Contents lists available at ScienceDirect

Science of the Total Environment

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

Windy events detection in big bioacoustics datasets using a pre-trained Convolutional Neural Network

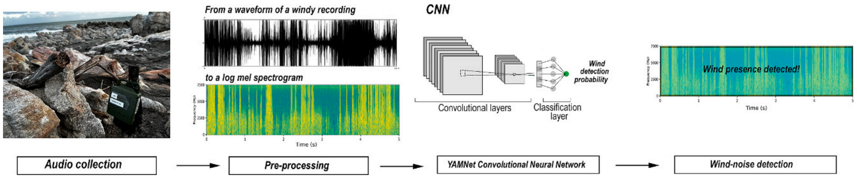
Francesca Terranova [a,*], Lorenzo Betti [b], Valeria Ferrario [a,c], Olivier Friard [a], Katrin Ludynia [d,e], Gavin Sean Petersen [d], Nicolas Mathevon [f,g,h], David Reby [f,g,1], Livio Favaro [a,i,1]

[a] Department of Life Sciences and Systems Biology, University of Turin, Turin, Italy
[b] Department of Network and Data Science, Central European University, Vienna, Austria
[c] Chester Zoo, Caughall Road, Chester, UK
[d] Southern African Foundation for the Conservation of Coastal Birds (SANCCOB), Cape Town, South Africa
[e] Department of Biodiversity and Conservation Biology, University of the Western Cape, Robert Sobukwe Road, Bellville, South Africa
[f] ENES Bioacoustics Research Lab, CRNL, University of Saint-Etienne, CNRS, Inserm, Saint-Etienne, France
[g] Institut universitaire de France, Ministry of Higher Education, Research and Innovation, 1 rue Descartes, CEDEX 05, Paris, France
[h] Ecole Pratique des Hautes Etudes, CHArt lab, PSL University, Paris, France
[i] Stazione Zoologica Anton Dohrn, Naples, Italy

HIGHLIGHTS

- Passive Acoustic Monitoring is increasingly utilised for studying wildlife.
- Wind-induced noise poses a significant challenge in sound processing.
- CNNs offer a cutting-edge approach for detecting acoustic events.
- YAMNet shows high-performance to detect wind-induced noise.
- Low computational needs can enable real-time analysis on portable devices.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Anastasia Paschalidou

Keywords:
Bioacoustics
Deep learning
Ecoacoustics
Passive Acoustic Monitoring
Soundscape ecology
Wind-noise

ABSTRACT

Passive Acoustic Monitoring (PAM), which involves using autonomous record units for studying wildlife behaviour and distribution, often requires handling big acoustic datasets collected over extended periods. While these data offer invaluable insights about wildlife, their analysis can present challenges in dealing with geophonic sources. A major issue in the process of detection of target sounds is represented by wind-induced noise. This can lead to false positive detections, i.e., energy peaks due to wind gusts misclassified as biological sounds, or false negative, i.e., the wind noise masks the presence of biological sounds. Acoustic data dominated by wind noise makes the analysis of vocal activity unreliable, thus compromising the detection of target sounds and, subsequently, the interpretation of the results. Our work introduces a straightforward approach for detecting recordings affected by windy events using a pre-trained convolutional neural network. This process facilitates identifying wind-compromised data. We consider this dataset pre-processing crucial for ensuring the reliable use of PAM data. We implemented this preprocessing by leveraging YAMNet, a deep learning model for sound classification tasks. We evaluated YAMNet *as-is* ability to detect wind-induced noise and tested its performance in a Transfer Learning scenario by using our annotated data from the Stony Point Penguin Colony in South Africa. While the classification of YAMNet *as-is* achieved a precision of 0.71, and recall of 0.66, those metrics strongly

[*] Corresponding author.
E-mail address: francesca.terranova@unito.it (F. Terranova).
[1] Co-senior author.

https://doi.org/10.1016/j.scitotenv.2024.174868
Received 6 March 2024; Received in revised form 16 July 2024; Accepted 16 July 2024
Available online 19 July 2024

improved after the training on our annotated dataset, reaching a precision of 0.91, and recall of 0.92, corresponding to a relative increment of >28 %.

Our study demonstrates the promising application of YAMNet in the bioacoustics and ecoacoustics fields, addressing the need for wind-noise-free acoustic data. We released an open-access code that, combined with the efficiency and peak performance of YAMNet, can be used on standard laptops for a broad user base.

1. Introduction

The use of Passive Acoustic Monitoring (PAM) is significantly growing for wildlife monitoring applications (Sugai et al., 2019; Ross et al., 2023). Two primary factors promote this methodology: technological advancements that enable cost-effective devices for recording extensive acoustic data (Szymański et al., 2021) and its non-invasive approach. This method allows for continuous, long-term wildlife monitoring without disturbing animals, by deploying multiple Autonomous Recording Units (ARUs) in an area over extended time periods (Sugai et al., 2019; Pérez-Granados and Traba, 2021). ARUs allow continuous recording under different environmental conditions with limited maintenance of the recorders (Rhinehart et al., 2020). However, they also present a challenge: do not allow for choosing the best conditions for recording the biological sounds beforehand, hence leading to the inclusion of environmental noise in the acoustic dataset. Indeed, other than biophony (wildlife vocalisations), the soundscape of natural environments is characterised by two more main components of the soundscape: anthrophony (anthropogenic activities) and geophony (weather-related phenomena) (Pijanowski et al., 2011). Geophony represents a distinctive and dynamic layer of the soundscape of a given area (e.g., strong wind or heavy rain can strongly characterise audio recordings). Windy events, in particular, pose a challenge in sound analysis because they could determine a persistent interference in the recordings (Depraetere et al., 2012; Stowell et al., 2018; Quinn et al., 2022).

In terrestrial environments, microphone wind noise is characterised by transient peaks, with particular emphasis on the lower frequency range (Bardeli et al., 2010; Walker and Hedlin, 2009; Nelke, 2016). To date, wind interference remains a significant issue in several PAM applications, reflected in the substantial number of published studies that highlight the negative impact of wind interference on PAM outcomes (e.g., Buxton and Jones, 2012; Digby et al., 2013; Gillam et al., 2009; Znidersic et al., 2021; Priyadarshani et al., 2018; Zhao et al., 2022; Hacker et al., 2023). Specifically, during the analysis for the detection of target sounds, wind noise can lead to the presence of false positives, i.e., the increasing peak energy of wind often produces false detection of target species (Zwart et al., 2014; Juodakis and Marsland, 2022) or false negatives, i.e., the presence of wind noise could mask the presence of the target sound (Zwart et al., 2014; Willacy et al., 2015; Salamon et al., 2016; Stowell et al., 2018). In sum, wind noise in the recordings poses a challenge, as its identification and filtering are necessary steps to achieve wind-free analysis which will improve the reliability of research relying on PAM techniques (Depraetere et al., 2012; Eldridge et al., 2018; Fairbrass et al., 2017; Stowell et al., 2018).

Researchers can take precautions to avoid environmental noise in the recordings, including the accurate choice of the recording area (Oswald et al., 2022) or post-filtering the data with bad weather conditions using meteorological data (Desjonquères et al., 2018). Nevertheless, the effectiveness of these techniques in mitigating wind-induced noise in recordings may vary depending on both the recording area and deployment methods. This variability is attributed to the dynamic nature of wind conditions, where extensive deployment areas may be exposed to diverse wind scenarios, subjecting ARUs to varying levels of exposure.

Finally, the volume of data generated through PAM has increased significantly in the last few years, leading to a critical need for fast wind detection methods. Long-term monitoring results in thousands of hours of recordings, rendering manual inspection impractical. Consequently, methods for identifying target species often rely on automated detection techniques, with widely used approaches including software applications. However, these software tools often struggle to distinguish whether energy peaks in the recording tools are caused by wind or animal species, as dynamic characteristics of windy events complicate template recognition.

Convolutional Neural Networks (CNNs) have proven remarkable performance in classification tasks involving different natural signals such as image, text, speech, and audio data (LeCun et al., 2015; Zhang et al., 2018; Norman et al., 2022). Their capabilities stem from processing the input signal through multiple layers that learn from data features at varying levels of abstraction, which allow CNNs to learn the hierarchical structure of the data (LeCun et al., 2015). When processing audio data, this property translates into CNNs being able to distinguish different sounds in a complex soundscape scenario. In the field of bioacoustics research, CNNs have been used for tasks such as classifying different species, identifying target species, and detecting individual cues encoded in vocal signals (Christin et al., 2019; Ruff et al., 2021; LeBien et al., 2020; Zhong et al., 2020; Kahl et al., 2021; Bedoya and Molles, 2021; Dufourq et al., 2022; Trapanotto et al., 2022; Ravaglia et al., 2023).

Despite their effectiveness, the widespread adoption of CNNs and deep neural networks is hindered by their high computational requirements and the necessity to be trained on massive, annotated datasets (Dufourq et al., 2022). Those limitations can be mitigated by developing efficient CNN architectures resulting in compact models that require limited computational complexity and small memory requirements at the cost of marginal performance loss compared to large-scale deep learning models (Howard et al., 2017; Sandler et al., 2018). This also allows the use of those models on mobile devices without the need for powerful hardware.

Additionally, Transfer Learning offers a solution for applications in which the availability of annotated data is limited. This approach involves leveraging neural networks pre-trained on large datasets and retraining them on smaller annotated datasets to achieve good performance in specialised tasks. By transferring knowledge learned from vast datasets to the specific domain of identification in audio recordings, Transfer Learning allows for efficient utilisation in settings with limited annotated data (Lu et al., 2021; Tsalera et al., 2021; Dufourq et al., 2022).

