Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2020 doi:10.20944/preprints202012.0152.v1

A machine learning approach for detecting wind farm noise amplitude modulation

Duc Phuc Nguyen,[1,a] Kristy Hansen,[1] Bastien Lechat,[1] Peter Catcheside,[2] and Branko Zajamsek[2]

[1] *College of Science and Engineering, Flinders University, Adelaide, SA 5042, Australia*

[2] *Adelaide Institute for Sleep Health, Flinders University, Adelaide, SA 5042, Australia*

Amplitude modulation (AM) is a characteristic feature of wind farm noise and has the potential to contribute to annoyance and sleep disturbance. This study aimed to develop an AM detection method using a random forest approach. The method was developed and validated on 6,000 10-second samples of wind farm noise manually classified by a scorer via a listening experiment. Comparison between the random forest method and other widely-used methods showed that the proposed method consistently demonstrated superior performance. This study also found that a combination of low-frequency content features and other unique characteristics of wind farm noise play an important role in enhancing AM detection performance. Taken together, these findings support that using machine learning-based detection of AM is well suited and effective for in-depth exploration of large wind farm noise data sets for potential legislative and research purposes.

[a] ducphuc.nguyen@flinders.edu.au

I. INTRODUCTION

Amplitude modulation (AM) of wind farm noise (WFN) is a unique feature known to contribute to annoyance (Ioannidou *et al.*, 2016; Lee *et al.*, 2011; Schäffer *et al.*, 2016) and possibly sleep disturbance (Bakker *et al.*, 2012; Liebich *et al.*, 2020; Micic *et al.*, 2018). AM in the context of WFN is defined as a periodic variation in sound pressure level (SPL) at the blade-pass frequency (Bass *et al.*, 2016; Hansen *et al.*, 2017), typically between 0.4 and 2 Hz, and is typically most prominent during the evening and night-time when environmental conditions tend to be more favourable for AM (Conrady *et al.*, 2020; Hansen *et al.*, 2019; Larsson and Öhlund, 2012). AM is a highly variable phenomenon, depending on meteorological conditions (Conrady *et al.*, 2020; Larsson and Öhlund, 2014; Paulraj and Välisuo, 2017), distance from the wind farm and wind farm operating conditions (Hansen *et al.*, 2019), making AM challenging to detect.

AM is commonly detected using simple engineering methods (Hansen *et al.*, 2017) using specific noise features (single predictor). For example, frequency domain-based methods (Larsson and Öhlund, 2014; Lundmark, 2011) detect and quantify AM using maximum spectral peaks between 0.6 Hz and 1.0 Hz. Time domain-based methods typically detect AM using SPL variations, where AM is classified as the difference between the 5^{th} and 95^{th} percentile of SPL greater than 2 dB (Fukushima *et al.*, 2013) or as a peak-to-trough difference of 3 dB or 5-6 dB (Bass, 2011; Cooper and Evans, 2013). Recently, the Institute of Acoustics UK has developed a hybrid method (Bass *et al.*, 2016), which is a combination of time and frequency domain methods. This method uses the prominence ratio, a ratio of peak and masking level, as a predictor of AM occurrence. The main advantage of these engineering methods is the ease of their implementation and computational speed, which makes them suitable for automated analysis of large data sets (Conrady *et al.*, 2020; Hansen *et al.*, 2019; Larsson and Öhlund, 2014). However, evaluation of the performance of these methods is currently limited to false positive rates alone, or to small data sets (Bass, 2011; Bass *et al.*, 2016; Larsson and Öhlund, 2014) or lacking is altogether (Fukushima *et al.*, 2013; Nordtest, 2002).

Machine learning methods are emerging in many acoustical applications (Bianco *et al.*, 2019) such as noise predictions (Valente, 2013), sound propagation (Hart *et al.*, 2016a,b) and sound classification (Nykaza *et al.*, 2017). These methods allow for combination of multiple, otherwise isolated noise features into one robust classifier. This overcomes one of the major issues associated with traditional AM detection methods, which is reliance on a single noise feature which poorly accounts for the highly variable and multifaceted phenomenon of AM (Hansen *et al.*, 2017). Here we present an AM detection method derived from a random forest classification algorithm (Breiman, 2001). We trained and tested this new method was trained and tested on human-scored data sets (hereafter referred to as the benchmark data set) followed by comparison against three widely-used AM detection methods (Bass *et al.*, 2016; Fukushima *et al.*, 2013; Larsson and Öhlund, 2014). Overall, the machine learning-based method outperformed current methods and is effective for exploration of large wind farm noise data sets.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2020 doi:10.20944/preprints202012.0152.v1

II. METHODS

A. Overview of data collection

The data set used for development and validation of the AM detection method contained WFN measured at four residences (H1-H4) located between 980 m and 3.5 km from the nearest wind turbine of South Australian wind farms (Supplementary Fig. S1). Residence H4 was unoccupied and located approximately 30 km from the nearest wind farm, and thus it was assumed that AM WFN did not exist at this location. Noise data were measured for one year at locations H1 and H2 and two weeks and five months at locations H3 and H4, respectively. The H3 data set also contained approximately three days of measurements of background noise when the wind farm was not operating. This data set together with the H4 data set were used for false positive rate evaluations.

At all measurement locations, acoustic data were acquired using a Bruel and Kajer LAN-XI Type 3050 data acquisition system with a sampling rate of 8,192 Hz and a G.R.A.S type 40 AZ microphone with a 26CG preamplifier, which has a noise floor of 16 dB(A) and a flat frequency response down to 0.5 Hz. Further details of the experimental setup are described in (Hansen *et al.*, 2014, 2019).

