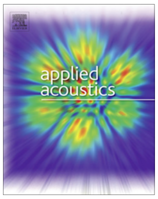
Applied Acoustics 183 (2021) 108286

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust



Benchmark characterisation and automated detection of wind farm noise amplitude modulation

Phuc D. Nguyen [a,*], Kristy L. Hansen [a], Bastien Lechat [a], Peter Catcheside [b], Branko Zajamsek [b], Colin H. Hansen [c]

[a] College of Science and Engineering, Flinders University, Adelaide, SA 5042, Australia
[b] Adelaide Institute for Sleep Health, Flinders University, Adelaide, SA 5042, Australia
[c] School of Mechanical Engineering, University of Adelaide, Adelaide, SA 5005, Australia

ARTICLE INFO

Article history:
Received 8 February 2021
Received in revised form 10 June 2021
Accepted 30 June 2021
Available online 20 July 2021

Keywords:
Amplitude modulation
Benchmark characterisation
Automated detection
Wind farm noise

ABSTRACT

Amplitude modulation (AM) is a characteristic feature of wind farm noise and has the potential to contribute to annoyance and sleep disturbance. Detection, quantification and characterisation of AM is relevant for regulatory bodies that seek to reduce adverse impacts of wind farm noise and for researchers and wind farm developers that aim to understand and account for this phenomenon. We here present an approach to detect and characterise AM in a comprehensive and long-term wind farm noise data set using human scoring. We established benchmark AM characteristics, which are important for validation and calibration of results obtained using automated methods. We further proposed an advanced AM detection method, which has a predictive power close to the practical limit set by human scoring. Human-based approaches should be considered as benchmark methods for characterising and detecting unique noise features.

1. Introduction

Amplitude modulation (AM) of wind farm noise (WFN) is a unique feature known to contribute to annoyance [1–3] and possibly sleep disturbance [4–6]. AM in the context of WFN is defined as a periodic variation in sound pressure level (SPL) at the blade-pass frequency [7,8], 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 [9–11]. AM is a highly variable phenomenon, depending on meteorological conditions [12,11,13], distance from the wind farm and wind farm operating conditions [9], making AM challenging to detect. Subsequently, characterising AM also becomes a challenging task because it depends on the performance of AM detectors.

Despite the difficulty in detecting AM, this noise phenomenon is commonly detected using simple engineering methods [8] using specific noise features (single predictors). For example, one of the first frequency domain-based methods, as proposed by Lundmark [14] detected and quantified AM using the AM spectrum of the time variation of instantaneous SPLs. To detect AM in field measurements of wind farm noise, this method was extended by specifying additional criteria such as a valid spectral peak frequency range of 0.6–1.0 Hz [12], and critical values of the maximum spectral peak of 0.4 dB [12] or 0.6 dB [11]. Time domain-based methods typically detect AM using SPL variations, where AM is classified as the difference between the 5th and 95th percentile of SPL greater than 2 dB [15] or as a peak-to-trough difference of 3 dB or 5–6 dB [16,17]. Recently, the UK Institute of Acoustics has developed a hybrid method [7], which is a combination of time and frequency domain methods. This method uses the prominence ratio, a ratio of the peak and masking noise levels, 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 [9,11,12]. However, evaluation of the performance of these methods is currently limited to false positive rates alone, or to small data sets [7,12,16] or is lacking altogether [15,18].

Detection and quantification of AM using automated detectors has been adopted in many previous studies [10,12,13,9,11]. This approach is practical and efficient as the analysis of AM is usually implemented on large data sets. In fact, using automated detectors, several unique AM features can be identified and possible associations between weather conditions [12], wind farm operation conditions, distances to wind farms [9], and the diurnal and seasonal variation of AM [11,13] can be identified. However, the above

[*] Corresponding author.
E-mail address: ducphuc.nguyen@flinders.edu.au (P.D. Nguyen).

https://doi.org/10.1016/j.apacoust.2021.108286
0003-682X/© 2021 Elsevier Ltd. All rights reserved.

AM characteristics and associated variables have been identified based on the assumption that the performance of currently available AM detectors is reasonable.

Human detectors are usually considered as a benchmark (or gold-standard) method for classification tasks which require unique skills to detect target features [19]. Although this approach is likely impractical to use for detecting AM in year-long data sets, it has some merits. A small subset can be extracted from a large data set using statistical sampling methods [20]. If AM samples in this subset are identified by skilled scorers, this information can be used to detect and quantify AM in the large data set. Additionally, this human-scored subset is useful for developing advanced AM detectors such as machine learning methods. In fact, machine learning methods are emerging in many acoustical applications [21] such as noise predictions [22,23], sound propagation [24] and noise source classification [25,26]. These methods allow for the 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 the reliance on a single noise feature, which poorly accounts for the highly variable and multifaceted phenomenon of AM [8].

The aims of this study were twofold: (1) to establish benchmark characteristics of AM based on the results of expert human detectors, and (2) to develop an advanced AM detection method based on the benchmark data set. To create the benchmark data set, 6,000 10-s audio files were randomly extracted from a database including 1 year measurements at two residences located near different wind farms. AM samples in this subset were then identified by a single scorer using a listening experiment under controlled conditions. Subsequently, the benchmark AM characteristics were established and compared with previous published findings. Finally, using the above benchmark data set, an advanced AM detection method was developed which is based on the random forest classification algorithm [27]. Three widely-used AM detection methods [12,15,7] were also evaluated. In particular, this study demonstrates a promising method to reliably establish AM characteristics. Also, the advanced method described in this paper, which is based on a state-of-the-art algorithm, outperformed current methods and is effective for exploration of large wind farm noise data sets.

2. Methods

2.1. Overview of study region and data collection

The acoustical data sets used in our study were measured at four residences (H1–H4) located 980 m (H1), 1.3 km (H2), 3.5 km (H3) and 30 km (H4) from the nearest wind turbine of South Australian wind farms (Fig. 1). These distances are relevant to wind farms in Australia where residences usually located greater than 1 km from wind farms. Residence H4 was unoccupied and located far away from wind farms, 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.

The data measured at H1 and H2 were used for establishing benchmark AM characteristics as well as training and validating the AM detection algorithm. The data measured at H3 and H4 were used for false positive rate validation of the proposed AM detection method and previously published methods. The characteristics of wind farms at the time of measurements are shown in Table 1.

A typical measurement setup included a microphone that was positioned at 1.5 m above ground level (except H1 where a ground level microphone was used) and protected using a secondary windshield with a diameter of 450 mm (See Hansen et al. [28] for details). The microphone was typically positioned at least 10 m away from the residence and surrounding vegetation to minimize façade reflections and wind-induced vegetation noise. 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 [9,28].

2.2. Benchmark data set generation

One benchmark data set contained 6,000 10-s audio files of WFN and the second one of equal size contained no WFN (environmental background noise only). The first data set was used for establishing benchmark AM characteristics and developing the AM detection method, while the latter data set was specifically constructed for testing false positive detection. These data sets were selected randomly from recorded data using the resampling without replacement technique (i.e., each 10-min sample has only one chance to be selected in the data set) (See Supplementary Fig. S1 for details).