Regardless of the methodologies employed in bioacoustic research, studies aimed at detecting species-specific vocalisations often assume either (i) animal sounds and geophony are mutually exclusive occupying different spaces within the frequency spectrum or (ii) environmental noise is negligible. However, they overlook the possibility of animals vocalising in windy conditions, leading to potential reliability issues. This oversight includes the risk of both underestimating animal sounds, where wind may mask vocal activity in the case of false negative presence, and overestimating animal sounds, where the wind may be erroneously classified as animal sounds, leading to a false-positive presence. Wind also poses a challenge for ecoacoustics indices, that often require wind-free recordings to avoid biases in the interpretation of indices (Farina, 2018; Fairbrass et al., 2018; Metcalf et al., 2021; Ross et al., 2021; Quinn et al., 2022). Wind could induce rapid broad-frequency interference, which demands careful consideration to ensure the accuracy and reliability of automated analyses in ecoacoustics research for soundscape analysis.

Here, we report a study whose primary objective was developing a methodology for pre-processing our big acoustic dataset, to reliably identify recordings affected by windy events. To do this we analysed an acoustic dataset gathered at the African penguin colony of Stony Point (South Africa), comprising recordings from the colony's terrestrial habitat. We employ a pre-trained CNN to classify windy events present in acoustic data and enhance the data pre-processing process for bioacoustics datasets. Specifically, we leverage a deep neural network known as "Yet Another Mobile Network (YAMNet)", a CNN with a lightweight architecture pre-trained on the Audioset Ontology. This ontology is a comprehensive collection of 2,084,320 human-labelled 10-s sound clips sourced from YouTube videos, encompassing 521 distinct audio classes, comprising a Wind class (Gemmeke et al., 2017). YAMNet[2] has proven effective in various fields for audio event classification (Mohammed et al., 2023; Tena et al., 2022; Hyun, 2023). We predict that CNN-based methods can be effective in a complex soundscape scenario characterised by multiple broadband signals with overlapping characteristic frequencies, such as wind and penguin vocalisations.

Initially, our approach involves utilising YAMNet *as-is* for classifying windy events. Subsequently, we applied Transfer Learning to YAMNet, retraining its last classification layer on our manually annotated dataset, based on the acoustic dataset of the African penguin colony of Stony Point. Accordingly, we created an efficient method to detect windy events from terrestrial soundscape recordings. Our overall goal was to improve the reliability and utility of long-term PAM research outcomes, contributing to a more accurate understanding of wildlife acoustic environments.

2. Methods

2.1. Acoustic dataset – case study

The acoustic dataset was collected at the Stony Point Penguin Colony (Betty's Bay, South Africa; Fig. 1) under the CapeNature research permit CN32-87-23209 and the South African Minister of Forestry, Fisheries and the Environment research permit RES2023-25. African penguins (*Spheniscus demersus*) exhibit a peak of vocal activity around sunrise and sunset (Favaro et al., 2021). Their vocalisations have a fundamental frequency (f_0) of approximately 200 Hz (Favaro et al., 2014), meaning that their detection is particularly challenging when wind gusts mask the low-frequency components of the soundscape.

We used the acoustic recordings collected from February to April 2023 over 76 consecutive days with eight Song Meter micro-omnidirectional recorders (Wildlife Acoustics, Inc.) mounted on stationary poles positioned 20 cm above ground level across eight distinct sampling sites, with a distance of 25–50 m from the shoreline (GPS coordinates are provided in Table S1). The deployment area was characterised by different types of substrates and vegetation, including an area blanketed by dense dune spinach (*Tetragonia decumbens*) bushes (Fig. 1a), another with dispersed *Baccharis halimifolia* bushes in a grassy region (Fig. 1b), and an area featuring a coastal landscape with a combination of rocks and sandy terrain (Fig. 1c).

We set up a duty cycle of 30-min recordings from 06:00 p.m. to 8:00 a.m. South African Standard Time (SAST), adjusted to local sunrise and sunset times. This time frame allows capturing the peak of vocal activity of African penguins, typically occurring between 4:00–8:00 a.m. and 5:30–9:30 p.m. (Favaro et al., 2021). For the resulting $n = 14{,}564$ 30-min recordings, corresponding to 7282 h), the gain of the recorders was set to 18 dB to balance optimal audio capture for animal recordings without being overly susceptible to environmental noise. The used microphones have Sensitivity: +2 dB FS +/− 4 dB re 1pa@1 kHz with +18 dB gain; Signal-to-Noise ratio: 73 dB Typ. at 1 kHz (1 Pa, A-weighted); Max input sound level: 104 dB SPL. Their output signals were digitised at a sampling rate of 48 kHz and saved into internal Secure Digital memory cards as .WAV files with 16-bit amplitude resolution.

2.2. Manual annotation of the spectrograms

We created the annotated dataset using a two-step process to ensure a diverse representation of soundscapes. Specifically, we aim to create a dataset comprising recordings with wind noise and recordings with African penguin vocalisation, the main species represented in our acoustic dataset. This is essential to provide the model with a variety of soundscapes of interest for our application. The Fig. 2 provides a schematic representation of the selection process.

In the first annotation step (Fig. 2a), we divided the daytime into two periods based on the vocal activity peaks of African penguins, which occur from 4:30 to 6:00 AM and from 6:30 to 8:00 PM (Favaro et al., 2021). For each period and each of the eight songmeters, we randomly selected 13 30-min recordings. This approach yielded a total of 208 recordings, which represent 1.4 % of the acoustic dataset. Subsequently, we segmented these recordings into 5-min intervals. Each segment was manually annotated to identify the presence of wind and biophonic events, i.e. animal calls, through audio and visual inspection of spectrograms using Praat v. 6.3.09 with the following spectrogram settings: View range (Hz): 0–5000; Window length (s): 0.05; Dynamic range (dB): 70 (Boersma and Weenink, 2024).

For the second annotation step (Fig. 2b), we employed the annotations from the previous step to refine our selection process. We randomly selected 18 30-minute recordings from each combination of wind and animal sounds. This strategy ensured a representation of all possible acoustic conditions and resulted in a total of 72 30-minute recordings, representing 0.5% of the acoustics dataset.

Two human observers (F.T. and V.F.) annotated this subset with a 5-second resolution (i.e., each recording was split into non-overlapping segments of 5 s) through visual inspection of the spectrograms (examples of the annotated spectrograms are provided in Fig. 3), by annotating for (i) the presence of wind, (ii) wind strength, (iii) presence of biophony, (iv) presence of rain. Wind strength was annotated into three categories: low, medium, and high. We defined *low wind* as the presence of at least 1 continuous second, extending up to 3 non-continuous seconds of windy events in the spectrogram, *medium wind* as the presence of over 3 s of wind, accompanied by clear, noise-free segments lasting >1 s above 2000 Hz in frequency range. Lastly, *high wind* is identified by continuous wind presence, masking the entire spectrogram throughout the 5-s segment (Fig. 3). We opted for a 5-second duration cutoff for these segments to effectively capture phenomena such as wind gusts, which often lack temporal discreteness, and considering that the longest African penguin vocalisation typically extends for approximately 5 s (Favaro et al., 2014; Quinn et al., 2022). This two-level labelling allows to have an annotated dataset that is well-balanced with respect to the distribution of wind and biophonic events, which is crucial to train YAMNet, thus ensuring that our sampling efforts yield a diverse range of soundscapes and avoiding scarcity in recordings with windy events. All the spectrograms were generated with Parselmouth (Jadoul et al., 2018), an open-source Python (Van Rossum and Drake Jr, 1995) interface to Praat's core functionality (Boersma and Weenink, 2024). The spectrogram settings are detailed in the caption of Fig. 3.

We used eight recordings (10 % of the annotated dataset) balanced for wind and animal sounds, resulting in 2880 5-second segments, to assess the inter-annotator agreement by employing Cohen's Kappa statistic for categorical data (i.e. binary label for the presence of wind, biophonic events, rain) (McHugh, 2012) and the weighted Cohen's Kappa for ordinal data (i.e. wind strength) (Cohen, 1968). The Cohen's Kappa for wind was 0.82, the Weighted Cohen's Kappa for wind strength was 0.82, Cohen's Kappa for rain was 0.94, and Cohen's Kappa for animal sound was 0.63. According to the interpretation of Kappa results (McHugh, 2012) the agreement was "almost perfect" for all the labels

[2] Official repository of the model: https://github.com/tensorflow/models/tree/master/research/audioset/yamnet.

