USING MACHINE LEARNING FOR THE DETECTION OF PROPELLER CAVITATION

D Smith QinetiQ Noise & Vibration, Rosyth, Scotland
B McKinlay QinetiQ Noise & Vibration, Rosyth, Scotland
R Potter QinetiQ Noise & Vibration, Rosyth, Scotland

1 INTRODUCTION

The noise generated by ships is recognized as having a significant detrimental impact on marine life[1]. This problem is further exacerbated with the growing number of vessels in operation. There is therefore a need to better understand and manage the noise radiated underwater by ships. Under normal operation, the propeller can contribute significantly to the overall platform noise. However, when cavitation is present on the propeller, the noise greatly increases and becomes the dominant noise source. Therefore, the impact of the noise radiated by a platform can be reduced if propeller cavitation can be avoided. This can be achieved if cavitation is promptly detected allowing for remedial action via the propeller controls to be taken.

In this contribution, we investigate the use of a range of readily available machine learning methods for the detection of propeller cavitation based on a number of different input features. Propeller cavitation detection is possible using a range of signal processing methods. Cyclostationarity is a recently proposed signal processing method for propeller cavitation detection[2]. It relies a number of frequency domain conversions, resulting in a cyclic spectrum. This spectrum is then searched for peaks, where peaks around the blade rate and its harmonics can indicate the presence of cavitation. Figure 1 compares the output at various stages of the cyclostationarity analysis for a cavitating and non-cavitating signal.

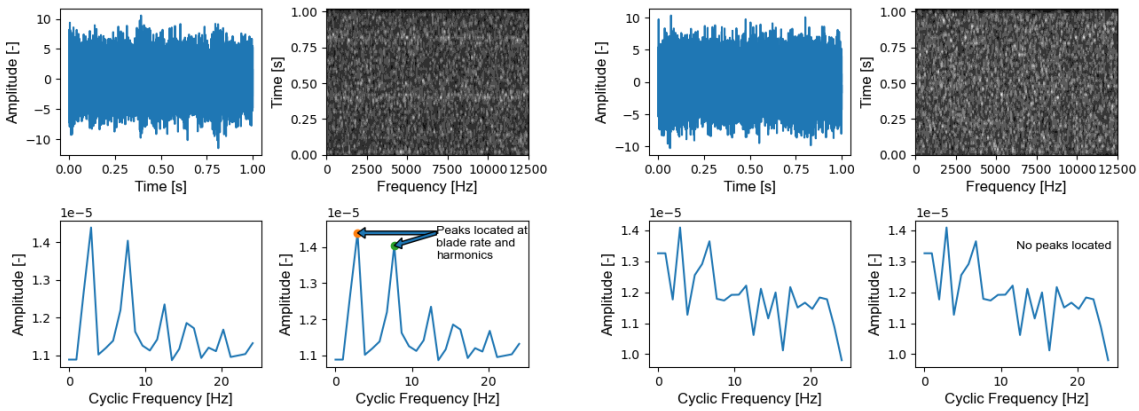


Figure 1: Comparison of cyclostationarity output at various steps for: (a) cavitating signal; and (b) non-cavitating signal.

Notable in Figure 1 is the inability to visually detect the presence of cavitation within the time waveform and frequency spectrum. We therefore investigate the use of machine learning methods on these two signal types in an effort to negate the significant signal processing required in the cyclostationarity analysis. A further motivation results from the reduction in capability of the cyclostationarity approach as the signal to noise ratio decreases. A number of further features, based on the statistics of the time waveform and frequency spectrum are also investigated in order to identify if these features provide further insights into the characterisation of cavitation.

2 METHODS

2.1 Cavitation Data

The intent of the present work is to investigate the potential utility offered by machine learning models for detecting cavitation on a propeller. In the absence of data suitable for publication, synthetic data has been created. Signals were generated to represent transducer measurements taken in close proximity to the propeller. Signals comprised a broadband signal, described by propeller characteristics, a continuous component, and where cavitation is present, a modulating component.

