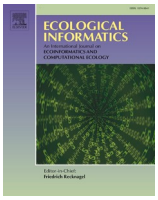
Ecological Informatics 90 (2025) 103222

Contents lists available at ScienceDirect

Ecological Informatics

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



Automated note annotation after bioacoustic classification: Unsupervised clustering of extracted acoustic features improves detection of a cryptic owl

Callan Alexander [a,*], Robert Clemens [b], Paul Roe [c], Susan Fuller [a]

[a] School of Biology & Environmental Science, Queensland University of Technology, 2 George Street, Brisbane, QLD 4000, Australia
[b] Research, Specialised and Data Foundations, Digital Solutions, Griffith University, 170 Kessels Rd, Nathan QLD 4111, Australia
[c] School of Computer Science, Queensland University of Technology, 2 George St, Brisbane, QLD 4000, Australia

ARTICLE INFO

Keywords:
Bioacoustics
UMAP
HDBSCAN
Powerful owl
Machine learning

ABSTRACT

Passive acoustic monitoring and machine learning are increasingly being used to survey threatened species. When automated detection models are applied to large novel datasets, false-positive detections are likely even for high-performing models, and arbitrary thresholds may result in missed detections. Manual validation of outputs is time consuming, and additional fine-scale annotation of individual notes is impractical for large datasets and difficult to automate when using passive field recordings. This research presents an acoustic monitoring pipeline which employs a multi-stage hybrid approach: initial detection using a convolutional neural network classifier, followed by segmentation and iterative unsupervised clustering of extracted acoustic features using UMAP and HDBSCAN to remove label noise. We applied the pipeline to a large acoustic dataset comprised of 2764 h of environmental recordings and test the utility of the approach on territorial calls of Australia's largest owl: the threatened Powerful Owl (*Ninox strenua*). The pipeline reduced the large acoustic dataset into 10,116 annotations, of which 9399 (93 %) were correctly annotated individual notes of the target species. The clustering process also eliminated 88 % of false positive detections while retaining 95 % true positives (F1 = 0.94). The approach is highly scalable, can be applied to very large acoustic datasets, and can rapidly collect note-level annotations from noisy field recordings. The acoustic features derived from this methodology identified population differences in our test dataset and enable further exploration of song structure, geographic variation, and vocal individuality. The clustering process also facilitates a semi-supervised learning approach, allowing rapid selection of uncertain examples for model improvement. The pipeline helps to address two key challenges in bioacoustic monitoring: the need for manual validation of automated detections and the difficulty of obtaining accurate note-level annotations in noisy field recordings. Adaptation of these methods to other species and vocalisations may facilitate improved detection and investigation of vocal characteristics across different populations or regions.

1. Introduction

Passive acoustic monitoring and machine learning are becoming essential tools for ecological research and conservation (Manzano-Rubio et al., 2022; Shonfield and Bayne, 2017a; Stowell, 2022; Teixeira et al., 2019). One of the most promising applications of this technology is the monitoring of vocal cryptic species, particularly for rare or nocturnal species, and those that inhabit remote areas (Duchac et al., 2020; Picciulin et al., 2019; Shonfield et al., 2018; Wood et al., 2019, 2024; Yan et al., 2019). Machine learning approaches for automated detection have become widely available, and tools like BirdNET or Google Perch are increasingly being applied to large datasets (Ghani et al., 2023; Kahl et al., 2021; Manzano-Rubio et al., 2022; Stowell, 2022; Znidersic et al., 2020). Even high performing models are still likely to exhibit false-positive detections when applied to large datasets and the use of arbitrary thresholds can result in missed detections and biased data (Lostanlen et al., 2019; Navine et al., 2024; Pérez-Granados, 2023). As such, a limiting factor of automated acoustic detection is the need for an expert to validate the output. Birdsong studies also often require vocalisations to be annotated in fine detail to allow elucidation of species-specific behaviours, population dynamics, or even individual vocal signatures (Backhouse et al., 2021; Kershenbaum et al., 2016;

* Corresponding author.
E-mail address: callan.alexander@gmail.com (C. Alexander).

https://doi.org/10.1016/j.ecoinf.2025.103222
Received 3 December 2024; Received in revised form 22 May 2025; Accepted 23 May 2025
Available online 25 May 2025

Terry et al., 2005). As the use of acoustic monitoring expands, so too does the need for efficient methods to handle the increasing volume of acoustic data without sacrificing accuracy or detail. Manual validation and annotation are time-consuming and resource-intensive, rendering large datasets impractical for detailed manual scrutiny (Shaw et al., 2022).