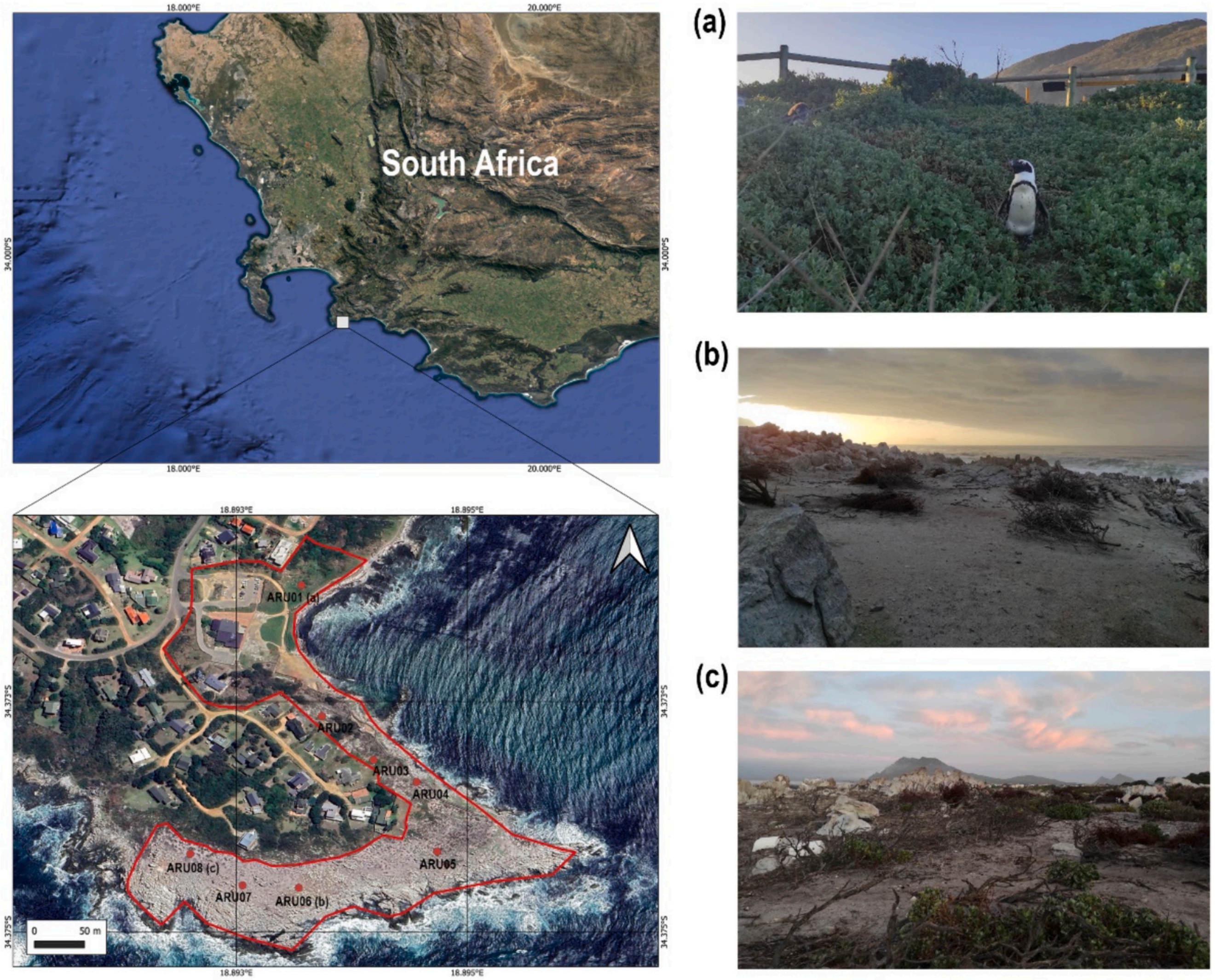


Fig. 1. Study area with the position of the ARUs (red dots). The right-hand panels represent the three different landscapes of the colony. (a) Area with dense dune spinach (*Tetragonia decumbens*); (b) coastal landscape with a combination of rocks and sandy terrain; (c) area with dispersed *Baccharis halimifolia* bushes.

but biophonic events, a case in which the agreement was "substantial".

To summarise, we employed a two-step annotation procedure to obtain an annotated dataset for the presence of biophonic events and noise sources (wind, rain) at a 5-second resolution. We reported the basic statistics of the annotated dataset in Table 1. By using data from two weather stations, we validated the annotations of the presence of wind and the wind strength categories (Fig. S1) and reported the distribution of wind speed and wind direction for each ARU (Fig. S2 and Fig. S3). Additional details on the validation are provided in the section "Validation of manual wind annotation" of the Supporting Information.

2.3. YAMNet convolutional neural network

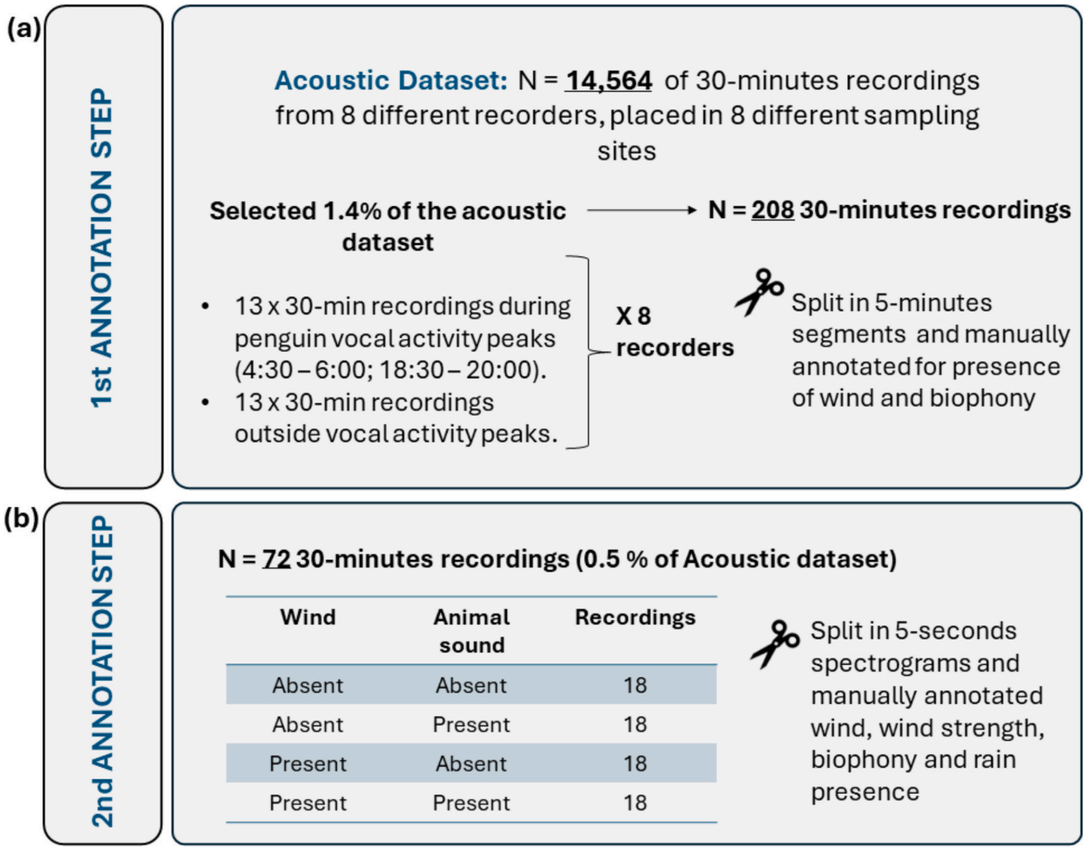
We employed YAMNet to predict the presence of wind in audio recordings. YAMNet is a Convolutional Neural Network that employs the MobileNets architecture (Howard et al., 2017), a lighter alternative to other state-of-the-art models that achieve competitive performance in computer vision tasks. Details of the architecture of the CNN are available in the release notes.[3] To make predictions from audio data, YAMNet requires an image-based representation of the audio obtained by preprocessing, segmenting, and transforming the original audio waveform into a spectrogram. YAMNet is pre-trained on the *AudioSet* dataset (Gemmeke et al., 2017), a large-scale dataset composed of around two million manually annotated audio events extracted from YouTube videos (top panel Fig. 4). Given an audio signal as input, the model outputs a score between 0 and 1 for 521 audio event classes belonging to the AudioSet hierarchical ontology, which covers a wide range of everyday sounds, from human and animal sounds to natural and environmental sounds, to musical and miscellaneous sounds.

2.4. YAMNet-*as is* wind detection model

Our first approach involved utilising the model '*as-is*' without any retraining (Fig. 4).

We pre-processed the original waveforms according to the release notes[3]. We resampled the audio signal at 16 kHz and scaled to obtain values in the range [−1.0, 1.0]. Next, we transformed the audio signal into a spectrogram using the magnitudes of the Short-Time Fourier Transform with a window size of 25 ms, a window hop of 10 ms, and a periodic Hann window. Subsequently, we computed a mel spectrogram by mapping the spectrogram to 64 mel bins covering the range 125–7500 Hz. Furthermore, we segmented the spectrogram into

[3] Official repository of the model: https://github.com/tensorflow/models/tree/master/research/audioset/yamnet.


(a)
1st ANNOTATION STEP
Acoustic Dataset: N = 14,564 of 30-minutes recordings from 8 different recorders, placed in 8 different sampling sites
Selected 1.4% of the acoustic dataset
N = 208 30-minutes recordings
• 13 x 30-min recordings during penguin vocal activity peaks (4:30 – 6:00; 18:30 – 20:00).
• 13 x 30-min recordings outside vocal activity peaks.
X 8 recorders
Split in 5-minutes segments and manually annotated for presence of wind and biophony
(b)
2nd ANNOTATION STEP
N = 72 30-minutes recordings (0.5 % of Acoustic dataset)
Split in 5-seconds spectrograms and manually annotated wind, wind strength, biophony and rain presence

Wind	Animal sound	Recordings
Absent	Absent	18
Absent	Present	18
Present	Absent	18
Present	Present	18

Fig. 2. Schematic representation of the two different selection processes.

segments of 0.96 s each, with a stride of 0.48 s, resulting in consecutive segments with a 50 % overlap. Ultimately, the model provided a vector of predictions for each segment.

Then, we used the pre-processed audio segment as the input of YAMNet, which returns two outputs for each audio segment. The first output is a feature vector that represents the input audio. The second is a set of predicted probabilities for each class defined in the AudioSet ontology, which YAMNet was pre-trained on. From these predictions, we selected three classes related to wind: 'Howl (wind)', 'Wind noise (microphone)' and 'Rustling Leaves'. We considered the highest probability score among these three classes as the measure of wind detection probability of each segment.

2.5. YAMNet transfer learning scenario

We used the features extracted from YAMNet *as-is* model to train a new classifier. This technique allowed us to exploit the pre-existing YAMNet pre-training phase and adapt it to our specific acoustic dataset. To accomplish this, we extracted the average-pooled output of the last convolutional layer of YAMNet *as-is* and used this vector as a feature representing the input audio segment. We then trained a single-layer neural network using those features as input. In other words, we trained the last classification layer of YAMNet on our annotated dataset from scratch. We utilised the Adam optimiser (Kingma and Ba, 2014) and a binary cross-entropy loss function during the training process. To prevent overfitting, we employed Early Stopping (Morgan and Bourlard, 1989) by using 20 % of the training set. A sigmoid activation function was utilised in the final layer to output scores ranging from 0 to 1, indicating the likelihood of wind presence in the audio recordings. The resulting model provided probability scores that indicate the confidence level in the presence of wind within the analysed audio segment. We refer to this second model as 'YAMNet Transfer Learning' (YAMNet TL) hereafter (Fig. 4). We reported the results obtained by setting the learning rate at 2×10^{-4}, batch size at 64, and number of epochs at 20. Other parameter choices led to only marginal differences in the performance (see Supporting information, Fig. S4). Finally, we replicated the experiment using half of the recordings (36 files = 18 h of recording) to assess the consistency of performance with a halved dataset.