The broadband signal is described by the propeller noise estimation formula after Ross[3], where the noise is characterised by the blade count and tip speed. The continuous component was characterised by Gaussian white noise at a specified signal to noise ratio. The modulating component was described by a summation of sinusoidal signals at frequencies corresponding to the blade rate and its harmonics up to the blade passing frequency. The signal is constructed by adding the continuous component to the product of the broadband and modulating components. Where cavitation is not present, the modulating part is excluded.

500 signals of 1 second duration have been generated for the analysis. The signals are characterised across a range of blade counts, rotational speeds and signal to noise ratios, with 50 % having cavitation present by means of the modulating component.

2.2 Machine Learning Models

The problem of cavitation detection is a binary classification one; that is, is cavitation present or not. Therefore, the machine learning models studied comprised a range of classification models. In particular, the following models were studied:

- random-forest;
- k-nearest neighbours
- support vector machines; and
- logistic regression.

These models were selected as they were readily available as part of the scikit-learn package[4]. Further, they allow for easy implementation for the present work. It is understood that other, more complex models, may be more suited for the current study. However, their implementation was beyond the scope of the preliminary investigation undertaken as part of this work. It is planned to investigate these in the near future.

The machine learning models described were investigated for their ability to detect cavitation across a number of different input feature groups. Here, we are investigating which features enable the greatest classification accuracy. That is, we are carrying out a feature selection investigation. The following feature groups have been studied:

- time series signal;
- frequency spectrum;
- cyclic spectrum; and
- signal statistics.

These groups have been selected in order to investigate if there is a dependence on signal pre-processing in the ability to accurately detect cavitation from the generated signals. These pre-processing steps related to a cyclostationarity analysis. A further feature group comprising a number of easy to compute signal statistics are evaluated. The following signal statistics are studied:

- mean;
- median;
- variance; and
- standard deviation;

For all feature groups, the model output was a binary flag denoting the presence of cavitation, assigned during the synthetic data generation. The data comprised 500 samples and was shuffled and split into training, 80 %, and testing, 20 % sets, with the testing data set used for blind evaluation of the different models. Across all investigations, hyperparameters for each model underwent tuning based on a random grid search.

3 RESULTS AND DISCUSSION

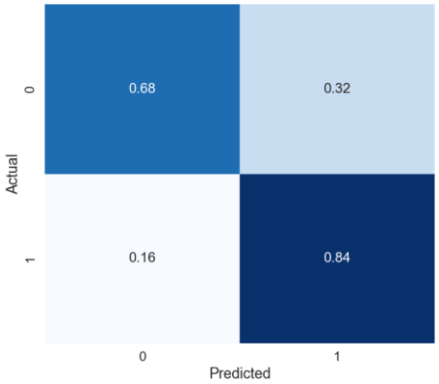
3.1 Time Series Classification

The first feature group studied was the raw time series signal. Time series signals present a particular challenge for machine learning models, with the order of the measurement points of primary importance. Further, time series data does not lend itself to easy interpretation of results or model decisions.

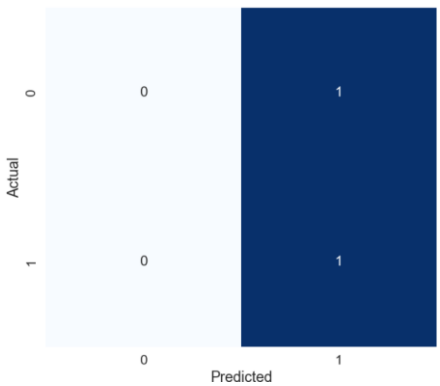
All four models were trained using the raw time series signal, with each discrete amplitude point representing a feature. The models were then evaluated against the previously unseen test data set. Table 1 presents the accuracy of each model, across all input feature groups. With reference to the time series results, the random forest shows the greatest accuracy, with all other models performing relatively poorly. To contextualise the accuracy, Figure 2 shows the confusion matrices for the evaluated models.