The WFN benchmark data set was primarily scored by a single scorer using a validated rating experiment procedure based on detection theory [29]. The scorer was an acoustician experienced with wind farm noise AM through both field measurements and listening tests. The scorer also was familiar with AM characteristics in the time and frequency domains. Acoustician scorers familiar with the acoustic features of AM were selected to avoid potential confounding and bias by other acoustic and non-acoustic features unrelated to AM through the use of non-acoustician scorers. Intra-scorer variability was validated in which the scorer re-scored a sub-set of the data (100 samples) in a blinded manner. To further evaluate inter-scorer agreement, another skilled scorer also rated a sub-sample of 100 randomly chosen audio samples. These scorers 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’). For this particular AM identification task, the modulated frequency and duration of AM presence were not identified by the scorer. A MATLAB GUI was designed for the experiment as shown in Fig. 2. To maximise the performance of detection task, the scorers were allowed to adjust headphone volume level and to listen the audio multiple times before rating. Therefore, AM samples, regardless of their audibility, were detected by the scorer. The visual characteristics of AM were also presented to the scorers, as shown in Fig. 2. This additional information was expected to further improve the scorer’s AM detection performance. The rating experiment was performed in a bedroom at the Adelaide Institute for Sleep Health. The noise reproduction system consisted of Bose Quite Comfort II headphones and a RME Babyface Pro sound card. The background noise in the headphone cavity was approximately 22 dBA during the experiment.

2.3. Automated AM detectors

The proposed AM detection method was compared against three previously published AM detection methods. The first method, labelled a1 [7], uses a “hybrid” approach involving analysis in both the time- and frequency-domains. The other two methods labelled a2 [12] and a3 [15] are implemented in the frequency- and time-domains, respectively. To make these

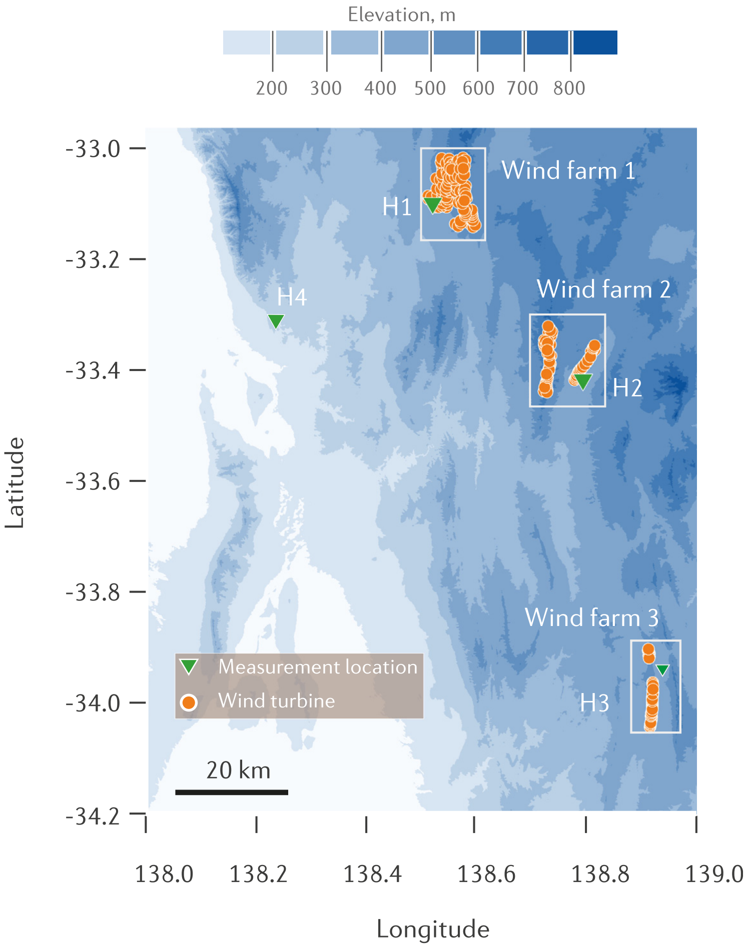


Fig. 1. (Color online). Study region.

methods consistent, all methods were implemented using audio samples with a 10 s period and a fast time weighting of 100 ms.

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 ratio (*PR*), the ratio between the spectral peak in the blade-pass frequency range and the noise floor, is greater than four [7].

Method a2 is implemented by first 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 [12].

Table 1
Wind farm characteristics.

Name	Wind farm 1	Wind farm 2	Wind farm3
Nominal capacity (MW)	315	148	131
Turbine size (MW)	3.2	2.1	(3.0 & 3.3)
Type	Siemens	Suzlon	Vestas
Number of turbines	99	70	37
Wind farm latitude	−33.058	−33.367	−33.983
Wind farm longitude	138.544	138.728	138.900
Annual output (mean ± s.d.) (MW)	125 ± 97	54 ± 44	45 ± 38

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 the detrended SPL via calculating the difference between statistical noise levels L_{95} and L_5, this value, referred to as *DAM*, is used as a threshold for detecting AM. The suggested threshold varies from 2 dB to 6 dB [15–17]. 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 [30] was re-implemented using MATLAB in our study (Supplementary Fig. S2).

2.4. Random forest classifier for AM detection

A random forest classifier [27] 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 [20]. To overcome these decision tree problems, the random forest classifier uses bootstrap sampling and random variable selection to build multiple trees, which are then combined into a random forest classifier as shown in Fig. 3. 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

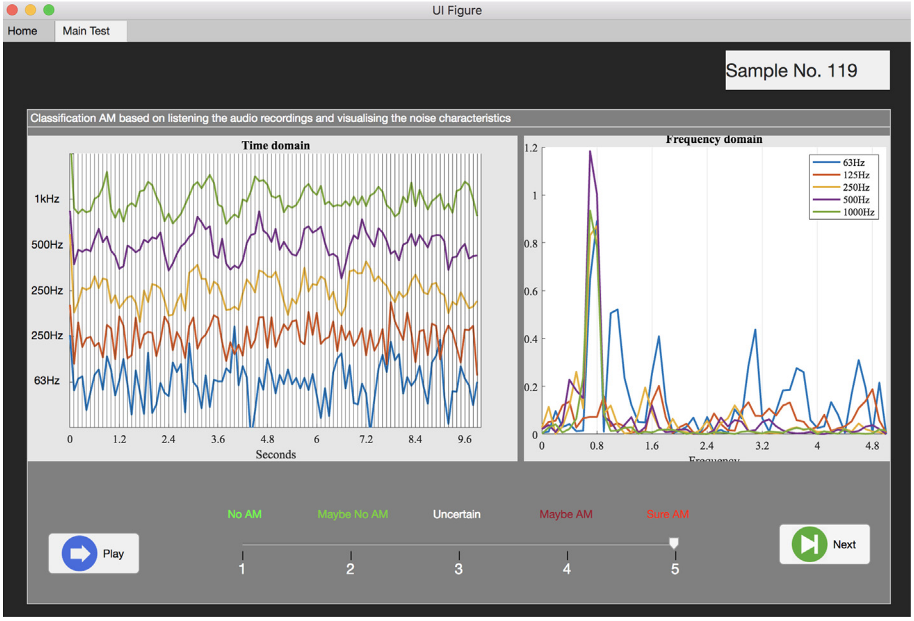


Fig. 2. (Color online). MATLAB GUI for rating the presence of amplitude modulation in the audio files.

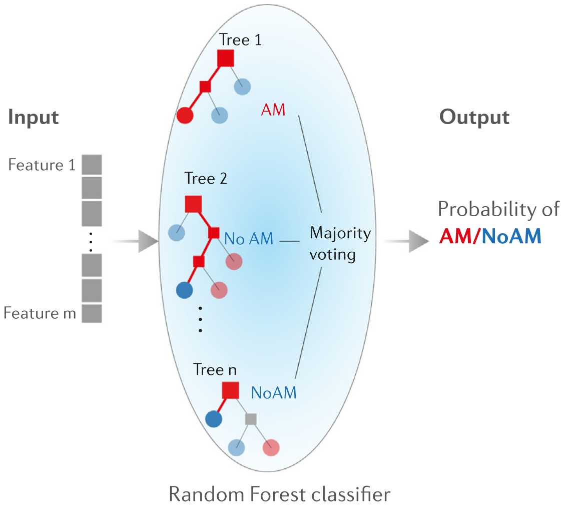


Fig. 3. (Color online). Random forest classifier.

number of trees voting “AM” out of the total tree population represents the probability of AM being present.