B. Benchmark data set generation

Two benchmark data sets were constructed, one containing 6,000 10-second audio files of WFN and the other one of equal size containing no WFN (environmental background noise only). The latter data set was specifically constructed for testing false positive detection.

These data sets were selected randomly from recorded data (Supplementary Fig. S2). The WFN benchmark data set was primarily scored by a single scorer using a validated rating experiment procedure based on detection theory (Macmillan and Creelman, 2004). To evaluate inter-scorer agreement, another expert scorer also rated a sub-sample of 100 randomly chosen audio samples. The scorers were acoustic engineers familiar with wind farm AM, who listened to the audio files and scored the presence versus absence of AM. AM presence was rated based on confidence level which varied from high confidence of AM absence (rating "1"), to uncertainty between AM presence/absence (rating "3"), to high confidence of AM presence (rating "5") (Supplementary Fig. S3). The rating experiment was performed in a bedroom at the Adelaide Institute for Sleep Health, which has a background noise level of 22 dBA. The noise reproduction system consisted of Bose Quite Comfort II headphones and a RME Babyface Pro sound card.

C. Automated AM detectors

The proposed AM detection method was compared against three previously published AM detection methods. The first method, labelled a1 (Bass *et al.*, 2016), uses a "hybrid" approach involving analysis in both the time- and frequency-domains. The other two methods labelled a2 (Larsson and Öhlund, 2014) and a3 (Fukushima *et al.*, 2013) are implemented in the frequency- and time-domains, respectively.

Method a1 band-pass filters the signal over the expected AM frequency range, calculates the fast-time weighted SPL time series, detrends the data, then transforms the detrended SPL time series data to the frequency-domain. AM is then detected where the prominence

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2020 doi:10.20944/preprints202012.0152.v1

ratio (PR), the ratio between the spectral peak in the blade-pass frequency range and the noise floor, is greater than four (Bass *et al.*, 2016).

Method a2 is implemented by firstly applying a low-pass filter at 1 kHz, calculating the fast-time weighted SPL and then transforming this time series into the frequency-domain. The $AMfactor$, the maximum spectrum amplitude between 0.6 Hz and 1 Hz, is then used to obtain the threshold for AM detection. The suggested threshold is 0.4 (Larsson and Öhlund, 2014).

Method a3 is implemented by applying a low-pass filter at 1 kHz and then detrending the fast-time weighted SPL. After quantifying the variation of detrended SPL via calculating the difference between statistical noise levels $L95$ and $L5$, this value, referred to as DAM, is used as a threshold for detecting AM. The suggested threshold varies from 2 dB to 6dB (Bass, 2011; Cooper and Evans, 2013; Fukushima *et al.*, 2013). More details regarding these methods are available as pseudo code provided in Supplementary Algorithm 1-3. Also, the source code for method a1, as provided by (Coles *et al.*, 2017) was reimplemented using MATLAB in our study (Supplementary Fig. S4).

D. Random Forest classifier for AM detection

A random forest classifier (Breiman, 2001) consists of decision trees, which represent possible outcome maps for a series of related choices. Decision trees are easy to use and generally work very well with the data used to create them, but are more problematic for predictive learning models requiring more flexibility for accurate classification of new data (Hastie *et al.*, 2009). To overcome these decision tree problems, the random forest classifier

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2020 doi:10.20944/preprints202012.0152.v1

uses bootstrap sampling and random variable selection to build multiple trees, which are then combined into a random forest classifier as shown in Fig. 1. To classify an input sample (i.e., AM or no AM), the relevant audio features are plugged into every predictor (tree) in the classifier. Then each predictor classifies the sample as "AM" or "no AM". Finally, a majority voting approach is used to decide if the input audio can be classified as containing "AM" or "no AM". This achieves a probabilistic classifier, where the ratio between the number of trees voting "AM" out of the total tree population represents the probability of AM being present.

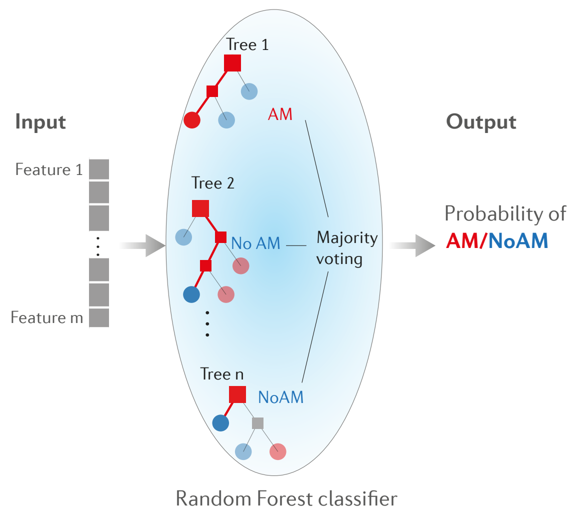

FIG. 1. (Color online). Random forest classifier.

Optimisation of hyperparameters, that is parameters which are set before the learning begins, was done using a random searching technique (Bergstra and Bengio, 2012). The following set of hyperparameters were adjusted: number of trees, number of features considered for splitting at each leaf node, maximum number of decision splits, and the minimum

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2020 doi:10.20944/preprints202012.0152.v1

number of data points allowed in a leaf node. The random searching technique utilises a range of realistic hyperparameters values, as shown in Tab. I.