Input Feature	Accuracy [%]			
	Random Forest	Nearest-neighbours	Logistic Regression	Support Vector
Time Series	76	50	49	50
Spectrum	63	60	74	67
Cyclic Spectrum	94	93	98	97
Signal Statistics	76	67	95	84

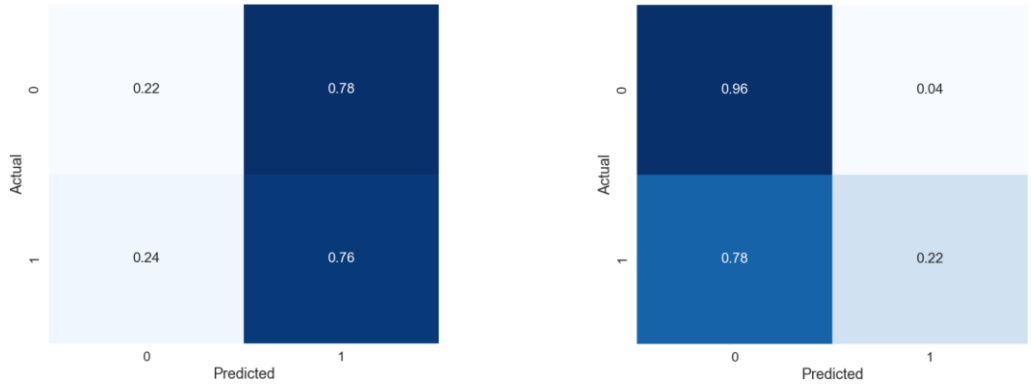
Table 1: Reported accuracy for all models across input feature groups.



(a)



(b)



(c) (d)

Figure 2: Confusion matrix obtained with time series as input for: (a) random forest; (b) nearest-neighbors; (c) logistic regression; and (d) support vector models.

The random forest model showed the greatest accuracy in predicting cavitation. This score was driven by a greater ability to classify presence of cavitation compared to when cavitation was not present. The nearest neighbours was not suitable for the current prediction – only predicting the presence of cavitation. Similarly, the logistic regression classification was biased towards the presence of cavitation. On the other hand, the support vector machine models was biased towards the negative classification, i.e. cavitation not present.

Overall, the models were not well suited for the time series classification. However, the time series classification is an inherently difficult task due to the nature of the input features, i.e. ordered discrete points. Therefore, moving forward, models focussed specifically on the classification of time series data should be evaluated. These models currently exist in the literature and typically focus on a number of pre-processing steps to break the data into a smaller number of features based on segments of the time series signal.

During the analysis, the signal to noise ratio was observed to have a significant impact on the results. Synthetic data was created over a range of different signal to noise ratio values, and this variation appears to make the classification more challenging. When the signal to noise ratio was constant across all input samples, even for high noise levels, the classification predictions improved. This highlights the importance of this parameter and the potential difficulties that may be faced when we transition to measured data where the signal to noise ratio may be quite variable depending on the measurement parameters.

3.2 Frequency Spectrum Classification

The frequency spectrum was obtained using a fast Fourier transform of the raw time series signals. The transform was applied over the entire signal duration resulting in a single amplitude spectrum for each time series signal. Whilst a cyclostationarity analysis typically uses short time Fourier transforms resulting in a spectrum over a discretised time, a single transform was used here to minimize the number and dimensions of the input features.

The discrete spectrum across the full bandwidth was used as the input feature for each training point. Table 1 shows the computed accuracy of the models when evaluated using the test data. Comparing with the time series performance, an increase in performance is observed for all models with exception of the random forest model. These results are contextualised in Figure 3 which shows the corresponding confusion matrices for each model.

With the exception of the logistic regression, accuracy is biased towards the negative classification, i.e. no cavitation. Similarly to the time series problem, due to the nature of the input features, results are difficult to interpret. However, the spectral data benefits from a reduction in the number of features. This may result in the ability of the models to generalise more effectively and hence the general observation of model improvement. Nonetheless, the results show that further work is required if the

spectrum is to be used as the principal feature for detection. The previous results related to the signal to noise ratio were also observed when the spectrum was used as the input feature.

