500 epochs were tested on the system with the conclusion that there was no significant difference when using 40 epochs compared to 500. All further experiments therefore use 40 epochs for training as it was much faster, although it is understood that the number of epochs used for training does not affect the testing time.

5 EXPERIMENTS

The aim of these experiments is to explore the accuracy of the CNN and to consider this in respect of the trade-off against processing requirements. The first test evaluates the CNN performance as the sampling frequency is reduced. Secondly, tests were run that altered both the temporal and spectral resolution of the input feature whilst maintaining a 50% window overlap. Later tests then maintained a frequency resolution of 15.6Hz and 3.9Hz whilst varying temporal resolutions.

Tests use a database of North Atlantic right whale up-calls that was obtained as part of the Marinexplore and Cornell University Whale Detection Challenge* where the audio is segmented into 2 second duration blocks. Each block is labelled as either containing a right whale or not, with annotations produced manually. A set of 10,934 audio blocks for are used for training, 1,122 for validation and 1,962 for testing. The training, validation and test sets are configured to contain equal numbers of blocks with and without right whales.

5.1 EFFECT OF SAMPLING FREQUENCY

Initially a set of baseline spectral parameters were selected to assess the effect that downsampling had on the detection accuracy. For the full sampling frequency of 2kHz on a 2 second signal, a width of 128 time-domain samples, overlap of 50% and 128 fast-Fourier transform (FFT) points were used. These parameters were scaled proportionally for both the 1kHz and 500Hz sampling frequencies in order to maintain a consistent frequency resolution.

Right whale calls typically rise to approximately 250Hz in frequency (see Figure 1) and so reducing the sampling frequency from 2kHz to 1kHz serves to remove the 500-1000Hz band which contains no whale tones. Figure 3 shows the achieved accuracy of the CNN across all frequencies. Both the 2kHz and 1kHz systems show little variation in accuracy, with the small improvement in the 1kHz system that we attribute to the removal of noise present in this band which may lead to false alarms. Downsampling further to 500Hz leaves the remaining signal bandwidth at 0-250Hz which is very close to the upper tone frequencies in right whale calls that may be the cause of a 1% drop in accuracy. Further experiments consequently focus on comparing 2kHz and 1kHz systems only.

* *https://www.kaggle.com/c/whale-detection-challenge/data*

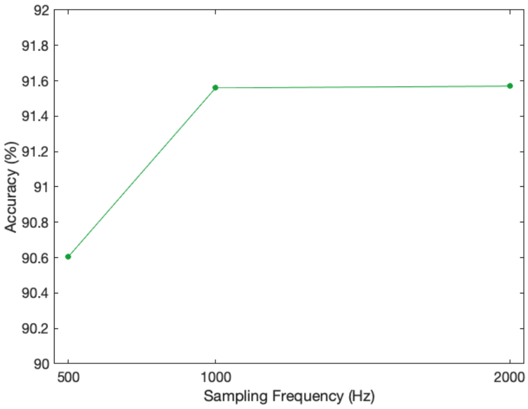

Figure 4: CNN accuracy across different sampling frequencies

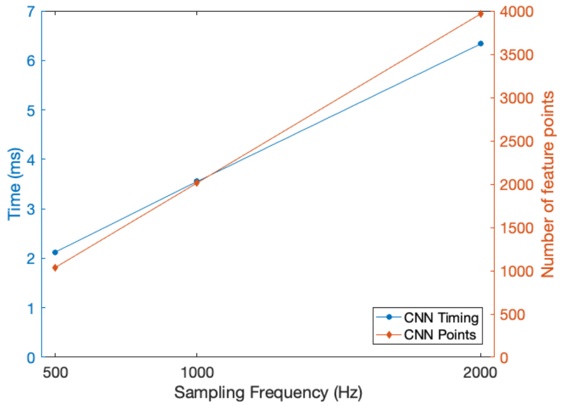

Figure 4: Processing time and number of time-frequency points across different sampling frequencies for a CNN

Given that an aim of this work is to deploy detection on low processing power devices situated on an ASV, we also measured processing times for the methods when run on a CPU. Figure 4 shows the processing time for the three sampling frequencies. Results show that the CNN is able to process each 2 second block in between 2-7ms which is substantially faster than real-time. These tests were performed on an Intel Core i7-870 CPU which is likely to be much faster than processors deployed on an ASV. Also shown on Figure 4 is the number of time-frequency points in the CNN input feature, which is seen to be linearly proportional to the processing time.