TABLE I. Value ranges of the hyperparameters used for random searching.

Hyperparameter	Range
Num tree	$\{2, 4, 8, ...1024\}$
Max num feature	$\{1, 2, 3, ...31\}$
Max num split	$\{2, 4, 8, ...4096\}$
Max leaf size	$\{2, 4, 8, ...1024\}$

E. Audio feature extraction

WFN spectra are dominated by lower-frequencies, particularly at distances greater than 1 km from a wind farm (Hansen *et al.*, 2017). Also, WFN can contain both tonal AM (Hansen *et al.*, 2019) and/or broadband AM. Furthermore, AM can occur at frequencies ranging from 30 Hz to more than 1 kHz, and the peak-to-trough magnitude can vary between each successive oscillation period (Larsson and Öhlund, 2014). To help capture the highly variable and evolving nature of WFN, which likely influences AM characteristics and consequently detection performance, a comprehensive range of 31 noise features were used in this study (Supplementary Table. S1). The noise features were divided into five categories, including spectral shape features, tonality features, overall noise features, time domain features and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2020 doi:10.20944/preprints202012.0152.v1

features extracted from the other automated AM detection methods described in Section C. Further details regarding the feature extraction process are provided in Supplementary Fig. S5.

F. Evaluation metrics

The performance of the automated AM detection methods was evaluated using both a precision-recall curve (PR) and the Matthews correlation coefficient (MCC), which are well suit to imbalanced data sets (Lever *et al.*, 2016). To construct the PR curve, pairs $(precision, recall)$ were calculated from the counts of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) as follows

$$recall = \frac{TP}{TP + FN}; \quad precision = \frac{TP}{TP + FP} \tag{1}$$

The aggregate metric of MCC is a more informative and faithful score of overall classification performance compared to common metrics such as accuracy or $F1$-score (Chicco and Jurman, 2020). The MCC ranges from -1 (classification is always wrong) to 0 (classification is no better than random guess) to 1 (classification is always correct), and it is calculated as follows

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2}$$

The use of a single metric, and even an aggregate metric like MCC, can be misleading without careful inspection of the underlying results. Thus, in this study, additional metrics

 doi:10.20944/preprints202012.0152.v1

including Cohen's kappa, accuracy, area under ROC curve, etc., (Lever *et al.*, 2016), were also calculated as secondary results (Supplementary Table. S2).

G. Data and statistical analysis

All signal, data and statistical analyses were implemented in MATLAB, in which the noise feature extraction was implemented using the Audio Toolbox. The random forest model was implemented using the Statistics and Machine learning Toolbox. The statistical significance threshold used was $\alpha = 0.05$. All data are reported as mean [95 % confidence interval], unless otherwise indicated. Pearson correlation coefficients were used to examine the strength of linear relationships between features.

H. Data availability

The MATLAB code used to extract features and build the random forest-based AM detection method can be found in the GitHub open repository together with the scored data set `https://github.com/ducphucnguyen/WFN_AM_Detection`.

III. RESULTS

A. Benchmark data set characteristics

The benchmark data set of 6,000 10-second audio files was unbalanced with around 40% of audio samples containing AM (Fig. 2a). The AM confidence rating was transformed into a binary score (AM vs. no AM) using a confidence rating threshold of three. Samples with

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2020 doi:10.20944/preprints202012.0152.v1

ratings greater than three were classified as AM, and all other samples were classified as no AM. Both positive and negative skewness was observed from the rating distribution, indicating high confidence in scorer rating. The MCC and $F1$-score for inter-scorer agreement were (mean [95% CI) 0.65 [0.49, 0.80] and 0.77 [0.66, 0.87], indicating a high degree of agreement (Warby *et al.*, 2014) (See Supplementary Table. S3 for other metrics). Distributions of scored audio files over months, hours and wind farm power output relative to capacity were also nearly uniform, consistent with ecological validity (Fig. 2b).

B. Random forest-based AM detection performance

Hyperparameters were estimated using the out-of-bag samples, which comprised approximately 37% of the total samples not used for training the classifier. The hyperparameters were chosen after 500 iterations by maximising the area under the precision-recall curve ($AUPRC$), (Breiman, 1996) (Fig. 3a). The optimal hyperparameter settings were: 1,024 trees, a maximum of 16 features, a maximum of 2,048 splits and a minimum of 4 samples in the leaf nodes. The precision-recall curve in Fig. 3b shows the optimal random forest classifier based on these hyperparameters with $AUPRC = 0.85$ [0.84, 0.86] (See Supplementary Table. S4 for other metrics).

Some selected features may not useful for AM prediction given a cluster of highly correlated variables in the dendrogram (showing the hierarchical relationship between features) and high Pearson correlation coefficient in Fig. 3c. The four most importance features for predicting AM are $AMfactor$, $SpectralCrest$, $diffLCLA$ and PR (Fig. 3d).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2020 doi:10.20944/preprints202012.0152.v1

