A comparison of machine learning methods for detecting right whales from autonomous surface vehicles

W. Vickers
School of Computing Sciences
University of East Anglia
Norwich, UK
w.vickers@uea.ac.uk

B. Milner
School of Computing Sciences
University of East Anglia
Norwich, UK
b.milner@uea.ac.uk

R. Lee
Gardline Environmental
Gardline Geosurvey Limited
Great Yarmouth, UK
robert.lee@gardline.com

J. Lines
School of Computing Sciences
University of East Anglia
Norwich, UK
j.lines@uea.ac.uk

***Abstract*—This work compares a range of machine learning methods applied to the problem of detecting right whales from autonomous surface vehicles (ASV). Maximising detection accuracy is vital as is minimising processing requirements given the limitations of an ASV. This leads to an examination of the trade-off between accuracy and processing requirements. Three broad types of machine learning methods are explored - convolution neural network (CNNs), time-domain methods and feature-based methods. CNNs are found to give best performance in terms of both detection accuracy and processing requirements. These were also tolerant to downsampling down to 1kHz which gave a slight improvement in accuracy as well as a significant reduction in processing time. This we attribute to the bandwidth of right whale calls which is around 250Hz and so downsampling is able to capture the sounds fully as well as removing unwanted noisy spectral regions.**

***Index Terms*—Cetacean detection, CNNs, machine learning, autonomous surface vehicles**

I. 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 potential harmful subsea activities. Detecting their presence before they enter a mitigation zone both protects the animal and avoids the shutdown of costly subsea operations.

Detection has traditionally been made by human observers on-board ships, but more recently ASVs have been used [1]. Using an ASV limits the detection to using just an acoustic signal, as opposed to visual with a human observer, however it provides a cheaper and more accessible alternative. The ASV employs 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 limited processing power whilst maximising detection accuracy. This work investigates a range of methods for right whale detection and considers both their accuracy and processing requirements.

A broad range of machine learning methods have been applied to cetacean detection in recent years. 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 [2]. Hidden Markov models (HMMs) have also been effective at recognising low frequency whale sounds using spectrogram features [3]. Comparisons have also been made between between artificial neural networks (ANNs) and spectrogram correlation for right whale detection [4]. Further to the use of ANNs, support vector machines (SVM) have also been applied effectively to odontocete classification [5]. SVMs have also been compared against Gaussian mixture models for classification of three types of odontocetes [6]. More recently, a convolution neural network (CNN) was applied to right whale classification [7].

The aim of this work is to apply a range of machine learning methods to the problem of right whale detection. Specifically, we evaluate techniques that include CNN, time-domain and feature-domain methods of detection. In addition to accuracy, we also measure processing requirements for detection and consider their suitability for deployment on an ASV. Training times are considered unimportant as this is carried out offline.

The remainder of the paper is organised as follows. Section II describes the sounds produced by right whales in both relatively clean and noisy conditions. Issues of detection from an ASV are highlighted in Section III. Sections IV, V and VI present the CNN, time-domain and feature-domain methods of detection that will be investigated. Finally, detection results are presented in Section VII.

II. Characteristics of right whales

Right whales are one of the most endangered marine mammals k [8] 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 and they can make a variety of vocalizations [9]. However, this work focuses on their most commonly