Optimisation of hyperparameters, that is parameters which are set before the learning begins, was done using a random searching technique [31]. 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 number of data points allowed in a leaf node. The random searching technique utilises a range of realistic hyperparameter values, as shown in Table 2.

Table 2
Value ranges of the hyperparameters used for random searching.

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

2.5. Audio feature extraction

WFN spectra are dominated by lower-frequencies, particularly at distances greater than 1 km from a wind farm [8]. Also, WFN can contain both tonal AM [9] 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 [12]. 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 as shown in Table 3 (See Supplementary Table S1 for full feature name details). The noise features were divided into four categories, including frequency domain features, overall noise features, time domain features and features extracted from the other automated AM detection methods described in Section 2.3.

The frequency domain feature categories (feature 1 to feature 13) have been explained in detail in previous reviews [33,32] and the pseudo code for extracting these features can also be found in [33]. Fig. 4A shows the process to extract these audio features. A hamming window of 125 ms (50% overlap) is applied to the input

Table 3
Feature descriptions.

Feature No.	Category	Description	Ref.
1–13	Frequency domain features	These features describe properties of the noise frequency content such as spectrum balance, spectrum shape and tonality.	[32,33]
14–17	Overall noise features	Overall A, C, G-weighted SPLs and its related features	[34]
18–27	Time domain features	Including features extracted from fast-time A-weighted, unweighted and octave-band unweighted SPLs.	Proposed
28–31	Published methods	*PR* (Prominence ratio)	[7]
		Fo (Fundamental frequency)	
		AMfactor	[12]
		DAM	[15]

signals which are then transformed to the frequency domain using an FFT. The signals are then filtered using bark scale critical bands and the spectral shape features are calculated for each hamming window. The outcome of the process is a matrix (No. of features x No. of windows). The mean values of the rows in this matrix were calculated, resulting in a single value for each feature. The overall noise feature category (feature 14 to feature 17) such as A, C and G-weighted SPLs were also extracted as shown in Fig. 4B. The selected features were $L_{Geq}/L_{Aeq}, L_{Ceq}/L_{Aeq}$, and $L_{Ceq} - L_{Aeq}$, as these measures are expected to be indicative of WFN presence and spectral balance [35,36,8]. The L_{Aeq} was selected as it has been used as a metric for analyzing AM in previous studies [12,37]. The time-domain feature category (Feature 18 to feature 27) was extracted as shown in Fig. 4C. The fast-time weighted SPL (125 ms overlapping 100 ms) was calculated, similar to the method for calculating the prominence of impulsive sounds outlined in Nordtest [18]. The derived SPL (40 Hz sampling frequency) was further smoothed using a moving average window of 5 samples. To estimate AM fundamental frequency of the smoothed SPL, the first derivative of the smoothed SPL was calculated and then transformed to the frequency domain. The highest peak (Feature 19) and its corresponding frequency (Feature 18) of the derivative in the frequency domain were obtained. Also, the average ramp-up and ramp-down of SPL were estimated by calculating mean values of positive and negative values of the derivative signals (Feature 20 and 21, respectively). Using the derivative signals is advantageous because the fluctuation frequency of the derivative signal is similar to the smoothed SPL, while its amplitude is less variant compared to the smoothed SPL. As a result, the blade-pass frequency peaks were clearer in the frequency domain. Feature 22 was calculated in a similar way to feature 18, except using the unweighted SPL. Features 23–27 are variations (calculated as $L_5 - L_{95}$) of the octave-band unweighted SPL for octave-band centre frequencies between 63 Hz to 1000 Hz. The automated methods (a1, a2 and a3) were also used as noise features (Feature 28 to 31).

2.6. 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 suited to imbalanced data sets [38]. To construct the PR curve, pairs (*precision*, *recall*) were calculated from the counts of true positives (*TP*), true negatives (*TN*), false positives (*FP*) and false negatives (*FN*) as follows

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

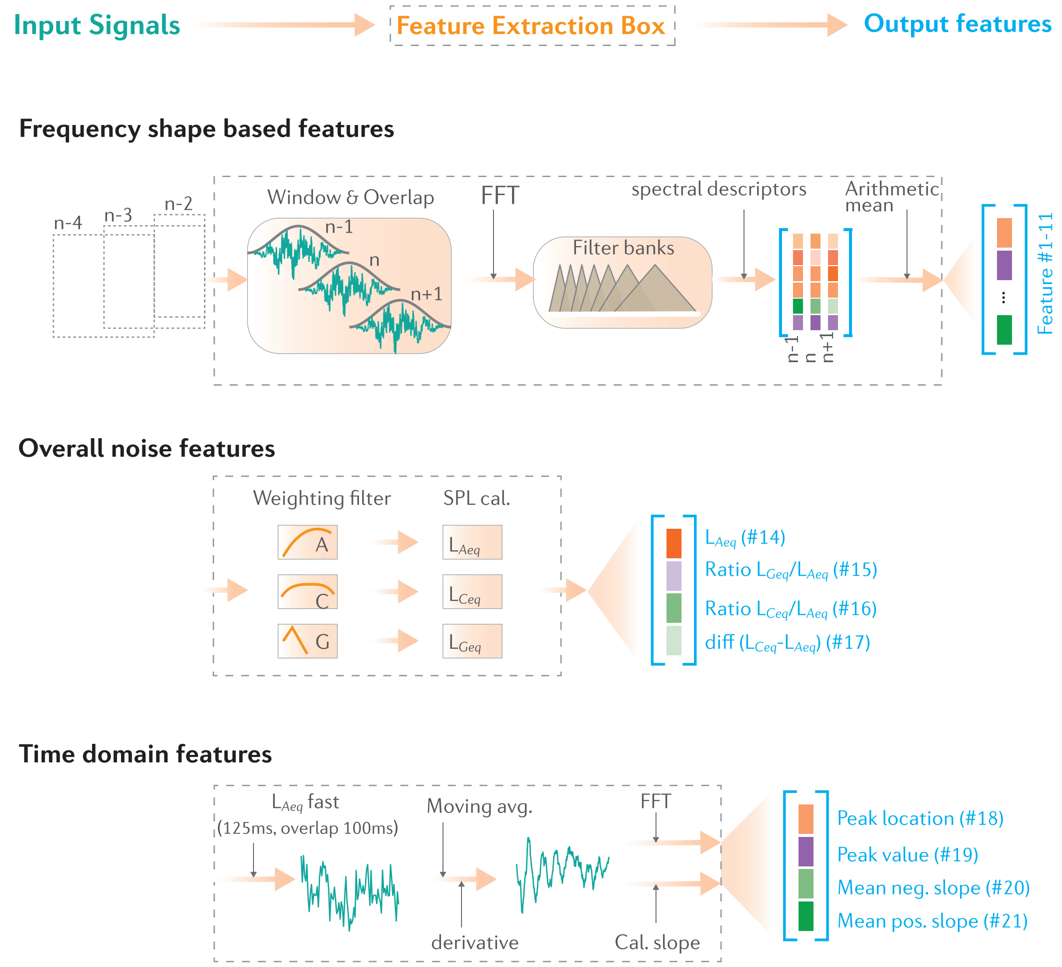


Fig. 4. (Color online). Feature extraction.

The aggregate metric of the *MCC* is a more informative and faithful score of overall classification performance compared to common metrics such as the accuracy or *F*1-score [39]. The *MCC* ranges from −1 (classification is always wrong) to 0 (classification is no better than random guessing) 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 including Cohen's kappa, accuracy, area under ROC curve, etc., [38], were also calculated as secondary results (Supplementary Table. S2).