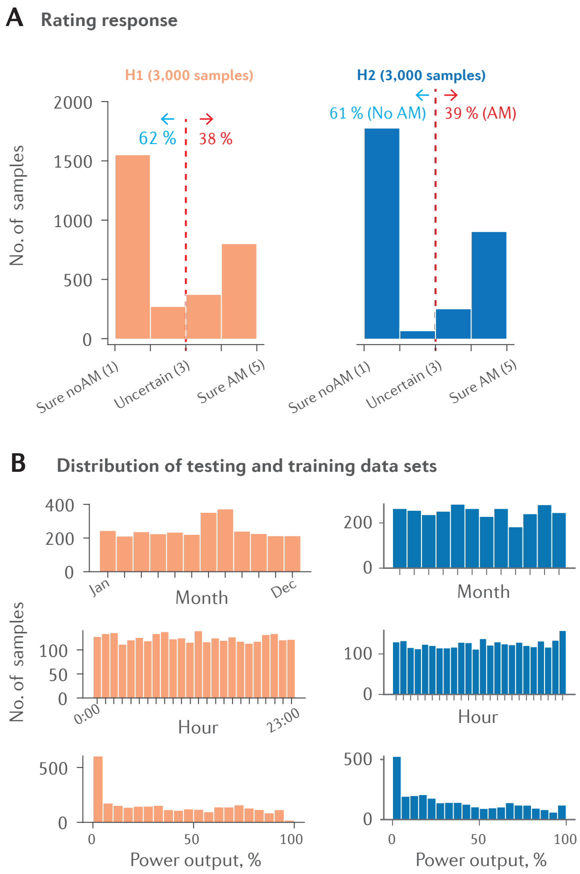

FIG. 2. (Color online). Characteristics of benchmark data sets. **A**, scorer ratings distribution with corresponding binary classification. **B**, distributions of audio files per month, hour and wind farm power percentage output relative to capacity.

C. Performance of the automated detectors

The performance of the random forest-based AM detection method was compared to three automated detectors (a1-a3) on precision-recall plots (Fig. 4a). The test set for detectors a1-a3 was all samples in the benchmark data set while the out-of-bag samples were used as the test set for the random forest detector. The random forest-based method outperformed the other methods (ANOVA P-value < 0.001), with an $AUPRC$ of 0.85. The

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2020 doi:10.20944/preprints202012.0152.v1

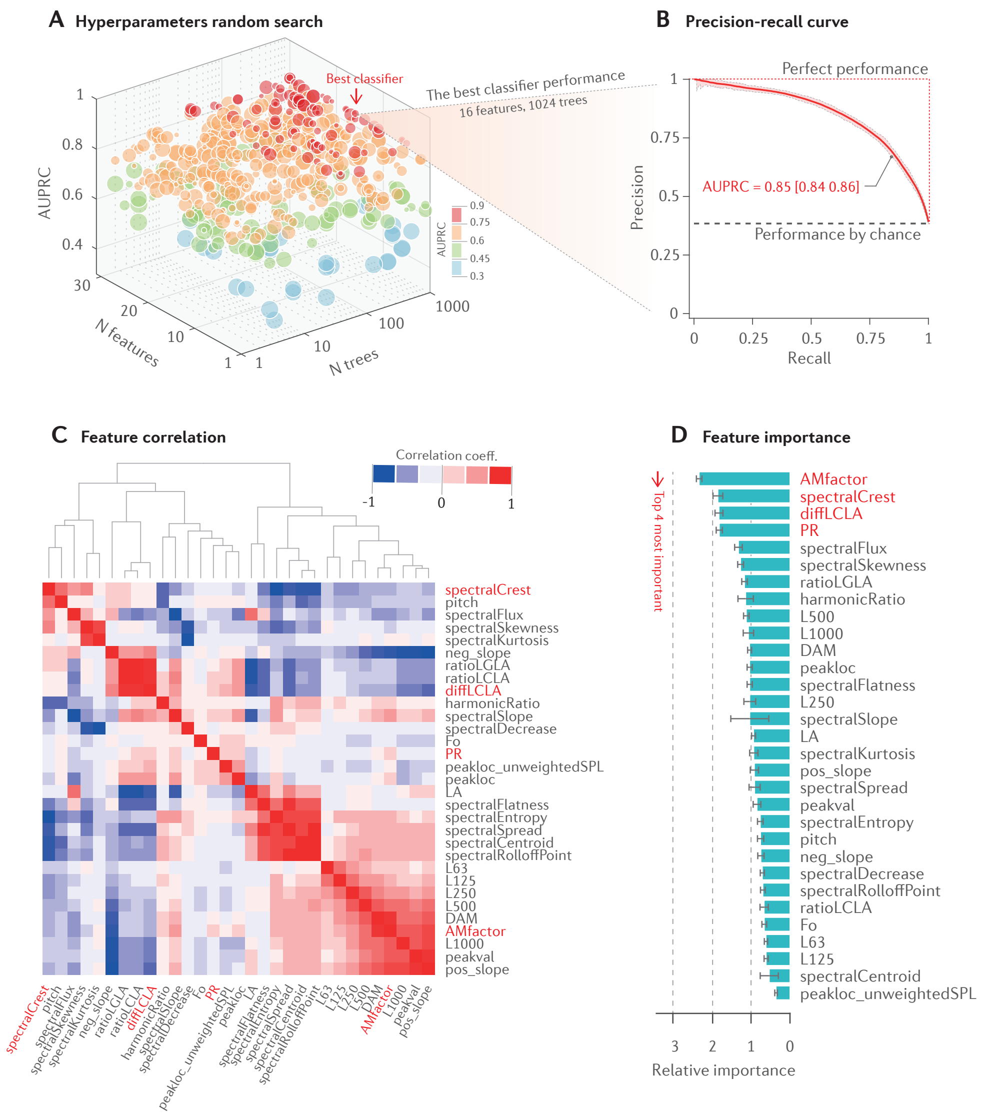

FIG. 3. (Color online). Random Forest classifier. **A**, hyperparameter tuning using a randomized search technique. The size of the circles represents the maximum splits. Minimum leaf node samples are not shown. B, the precision-recall curve of the best random forest classifier. The shaded area indicates 95% CI. **C**, Pearson correlation coefficient (Pearson's r) map with dendrogram for illustrating clusters. **D**, feature importance in descending order from top to bottom. Error bars indicate 95% CI.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2020 doi:10.20944/preprints202012.0152.v1

performance of a1-a3 was poor with the mean $AUPRC$ ranging from 0.43 to 0.55 (Table II). The performance of a1 was better than a2 and a3 (all $P < 0.001$), and a2 performed better than a3 ($P < 0.001$).