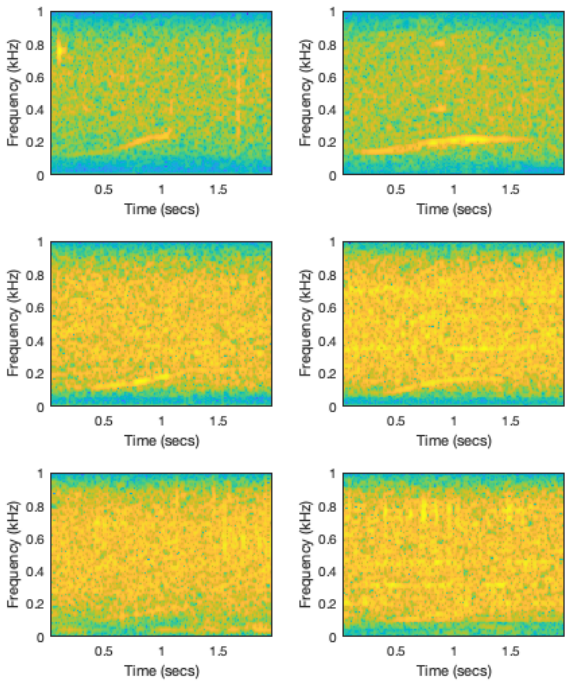


Fig. 1. *Example spectrograms showing up-sweep calls from right whales for: top row, high SNR; middle row, medium SNR; bottom row, low SNR.*

documented sound, an up-sweep tone from approximately 60Hz to 250Hz typically lasting 1 second. Examples of this 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 and frequency range, and by time of day, season and age of the animal [10]. Right whale vocalization patterns are also extremely variable with periods of silence regularly spanning many hours [11].

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 [12]. 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. For right whale detection, their low frequency calls allow sampling frequencies down to 2kHz to be used, although typically much higher sampling frequencies are used initially before being downsampled.

III. Detection from 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 [1]. 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. 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 resulting 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.

IV. CNN-based detection

CNN-based detection is based on first extracting a time-frequency spectral feature from the audio signal and then inputting this into a CNN to predict presence of a whale. The time-frequency feature, $\mathbf{X}$, 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 T samples, a total of $\lceil \frac{T-N+1}{S} \rceil$ spectral vectors are computed. This gives the total number of time-frequency points, D, as

$$D = \lceil \frac{T-N+1}{S} \rceil \times \frac{N}{2} \quad (1)$$

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

A number of CNN architectures were considered with highest accuracy found using three convolutional layers. Each of these is followed by 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 VII. 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. Again, other filter sizes and numbers of filters in each layer were tested, with highest accuracy attained with this configuration. The final dense layer uses a sigmoid activation function to give a probability of whale detection.

V. Time-domain detection

An alternative approach is to use the audio signal directly to form a time series classification (TSC) problem. The vast majority of TSC algorithms operate on time domain data as, until recently, the consensus was that '*simple nearest neighbour classification is very difficult to beat*' [13]. As such, much emphasis has been placed on developing approaches for solving problems in the time-domain using alternative elastic distance measures with nearest neighbour classifiers [14]–[16]. The most popular of these approaches is dynamic time warping (DTW) with a warping window set through cross-validation and 1-nearest neighbour (DTW 1NN). While it has been shown that ensembling different elastic nearest neighbour classifiers can be significantly more accurate [17], combining such *lazy* classifiers increases test classification run-time. With a large amount of training data, necessary for capturing the range of whale signals and background noises, detection processing times are likely prohibitive for real-time deployment on an ASV, so for this application we use DTW 1NN as a benchmark for time-domain approaches.

VI. Feature-based detection

Feature-based methods operate by transforming the time-domain signals into an alternative representation where discriminatory information is more easily detected. A recent comparison of approaches [18] demonstrated that best performance is obtained by combining ensemble classifiers built over various representations of a problem to produce combined predictions from a meta-ensemble [19]. However, given the processing limitations in implementing detection on ASVs, this would not be practical but suitable constituent transformation-based approaches may produce fast, accurate results. In particular, three such constituents are considered: 1) *time series forest* which is built on summary features from phase-dependent intervals [20]; 2) *shapelet transform* which is a heterogeneous ensemble using data transformed by similarity to phase-independent discriminatory subsequences [21]; 3) *RISE*, random interval spectral ensemble which is a forest-based ensemble classifier that builds constituents using features extracted from the auto-correlation and power spectral domains [19].

VII. Experimental results

The aim of these experiments is to explore the accuracy of the detection methods and to consider these in respect of the trade-off against processing requirements. The first test compares the techniques as the sampling frequency is reduced and shows how processing time is affected. Secondly, parameters of detection methods are adjusted to vary the amount of processing required and to see its effect on accuracy.