2.6. Classification and threshold optimization

Classification within machine learning involves training a model to accurately predict the target class(es) to which data belongs. By defining a *positive sample* as an audio segment containing wind, the comparison between the YAMNet output and target label could produce four different results: true positive (TP) when a positive sample is correctly identified; false negative (FN) when a positive sample is misclassified as negative; true negative (TN), when a negative sample is correctly identified; and false positive (FP) when a negative sample is mistakenly classified as positive.

The evaluation of our models incorporated several metrics, including Recall (also known as Sensitivity), Specificity, Precision, and F(β)-score.

Precision measures the proportion of samples that the model correctly identified as positive out of all samples it predicted as positive:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Recall assesses the ability to capture as many true positive instances as possible, minimising false negatives.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

The interplay between precision and recall often presents a trade-off that depends on the use case. The F(β)-score can be used to optimise the

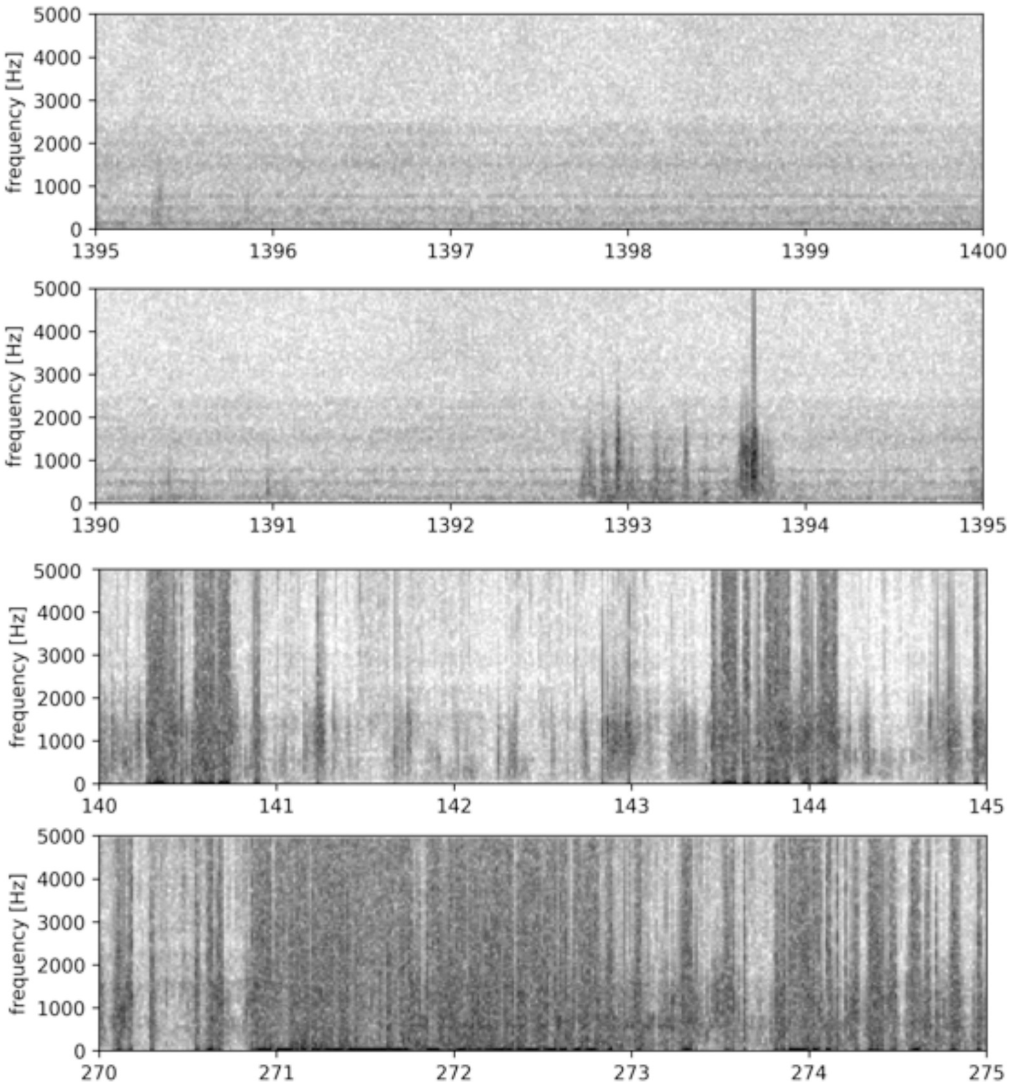


Fig. 3. Examples of 5-s spectrograms used for the second annotation step, taken from different recordings, showing four different levels of wind strengths. From top to bottom: absent wind, low wind, medium wind, high wind. Parselmouth settings used to generate the annotation pages: time interval = 5 s; dynamic range = 70 dB, Window length = 0.03 s, maximum frequency = 5000 Hz; Time step (overlap) = 0.015 s, Frequency step = 50 Hz, NFFT = 960.

Table 1
Soundscape components of the annotated dataset. Occurrences of 5-second segments annotated for wind strength, rain and biophonic events. The percentages indicate the proportion of each element within specific wind strength categories.

Wind strength	Geophony (rain)	Biophonic events	Total
Absent	2903 (17 %)	2613 (16 %)	16.627
Low	990 (13 %)	748 (10 %)	7494
Medium	3 (0 %)	66 (5 %)	1283
High	0 (0 %)	0 (0 %)	446
Total	3896	3427	–

classification threshold depending on the desired relative weight for precision and recall (Sokolova et al., 2006). The F(β)-score is defined as:

$$F(\beta) = \frac{(1+\beta^2)*precision*recall}{(\beta^2*precision)+recall}$$

where β is a parameter favouring precision (β < 1) or recall (β > 1). We tested β values of 1 and 2, covering two scenarios in which we give to precision and recall the same weight (β = 1) and more weight to recall (β = 2). Across these values, classification thresholds are tested incrementally from 0 to 1 in 0.01 steps. The optimal threshold is then the threshold that maximises the F(β)-score. We reported results optimised for the F_1-score unless specified otherwise. We also evaluated Specificity—a key measure assessing the model's ability to correctly identify true negatives (Tharwat, 2020):

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

Overall performance was evaluated using the area under the receiver operating characteristic curve (AUC-ROC). ROC curve (Receiver Operating Characteristic curve) is a graph showing the performance of a classification model at all classification thresholds: it plots True Positive and False Positive rates (Tharwat, 2020). AUC stands for “Area under the ROC Curve”. That is, AUC (Area Under the Curve) measures the two-dimensional area underneath the ROC curve from (0,0) to (1,1). AUC provides an aggregate measure of performance across all possible classification thresholds. Similarly, the Precision-Recall curve plots the values of Precision and Recall for all possible classification thresholds.

We followed the recommendations outlined by Knight et al. (2017) to ensure robust and reliable performance evaluations.