5.2 ANALYSIS AND OPTIMISATION OF THE CNN

Tests now concentrate on the CNN using only 2kHz and 1kHz features and examine further the trade-off between accuracy and processing time by examining the time and frequency resolution of the input feature.

Frame widths between 256ms and 16ms are considered first with a fixed 50% overlap of frames which gives a time resolution, Δt, between 128ms and 8ms. In terms of the frequency resolution, Δf, this varies between 3.9Hz and 62.5Hz, depending on the window size and sampling frequency. The number of time-frequency points, D, for each configuration is computed using Equation 1. For each time resolution, Table 1 shows the resulting frequency resolution, number of time-frequency points and detection accuracy, for sampling frequencies of 2kHz and 1kHz - we chose not to pursue the 500Hz system as accuracy had reduced slightly.

Table 1: Detection accuracy and number of points for varying time and frequency resolution features with 50% frame overlap.

	Δt	128ms	64ms	32ms	16ms	8ms
2kHz	Δf D Accuracy	3.9Hz 3584 91.4%	7.8Hz 3840 **92.1%**	15.6Hz 4032 91.6%	31.3Hz 3968 90.2%	62.5Hz 3984 89.9%
1kHz	Δf D Accuracy	3.9Hz 1792 91.2%	7.8Hz 1920 **92.0%**	15.6Hz 1952 91.6%	31.3Hz 1984 90.6%	62.5Hz 1992 90.0%

Highest accuracy for both sampling frequencies is found with the 64ms-7.8Hz time-frequency resolution, with 92.1% for 2kHz and 92.0% for 1kHz. Considering the number of points, and hence processing time, the 1kHz system requires half the computations and gives almost equal performance to the 2kHz system.

Table 2: Detection accuracy and number of points for varying time resolutions and frequency resolutions of 15.6Hz and 3.9Hz.

	Δt	64ms	32ms	16ms	8ms
2kHz	Δf	15.6Hz	15.6Hz	15.6Hz	15.6Hz
	D	1984	3904	7808	15552
	Accuracy	91.1%	**91.6%**	91.0%	90.0%
	Δf	3.9Hz	3.9Hz	3.9Hz	-
	D	7168	14080	28160	-
	Accuracy	92.1%	**92.3%**	91.3%	-
1kHz	Δf	15.6Hz	15.6Hz	15.6Hz	15.6Hz
	D	992	1952	3904	7776
	Accuracy	91.0%	**91.6%**	91.5%	91.0%
	Δf	3.9Hz	3.9Hz	3.9Hz	3.9Hz
	D	3584	7040	14080	28032
	Accuracy	92.3%	**92.5%**	91.6%	91.0%

The tests in Table 1 were performed with 50% frame overlap which means that frequency resolution deteriorates as time resolution improves. We now consider these independently by allowing the frame overlap, S, to vary while keeping the frame width fixed. Specifically, we consider two fixed frame widths to give high and low frequency resolutions of $\Delta f = \{3.9Hz, 15.6Hz\}$ and adjust the frame slide to give varying time resolutions Δt, from 64ms to 8ms. The resulting accuracy and number of time-frequency points are shown in Table 2 for 2kHz and 1kHz sampling frequencies.

For both frequency resolutions and both sampling frequencies the time resolution has relatively little effect between 64ms and 16ms, with highest accuracy at 32ms. In terms of frequency resolution, the finer resolution gives higher accuracy across all configurations tested, although this comes at the cost of increased processing time. For example, highest performance of 92.5%, with 1kHz sampling frequency, 3.9Hz frequency resolution and 32ms time resolution used 7,040 points. This could be reduced to 1,952 points (corresponding to a processing time three times faster) by using a wider frequency resolution but with a reduction in accuracy to 91.6%.

6 CONCLUSION

A CNN has been applied to right whale detection within audio, using a range of parameters for feature extraction. The feature extraction parameters have been specifically chosen in order to reduce the resolution of the feature and subsequently its computational complexity. The audio has been converted from the time-domain into the time-frequency domain and inputted as features for the CNN to classify. Downsampling the audio leaves accuracy almost unchanged but gives a substantial reduction in processing time which is advantageous for ASVs. Considering time and frequency resolutions reveals that a wide resolution of 32ms gives good accuracy whilst higher frequency resolutions are marginally better, albeit at an increased processing cost.