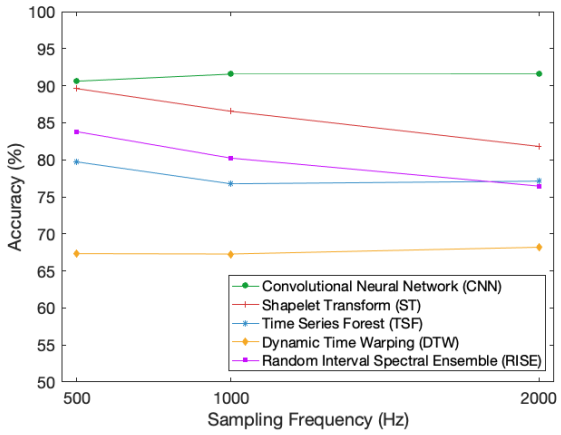
Tests use part of the Marinexplore and Cornell University Whale Detection Challenge[1] database of North Atlantic right whale up-calls. The audio is segmented into 2 second duration blocks with each labelled manually as containing a right whale or not. Specifically, we use 10,934 audio blocks for training, 1,122 for validation and 1,962 for testing and these are balanced to contain equal numbers of segments with and without whales present.

A. Effect of sampling frequency

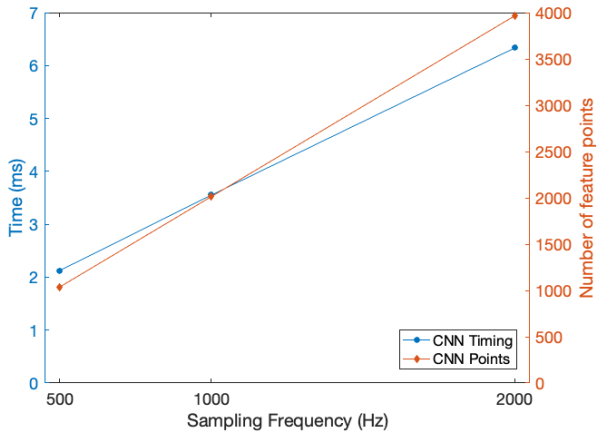
These tests examine the effect that reducing the sampling frequency has on detection accuracy and on processing time and aim to identify methods with a good compromise that can be taken forward for further consideration. Tests are performed at the original recording sampling frequency of 2kHz and then downsampled to 1kHz and then to 500Hz. Figure 2 shows detection accuracy and processing time for the different sampling frequencies using the CNN, shapelet transform, time series forest, DTW and RISE detection methods. Highest accuracy is obtained with the CNN at all sampling frequencies – this configuration used a window width of 64ms with 50% frame overlap. With the CNN, sampling frequency has some effect on detection accuracy with a slight peak at 1kHz. Right whale calls typically rise to around 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. This does not degrade performance and the small improvement 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 which may cause the small reduction in accuracy. Considering now the feature-domain methods, these perform worse with highest accuracy achieved by the shapelet transform. Interestingly, these methods tend to perform better at lower sampling frequencies. Worst performance is with the DTW time-domain method.

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 2 shows the processing time for the three sampling frequencies. Results are shown only for the CNN which is able to process each 2 second block in between 2-7ms which is substantially faster than real-time. These tests were performed on a Intel Core i7-870 CPU which is likely to be much faster than processors deployed on an ASV. The feature-domain and time-domain methods were found to be much slower (ranging from approximately 1 second for the shapelet transform to 7 seconds for DTW) making them unsuitable for deployment. Also shown on Figure 2 is the number of time-frequency points in the CNN input feature, which is seen to be linearly proportional to the processing time.