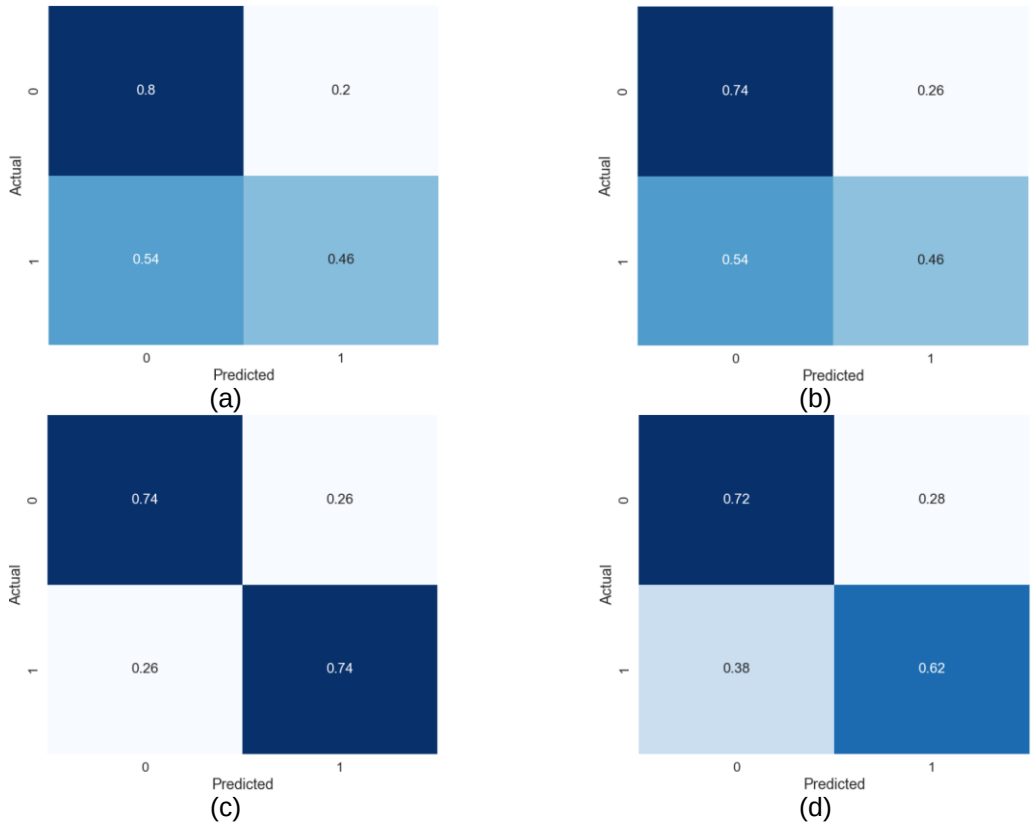


Figure 3: Confusion matrix obtained with frequency spectrum as input for: (a) random forest; (b) nearest-neighbors; (c) logistic regression; and (d) support vector models.

3.3 Cyclic Spectrum Classification

Models were subsequently evaluated against the cyclic spectrum. With reference to Figure 1, the classification is based on distinguishing between a spectrum of uncorrelated noise, no cavitation, and a spectrum dominated by peaks at the blade rate and its harmonics, the cavitating case. However, this is made more difficult due to the differing blade counts and rotational speeds, resulting in peaks being at different indices in the input feature data.

Table 1 shows high levels of accuracy across all models where the cyclic spectrum is used as the input feature. Figure 4 shows the corresponding confusion matrix for each model. The confusion matrices show good performance across all models for both positive and negative cavitation classification. The nearest-neighbour model showed the lowest performance, related to positive classification of cavitation.