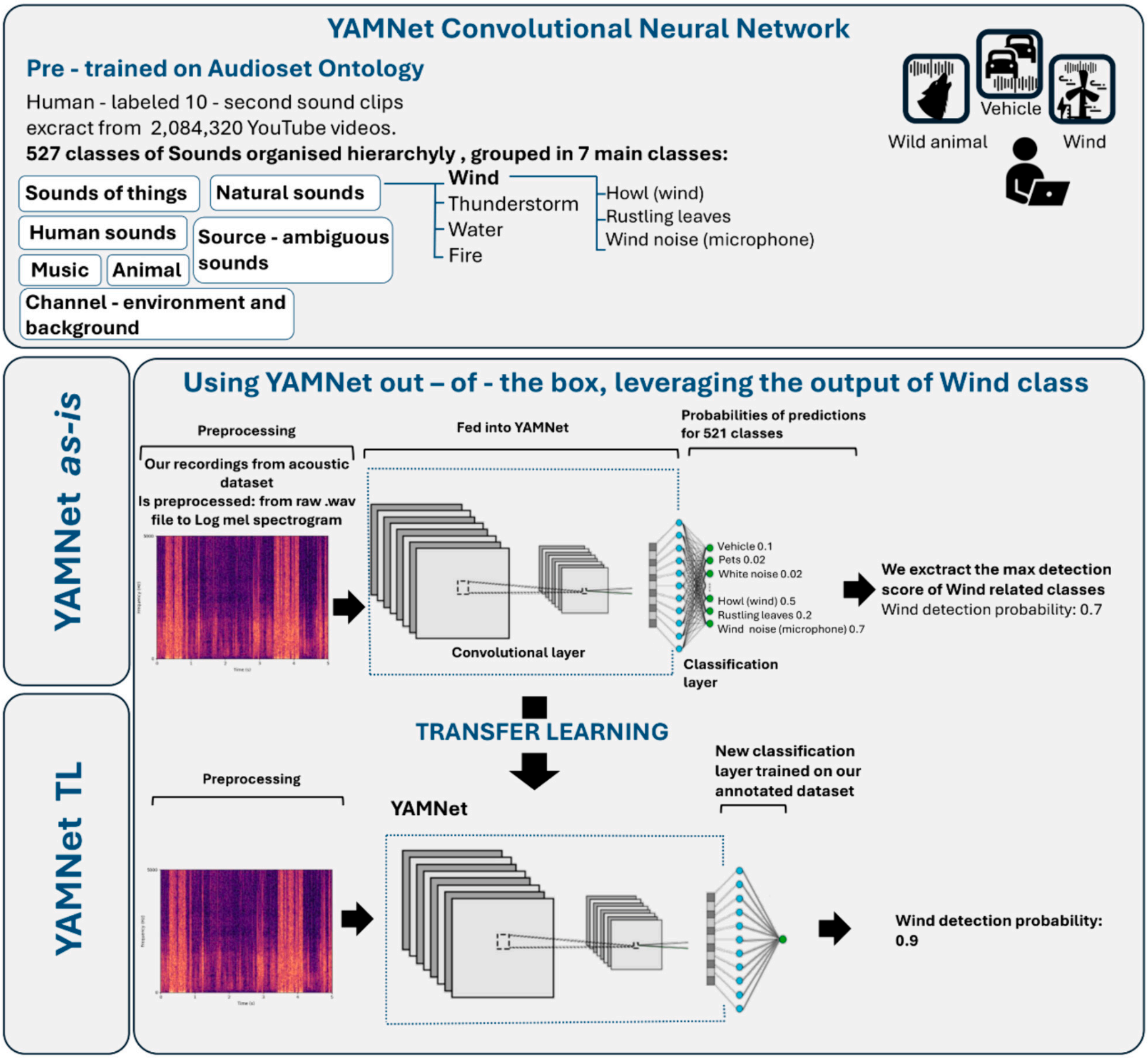

YAMNet Convolutional Neural Network
Pre - trained on Audioset Ontology
Human - labeled 10 - second sound clips
excract from 2,084,320 YouTube videos.
527 classes of Sounds organised hierarchyly , grouped in 7 main classes:
Sounds of things
Natural sounds
Human sounds
Source - ambiguous sounds
Music
Animal
Channel - environment and background
Wind
Thunderstorm
Water
Fire
Howl (wind)
Rustling leaves
Wind noise (microphone)
Vehicle
Wild animal
Wind
YAMNet as-is
Using YAMNet out – of - the box, leveraging the output of Wind class
Preprocessing
Our recordings from acoustic dataset
Is preprocessed: from raw .wav file to Log mel spectrogram
Fed into YAMNet
Convolutional layer
Probabilities of predictions for 521 classes
Vehicle 0.1
Pets 0.02
White noise 0.02
Howl (wind) 0.5
Rustling leaves 0.2
Wind noise (microphone) 0.7
Classification layer
We exctract the max detection score of Wind related classes
Wind detection probability: 0.7
TRANSFER LEARNING
YAMNet TL
Preprocessing
YAMNet
New classification layer trained on our annotated dataset
Wind detection probability: 0.9

Fig. 4. Schematic Representation of the CNN Pipeline. The top section outlines the general CNN model structure. The middle section details the YAMNet model used '*as-is*' for predicting wind sounds from audio recordings, utilising pre-trained classes related to wind. The bottom section illustrates the transfer learning scenario, where features extracted from YAMNet are used to train a new classifier specifically for wind sound detection.

2.7. *Cross-validation for model evaluation*

We evaluated the performance of the proposed models using cross-validation. This procedure consists of partitioning at random the dataset into k folds. Then, for each iteration, k-1 folds are chosen to train the model and the remaining fold is used to evaluate the model's performance. In such a way, the model can be evaluated on a held-out test set in an iterative fashion. However, since our annotated audio segments belong to 30-minute-long audio recordings, the presence of wind in an audio segment may not be independent from the one in another segment belonging to the same audio recording. Such temporal autocorrelations in the target label may lead to overestimating the performance. This is why we opted for Block cross-validation to ensure a robust assessment of our model's performance (Roberts et al., 2017). Instead of randomly partitioning the segments into k equal-sized folds as for the classic k-fold cross-validation, we split each audio recording into 100-second-long blocks and then partition each audio recording separately into k folds (Fig. S5). This cross-validation strategy limits the risk that temporal autocorrelations in the target label within the same audio recording can overestimate the model's performance (Roberts et al., 2017).

For our experiments, we used k = 5. We considered the average of the k scores to estimate the model's performance and its standard deviation as a measure of its variability across folds.

2.8. *Investigating the effect of non-wind components on model's performance*

We assessed our models' performance across different soundscapes characterised by the presence of non-wind components. This analysis aimed to explore to what extent other soundscapes may affect the ability of the models to detect wind-induced noise. To do that, we compared precision and recall against three distinct subsets of the annotated dataset, each excluding segments affected by rain and biophony. We compared such scores against the scores obtained on the whole annotated dataset and discussed their relative differences concerning one of the whole annotated datasets. A positive relative difference indicates that the performance would increase by discarding other soundscapes, thus pointing to the misclassification of such soundscapes into wind-induced noise.

2.9. Effect of wind-strength on YAMNet scores

To test whether the distributions of the probabilities predicted by the two models differ for different wind strengths, we first employed a Kruskal–Wallis test due to the heterogeneity of sample sizes and non-homogeneity of variances among wind strength categories (Kruskal and Wallis, 1952). Significance from the Kruskal-Wallis test ($p < 0.05$) would indicate significant differences among the distributions. Subsequently, we employed Dunn's simultaneous multiple-comparison test (Dunn, 1964) with a Bonferroni correction to address multiple comparisons to distinguish which score distributions significantly differed. However, these tests do not provide information about the correlation between wind strength and probability scores. To investigate the potential correlation, we computed the Kendall Tau-b correlation coefficient between the probability scores and the low, medium, and absent wind strength classes. This additional analysis allows us to explore whether a monotonic relationship exists between YAMNet scores and wind strength, even though this information was unavailable during the training. The two YAMNet models were tested separately, and the score distributions were obtained by concatenating the test sets of the 5-folds used for cross-validation.

3. Results

3.1. Performance of the models

The discriminatory performance of the two models was quantified through ROC and Precision-Recall (PR) curves displayed in Fig. 5. Those two metrics are independent of the threshold used for classifying the wind. YAMNet *as-is* achieved a ROC-AUC score of 0.81, while YAMNet TL exhibited superior performance with a score of 0.98. Similarly, YAMNet TL outperforms the pre-trained version concerning the PR-AUC score, achieving 0.97 and 0.77, respectively.

After optimising the classification threshold, we evaluated precision, recall, specificity, and F_1-score for two possible thresholds: one optimising the F_1-score and one optimising the F_2-score. YAMNet TL outperforms YAMNet *as-is* in all the metrics. The only exception occurs for the recall when optimising for the F_2-score, which gives more weight to recall than to precision, in which they performed similarly (Table 2). Indeed, both YAMNet *as-is* and YAMNet TL achieve a recall of 0.99 and 0.97 respectively. Furthermore, repeating the experiment with half of the recordings (36 files = 18 h of recording) yielded identical performance (Fig. S6).

Table 2
Classification metrics for the YAMNet *as-is* and YAMNet TL models. Values refer to the score averaged across the 5-fold cross-validation. Metrics are reported for both settings where we optimised the classification threshold to maximise the F_1-score and F_2-score. Standard deviations are smaller than 0.06 for all entries.

	Precision		Recall		Specificity		F_1	
Optimised for	F_1	F_2	F_1	F_2	F_1	F_2	F_1	F_2
YAMNet *as-is*	0.71	0.36	0.66	0.99	0.85	0.03	0.68	0.53
YAMNet TL	0.91	0.82	0.92	0.97	0.95	0.88	0.91	0.89

3.2. Impact of other environmental sounds on the performance of the models

We assessed the performance of our models across different components of the soundscape, incorporating annotations for rain and biophonic events. The evaluation involved comparing the metrics on the entire acoustic dataset ("All") and the subsets excluding audio segments with rain ("No rain") and segments with animal sounds ("No biophony"). Results are shown in Fig. 6. Biophonic events and rain had marginal effects (absolute relative difference within 6 %), except for the recall of YAMNet *as-is* for rain (relative increase of 11 %). Instead, YAMNet TL proved to be more robust to the presence of rain (relative increase of 1 %). YAMNet TL demonstrated increased overall power (scores increased) and enhanced robustness, with smaller relative differences. All relative differences for these assessments are present in the Supporting information (Table S2).

3.3. Effect of wind strength on YAMNet scores

We tested the distributions of the probabilities predicted by the two models for different wind strengths, plotted in Fig. 7. Kruskal-Wallis revealed significant differences among wind strengths for both YAMNet *as-is* (H = 7063, $p < 0.0001$) and YAMNet TL (H = 17,181, $p < 0.0001$). All the pairwise comparisons resulted in statistically significant differences according to Dunn's test ($p < 0.0001$), except in YAMNet TL between medium and high wind strengths ($p = 0.54$). The correlation between wind strengths and predicted probabilities resulted positive and significant in both cases: YAMNet *as-is* τ =0.20 and YAMNet TL τ =0.47, both with $p < 0.0001$.

4. Discussion

In this study, we illustrate the efficacy of using YAMNet, a pre-

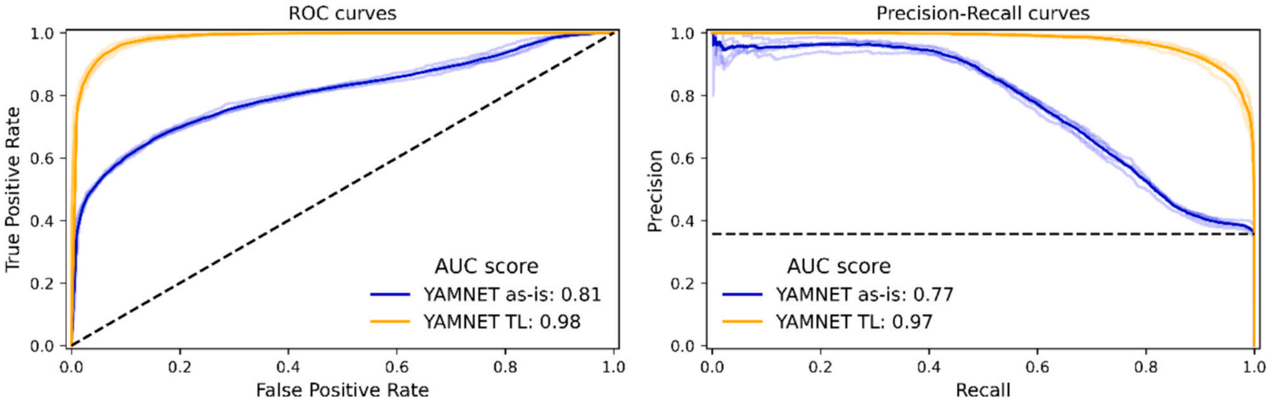


Fig. 5. ROC and Precision-Recall curves for YAMNet *as-is* and YAMNet TL. Transparent lines refer to the 5-fold cross-validation predictions, while the opaque lines are their averages. Dashed lines refer to the performance of dummy classifiers. Average ROC and Precision-Recall AUC are reported in the caption. Standard deviations are <0.01.