TABLE II. Area under the precision-recall curves and optimal MCC of four methods.

Method	$AUPRC$	Max MCC
Random forest	0.85 [0.84 0.86]	0.62
a1	0.55 [0.52 0.58]	0.29
a2	0.47 [0.45 0.49]	0.32
a3	0.43 [0.40 0.44]	0.28

The performance of AM detection algorithms has previously been described in terms of the false positive rate (FPR) (Bass *et al.*, 2016; Larsson and Öhlund, 2014), and thus this metric was also examined (Fig. 4b). As the random forest classifier is based on probabilistic values, a threshold of 0.5 was used for binary classification of AM. Thus, if more than 50% of trees in the classifier voted for "AM", the sample was classified as an AM sample, otherwise "no AM" was declared. The cut-of values for method a1-a3 were 4, 0.2 and 2, respectively (See Methods section). The false positive rate of the random forest classifier was low (1.6%) compared to methods a1-a3 (50%, 19% and 62%, respectively). The false positive rate of methods a1 and a3 was not reported in the original descriptions of these methods (Bass

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2020 doi:10.20944/preprints202012.0152.v1

et al., 2016; Fukushima *et al.*, 2013), but was reported to be 2.6% for method a2 (Larsson and Öhlund, 2014), and thus substantially lower than in our data set analysed in this study.

To evaluate if the performance of all detectors could be improved using different threshold values, thresholds for each method were varied systematically to find the highest MCC values as shown in Fig. 4c. The optimal threshold for the random forest classifier was 0.44 (44% of trees voted "AM"). The optimal threshold for method a1 was PR=6.7, which is higher than the original reported value of $PR = 4$ in (Bass *et al.*, 2016) and the value obtained using a Receiver Operating Characteristic curve (PR=3) in (Hansen *et al.*, 2019). In contrast, the optimal thresholds for method a2 and a3 were lower than original suggested values (Fukushima *et al.*, 2013; Larsson and Öhlund, 2014). For comparison, the MCC between two scorers was calculated and considered as the ceiling value for the AM detection task ($MCC = 0.65$), supporting that the performance of the random forest classifier was remarkably close to human performance.

D. Interpretable predictor

The random forest classifier with 31 features and 1,024 trees outperformed traditional detection methods and showed performance comparable with human classifiers. However, random forest classifiers work much like a black box, which is difficult to interpret. The classifier also requires skilled human and computer resources to implement. Given the feature importance findings of the importance of $AMfactor$, $diffLCLA$, $SpectralCrest$ and PR features, we thus aimed to build a simplified classifier, which can be used as a simpler and more portable classifier for AM detection. This simplified classifier was a single

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2020 doi:10.20944/preprints202012.0152.v1