The cyclic spectrum comprises significantly less input features when compared to the time series or frequency spectrum cases and this may have helped models generalise better on the test data. Further, there is an objectively much clearer distinction between the two classifications when using the cyclic spectrum relative to the previous features and this may have also helped. The effect of signal to noise ratio is also less apparent on the cyclic spectrum data, presenting a less challenging classification problem.

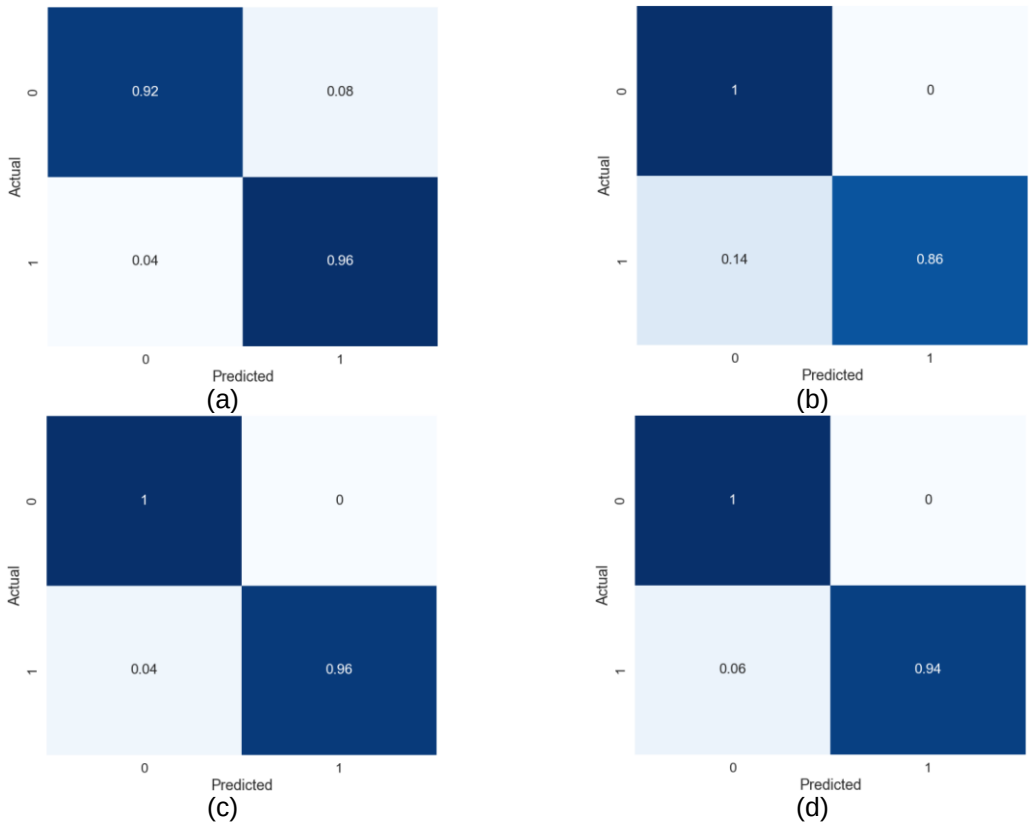


Figure 4: Confusion matrix obtained with cyclic spectrum as input for: (a) random forest; (b) nearest-neighbors; (c) logistic regression; and (d) support vector models.

3.4 Signal Statistics Classification

A number of signal statistics were computed and used as input features. The purpose of this investigation was to identify if there were any easy to compute statistical features of the signal that could be used to identify the presence of cavitation.

Table 1 shows mixed results between the different models. The nearest-neighbours models performed the worst, with the random forest performing only slightly better. Both the support vector and logistic regression showed very promising results, with the logistic regression performing the best.

Figure 5 shows the corresponding confusion matrix for each model. The random forest and nearest neighbour show similar trends with relatively poor ability to predict both classifications. The support vector performed better, but results were biased in its improved ability to classify the where cavitation was not present. The logistic regression showed high accuracy in both states.