Birdsong analyses have greatly benefited from advanced computational techniques, with unsupervised classification emerging as a useful tool. UMAP (Uniform Manifold Approximation and Projection) and HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) are becoming recognised as powerful tools for birdsong analysis. UMAP is a dimensionality reduction algorithm that can effectively capture the complex, non-linear relationships often present in birdsong data, allowing researchers to visualise high-dimensional acoustic features in a lower-dimensional space (McInnes et al., 2020). HDBSCAN is a clustering algorithm that excels at identifying meaningful groups within the data, even when these groups have varying densities or irregular shapes (Campello et al., 2015; McInnes et al., 2017). When used together, these techniques can reveal patterns in birdsong repertoires, help identify distinct song types and assist in tracking changes in vocalisations across populations and individuals (Best et al., 2023; Bravo Sanchez et al., 2024; Zhao et al., 2023). Studies are increasingly demonstrating the effectiveness of combining UMAP and HDBSCAN for birdsong analysis. UMAP in particular has been shown to substantially improve the performance of clustering algorithms (Allaoui et al., 2020). Sainburg et al. (2020) noted that UMAP effectively produces "more meaningful data representations" across natural science datasets when compared with other counterparts such as t-distributed stochastic neighbor embedding. This study also reported that HDBSCAN clusters most closely matched human labelling. Similar reports have been found in multiple studies applying HDBSCAN to UMAP projections (Best et al., 2023; Blanco-Portals et al., 2022; Koch et al., 2024).

The automated segmentation and annotation of birdsong has advanced significantly in recent years, driven by the application of machine learning techniques and the development of specialised software tools. Traditional methods rely heavily on manual spectrogram analysis, but these have gradually been supplanted by more efficient automated approaches (Neal et al., 2011). Improved segmentation methods and the introduction of deep learning models has markedly improved the accuracy and efficiency of birdsong annotation (Cohen et al., 2022). Notable contributions include the *scikit-maad* package, *TweetyNET*, *Deep Audio Segmenter* and *AVN* (Cohen et al., 2022; Koch et al., 2024; Steinfath et al., 2021; Ulloa et al., 2021). Despite advancements, challenges persist in areas such as generalisability across species, robustness to background noise, and adaptation to varying recording conditions (Wood et al., 2023). Segmentation and annotation pipelines based on deep neural networks have already been shown to work well in laboratory settings, but often struggle to translate effectively to field conditions and natural populations (Coffey et al., 2019; Cohen et al., 2022; Recalde, 2023; Steinfath et al., 2021). 'Label noise' is a notable issue for segmentation methodologies, which refers to when a sound other than the sound of interest is annotated (Denton et al., 2022; Henkel et al., 2021). Michaud et al. (2023) combat this issue by segmenting the sound, computing the acoustic features of each sound unit, and then applying an unsupervised DBSCAN algorithm. This approach has been applied to Xeno Canto data and demonstrated significant reduction in the initial label noise present in the dataset but the authors note that "further developments are still required to adapt such facilities to soundscape recordings where the sounds of interest of several species are mixed". There are relatively few studies using similar approaches and automatically extracting acoustic features from field data. Denton et al. (2022) demonstrate the utility of a hybrid approach, applying an unsupervised sound separation approach prior to bird classification. Deep embeddings from neural network outputs are also increasingly being used in unsupervised approaches following birdsong classification, and are used to facilitate active learning, remove false-positives and improve datasets (Bravo Sanchez et al., 2024; McGinn et al., 2023; Tolkova et al., 2021). High background noise in recordings, however, can impact the quality of unsupervised clustering methods and it has been suggested that signal-aware methods for reducing noise before projecting the data could be beneficial (Sainburg et al., 2020). Stowell (2022) notes that the use of deep learning to drive clustering is not heavily studied.