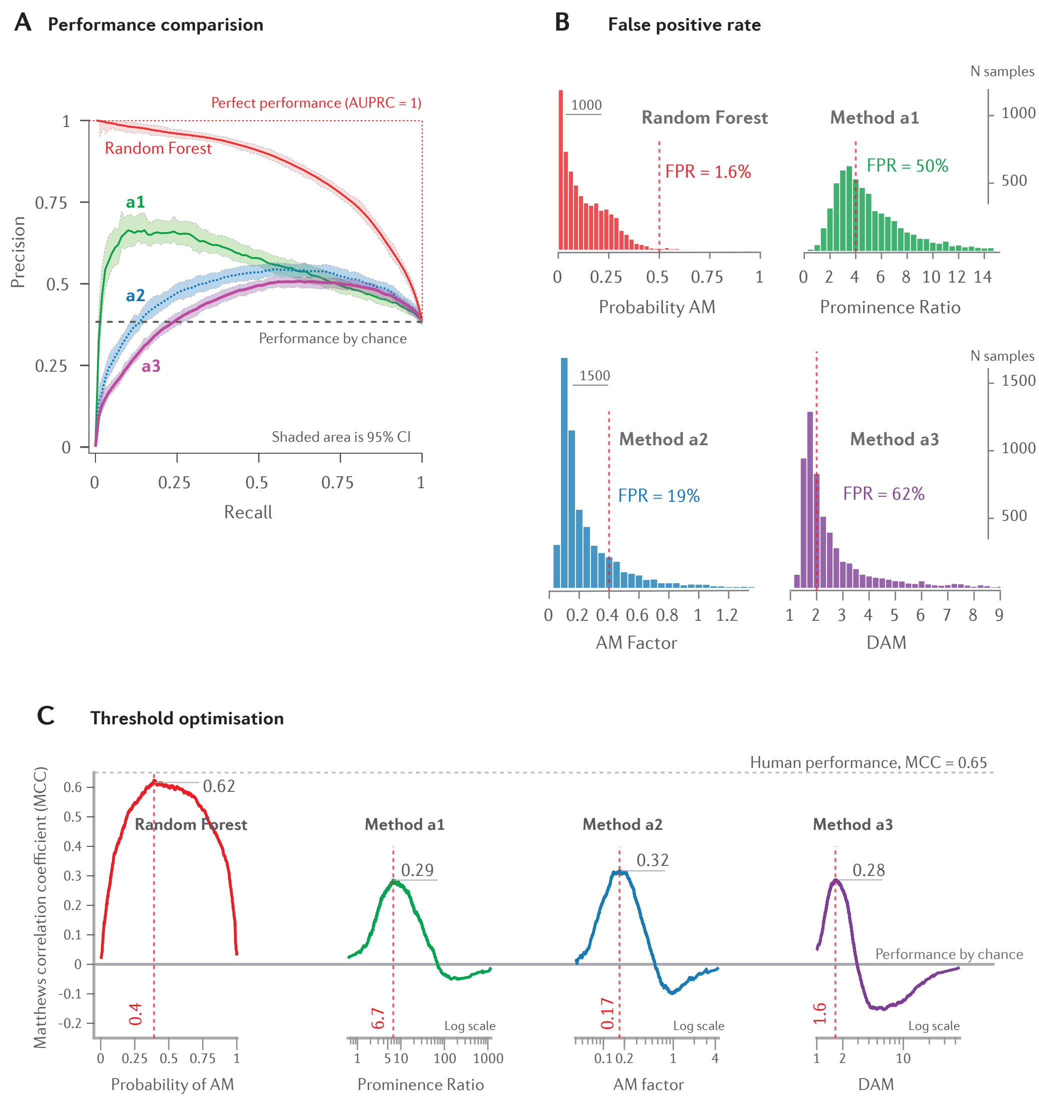

FIG. 4. (Color online). Performance of automated detectors. **A**, performance using the benchmark data set, where the values associated with each curve are mean [95% confidence interval]. The shaded area is the 95% CI. **B**, false positive rate of each detection method estimated from the no wind farm noise data set. The dashed lines indicate the AM classification threshold. **C**, optimal AM detection threshold according to MCC, where negative values indicate performance worse than by chance

 doi:10.20944/preprints202012.0152.v1

decision tree built from four features, as shown in Fig. 5. The performance of the single decision tree showed $AUCPR = 0.68$ [0.64, 0.71], which is lower than the random forest classifier, yet still higher than methods a1-a3. These results further illustrate that a simple combination of several features outperforms traditional single feature detection methods.

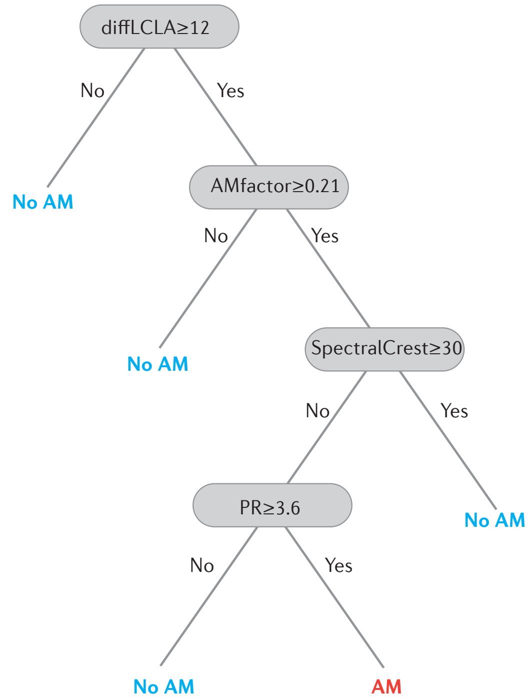

FIG. 5. (Color online). A simplified single tree classifier utilising the four most important features for identified by the random forest classifier AM detection.

IV. DISCUSSION

A validated and high-performing WFN AM classifier based on random forest machine learning technique was presented. This classifier substantially outperforms currently available classifiers, with a predictive power close to its practical limit set by human scoring.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2020 doi:10.20944/preprints202012.0152.v1

This approach shows major promise as an effective automated tool which could be used for detecting WFN AM presence in large data sets, such as for research or to support regulatory purposes. This approach also reveals new insights into the nature of AM itself, as it shows that other acoustical parameters apart from noise level variations are important for AM detection.

AM is a challenging signal to detect as its characteristics vary depending on meteorological conditions. As a result, the spectral content and time varying features are not constant. Despite these changes, the auditory system can still recognize the presence of wind farm AM. Thus, our presented algorithm sought to incorporate the most important acoustical features predictive of human scored AM. The selected features cover the whole range of the most dominant WFN characteristics, including noise level variation (or AM), tonality and low-frequency content. Two features incorporate noise level variations ($AMfactor$ and PR), the difference between $LCeq$ and $LAeq$ is an indicator of low-frequency noise presence and the spectral crest provides a simple measure of tonality. These findings support that human perception of AM is more complex than assumed by previous AM detection methods which are based on noise level variations alone. Hence, it is not surprising that the method presented here achieved substantial improvements in performance compared to previous methods.