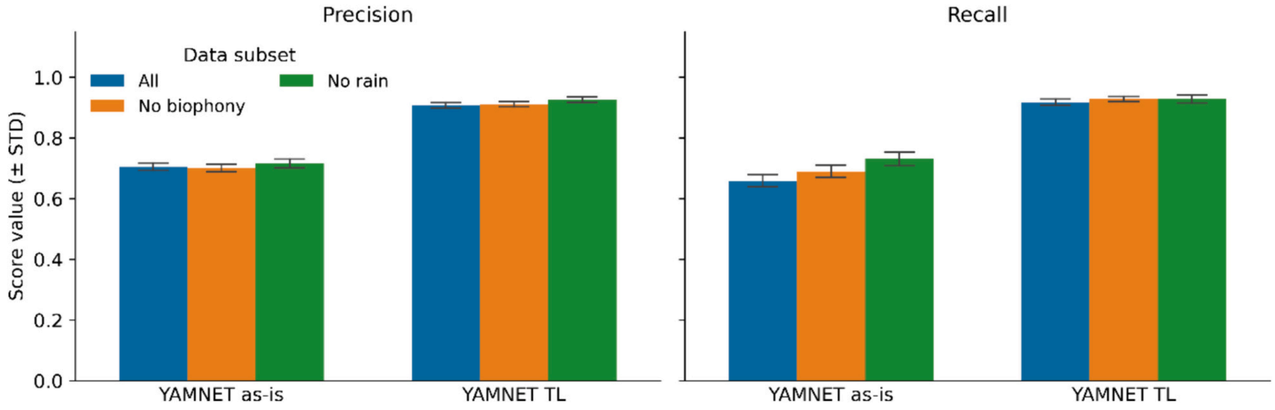


Fig. 6. Comparative results for YAMNet models across different soundscapes. The model's performance across three subsets of the dataset, showing the variations in precision scores (a) and recall scores (b) within each distinct subset for YAMNet *as-is* and YAMNet TL.

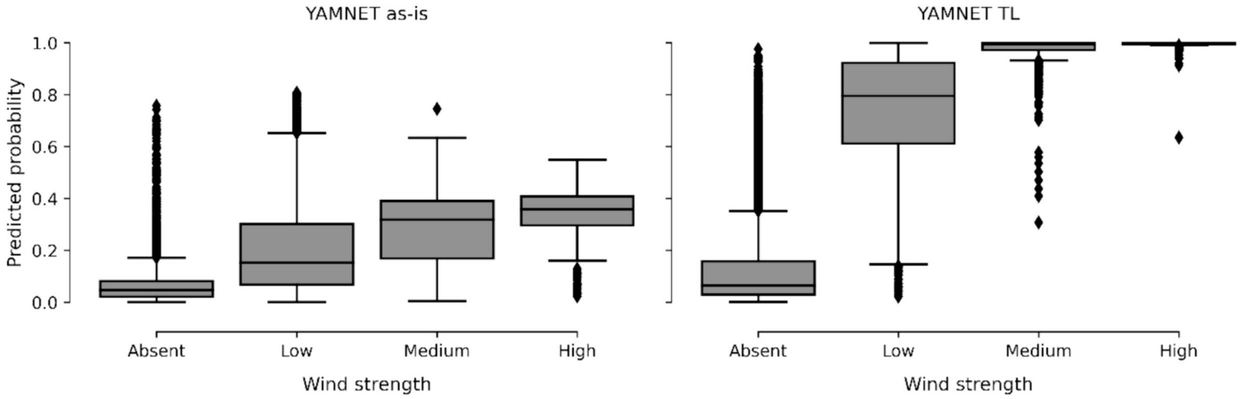


Fig. 7. Distribution of YAMNet *as-is* and YAMNET TL predicted probabilities for different wind strengths. The boxplots show the distribution of the probability scores predicted by each model for the wind strength classes. All distributions within each model differ from each other according to Dunn's test ($p < 0.0001$), except for YAMNET TL between medium and high wind strengths.

trained audio event classifier, in accurately detecting windy recordings over a large dataset of soundscape recordings. Additionally, we explored methods to enhance its performance through Transfer Learning. Our results emphasise the high-performance capabilities of YAMNet-like models, especially with application of Transfer Learning, serving as a valuable tool in the pre-processing bioacoustics and ecoacoustics data.

We showed that YAMNet *as-is* is capable of identifying windy events in recordings from a penguin colony. This capability arises from the presence of wind-related classes in its pre-training dataset, even though it was not directly exposed to the specific soundscapes we tested it on. While the performance being modest, it highlights that YAMNet can prove useful in diverse contexts for preliminary data explorations where external annotations for the presence of wind are absent (e.g., weather station or manual annotations). However, a relative increase larger than 28 % in precision and recall is achieved by training a classifier (YAMNet TL) using the features extracted from YAMNet *as-is* as input at the cost of a small manual annotation of the dataset, in our case 0.5 % of the dataset corresponding to 36 h of recordings. This indicates that ad-hoc re-training can adapt a generic model to the environment under study. Moreover, our findings illustrate how YAMNet *as-is* classifications are robust to the presence of other components of the soundscapes, such as biophonic events (e.g., African penguin vocalisations) and geophonic sounds (i.e., rain). However, we showed that ad-hoc retraining could make the model even more robust to the presence of such additional components of soundscapes. Overall, this study demonstrates that YAMNet *as-is* maintains the ability to detect windy events in dynamic acoustic environments. Moreover, the re-training achieved through YAMNet Transfer Learning enhances the model's robustness, pointing to its usefulness in bioacoustics and ecoacoustics studies. While our data collection focused on an African penguin colony, and thus designed for terrestrial environments, the effectiveness of this methodology could be extended beyond our specific study, functioning as a valuable tool for the pre-processing of sound data in a variety of different environments. Indeed, the strength of YAMNet lies in its extensive pre-training on 521 different audio classes, with a collection of 2,084,320 human-labelled 10-s sound clips drawn from YouTube videos. Thus, starting from this, transfer learning scenarios could be applicable in many different contexts, both terrestrial and aquatic. For example, the Audioset ontology includes categories such as the 'vehicle class.' Therefore, future works could explore employing the same methodology for anthropogenic noise classification.

Despite their widespread and growing usage in bioacoustics and ecoacoustics studies, automatic analysis methods too often overlook the crucial data pre-processing stage, rendering our contribution particularly timely. Within Passive Acoustic Monitoring (PAM) studies, the prevalence of geophony reflecting predominantly wind patterns, highlighted in recent investigations (Quinn et al., 2022), mirrors our case study, where approximately 30 % of annotated data reveals wind presence, coinciding often with an overlap between wind frequencies and low-pitched African penguin vocalisations. This overlap poses a significant challenge to downstream analyses. Despite the recent denoising methods proposed for wind-robust detection of the biophonic sounds (Juodakis and Marsland, 2022), the varying levels of wind strengths in windy-exposed environments present challenges, as differing gusts can entirely mask the signal of interest in the spectrogram, rendering conventional denoising inadequate. Consequently, our approach advocates for removing irrevocably wind-corrupted audio segments alongside potentially identifying moderately affected segments for analysis using denoising-based methods. This would greatly reduce the computational requirements for denoising and, at the same time, lower the risk of false positives and false negatives for biophonic events detection.

Moreover, our results suggest that YAMNet can be tailored for different applications and adapted to accommodate diverse data collection scenarios. Indeed, YAMNet *as-is* can be used to recommend a subset of the acoustic dataset that likely contains windy events. This can speed up a full manual dataset annotation by selecting audio segments worth annotating. This application is supported by the 0.99 recall obtained by YAMNet *as-is* optimising for F_2-score. Once the annotated dataset containing all the acoustic scenarios specific to the studied environment has been established, YAMNet TL can be used for an automatic data cleaning, as shown by the 0.91 precision and 0.92 recall obtained by optimising for the F_1-score. This can be achieved at the lower cost of manually annotating only a small subset of the acoustic dataset. For instance, we have reached such a result by manually annotating only the 0.5 % of the whole acoustic dataset of our experiments. However, we also show that similar performances can be obtained by annotating an even smaller fraction of the acoustic dataset. Indeed, repeating the experiment with half of the recordings (36 files = 18 h of recording, in our case) yielded to identical performances. The similar performance achieved on a smaller annotated dataset demonstrates the feasibility of also applying our approach to acoustic datasets collected under diverse scenarios, such as from different regions with diverse soundscapes or using different types of acoustic recorders. In these contexts, recordings may exhibit heterogeneous sound characteristics, making it beneficial to apply our approach separately to each acoustic scenario to better capture their peculiarities. Taken together, our work points to YAMNet along with its TL specialised versions being a versatile tool for wind noise detection.

Since our main goal with this study was to identify recordings impacted by wind, a predominant challenge in our study area, we did not encompass the utilisation of YAMNet for other noise sources. We encourage further research exploring leveraging Transfer Learning techniques on the YAMNet architecture to create new methods for classifying additional noise sources. These could include categorising rain sounds or differentiating the noise produced by traffic vehicles, providing a more comprehensive and adaptable solution to environmental noise challenges.