2.7. Benchmark AM characterisation

The diurnal and seasonal variation of AM prevalence were compared against previously published AM characteristics obtained using WebPlotDigitizer (https://automeris.io/WebPlotDigitizer/) [40]. Specifically, diurnal variation of AM prevalence was extracted from Figures 7 and 8 in [10], Figure 12e in [9] and a mean value of the data in Figures 4 a, c and e in [11]. The seasonal variation data were extracted from Figure 3 ($AM_{0.4}$) [11] and Table 1 [13].

2.8. Data and statistical analysis

Audio signal 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. Statistical analysis and visualisation were implemented in R (https://www.r-project.org). The statistical significance threshold used was $\alpha = 0.05$. All data are reported as mean [95 % confidence interval], unless otherwise indicated. The 95% CI range of performance metrics was estimated using a bootstrapping method with 2,000 simulations (See Supplementary Fig. S4 and Supplementary Algorithm 4 for details). Pearson correlation coefficients were used to examine the strength of linear relationships between features and AM quantification metrics.

2.9. 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.

3. Results

3.1. Benchmark data set

The benchmark data set of 6,000 10-s audio files was unbalanced with around 40% of audio samples containing AM (Fig. 5A). The AM confidence rating was transformed into a binary score (AM vs. no AM) using a confidence rating threshold of three. Samples with 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*, Cohen's kappa coefficient (κ) and *F*1-score for inter-scorer agreement were (0.65 [0.49, 0.80], 0.64 [0.48, 0.8] and 0.77 [0.66, 0.87], indicating a high degree of agreement [19] (See Supplementary Table. S3 for other metrics). Also, intra-scorer agreement was higher than inter-scorer agreement (MCC = 0.71 [0.56, 0.85], κ = 0.7 [0.56, 0.85], *F*1-score = 0.82 [0.71, 0.91]; See supplementary Table S4 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. 5B).

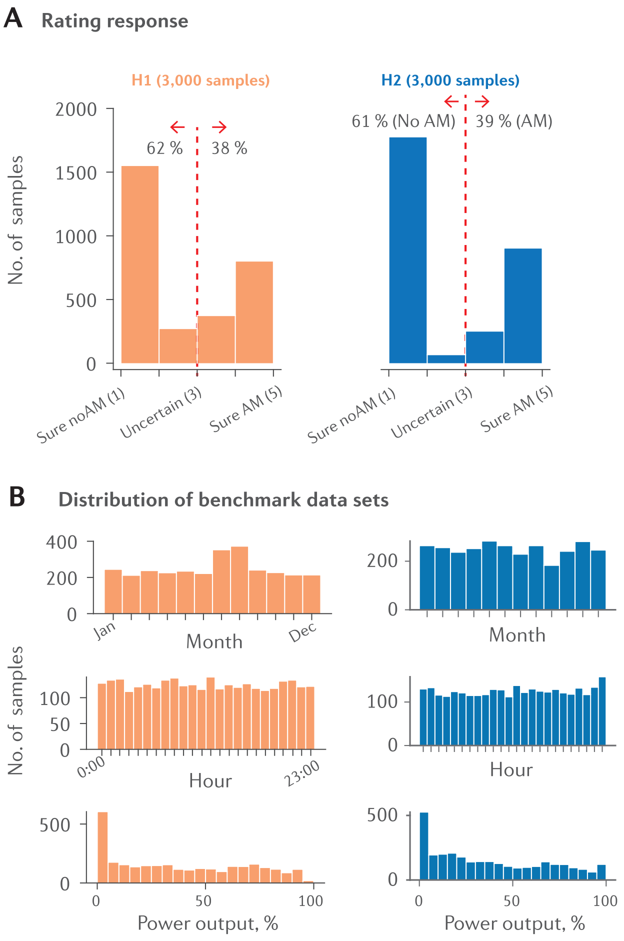


Fig. 5. (Color online). Characteristics of benchmark data sets. **A**, scorer ratings distribution with corresponding binary classification. **B**, distributions of audio files per month, hour and wind farm power percentage output relative to capacity.

3.2. Benchmark AM characteristics

At the residential locations investigated, which were approximately 1 km from the nearest wind turbine, approximately 90% of AM samples in the benchmark data set had an associated A-weighted SPL between 30 and 50 dBA (Fig. 6A). This supports the feasibility of using a threshold of 30 dBA to trigger AM analysis [12], at least for data recorded at similar distances from the wind farm. This could thus reduce false positive rates and/or exclude samples with low SPLs which are likely to be less relevant for assessing community annoyance. We noted that our results can be considered as an upper bound of AM prevalence as both audible and inaudible AM samples were quantified. The audible AM is more relevant to human response to the noise such as annoyance response. The prevalence of audible AM can be determined using the approach proposed by Hansen et al. [9] by considering the normal hearing threshold curve.

There are three common metrics (i.e., *AMdepth*, *AMfactor* and *DAM*) to quantify the strength of SPL variations (See Methods and Supplementary Algorithm 1–3 for calculation details). The magnitude of AM hereafter is referred to as the AM depth,