Unlike the other feature groups, the statistics features are much more interpretable. To this end, Figure 6 shows the feature importance computed for the logistic regression model. High positive values indicate importance for positive detections, whilst high negative values are important for negative classification. The standard deviation of the time series signal is highlighted as the most important feature for the positive detection of cavitation. On the other hand, the mean value of the frequency spectrum is identified as the most important feature for classifying when cavitation is not present. Upon analysing these features, no clear trend was observed, therefore, further work is required to better understand the importance of these features and how they influence the classification. It is also interesting to note that the mean and median values of the time series signal

was not important in the classification. Overall, breaking down the signal into a number of statistics proved successful in enabling accurate detection of cavitation when using a logistic regression model.

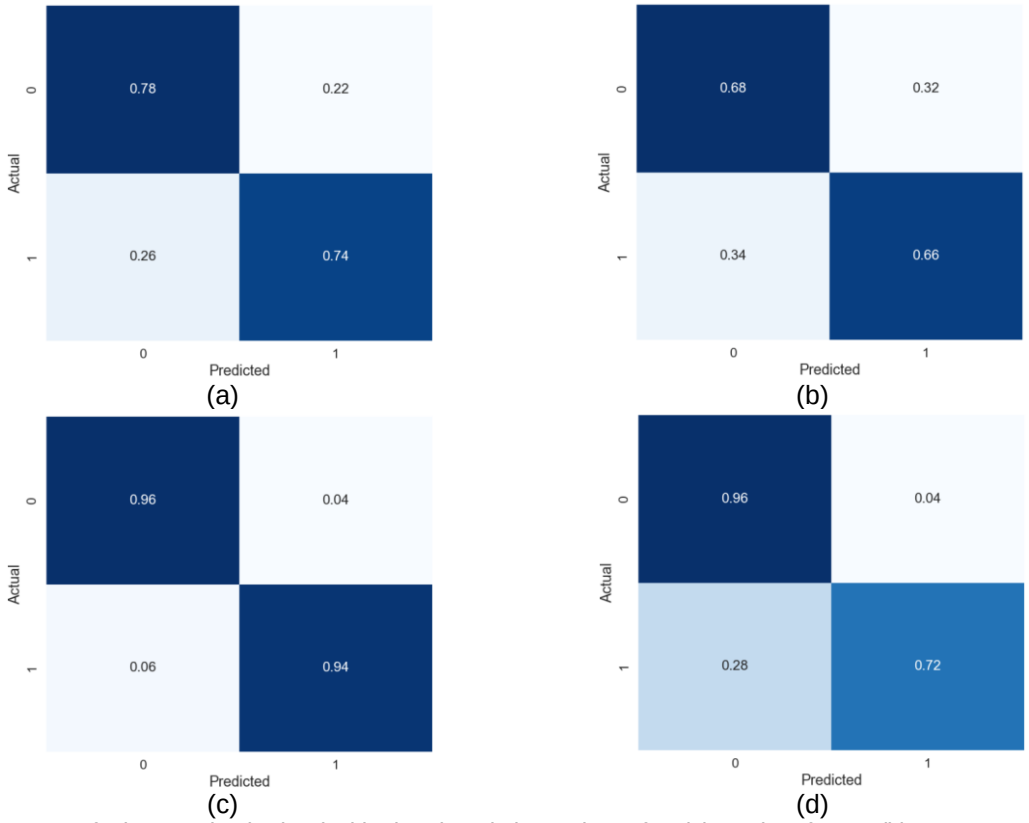


Figure 5: Confusion matrix obtained with signal statistics as input for: (a) random forest; (b) nearest-neighbors; (c) logistic regression; and (d) support vector models.