Big data acquisition drives a rising demand within the bioacoustics community to integrate Machine Learning (ML) techniques. We demonstrate that it is possible to bridge this gap by introducing an affordable and cutting-edge methodology available even to those unfamiliar with ML. Notably, the compact size of YAMNet, attributed to its employment of the MobileNet architecture family (Howard et al., 2017; Sandler et al., 2018) eliminates the requirement for high-cost hardware, thereby ensuring accessibility for researchers employing standard laptops.

Furthermore, the promising performance of YAMNet with its low demand for computational resources, paves the way for its integration into portable devices for real-time data cleaning and analysis of soundscapes. This possibility can complement ARUs' ability to record prolonged recordings of natural habitats while pre-filtering environmental noise that may affect the downstream analysis. Recent research has made noteworthy advancements, such as implementing YAMNet sound detection in cost-effective technologies like a Raspberry Pi-connected microphone (Hyun, 2023). This forward-looking capability ensures instantaneous access to bioacoustics insights and judiciously optimises data storage by avoiding recordings in less-than-optimal environmental conditions.

To conclude, in the realm of sound detection, while previous studies have utilised YAMNet to identify target sounds (Mohammed et al., 2023; Tena et al., 2022; Hyun, 2023), our work contributes to advancing the application of machine learning in bioacoustics, representing the first instance of using YAMNet to pre-process bioacoustics data. Combining YAMNet and Transfer Learning provides a valid solution to address the challenges of windy environments, opening avenues for improved audio classification in bioacoustics and ecoacoustics studies.

Funding

F. T. was supported by a PON PhD research scholarship (REACT-EU FSE DM 1061). D.R. was partially supported by the National Geographic Society (grant no. NGS-67401R-20). D.R. and N.M. were also supported by the Centre National de la Recherche Scientifique (CNRS) of the University of Saint-Etienne, the Institut universitaire de France, and the Labex CeLyA. SANCCOB (K.L. and G.S.P.) is financially supported by a wide range of donors. L.F. was supported by the EMBRC project n° IR0000035, CUP C63C22000570001 entitled “Unlocking the Potential for health and food from the seas".

CRediT authorship contribution statement

Francesca Terranova: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lorenzo Betti:** Writing – original draft, Validation, Methodology, Investigation, Formal analysis. **Valeria Ferrario:** Writing – original draft, Validation, Methodology, Investigation. **Olivier Friard:** Writing – review & editing, Methodology, Investigation. **Katrin Ludynia:** Writing – review & editing, Supervision, Resources. **Gavin Sean Petersen:** Writing – review & editing, Investigation, Data curation. **Nicolas Mathevon:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **David Reby:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Livio Favaro:** Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization.

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.

Data availability

The data code and raw data utilised in this research article are archived on Zenodo (https://doi.org/10.5281/zenodo.11220741).

Acknowledgments

We are grateful to Cape Nature and particularly to the rangers of the Stony Point Penguin Colony for their help during the collection of the acoustic recordings. We also thank the users of the Bioacoustics Stackexchange that inspired this work, with particular acknowledgment of Mr. Jon Nordby, whose contributions served as a source of inspiration for this endeavour. We express our gratitude to Francesco Varrato for his insightful suggestions during the manuscript writing process. Finally, we thank Ines Leclerc for her continuous and invaluable support throughout the fieldwork.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2024.174868.

References

Bardeli, R., Wolff, D., Kurth, F., Koch, M., Tauchert, K.H., Frommolt, K.H., 2010. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. Pattern Recogn. Lett. 31 (12), 1524–1534. https://doi.org/10.1016/j.patrec.2009.09.014.

Bedoya, C.L., Molles, L.E., 2021. Acoustic Censusing and Individual Identification of Birds in the Wild. https://doi.org/10.1101/2021.10.29.466450.

Boersma, P., Weenink, D., 2024. Praat: Doing Phonetics by Computer [Computer Program]. Version 6.3.09. from. http://www.praat.org/.

Buxton, R.T., Jones, I.L., 2012. Measuring nocturnal seabird activity and status using acoustic recording devices: applications for island restoration. J. Field Ornithol. 83 (1), 47–60. https://doi.org/10.1111/j.1557-9263.2011.00355.x.

Christin, S., Hervet, E., Lecomte, N., 2019. Applications for deep learning in ecology. Methods Ecol. Evol. 10 (10), 1632–1644. https://doi.org/10.1111/2041-210X.13256.

Cohen, J., 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol. Bull. 70 (4), 213. https://doi.org/10.1037/h0026256.

Depraetere, M., Pavoine, S., Jiguet, F., Gasc, A., Duvail, S., Sueur, J., 2012. Monitoring animal diversity using acoustic indices: implementation in a temperate woodland. Ecol. Indic. 13 (1), 46–54. https://doi.org/10.1016/j.ecolind.2011.05.006.

Desjonquères, C., Rybak, F., Castella, E., Llusia, D., Sueur, J., 2018. Acoustic communities reflects lateral hydrological connectivity in riverine floodplain similarly to macroinvertebrate communities. Sci. Rep. 8 (1), 14387. https://doi.org/10.1038/s41598-018-31798-4.

Digby, A., Towsey, M., Bell, B.D., Teal, P.D., 2013. A practical comparison of manual and autonomous methods for acoustic monitoring. Methods Ecol. Evol. 4 (7), 675–683. https://doi.org/10.1111/2041-210X.12060.

Dufourq, E., Batist, C., Foquet, R., Durbach, I., 2022. Passive acoustic monitoring of animal populations with transfer learning. Eco. Inform. 70, 101688 https://doi.org/10.1016/j.ecoinf.2022.101688.

Dunn, O.J., 1964. Multiple comparisons using rank sums. Technometrics 6 (3), 241–252. https://doi.org/10.1080/00401706.1964.10490181.

Eldridge, A., Guyot, P., Moscoso, P., Johnston, A., Eyre-Walker, Y., Peck, M., 2018. Sounding out ecoacoustic metrics: Avian species richness is predicted by acoustic indices in temperate but not tropical habitats. Ecol. Indic. 95, 939–952. https://doi.org/10.1016/j.ecolind.2018.06.012.

Fairbrass, A.J., Rennert, P., Williams, C., Titheridge, H., Jones, K.E., 2017. Biases of acoustic indices measuring biodiversity in urban areas. Ecol. Indic. 83, 169–177. https://doi.org/10.1016/j.ecolind.2017.07.064.

Fairbrass, A.J., Firman, M., Williams, C., Brostow, G.J., Titheridge, H., Jones, K.E., 2018. CityNet—deep learning tools for urban ecoacoustic assessment. Methods Ecol. Evol. 10 (2), 186–197. https://doi.org/10.1111/2041-210x.13114.

Farina, A., 2018. Ecoacoustics: a quantitative approach to investigate the ecological role of environmental sounds. Mathematics 7 (1), 21. https://doi.org/10.3390/math7010021.

Favaro, L., Ozella, L., Pessani, D., 2014. The vocal repertoire of the African penguin (*Spheniscus demersus*): structure and function of calls. PLoS One 9 (7), e103460. https://doi.org/10.1371/journal.pone.0103460.

Favaro, L., Cresta, E., Friard, O., Ludynia, K., Mathevon, N., Pichegru, L., Reby, D., Gamba, M., 2021. Passive acoustic monitoring of the endangered African Penguin (*Spheniscus demersus*) using autonomous recording units and ecoacoustic indices. Ibis 163 (4), 1472–1480. https://doi.org/10.1111/ibi.12970.

Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M., 2017. Audio set: an ontology and human-labeled dataset for audio events. In: IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 776–780. https://doi.org/10.1109/ICASSP.2017.7952261.

Gillam, E.H., McCracken, G.F., Westbrook, J.K., Lee, Y.-F., Jensen, M.L., Balsley, B.B., 2009. Bats aloft: variability in echolocation call structure at high altitudes. Behav. Ecol. Sociobiol. 64 (1), 69–79. https://doi.org/10.1007/s00265-009-0819-1.

Hacker, F., Terranova, F., Petersen, G.S., Tourtigues, E., Friard, O., Gamba, M., Ludynia, K., Gridley, T., Pichegru, L., Mathevon, N., Reby, D., Favaro, L., 2023. Effect of environmental variables on African penguin vocal activity: implications for acoustic censusing. Biology 12 (9), 1191. https://doi.org/10.3390/biology12091191.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. ArXiv, abs/1704.04861. https://doi.org/10.48550/arXiv.1704.04861.

Hyun, S.H., 2023. Sound-event detection of water-usage activities using transfer learning. Sensors 24 (1), 22. https://doi.org/10.3390/s24010022.

Jadoul, Y., Thompson, B., De Boer, B., 2018. Introducing parselmouth: a python interface to praat. J. Phon. 71, 1–15. https://doi.org/10.1016/j.wocn.2018.07.001.

Juodakis, J., Marsland, S., 2022. Wind-robust sound event detection and denoising for bioacoustics. Methods Ecol. Evol. 13 (9), 2005–2017. https://doi.org/10.1111/2041-210x.13928.

Kahl, S., Wood, C.M., Eibl, M., Klinck, H., 2021. BirdNET: a deep learning solution for avian diversity monitoring. Eco. Inform. 61, 101236 https://doi.org/10.1016/j.ecoinf.2021.101236.

Kingma, D., Ba, J., 2014. Adam: A Method for Stochastic Optimization. ArXiv preprint arXiv, 1412, p. 6980. https://doi.org/10.48550/arXiv.1412.6980.

Knight, E.C., Hannah, K.C., Foley, G.J., Scott, C.D., Brigham, R.M., Bayne, E., 2017. Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. Avian Conservation and Ecology 12 (2). https://doi.org/10.5751/ace-01114-120214.