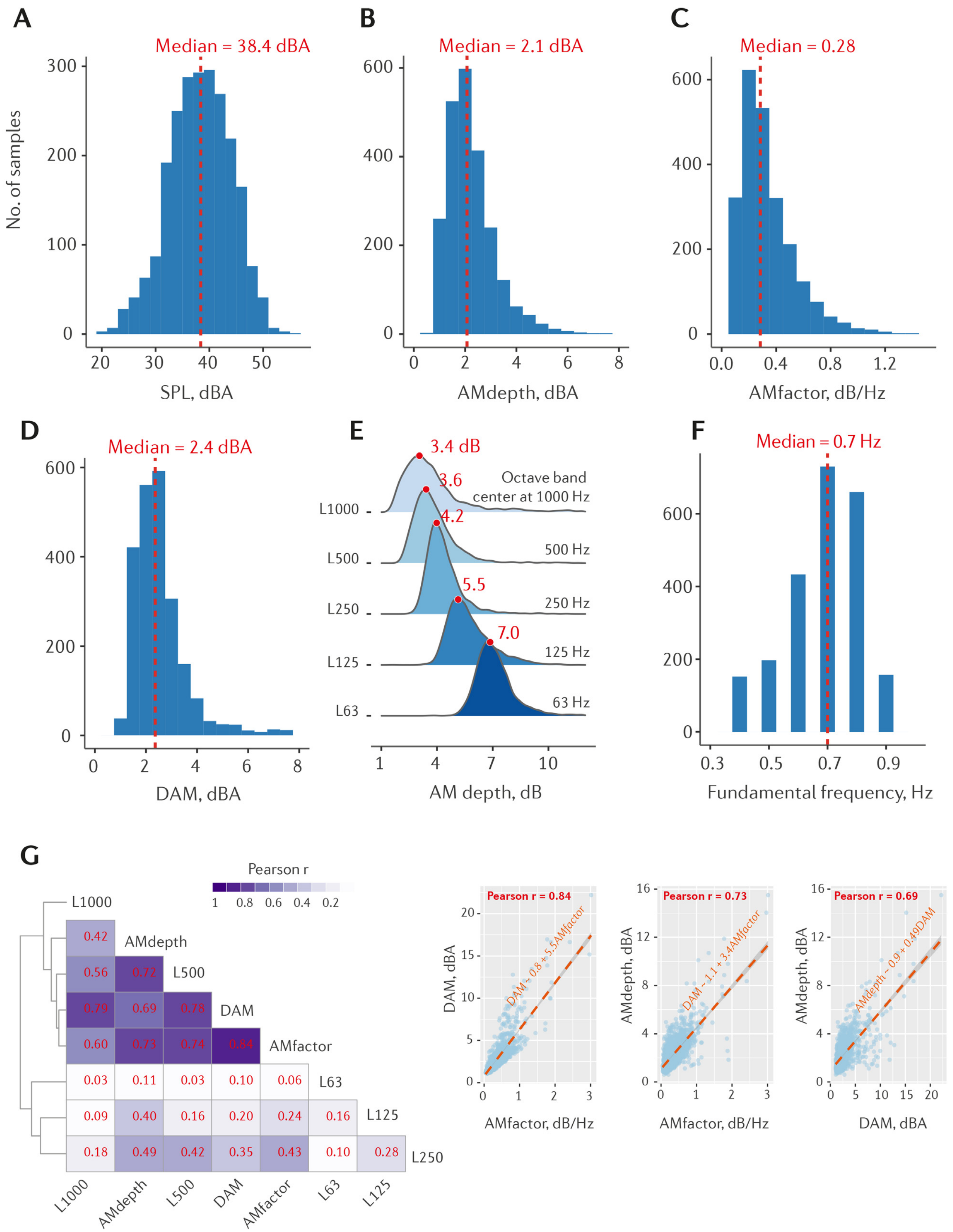


Fig. 6. (Color online). Characteristics of 2,329 AM samples in the benchmark data set. **A**, A-weighted SPL. **B-D**, AM depth as quantified using three common metrics. **E**, Distribution of AM depth in each octave band. **F**, Fundamental frequency of AM. **G**, Correlation between AM depth metrics.

despite differences in AM depth values obtained using each metric. The distributions of AM depth as quantified by the three metrics are shown in Fig. 6B–D. More than 50% of the AM samples had an AM depth greater than 2 dBA using the *AMdepth* and *DAM* metrics, which is the fluctuation sensation threshold [41]. All three above metrics evaluated AM depth using A-weighted

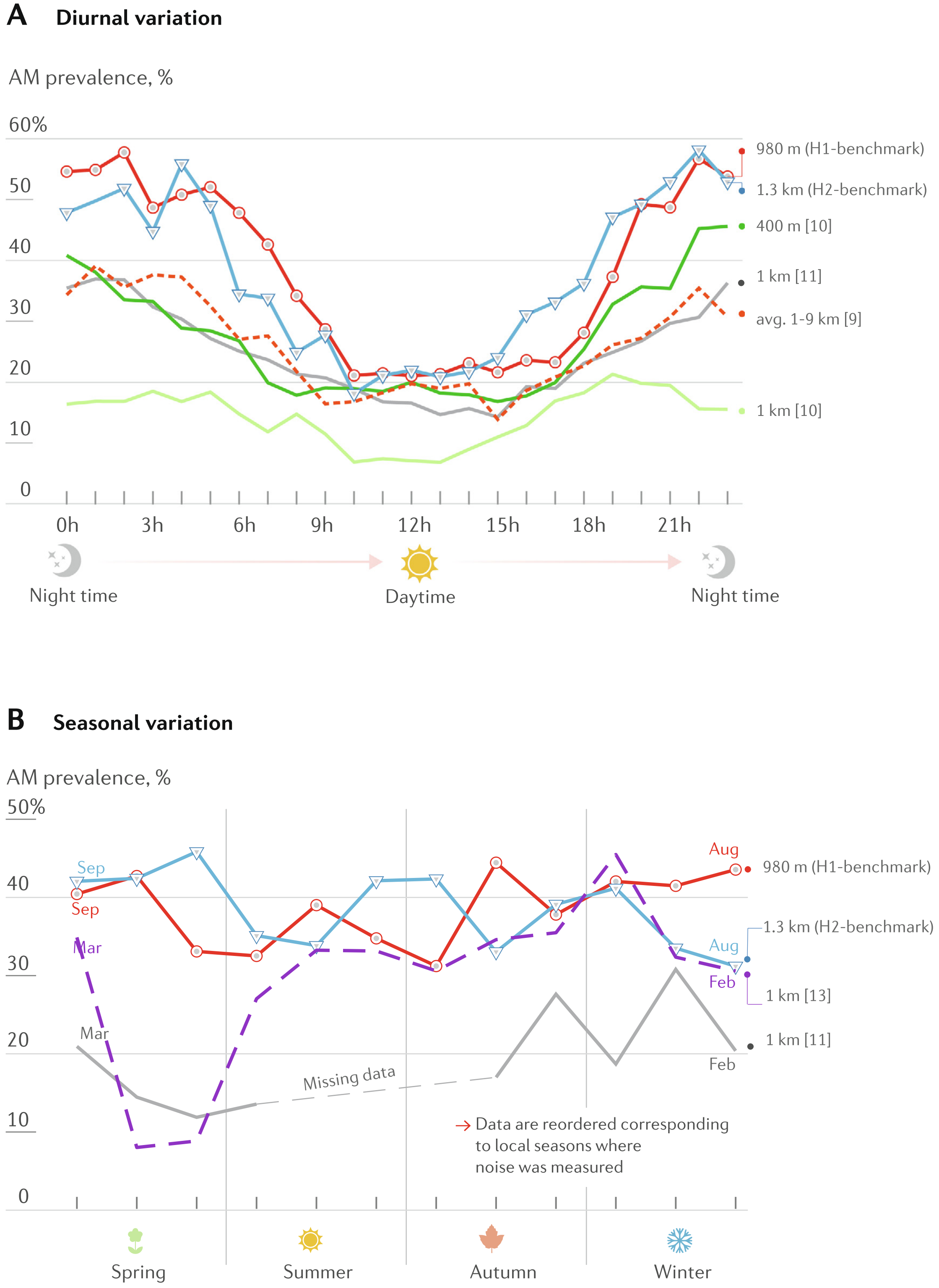


Fig. 7. (Color online). Variation of AM prevalence. **A**, diurnal variation. **B**, Seasonal variation. Previously published data are from Australia [9], Sweden [10,11] and Finland [13].

SPLs, resulting in underestimation of the AM depth occurring at low frequencies. The distributions of AM depth as quantified in each octave band from 63 Hz to 1000 Hz are shown in Fig. 6E, where it can be seen that the AM depth increased for low-frequency bands. The modulation frequency was dominant between 0.6 Hz and 0.8 Hz, accounting for approximately 80% of AM samples. This frequency corresponds to the expected blade-pass frequency when the wind turbines are operating at their nominal speed of 14 to 16 rpm.

The AM depth is one of the most important characteristics of AM, as its magnitude is directly related to the levels of annoyance. Thus, to further characterise AM depth, the Pearson correlation