[1] https://www.kaggle.com/c/whale-detection-challenge/data

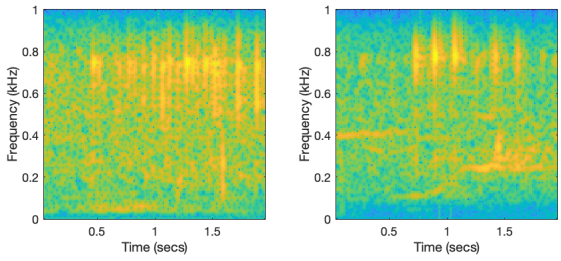


(a) Detection accuracy across different sampling frequencies

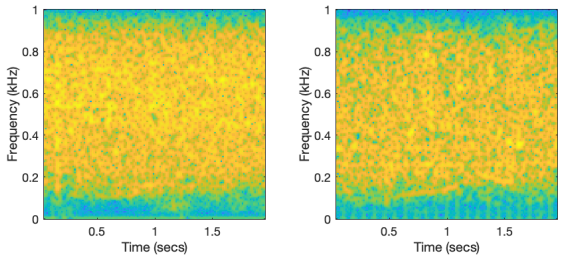


(b) Processing time and number of time-frequency points across different sampling frequencies

Fig. 2. *Effect of sampling frequency on a) detection accuracy and b) processing time and number of time-frequency points.*



(a) False negatives



(b) False positives

Fig. 3. *Example spectrograms that have been classified as a) false negative or b) false positive.*

TABLE I
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	3.9Hz	7.8Hz	15.6Hz	31.3Hz	62.5Hz
2kHz	D	3584	3840	4032	3968	3984
2kHz	Accuracy	91.4%	**92.1%**	91.6%	90.2%	89.9%
1kHz	Δf	3.9Hz	7.8Hz	15.6Hz	31.3Hz	62.5 Hz
1kHz	D	1792	1920	1952	1984	1992
1kHz	Accuracy	91.2%	**92.0%**	91.6%	90.6%	90.0%

To investigate false alarms and missed detections, Figure 3 shows spectrograms of two false negatives and two false positives, produced by the CNN operating at 2kHz. These are typical of both types of error and for the false negatives show much background noise to be present that has largely masked the right whale call. False positives are also more likely in high noise conditions, where the characteristic of noise creates spectral energies similar to the right whale call leading to the false alarms.

Given the higher accuracy and substantially faster processing time found with the CNN, we take this forward for further analysis and optimisation.

B. Analysis and optimisation of CNN detection

Tests now concentrate on the CNN method of detection 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 (1). For each time resolution, Table I 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. 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.

The tests in Table I 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 Δf=$\{$3.9Hz, 15.6Hz$\}$ 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 II

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

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%.

VIII. CONCLUSION

A range of time-series, feature-domain and CNN methods have been applied to the detection of right whales with best performance, in terms of both accuracy and processing time, given by the CNN. 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 better, albeit at increased processing cost.

Analysis of errors, both false negatives and false positives, has shown these to occur most often in low SNRs. Methods such as prefiltering to remove noise prior to detection or the use of more noisy training data may alleviate some of these errors. 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.

ACKNOWLEDGMENT

We acknowledge the support of the Next Generation Unmanned Systems Science (NEXUSS) Centre for Doctoral Training, Gardline Geosurvey Limited and NVIDIA.

REFERENCES

[1] U. K. Verfuss, A. S. Aniceto, D. V. Harris, D. Gillespie, S. Fielding, G. Jiménez, P. Johnston, R. R. Sinclair, A. Sivertsen, S. A. Solbø *et al.*, "A review of unmanned vehicles for the detection and monitoring of marine fauna," *Marine Pollution Bulletin*, vol. 140, pp. 17–29, 2019.

[2] X. Mouy, 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*, vol. 126, pp. 2918–28, 12 2009.

[3] D. K. Mellinger and C. W. Clark, "Recognizing transient low-frequency whale sounds by spectrogram correlation," *The Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3518–3529, May 2000.

[4] D. K. Mellinger, "A comparison of methods for detecting right whale calls," *Canadian Acoustics*, vol. 32, no. 2, pp. 55–65, Jun. 2004.