In this study, we focus on detecting the calls of the Powerful Owl (*Ninox strenua*), a key predator species in Australia. Owls are an ideal candidate for passive acoustic monitoring. They occur in low densities, are highly cryptic, and are difficult to locate in the wild (Duchac et al., 2020; Johnsgard, 1988). Powerful Owls are the largest owl found in Australia, are highly cryptic and can take up to 20 visits using traditional survey approaches to detect (Loyn et al., 2001). They are found in eucalypt forests along the east coast of Australia and generally roost in dense riparian forest vegetation (Bradsworth et al., 2017). They are listed as threatened in three Australian states and are highly reliant on old-growth trees for breeding hollows (Fauna and Flora Guarantee Act, 1988 (Vic), *Queensland Nature Conservation Act 1992* (Qld), *Biodiversity Conservation Act 2016* (NSW)). The typical adult vocalisation is a 'double hoot' which typically occurs between 200 and 600 Hz (Fig. 1; Alexander, 2022).

Passive acoustic monitoring provides an alternative survey methodology that is non-invasive and has the potential to significantly reduce survey efforts, requiring far fewer trips to a site to obtain multiple nights of data. Bioacoustic recorders have been applied successfully for large-scale owl monitoring programs, and are even being used for invasion surveillance in instances where owl populations are expanding in an undesired manner threatening other native fauna (Rognan et al., 2012; Wood et al., 2019, 2024). Owls have been suggested as a taxon that may exhibit vocal individuality, with evidence of individually distinct vocalisations (Madhavan and Linhart, 2024). It is possible that vocal individuality and geographic variation studies could allow individual owls or owls from certain locations to be automatically detected in passive audio (Grava et al., 2008; Tseng et al., 2024).

In this paper, we present a machine learning pipeline that addresses the challenges of extracting and annotating vocalisations from noisy field recordings. Our research explores the integration of neural network classification with segmentation, acoustic feature extraction and unsupervised clustering as a post-processing approach. This methodology aims to serve two key purposes: first, to efficiently filter out false-positive detections from classifier output (and equally to investigate potential missed detections 'underneath' the chosen threshold) and second, to enable rapid, accurate extraction of individual note annotations from field recordings. This approach has broader implications for acoustic monitoring of other cryptic species, potentially offering a more scalable solution for analysing large acoustic datasets and rapidly collecting note annotations for geographic call variation or vocal individuality studies.

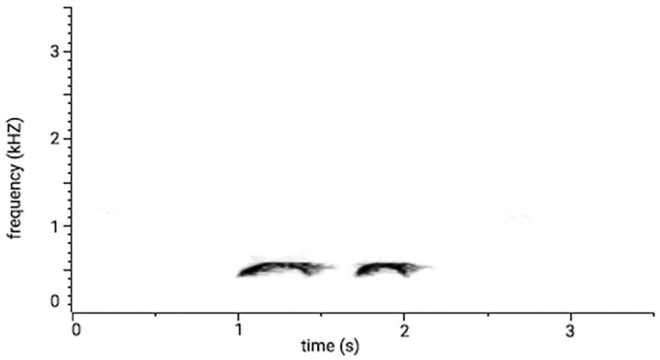


Fig. 1. Typical spectrogram of a Powerful Owl Territorial Call.

2. Materials and methods

2.1. Data acquisition and pre-processing

Seven study sites in south-east Queensland, Australia (Fig. 2), were selected based on historical records and survey data provided by Birdlife Australia's 'Powerful Owl Citizen Science Program'. All sites were located in dry sclerophyll forest and most sites featured riparian habitat sections with old growth forest in the vicinity. Initial listening surveys were conducted to determine approximate roosting locations of the owls where possible. These surveys were undertaken in late March and early April, periods which are known to have high calling activity at dawn and dusk (Debus, 1995). These surveys were conducted approximately one hour prior to dusk and continued until an owl vocalisation was heard, or until approximately one hour after dusk. If owl calling was detected at a site, daytime visits were conducted to ascertain the roost location. This was determined by walking and searching for whitewash (owl faeces) and pellets. All surveys were conducted by the same observer.