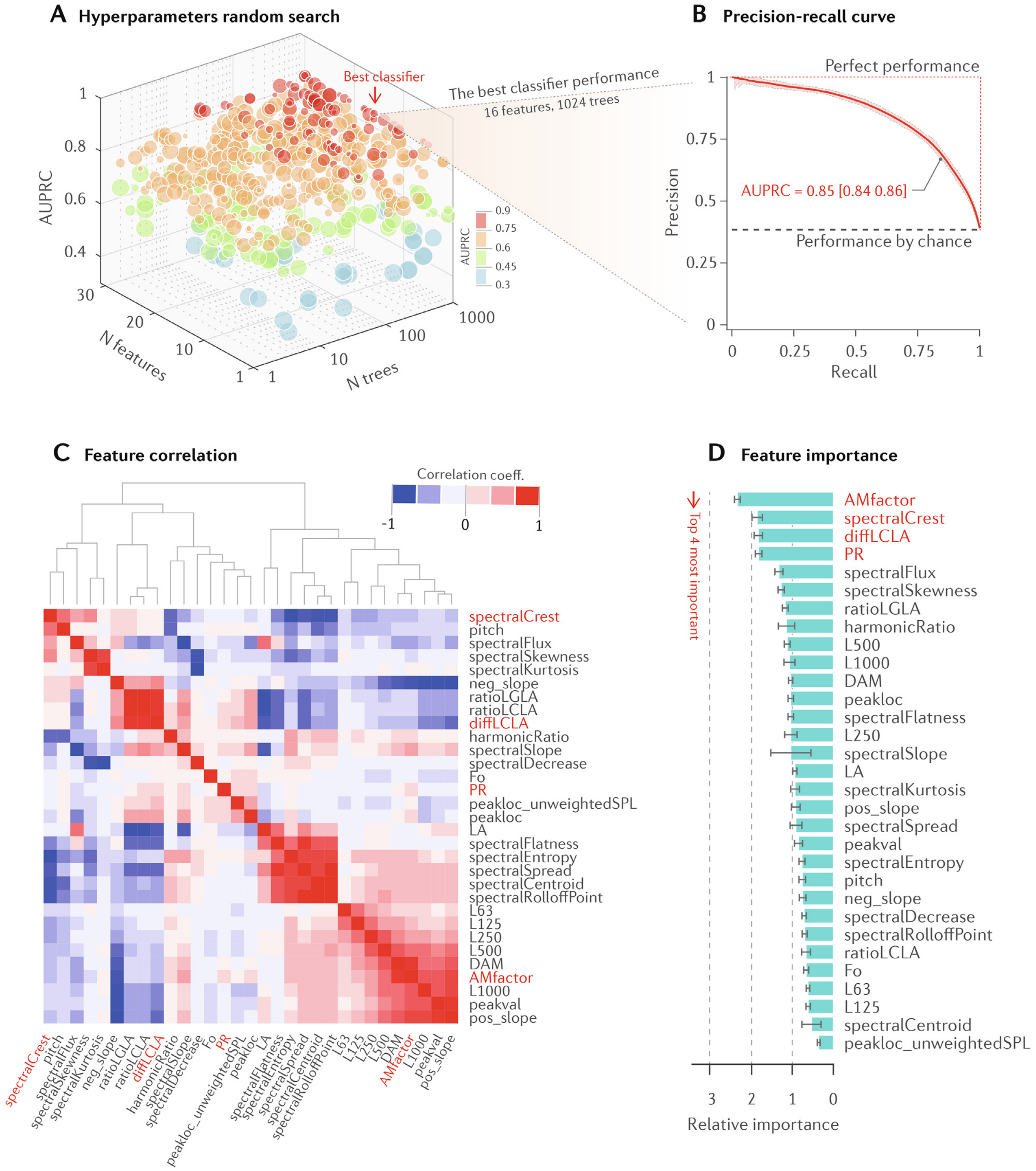


Fig. 8. (Color online). Random Forest classifier. **A**, hyperparameter tuning using a randomized search technique. The size of the circles represents the maximum splits. Minimum leaf node samples are not shown. B, the precision-recall curve of the best random forest classifier. The shaded area indicates 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.

coefficients (Pearson's r) between pair metrics are shown in Fig. 6G. Two clusters were observed from pairs as shown in the dendrogram. The first cluster included the three above metrics with AM depth quantified for mid- to high-frequency bands (i.e., 500 and 1000 Hz). The second cluster included the metrics used to quantify AM depth for low-frequency bands (i.e., 63, 125 and 250 Hz). Additionally, a linear relationship between three common metrics is shown in 6G on the left. A strong correlation between these metrics was observed, especially between the *AMfactor*, *DAM* pair, followed by the *AMfactor*, *AMdepth* and *DAM*, *AMdepth* pairs, respectively.

3.3. Diurnal and seasonal AM variation

AM appeared to be more prevalent during the evening and night (Fig. 7A). Previous studies [9,12,11] showed that AM occurs approximately 20% to 40% of the nighttime (defined on the basis of 22:00 to 6:00) and around 20% during the daytime. Amplitude modulation was detected using method a1 (for the study of [9]) and method a2 (for the studies of [10,11,13]). From the comparison of daytime and nighttime, it appears that although the automated detection methods can capture a general pattern of diurnal variation, AM prevalence was lower compared with the benchmark data set, especially during the nighttime. Note that AM prevalence is also substantially affected by the difference in meteorological conditions, distance to wind farms, geographical conditions and wind farm layout. On the other hand, seasonal variation is likely to have a negligible effect on AM prevalence, as shown in Fig. 7B.

3.4. Random forest-based AM detection

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

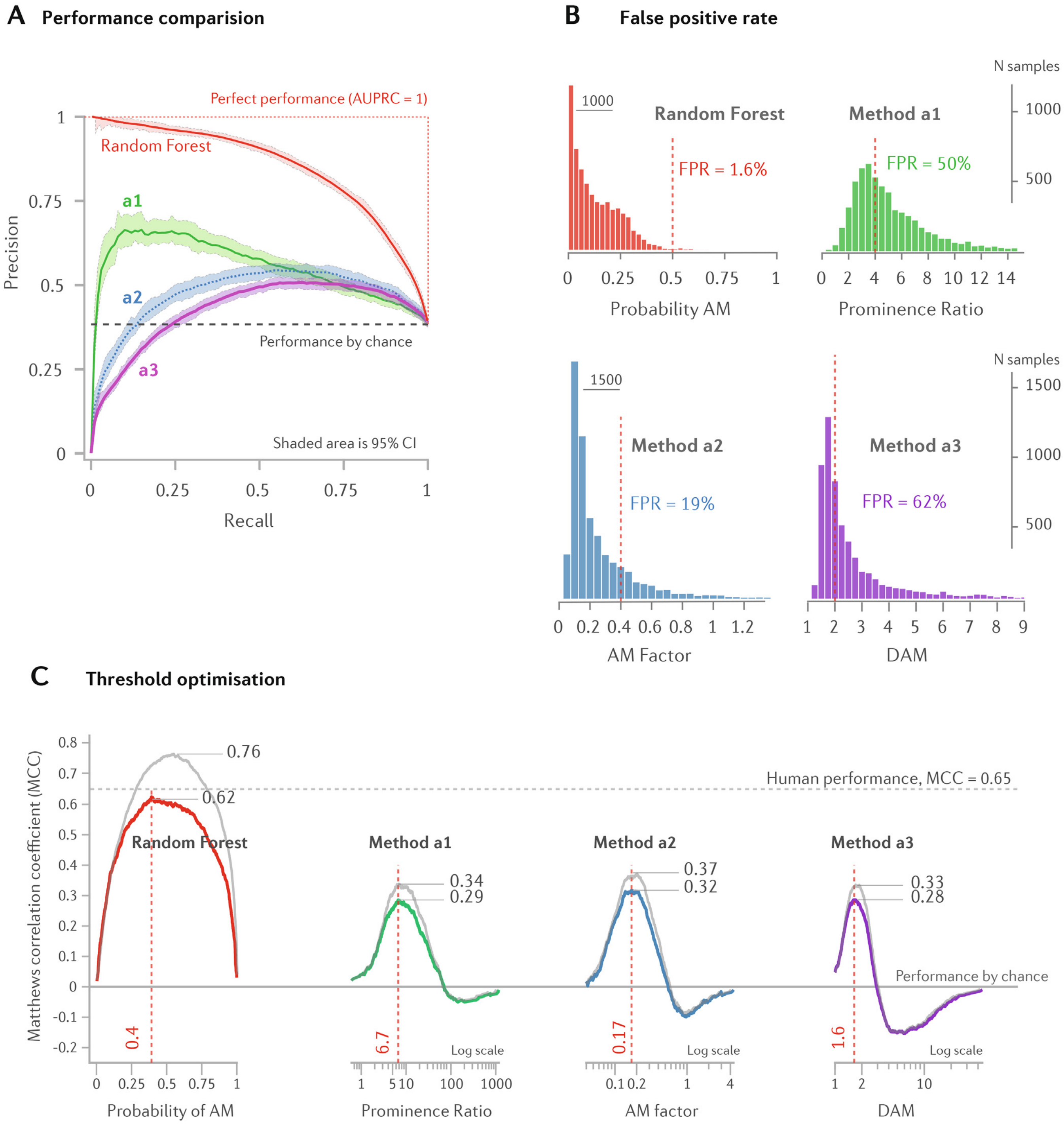


Fig. 9. (Color online). Performance of automated detectors. **A**, performance using the benchmark data set, where the values associated with each curve are mean [95% confidence interval]. The shaded area is the 95% CI. **B**, false positive rate of each detection method estimated from the no wind farm 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.

curve (*AUPRC*), [42] (Fig. 8A). 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. 8B 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. 8C. The four most important features for predicting AM are *AMfactor*, *SpectralCrest*, *diffLCLA* and *PR* (Fig. 8D).