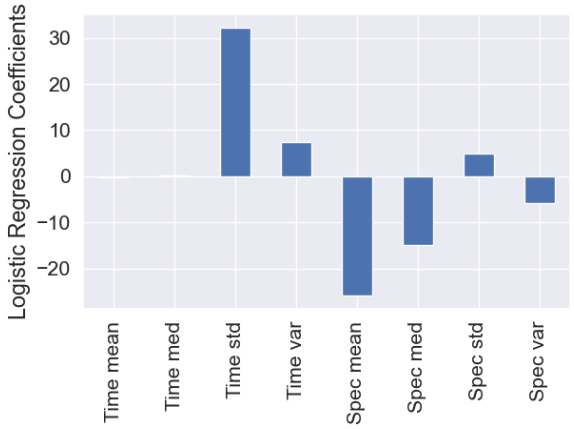


Figure 6: Feature importance based on logistic regression coefficients.

4 CONCLUSIONS

The present work has described the outcomes of a preliminary investigation of machine learning models for the detection of propeller cavitation. Four widely available models have been investigated, without modification, to the problem of binary classification across a range of input feature groups. The investigation focussed on identifying which features and machine learning models showed the greatest promise to direct work towards the development of a cavitation detection system. The investigation was based on synthetic data comprising cavitating and non-cavitating states.

The classification of raw time series data proved the most challenging, with most models performing poorly. The random forest model showed the best performance, but due to the nature of the input features – time series data, interpreting the results proved difficult. Transforming the time series signals to the frequency domain resulted in an improvement in the ability of all models to classify accurately. However, overall performance was still relatively poor. The reason for the poor performance of the time series and spectrum input features is believed to be due to the nature of the input data, seemingly unrelated data points, where the order is important. More advanced models designed specifically for this data should be investigated in order to evaluate their capability for cavitation detection.

The greatest performance was observed when the cyclic spectrum was used as the input feature. Unlike the time and spectrum features, the cyclic spectrum is observed to have a very noticeable contrast in features when cavitation is present and when it is not and is believed to be responsible for the improved performance. However, obtaining the cyclic spectrum requires a number of signal processing steps, with peak-picking being the final stage of the cyclostationary approach to cavitation detection. Therefore, unless there is significant difficulty with the peak-picking, the use of the machine learning approach may not be justified.

The final input features was a set of easy to compute statistics based on the time and frequency signals. Results across models were mixed. However, the logistic regression model showed very promising results. This input feature set shows that accurate cavitation detection may be possible with little processing of the time series data. The use of these statistics indicates the potential of minimal pre-processing of the raw signal to obtain accurate cavitation detection and approaches utilising some dissection of the time signal using statistical measures should be investigated further.

Across most input feature groups, the signal to noise ratio was found to have a significant impact on the accuracy of models. In particular, it is believed the variation in signal to noise ratio across input samples proved to make classification more challenging. This may provide a significant challenge when investigations are undertaken with real data. Therefore, work is required in order to investigate the challenges resulting from the variation in signal to noise ratio more thoroughly. Further, it is necessary to verify if the machine learning models are limited to the same extent as the cyclostationarity analysis where signal to noise is concerned.

This work has provided some insight into the use of machine learning models for the detection of cavitation. Whilst the present work has used synthetic data, future work will apply the same investigations using real measured data. Further, more complex models, specifically suited to the challenges associated with the described input features will be explored.

5 REFERENCES

1. C. Erbe, S.A. Marley, R.P. Schoeman, J.N. Smith, L.E. Trigg and C.B. Embling. “The effects of ship noise on marine mammals – a review”. Frontiers in Marine Science, vol. 6, 2019. doi: 10.3389/fmars.2019.00606

2. J. Antoni and D. Hanson, "Detection of Surface Ships From Interception of Cyclostationary Signature With the Cyclic Modulation Coherence," in IEEE Journal of Oceanic Engineering, vol. 37, no. 3, pp. 478-493, 2012.

3. D. Ross. “Propeller Cavitation Noise”, in: Mechanics of Underwater Noise. Pergamon, 1976.

4. F. Pedregosa *et al*. “Scikit-learn: Machine Learning in Python”, Journal of Machine Learning Research. vol.12, pp. 2825-2830, 2011.