Possible pre-processing steps such as prefiltering to remove noise or using more noisy training data may result in a more accurate system. We have set the decision boundary at a probability threshold of 0.5 which gives close to an equal error rate. This could be adjusted to bias detections and it may be useful to reduce false alarms as whales typically exhibit long periods of calls, which makes it unlikely to miss all of them at times when mitigation alerts are necessary.

7 REFERENCES

1. Cox TM, Ragen TJ, Read AJ, Vos E et al. Understanding the impacts of anthropogenic sound on beaked whales. Journal of Cetacean Research and Management. 2006;7:177-187.

2. Verfuss UK, Aniceto AS, Harris DV et al. A review of unmanned vehicles for the detection and monitoring of marine fauna. Marine Pollution Bulletin. 2019;140:17-29.
3. Mouy X, Bahoura M, Simard Y. 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: 2918-28
4. Mellinger D, Clark C. Recognizing transient low-frequency whale sounds by spectrogram correlation. The Journal of the Acoustical Society of America. 2000;107(6):3518-3529.
5. Brown J, Smaragdis P. Hidden Markov and Gaussian mixture models for automatic call classification. The Journal of the Acoustical Society of America. 2009;125(6):EL221-EL224.
6. Mellinger DK. A comparison of methods for detecting right whale calls. Canadian Acoustics. 2004 Jun 1;32(2):55-65.
7. Mellinger DK. A neural network for classifying clicks of Blainville's beaked whales (Mesoplodon densirostris). Canadian Acoustics. 2008 Mar 1;36(1):55-9.
8. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 2012 (pp. 1097-1105).
9. Deng L, Hinton G, Kingsbury B. New types of deep neural network learning for speech recognition and related applications: An overview. IEEE International Conference on Acoustics, Speech and Signal Processing 2013 (pp. 8599-8603).
10. Lai S, Xu L, Liu K, Zhao J. Recurrent convolutional neural networks for text classification. 29th AAAI conference on artificial intelligence 2015.
11. Smirnov E. North Atlantic right whale call detection with convolutional neural networks. InProc. Int. Conf. on Machine Learning, Atlanta, USA 2013 (pp. 78-79).
12. Thomas PO, Taber SM. Mother-infant interaction and behavioral development in southern right whales, Eubalaena australis. Behaviour. 1984 Jan 1:42-60.
13. Clark CW. Acoustic communication and behavior of the southern right whale (Eubalaena australis). Communication and behavior of whales. 1983:163-98.
14. Kraus SD, Brown MW, Caswell H, Clark CW et al. North Atlantic right whales in crisis. Science. 2005 Jul 22;309(5734):561-2.
15. Clark CW. The acoustic repertoire of the southern right whale, a quantitative analysis. Animal Behaviour. 1982 Nov 1;30(4):1060-71.
16. Mussoline SE, Risch D, Hatch LT et al. Seasonal and diel variation in North Atlantic right whale up-calls: implications for management and conservation in the northwestern Atlantic Ocean. Endangered Species Research. 2012 Apr 12;17(1):17-26.
17. Parks SE, Clark CW, Tyack PL. Short-and long-term changes in right whale calling behavior: The potential effects of noise on acoustic communication. The Journal of the Acoustical Society of America. 2007 Dec;122(6):3725-31.
18. Pylypenko K. Right whale detection using artificial neural network and principal component analysis. IEEE 35th International Conference on Electronics and Nanotechnology 2015 (pp. 370-373)
19. Matthews JN, Brown S, Gillespie D et al. Vocalisation rates of the North Atlantic right whale (Eubalaena glacialis). Journal of Cetacean Research and Management. 2001;3(3):271-82.
20. Gillespie D. Detection and classification of right whale calls using an 'edge' detector operating on a smoothed spectrogram. Canadian Acoustics. 2004 Jun 1;32(2):39-47.
21. Baumgartner MF, Mussoline SE. A generalized baleen whale call detection and classification system. The Journal of the Acoustical Society of America. 2011;129(5):2889-902.
22. Russakovsky O, Deng J, Su H et al. Imagenet large scale visual recognition challenge. International journal of computer vision. 2015 Dec 1;115(3):211-52.
23. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998 Nov 11;86(11):2278-324.