3.5. 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. 9A). 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 performance of a1–a3 was poor with the mean *AUPRC* ranging from 0.43 to 0.55 (Table 4). The performance of a1 was better than a2 and a3 (all $P < 0.001$), and a2 performed better than a3 ($P < 0.001$).

The performance of AM detection algorithms has previously been described in terms of the false positive rate (*FPR*) [12,7], and thus this metric was also examined (Fig. 9B). 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 [7,15], but was reported to be 2.6% for method a2 [12], 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. 9C. 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 [7] and the value obtained using a Receiver Operating Characteristic curve (PR = 3) in [9]. In contrast, the optimal thresholds for method a2 and a3 were lower than the original suggested values [12,15]. 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. We further investigated if the performance of automated methods could be improved when using only samples corresponding with certain responses of the scorer (i.e., sure AM with responses > 4.5 vs sure no AM with responses < 1.5). The performance of all automated methods increased, especially the Random Forest based method, which showed an approximately 22% increase in performance (MCC = 0.76, Fig. 9C). This was expected as clearer AM or no AM events were likely detected with higher confidence.

Table 4
Area under the precision-recall curves and optimal MCC for the 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

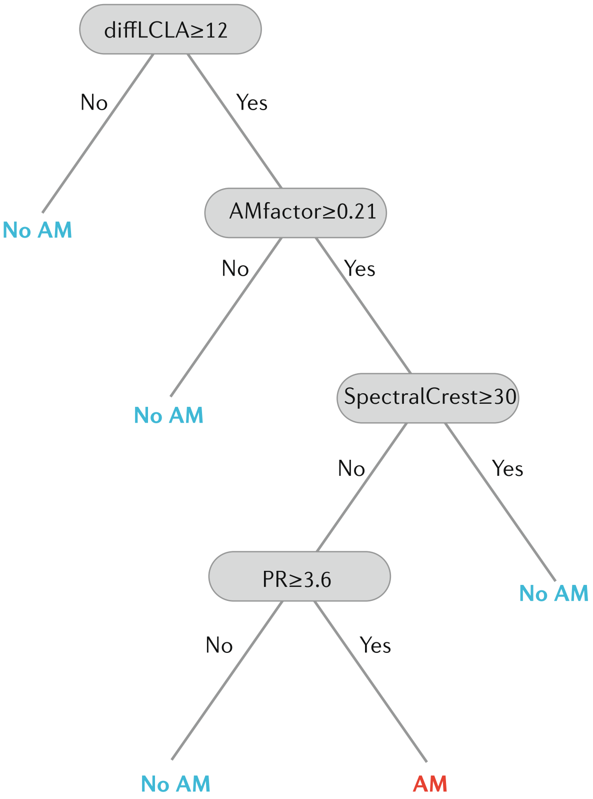


Fig. 10. (Color online). A simplified single tree classifier utilising the four most important features identified by the random forest classifier for AM detection.

3.6. 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, making them difficult to interpret. The classifier also requires skilled human and computer resources to implement. Given the findings of the importance of *AMfactor*, *diffLCLA*, *SpectralCrest* and *PR* features, this study 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 decision tree built from four features, as shown in Fig. 10. 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.

4. Discussion

In summary, we presented a new and promising approach to characterise AM in a large data set using an expert human scoring method. The resulting estimates of benchmark AM characteristics such as AM depth, frequency, and diurnal and seasonal variations are important for validation and calibration of the results using automated methods. We further show that it is possible to develop an advanced AM detection method with a predictive power close to the practical limit set by human scoring. 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 wind farm noise regulations.

Although AM identification by humans was a benchmark approach to establish high quality scored data, it is striking to find that an advanced machine learning algorithm performed close to the human limit. In fact, 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 human 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 of the 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 the idea that human perception of AM is more complex than assumed by previous AM detection methods that 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 [12,7]. However, it is worth noting that method a1 was originally designed and evaluated on 10-min samples, as opposed to the 10-s samples used in our work, and method a1 classifies AM if more than 50% of 10-s blocks within 10 min contain AM. By introducing the above criterion, the false positive rate may be substantially reduced, as reported in [7]. However, 10-s long samples appear to have higher validity, as typical AM events usually last around 10–15 s [12]. 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% [12] 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 [19], 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. As a result, the benchmark AM characteristics are not relevant at other distances. The proposed classifier also may not work well for detecting AM measured several kilometers from the nearest wind turbine, where AM may have different characteristics [9]. The classifier could not be tested on data sets measured outside of South Australia, where weather conditions and topography near wind farms will inevitably vary. Although the reliability of human scoring has been tested, using a single scorer to classify the AM is not ideal. Human scoring is a subjective process, for which intra- and inter-scorer variability should be expected [43]. We used a single scorer to identify the presence of AM to minimise inter-scorer variability effects which are typically higher than intra-scorer variability. Nevertheless, it remains unclear how generalizable these findings may be to AM more broadly, for which inter-scorer differences as well as noise source and climatic effects could be important. As suggested by Wendt et al. [43], two or more scorers and a consensus scoring approach may be preferable to a single scorer to help ensure broader generalisability. Future studies should examine if residents living near wind farms identify AM similarly to acoustician and algorithm scored AM, and how strongly AM identification ratings are related to annoyance ratings. Nevertheless, a single scorer is more practical and avoids the 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 [30] 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. [44]. 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.

5. Conclusions

In conclusion, this study demonstrated that human scoring is a feasible and promising approach to identify AM. This approach is invaluable for detecting unique characteristics of wind farm noise in cases where the performance of automated detectors is low or not validated. The advanced AM detector based on the random forest approach demonstrated high performance, and substantially outperformed traditional AM detection methods to achieve a 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 achieved good classification performance. This approach is readily interpretable and easy to implement without the need for extensive computer resources. We hope that, in the future, further insight into the prevalence of AM and associated meteorological conditions, and impacts on humans will help to explain underlying noise generation mechanisms relevant to human perception. Ultimately, this will improve the design of wind turbines such that they are less disturbing and hence, more acceptable to surrounding communities.

CRediT authorship contribution statement

Phuc D. Nguyen: Conceptualization, Methodology, Data curation, Writing - original draft, Formal analysis, Visualization, Writing - review & editing. **Kristy L. Hansen:** Conceptualization, Methodology, Data curation, Supervision, Writing - review & editing. **Bastien Lechat:** Supervision, Writing - review & editing. **Peter Catcheside:** Supervision, Writing - review & editing. **Branko Zajamsek:** Supervision, Writing - review & editing. **Colin H. Hansen:** Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

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.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.apacoust.2021.108286.