Very high false positive rates were found for methods a1-a3, which is inconsistent with previous reports in (Bass *et al.*, 2016; Larsson and Öhlund, 2014). However, it is worth noting that method a1 was originally designed and evaluated on 10-minute samples, as opposed to the 10-second samples used in our work, and method a1 classifies AM if more than 50% of 10-second blocks within 10 minutes contain AM. By introducing the above criterion, the false positive rate may be substantially reduced, as reported in (Bass *et al.*, 2016). However, 10-second long samples appear to have higher ecological validity, as typical AM events usually last around 10-15 seconds (Larsson and Öhlund, 2014). With regards to the false positive rate for method a2, an arbitrary 30 dBA L_{Aeq} cut-off was imposed in the original evaluation, which was not used in our study, and likely helps to explain the large discrepancy between the originally reported 2.6% (Larsson and Öhlund, 2014) and the 19% false positive rate in our study. If the 30 dBA cut-off is applied to our data before method a2 is used to detect AM, the false positive rate is reduced from 19% to 9%. This number is expected to further reduce if data were measured in a quiet area, where many samples would have associated noise levels less than 30 dBA. Therefore, these findings further support that false positive rate metrics are problematic for evaluating detection performance (Warby *et al.*, 2014), as this only represents one parameter in a confusion matrix.

A limitation of the present study is the under-representation of noise data measured greater than 1 km used for training and testing the random forest classifier. As a result, the proposed classifier may not work well for detecting AM measured several kilometers from the nearest wind turbine, where AM may have different characteristics (Hansen *et al.*, 2019). The classifier could not be tested on data sets measured outside of South Australia, where weather conditions and topography near wind farms will inevitably to vary. Although the reliability of human scoring has been tested, using a single scorer to classify the AM is not ideal. As suggested by Wendt *et al.* (2015), two or more scorers and a consensus scoring approach may be preferable to a single scorer to help ensure broader generalisability.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2020 doi:10.20944/preprints202012.0152.v1

Nevertheless, a single scorer is more practical and avoids potential effects of poor inter-scorer agreement. Also, good inter-scorer agreement was found in a smaller subset of the data, supporting this approach.

Although detector a1 clearly warrants improvements in order to increase accuracy, the source code (Coles *et al.*, 2017) is readily available, making it easy to understand the methodology and to implement the method. Although the other methods were reproduced as closely as possible, our codes may be different from the original codes. This is a similar problem previously identified for the reproduction of the tonality assessment code in Søndergaard *et al.* (2019) . Thus, depositing source code to open source repositories, together with relevant data sets would greatly advance the development of practical and robust amplitude modulation detection methods.

V. CONCLUSIONS

In summary, this study demonstrates that random forest-based AM detection is a good approach for AM classification, and substantially outperforms traditional AM detection methods to achieve classification performance close to that of humans. It was also shown that a simplified classifier based on a single decision tree using the four main features identified through the random forest approach also achieves good classification performance. This approach is readily interpretable and easy to implement without the need for extensive computer resources. Finally, it is important to stress that the main aim for developing an improved AM detection algorithm was to better understand the characteristics of this phenomenon, and thus algorithm performance was prioritized above algorithm simplicity and low computational time. We hope that, in the future, further insight into the prevalence of AM, associated meteorological conditions, and impacts on humans will help to explain underlying noise generation mechanisms. Ultimately, this will improve the design of wind turbines such that they are less disturbing and hence, more acceptable to surrounding communities.

ACKNOWLEDGMENTS

The authors gratefully acknowledge financial support from the Australian Research Council, projects DP120102185 and DE180100022 and the National Health and Medical Research Council, Project 1113571. The author DPN was supported by a Flinders University Research Scholarship (FURS) for this work.

Bakker, R., Pedersen, E., van den Berg, G., Stewart, R., Lok, W., and Bouma, J. (**2012**). "Impact of wind turbine sound on annoyance, self-reported sleep disturbance and psychological distress," The Science of the total environment **425**, 42–51.

Bass, J. (**2011**). "Investigation of the "den brook" amplitude modulation methodology for wind turbine noise," Acoustics Bulletin **36**(6), 18–24.

Bass, J., Cand, M., Coles, D., Davis, R., Irvine, G., Leventhall, G., Levet, T., Miller, S., Sexton, D., and Shelton, J. (**2016**). "Institute of acoustics ioa noise working group (wind turbine noise) amplitude modulation working group final report a method for rating amplitude modulation in wind turbine noise version 1," .

Bergstra, J., and Bengio, Y. (**2012**). "Random search for hyper-parameter optimization," The Journal of Machine Learning Research **13**(1), 281–305.

Bianco, M. J., Gerstoft, P., Traer, J., Ozanich, E., Roch, M. A., Gannot, S., and Deledalle, C.-A. (**2019**). "Machine learning in acoustics: Theory and applications," The Journal of the Acoustical Society of America **146**(5), 3590–3628.

Breiman, L. (**1996**). "Out-of-bag estimation," Technical report, Statistics Department, University of California Berkeley .

Breiman, L. (**2001**). "Random forests," Machine learning **45**(1), 5–32.

Chicco, D., and Jurman, G. (**2020**). "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," BMC genomics **21**(1), 6.

Coles, D., Bass, H. J., and Cand, M. (**2017**). "Ioa am code - implementation of the core routine for am analysis from the ioa amwg" https://sourceforge.net/projects/ioa-am-code/.

Conrady, K., Bolin, K., Sjöblom, A., and Rutgersson, A. (**2020**). "Amplitude modulation of wind turbine sound in cold climates," Applied Acoustics **158**, 107024.

Cooper, J., and Evans, T. (**2013**). "Automated detection and analysis of amplitude modulation at a residence and wind turbine," in *Acoustics 2013*, Victor Harbor, Australia.

Fukushima, A., Yamamoto, K., Uchida, H., Sueoka, S., Kobayashi, T., and Tachibana, H. (**2013**). "Study on the amplitude modulation of wind turbine noise: Part 1–physical investigation," in *Internoise 2013*.