[5] S. Jarvis, N. DiMarzio, R. Morrissey, and D. Morretti, "Automated Classification of Beaked Whales and Other Small Odontocetes in the Tongue of the Ocean, Bahamas," in *OCEANS 2006*, Sep. 2006, pp. 1–6.

[6] M. A. Roch, M. S. Soldevilla, R. Hoenigman, S. M. Wiggins, and J. A. Hildebrand, "Comparison of machine learning techniques for the classification of echolocation clicks from three species of odontocetes," *Canadian Acoustics*, vol. 36, no. 1, pp. 41–47, Mar. 2008. [Online]. Available: https://jcaa.caa-aca.ca/index.php/jcaa/article/view/1989

[7] E. Smirnov, "North atlantic right whale call detection with convolutional neural networks," in *Proc. Int. Conf. on Machine Learning, Atlanta, USA*. Citeseer, 2013, pp. 78–79.

[8] S. D. Kraus, M. W. Brown, H. Caswell, C. W. Clark, M. Fujiwara, P. K. Hamilton, R. D. Kenney, A. R. Knowlton, S. Landry, C. A. Mayo, W. A. McLellan, M. J. Moore, D. P. Nowacek, D. A. Pabst, A. J. Read, and R. M. Rolland, "North Atlantic Right Whales in Crisis," *Science*, vol. 309, no. 5734, pp. 561–562, Jul. 2005. [Online]. Available: http://science.sciencemag.org/content/309/5734/561

[9] C. W. Clark, "Acoustic communication and behavior of the southern right whale (eubalaena australis)," *Communication and behavior of whales*, pp. 163–198, 1983.

[10] K. Pylypenko, "Right whale detection using artificial neural network and principal component analysis," Apr. 2015, pp. 370–373.

[11] J. N. Matthews, S. Brown, D. Gillespie, M. Johnson, R. McLanaghan, A. Moscrop, D. Nowacek, R. Leaper, T. Lewis, and P. Tyack, "Vocalisation rates of the North Atlantic right whale (Eubalaena glacialis)," p. 12, 2001.

[12] D. Gillespie, "Detection and classification of right whale calls using an 'edge' detector operating on a smoothed spectrogram," *Canadian Acoustics*, vol. 32, no. 2, pp. 39–47, Jun. 2004. [Online]. Available: https://jcaa.caa-aca.ca/index.php/jcaa/article/view/1586

[13] G. E. Batista, X. Wang, and E. J. Keogh, "A complexity-invariant distance measure for time series," in *Proceedings of the 2011 SIAM international conference on data mining*. SIAM, 2011, pp. 699–710.

[14] A. Stefan, V. Athitsos, and G. Das, "The move-split-merge metric for time series," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, pp. 1425–1438, 2013.

[15] E. J. Keogh and M. J. Pazzani, "Derivative dynamic time warping," in *Proceedings of the 2001 SIAM International Conference on Data Mining*. SIAM, 2001, pp. 1–11.

[16] P.-F. Marteau, "Time warp edit distance with stiffness adjustment for time series matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 306–318, 2009.

[17] J. Lines and A. Bagnall, "Time series classification with ensembles of elastic distance measures," *Data Mining and Knowledge Discovery*, vol. 29, no. 3, pp. 565–592, 2015.

[18] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh, "The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances," *Data Mining and Knowledge Discovery*, vol. 31, no. 3, pp. 606–660, 2017.

[19] J. Lines, S. Taylor, and A. Bagnall, "Time series classification with hive-cote: The hierarchical vote collective of transformation-based ensembles," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 12, no. 5, p. 52, 2018.

[20] H. Deng, G. Runger, E. Tuv, and M. Vladimir, "A time series forest for classification and feature extraction," *Information Sciences*, vol. 239, pp. 142–153, 2013.

[21] J. Hills, J. Lines, E. Baranauskas, J. Mapp, and A. Bagnall, "Classification of time series by shapelet transformation," *Data Mining and Knowledge Discovery*, vol. 28, no. 4, pp. 851–881, 2014.