References

[1] Lee S, Kim K, Choi W, Lee S. Annoyance caused by amplitude modulation of wind turbine noise. Noise Contr Eng J 2011;59:38. https://doi.org/10.3397/1.3531797.

[2] Schäffer B, Schlittmeier SJ, Pieren R, Heutschi K, Brink M, Graf R, Hellbrück J. Short-term annoyance reactions to stationary and time-varying wind turbine and road traffic noise: A laboratory study. J Acoust Soc Am 2016:139;2949–2963. URL: http://asa.scitation.org/doi/10.1121/1.4949566. doi: 10.1121/1.4949566.

[3] Ioannidou C, Santurette S, Jeong C-H. Effect of modulation depth, frequency, and intermittence on wind turbine noise annoyance. J Acoust Soc Am 2016:139;1241–1251. URL: http://asa.scitation.org/doi/10.1121/1.4944570. doi: 10.1121/1.4944570.

[4] Micic G, Zajamsek B, Lack tL, Hansen K, Doolan C, Hansen C, et al., A review of the potential impacts of wind farm noise on sleep. Acoust Austr (2018). URL: https://link.springer.com/content/pdf/10.1007%2Fs40857-017-0120-9.pdf. doi: 10.1007/s40857-017-0120-9.

[5] Bakker R, Pedersen E, van den Berg G, Stewart R, Lok W, Bouma J. Impact of wind turbine sound on annoyance, self-reported sleep disturbance and psychological distress. Sci Total Environ 2012;425:42–51.

[6] Liebich T, Lack L, Hansen K, Zajamšek B, Lovato N, Catcheside P, Micic G. A systematic review and meta-analysis of wind turbine noise effects on sleep using validated objective and subjective sleep assessments. J Sleep Res 2020: e13228.

[7] Bass J, Cand M, Coles D, Davis R, Irvine G, Leventhall G, et al., 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, Institute of Acoustics (2016).

[8] Hansen CH, Doolan CJ, Hansen KL. Wind farm noise: measurement, assessment and control. first ed. John Wiley & Sons Ltd; 2017.

[9] Hansen KL, Nguyen P, Zajamšek B, Catcheside P, Hansen CH. Prevalence of wind farm amplitude modulation at long-range residential locations. J Sound Vib 2019;455:136–49.

[10] Larsson C, Öhlund O. Variations of sound from wind turbines during different weather conditions. Inter Noise 2012;2012.

[11] Conrady K, Bolin K, Sjöblom A, Rutgersson A. Amplitude modulation of wind turbine sound in cold climates. Appl Acoust 2020;158:107024.

[12] Larsson C, Öhlund O. Amplitude modulation of sound from wind turbines under various meteorological conditions. J Acoust Soc Am 2014;135:67–73.

[13] Paulraj T, Välisuo P. Effect of wind speed and wind direction on amplitude modulation of wind turbine noise. In: INTER-NOISE and NOISE-CON congress and conference proceedings, vol. 255, Institute of Noise Control Engineering; 2017, p. 5479–89.

[14] Lundmark G. Measurement of swish noise: a new method. In: Fourth international meeting on wind turbine noise. Rome: Italy; 2011. p. 2011.

[15] Fukushima A, Yamamoto K, Uchida H, Sueoka S, Kobayashi T, Tachibana H. Study on the amplitude modulation of wind turbine noise: Part 1–physical investigation, in: Internoise 2013; 2013.

[16] Bass J. Investigation of the "den brook" amplitude modulation methodology for wind turbine noise. Acoust Bullet 2011;36:18–24.

[17] Cooper J, Evans T. Automated detection and analysis of amplitude modulation at a residence and wind turbine. In: Acoustics 2013, Victor Harbor, Australia; 2013.

[18] Nordtest, Nt-acou-112: Prominence of impulsive sounds and for adjustment of laeq; 2002.

[19] Warby SC, Wendt SL, Welinder P, Munk EG, Carrillo O, Sorensen HB, Jennum P, Peppard PE, Perona P, Mignot E. Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. Nat Meth 2014;11:385.

[20] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media; 2009.

[21] Bianco MJ, Gerstoft P, Traer J, Ozanich E, Roch MA, Gannot S, Deledalle C-A. Machine learning in acoustics: Theory and applications. J Acoust Soc Am 2019;146:3590–628.

[22] Valente D. Data-driven prediction of peak sound levels at long range using sparse, ground-level meteorological measurements and a random forest. J Acoust Soc Am 2013;134. 4159–4159.

[23] Iannace G, Ciaburro G, Trematerra A. Wind turbine noise prediction using random forest regression. Machines 2019;7:69.

[24] Hart CR, Reznicek NJ, Wilson DK, Pettit CL, Nykaza ET. Comparisons between physics-based, engineering, and statistical learning models for outdoor sound propagation. J Acoust Soc Am 2016;139:2640–55.

[25] Paulraj T, Välisuo P. A method to generate a database of source labelled environmental noise samples using open noise data and to quantify wind turbine noise in it., in. In: INTER-NOISE and NOISE-CON congress and conference proceedings, vol. 261, Institute of Noise Control Engineering. p. 2145–56.

[26] Välisuo PO. Automated wind turbine noise analysis by machine learning. In: INTER-NOISE and NOISE-CON congress and conference proceedings, vol. 255, Institute of Noise Control Engineering; 2017. p. 5667–78.

[27] Breiman L. Random forests. Mach Learn 2001;45:5–32.

[28] Hansen K, Zajamsek B, Hansen C. Identification of low frequency wind turbine noise using secondary windscreens of various geometries. Noise Contr Eng J 2014;62:69–82.

[29] Macmillan NA, Creelman CD. Detection theory: A user's guide. Psychology Press; 2004.

[30] Coles D, Bass HJ, Cand M, Ioa am code – implementation of the core routine for am analysis from the ioa amwg (2017). URL: https://sourceforge.net/projects/ioa-am-code/.

[31] Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J. Mach Learn Res 2012;13:281–305.

[32] Alías F, Socoró JC, Sevillano X. A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. Appl Sci 2016;6:143.

[33] Sharma G, Umapathy K, Krishnan S. Trends in audio signal feature extraction methods. Appl Acoust 2020;158:107020.

[34] Bies DA, Hansen C, Howard C. Engineering noise control. CRC Press; 2017.

[35] Kelley N. A proposed metric for assessing the potential of community annoyance from wind turbine low-frequency noise emissions, Technical Report. Golden, CO (USA): Solar Energy Research Inst; 1987.

[36] Y. Tokita, A. Oda, K. Shimizu, On the frequency weighting characteristics for evaluation of infra and low frequency noise, in: Proceedings of INTER-NOISE 84 and NOISE-CON 84, Institute of Noise Control Engineering, 1984, pp. 917–920.

[37] Developments R. Written scheme relating to Condition 21 Den Brook wind farm. Implementation of condition 20 for the identification of greater than expected amplitude modulation. Technical Report, RES Developments Ltd. 2014.

[38] Lever J, Krzywinski M, Altman N. Classification evaluation 2016.

[39] Chicco D, Jurman G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC Genomics 2020;21:6.

[40] Rohatgi A, Webplotdigitizer; 2017.

[41] Bowdler R, Leventhall G. Wind Turbine Noise, Multi-Science 2011.

[42] Breiman L. Out-of-bag estimation, Technical report, Statistics Department. University of California Berkeley; 1996.

[43] Wendt SL, Welinder P, Sorensen HB, Peppard PE, Jennum P, Perona P, Mignot E, Warby SC. Inter-expert and intra-expert reliability in sleep spindle scoring. Clin Neurophysiol 2015;126:1548–56.

[44] Søndergaard LS, Thomsen C, Pedersen TH, 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; 2019.