Kruskal, W.H., Wallis, W.A., 1952. Use of ranks in one-criterion variance analysis. J. Am. Stat. Assoc. 47, 583–621. https://doi.org/10.1080/01621459.1952.10483.

LeBien, J., Zhong, M., Campos-Cerqueira, M., Velev, J.P., Dodhia, R., Ferres, J.L., Aide, T. M., 2020. A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. Eco. Inform. 59, 101113 https://doi.org/10.1016/j.ecoinf.2020.101113.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. https://doi.org/10.1038/nature1453.

Lu, T., Han, B., Yu, F., 2021. Detection and classification of marine mammal sounds using AlexNet with transfer learning. Eco. Inform. 62, 101277 https://doi.org/10.1016/j.ecoinf.2021.101277.

McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochem. Med. 276-282 https://doi.org/10.11613/bm.2012.031.

Metcalf, O.C., Barlow, J., Devenish, C., Marsden, S., Berenguer, E., Lees, A.C., 2021. Acoustic indices perform better when applied at ecologically meaningful time and frequency scales. Methods Ecol. Evol. 12 (3), 421–431. https://doi.org/10.1111/2041-210X.13521.

Mohammed, K.K., El-Latif, E.I.A., El-Sayad, N.E., Darwish, A., Hassanien, A.E., 2023. Radio frequency fingerprint-based drone identification and classification using Mel spectrograms and pre-trained YAMNet neural. Internet of Things 23, 100879. https://doi.org/10.1016/j.iot.2023.100879.

Morgan, N., Bourlard, H., 1989. Generalization and parameter estimation in feedforward nets: some experiments. In: Touretzky, D. (Ed.), Advances in Neural Information Processing Systems, Vol. 2, pp. 630–670.

Nelke, C.M., 2016. Wind Noise Reduction: Signal Processing Concepts. Wissenschaftsverlag Mainz.

Norman, D.L., Bischoff, P.H., Wearn, O.R., Ewers, R.M., Rowcliffe, J.M., Evans, B., Sethi, S., Chapman, P.M., Freeman, R., 2022. Can CNN-based species classification generalise across variation in habitat within a camera trap survey? Methods Ecol. Evol. 14 (1), 242–251. https://doi.org/10.1111/2041-210x.14031.

Oswald, J.N., Van Cise, A.M., Dassow, A., Elliott, T., Johnson, M.T., Ravignani, A., Podos, J., 2022. A collection of best practices for the collection and analysis of bioacoustic data. Appl. Sci. 12 (23), 12046. https://doi.org/10.3390/app122312046.

Pérez-Granados, C., Traba, J., 2021. Estimating bird density using passive acoustic monitoring: a review of methods and suggestions for further research. Ibis 163 (3), 765–783. https://doi.org/10.1111/ibi.12944.

Pijanowski, B.C., Villanueva-Rivera, L.J., Dumyahn, S.L., Farina, A., Krause, B.L., Napoletano, B.M., Gage, S.H., Pieretti, N., 2011. Soundscape ecology: the science of sound in the landscape. BioScience 61 (3), 203–216. https://doi.org/10.1525/bio.2011.61.3.6.

Priyadarshani, N., Castro, I., Marsland, S., 2018. The impact of environmental factors in birdsong acquisition using automated recorders. Ecol. Evol. 8 (10), 5016–5033. https://doi.org/10.1002/ece3.3889.

Quinn, C.A., Burns, P., Gill, G., Baligar, S., Snyder, R.L., Salas, L., Goetz, S.J., Clark, M.L., 2022. Soundscape classification with convolutional neural networks reveals temporal and geographic patterns in ecoacoustic data. Ecol. Indic. 138, 108831 https://doi.org/10.1016/j.ecolind.2022.108831.

Ravaglia, D., Ferrario, V., De Gregorio, C., Carugati, F., Raimondi, T., Cristiano, W., Torti, V., Von Hardenberg, A., Ratzymbazafy, J., Valente, D., Giacoma, C., Gamba, M., 2023. There you are! Automated detection of Indris' songs on features extracted from passive acoustic recordings. Animals 13 (2), 241. https://doi.org/10.3390/ani13020241.

Rhinehart, T.A., Chronister, L.M., Devlin, T., Kitzes, J., 2020. Acoustic localization of terrestrial wildlife: current practices and future opportunities. Ecol. Evol. 10 (13), 6794–6818. https://doi.org/10.1002/ece3.6216.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40 (8), 913–929. https://doi.org/10.1111/ecog.02881.

Ross, S.R.P.J., Friedman, N.R., Yoshimura, M., Yoshida, T., Donohue, I., Economo, E.P., 2021. Utility of acoustic indices for ecological monitoring in complex sonic environments. Ecol. Indic. 121, 107114 https://doi.org/10.1016/j.ecolind.2020.107114.

Ross, S.R.P.J., O'Connell, D.P., Deichmann, J.L., Desjonquères, C., Gasc, A., Phillips, J.N., Sethi, S.S., Wood, C.M., Burivalova, Z., 2023. Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions. Funct. Ecol. 37 (4), 959–975. https://doi.org/10.1111/1365-2435.14275.

Ruff, Z.J., Lesmeister, D.B., Appel, C.L., Sullivan, C.M., 2021. Workflow and convolutional neural network for automated identification of animal sounds. Ecol. Indic. 124, 107419 https://doi.org/10.1016/j.ecolind.2021.107419.

Salamon, J., Bello, J.P., Farnsworth, A., Robbins, M., Keen, S., Klinck, H., Kelling, S., 2016. Towards the automatic classification of avian flight calls for bioacoustic monitoring. PLoS One 11, e0166866. https://doi.org/10.1371/journal.pone.0166866.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C., 2018. MobileNetV2: inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.2018.00474.

Sokolova, M., Japkowicz, N., Szpakowicz., S., 2006. Beyond accuracy, F-score, and ROC: a family of discriminant measures for performance evaluation. In: Sattar, A., Kang, B. (Eds.), Advances in Artificial Intelligence, Lecture Notes in Computer Science, vol. 4304. Springer, pp. 1015–1021. https://doi.org/10.1007/11941439.

Stowell, D., Wood, M.D., Pamuła, H., Stylianou, Y., Glotin, H., 2018. Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. Methods Ecol. Evol. 10, 368–380. https://doi.org/10.1111/2041-210X.13103.

Sugai, L.S.M., Silva, T.S.F., Ribeiro Jr., J.W., Llusia, D., 2019. Terrestrial passive acoustic monitoring: review and perspectives. BioScience 69 (1), 15–25. https://doi.org/10.1093/biosci/biy147.

Szymański, P., Olszowiak, K., Wheeldon, A., Budka, M., Osiejuk, T.S., 2021. Passive acoustic monitoring gives new insight into year-round duetting behaviour of a tropical songbird. Ecol. Indic. 122, 107271 https://doi.org/10.1016/j.ecolind.2020.107271.

Tena, A., Clarià, F., Solsona, F., 2022. Automated detection of COVID-19 cough. Biomedical Signal Processing and Control 71, 103175. https://doi.org/10.1016/j.bspc.2021.103175.

Tharwat, A., 2020. Classification assessment methods. Applied Computing and Informatics 17 (1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003.

Trapanotto, M., Nanni, L., Brahnam, S., Guo, X., 2022. Convolutional neural networks for the identification of African lions from individual vocalizations. Journal of Imaging 8 (4), 96. https://doi.org/10.3390/jimaging8040096.

Tsalera, E., Papadakis, A., Samarakou, M., 2021. Comparison of pre-trained CNNs for audio classification using transfer learning. J. Sens. Actuator Netw. 10, 72. https://doi.org/10.3390/jsan10040072.

Van Rossum, G., Drake Jr., F.L., 1995. Python Reference Manual. Centrum voor Wiskunde en Informatica Amsterdam.

Walker, K.T., Hedlin, M.A., 2009. A Review of Wind-noise Reduction Methodologies. Infrasound Monitoring for Atmospheric Studies, 141-182. https://doi.org/10.1007/978-1-4020-9508-5_5.

Willacy, R.J., Mahony, M., Newell, D.A., 2015. If a frog calls in the forest: bioacoustic monitoring reveals the breeding phenology of the endangered Richmond range mountain frog (*Philoria richmondensis*). Austral Ecol. 40 (6), 625–633. https://doi.org/10.1111/aec.12228.

Zhang, Z., Geiger, J., Pohjalainen, J., Mousa, A.E.D., Jin, W., Schuller, B., 2018. Deep learning for environmentally robust speech recognition: an overview of recent developments. ACM Transactions on Intelligent Systems and Technology (TIST) 9 (5), 1–28. https://doi.org/10.48550/arXiv.1705.10874.

Zhao, K., Chen, G., Liu, Y., Møller, A.P., Zhang, Y., 2022. Population size assessment of Adélie penguin (Pygoscelis adeliae) chicks based on vocal activity rate index. Global Ecology and Conservation 38, e02263. https://doi.org/10.1016/j.gecco.2022.e02263.

Zhong, M., LeBien, J., Campos-Cerqueira, M., Dodhia, R., Lavista Ferres, J., Velev, J., Aide, T.M., 2020. Multispecies bioacoustics classification using transfer learning of deep convolutional neural networks with pseudo-labeling. J. Acoust. Soc. Am. 148 (4_Supplement), 2442. https://doi.org/10.1121/1.5146738.

Znidersic, E., Towsey, M.W., Hand, C., Watson, D.M., 2021. Eastern Black Rail detection using semi-automated analysis of long-duration acoustic recordings. Avian Conservation and Ecology 16 (1). https://doi.org/10.5751/ace-01773-160109.

Zwart, M.C., Baker, A., McGowan, P.J.K., Whittingham, M.J., 2014. The use of automated bioacoustic recorders to replace human wildlife surveys: an example using nightjars. PLoS One 9 (7), e102770. https://doi.org/10.1371/journal.pone.0102770.