Two acoustic recorders (Audiomoth v1.0, Hill et al., 2018, Songmeter SM2) were placed together within a potential Powerful Owl territory (see Fig. 2). The recorders were strapped to a tree at head height. The devices were programmed to record from one hour before dusk until dawn and produced .wav files of one-hour duration. The Audiomoths were programmed to record at 48 kHz at medium-high gain. The Songmeters were set to 22.1 kHz and default gain settings. Two recorders were used to function as a backup in case of failure, and to provide a mixture of sample rates and noise-floors. The recorders were placed within 100 m of the most recently located roost spot where possible. If owls could not be located, the recorders were deployed near the last known roosting location, or near fresh whitewash, or failing either of those options were placed in likely habitat. This was repeated at seven study sites, batteries and SD cards were replaced every three weeks.

2.2. Manual processing

Manual processing of the audio data was undertaken to identify Powerful Owl vocalisations and build an initial training dataset. 100 h of dawn and dusk recordings from multiple sites were aurally verified using random sampling. One-hour files from each site were selected and manually annotated in Raven Pro v1.6.1 (Charif et al., 2010) and 5 s audio snippets were generated for each annotation. The recording time before the annotation begins was randomised within the snippets, to avoid the vocalisation being at the exact start of each file. Random 5 s snippets of negative data (not containing owl vocalisations) were also selected from recordings not containing any owl vocalisations.

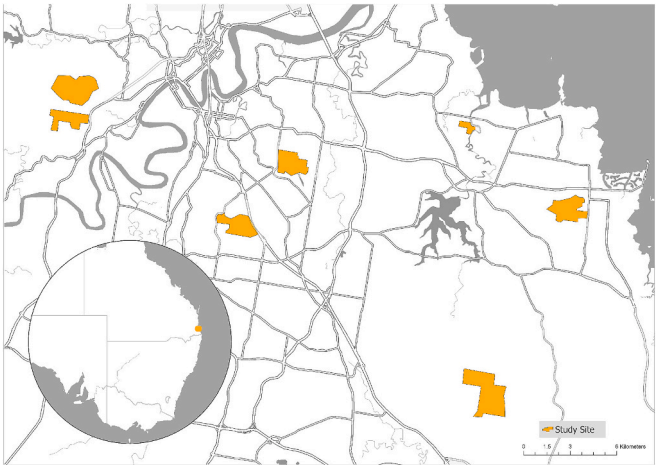


Fig. 2. Map of study locations in South-east Queensland.

Annotated files and negative snippets were transformed into 5 s spectrograms by applying a Fast Fourier Transform (FFT) with a window size of 4096 and a hop length of 512, which were then mapped onto the mel scale. These spectrograms were subsequently enhanced for brightness and contrast and resized to 224 × 224 pixels.

2.3. Model configuration and training

The MobileNetV2 architecture, initially pre-trained on the ImageNet dataset, was adapted for binary classification of the owl vocalisations. The model was optimised using the Adam optimiser with a learning rate of 0.0005, trained for 50 epochs with binary cross-entropy loss. Data augmentation techniques were employed to enhance the generalisability of the model including time and frequency shifting (see supplementary materials for full code). An iterative process was applied, similar to that used in Eichinski et al. (2022), whereby several versions of the model were produced using different subsets of the training data and collecting more negative examples until training performance was satisfactory. The initial model consisted of 4003 positive examples and 4173 negative examples. The model was evaluated and tested on two datasets. The first consisted of ~10 % of the training data that was randomly excluded from training (half positive and half negative examples). The second dataset was intended to replicate field conditions and comprised 10 h of field audio. This audio contained 5-min chunks of randomly selected audio from a separate Powerful Owl study. These recordings were taken from the same locations, but at different recording points and at a different time of year (Spring 2020). This dataset contained 263 vocalisations, many of them faint and obscured by road, wind and rain noise.

2.4. Inference

The trained model made predictions on 2764 h of acoustic data spanning seven sites. The number of hours per site differed slightly due to battery variability but all recordings were taken from April and May 2020 between dusk and dawn (Table 1). All recordings were from the SM2 recorders (largely due to longer battery life resulting in more hours available) except for the 'whites hill' site where Audiomoth recordings were also used due to a SM2 recorder failure.