Hansen, C. H., Doolan, C. J., and Hansen, K. L. (**2017**). *Wind Farm Noise: Measurement, Assessment and Control*, 1 ed. (John Wiley Sons Ltd).

Hansen, K., Zajamsek, B., and Hansen, C. (**2014**). "Identification of low frequency wind turbine noise using secondary windscreens of various geometries," Noise Control Engineering Journal **62**(2), 69–82.

Hansen, K. L., Nguyen, P., Zajamšek, B., Catcheside, P., and Hansen, C. H. (**2019**). "Prevalence of wind farm amplitude modulation at long-range residential locations," Journal of Sound and Vibration **455**, 136–149.

Hart, C. R., Reznicek, N. J., Wilson, D. K., Pettit, C. L., and Nykaza, E. T. (**2016**a). "Comparisons between physics-based, engineering, and statistical learning models for outdoor sound propagation," The Journal of the Acoustical Society of America **139**(5), 2640–2655.

Hart, C. R., Wilson, D., Pettit, C. L., and Nykaza, E. T. (**2016**b). "Adaptive statistical learning models for long-range sound propagation," The Journal of the Acoustical Society of America **140**(4), 3088–3088.

Hastie, T., Tibshirani, R., and Friedman, J. (**2009**). *The elements of statistical learning: data mining, inference, and prediction* (Springer Science & Business Media).

Ioannidou, C., Santurette, S., and Jeong, C.-H. (**2016**). "Effect of modulation depth, frequency, and intermittence on wind turbine noise annoyance," The Journal of the Acoustical Society of America **139**(3), 1241–1251, http://asa.scitation.org/doi/10.1121/1.4944570, doi: 10.1121/1.4944570.

Larsson, C., and Öhlund, O. (**2012**). "Variations of sound from wind turbines during different weather conditions," Inter Noise 2012 .

Larsson, C., and Öhlund, O. (**2014**). "Amplitude modulation of sound from wind turbines under various meteorological conditions," Journal of the Acoustical Society of America **135**(1), 67–73.

Lee, S., Kim, K., Choi, W., and Lee, S. (**2011**). "Annoyance caused by amplitude modulation of wind turbine noise," Noise Control Engineering Journal **59**(1), 38, http://www.ingentaconnect.com/content/ince/ncej/2011/00000059/00000001/art00005, doi: 10.3397/1.3531797.

Lever, J., Krzywinski, M., and Altman, N. (**2016**). "Classification evaluation" .

Liebich, T., Lack, L., Hansen, K., Zajamšek, B., Lovato, N., Catcheside, P., and Micic, G. (**2020**). "A systematic review and meta-analysis of wind turbine noise effects on sleep using validated objective and subjective sleep assessments," Journal of Sleep Research e13228.

Lundmark, G. (**2011**). "Measurement of swish noise: a new method," in *Fourth International Meeting on Wind Turbine Noise*, Rome, Italy.

Macmillan, N. A., and Creelman, C. D. (**2004**). *Detection theory: A user's guide* (Psychology press).

Micic, G., Zajamsek, B., Lack, t. L., Hansen, K., Doolan, C., Hansen, C., Vakulin, A., Lovato, N., Bruck, D., Ching, t., Chai-Coetzer, L., Mercer, J., and Catcheside, P. (**2018**). "A review of the potential impacts of wind farm noise on sleep," Acoustics Australia

Nordtest (**2002**). "Nt-acou-112: Prominence of impulsive sounds and for adjustment of laeq" .

Nykaza, E. T., Blevins, M. G., Hart, C. R., and Netchaev, A. (**2017**). "Bayesian classification of environmental noise sources," The Journal of the Acoustical Society of America **141**(5),

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2020 doi:10.20944/preprints202012.0152.v1

3522–3522.

Paulraj, T., and Välisuo, P. (**2017**). "Effect of wind speed and wind direction on amplitude modulation of wind turbine noise," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, Institute of Noise Control Engineering, Vol. 255, pp. 5479–5489.

Schäffer, B., Schlittmeier, S. J., Pieren, R., Heutschi, K., Brink, M., Graf, R., and Hellbrück, J. (**2016**). "Short-term annoyance reactions to stationary and time-varying wind turbine and road traffic noise: A laboratory study," The Journal of the Acoustical Society of America **139**(5), 2949–2963, http://asa.scitation.org/doi/10.1121/1.4949566, doi: 10.1121/1.4949566.

Søndergaard, L. S., Thomsen, C., and Pedersen, T. H. (**2019**). "Prominent tones in wind turbine noise - round-robin test of the iec 61400-11 and iso/pas 20065 methods for analysing tonality content," in *8th International Conference on Wind Turbine Noise*, Lisbon, Portugal.

Valente, D. (**2013**). "Data-driven prediction of peak sound levels at long range using sparse, ground-level meteorological measurements and a random forest," The Journal of the Acoustical Society of America **134**(5), 4159–4159.

Warby, S. C., Wendt, S. L., Welinder, P., Munk, E. G., Carrillo, O., Sorensen, H. B., Jennum, P., Peppard, P. E., Perona, P., and Mignot, E. (**2014**). "Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods," Nature methods **11**(4), 385.

Wendt, S. L., Welinder, P., Sorensen, H. B., Peppard, P. E., Jennum, P., Perona, P., Mignot, E., and Warby, S. C. (**2015**). "Inter-expert and intra-expert reliability in sleep spindle

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2020 doi:10.20944/preprints202012.0152.v1

scoring," Clinical Neurophysiology **126**(8), 1548–1556.