Hour-long audio recordings were processed using the Python *librosa* library (McFee et al., 2015). Each recording was divided into 5 min chunks, which were further segmented into 5 s intervals with a 2.5 s overlap. Mel-spectrograms were generated for each segment using a 2048-sample window size and 256 mel bands. The spectrograms were converted into 224 × 224 pixel images and were then input into the trained model for classification. Predictions were made in batches, with each segment receiving a score indicating the confidence of the model in the presence of an owl call. A sigmoid activation function was applied at the end of the network to produce an uncalibrated confidence score between 0 and 1 (see Wood and Kahl, 2024).

2.5. Manual validation of output

All classified outputs with a confidence greater than 0.1 were extracted. This was deemed to be the lowest confidence score for the dataset at which manual ground-truthing of all detections was

Table 1
Number of recording hours per site.

Subfolder	# days	hours
chapelhill	34	452
hilliards	34	439
slaughter	22	291
tingalpa	31	417
toohey	29	389
venman	31	411
whiteshill	31	365

logistically feasible. Each of these segments was exported as a 5 s .wav file containing the site name, data, time and confidence score in the file name. The segments were then manually verified in Raven Pro (approx. 14 h of recordings). The segments were verified by importing all segments concurrently into Raven Pro and annotating all the false-positive segments with the 'Begin File' measurement included in the annotation table. The resulting .csv file then contained a list of incorrectly classified .wav files. The files are then programmatically sorted into true-positive and false-positive folders. This provides a 'segment ground-truth' dataset, detailing which outputs from the classifier were correctly classified across all confidence thresholds.

2.6. Feature extraction and clustering with ROI segmentation

The *sci-kit maad* Python package (Ulloa et al., 2021) was used to automate a 'region of interest' (ROI) segmentation process to isolate relevant acoustic features from the background noise within the outputs of the binary owl classifier. This process automatically tagged any vocalisations or other noise in segments above a selected confidence level. This was achieved by first applying a low-pass filter to the audio at 1000hz, removing the background of the image using a median filtering approach, followed by applying a binary mask that thresholds the spectrogram based on relative energy levels. Identified ROIs were then processed to extract 31 acoustic features, including frequency, duration and shape attributes.

2.7. ROI or segment label

For clarity we will define 'segment' vs 'ROI. 'Segment' refers to the entire five second .wav file and corresponding spectrogram that has been assigned a confidence score by the classifier. 'ROI' refers to the region of interest extracted by the segmentation process. Each segment may contain multiple ROIs. Each ROI consists of a value for each of the extracted acoustic features (see Figs. 3 and 7 for ROI examples).

2.8. Dimensionality reduction and clustering

UMAP was chosen for dimensionality reduction, combined with HDBSCAN for unsupervised clustering (McInnes et al., 2017; 2018). An initial clustering grid-search was conducted on the ROI dataset using both silhouette score and DBCV to broadly tune the UMAP and HDBSCAN clustering parameters. After high-scoring parameters were found, adjustments were made alongside manual inspection of datapoints to determine final cluster settings. All clustering was undertaken on the 31 extracted acoustic features, with each ROI represented as a single point on the projection.

2.9. Manual validation

100 segments from each cluster were manually inspected using a simple cluster-inspection tool GUI (available in the supplementary materials) which allows the user to view and listen to segments from a selected cluster. Clusters were labelled either positive or negative depending on whether they contained primarily (>50 %) owl vocalisation ROIs (see Fig. 3). The threshold and number of ROIS to validate was arbitrarily selected and could be adjusted in future work, however in practice for this dataset each cluster contained very high percentages of either TP or FP segments.

Segments were only retained if they contained at least one ROI in a positive cluster. Any segments containing only ROIs from negative clusters were removed. This removal was then compared with the 'segment ground-truth' dataset, to determine how accurately true and false positive detections were separated by the clustering process. After the negative clusters were removed, a second iteration of clustering was conducted on the remaining ROIs (see Fig. 4).

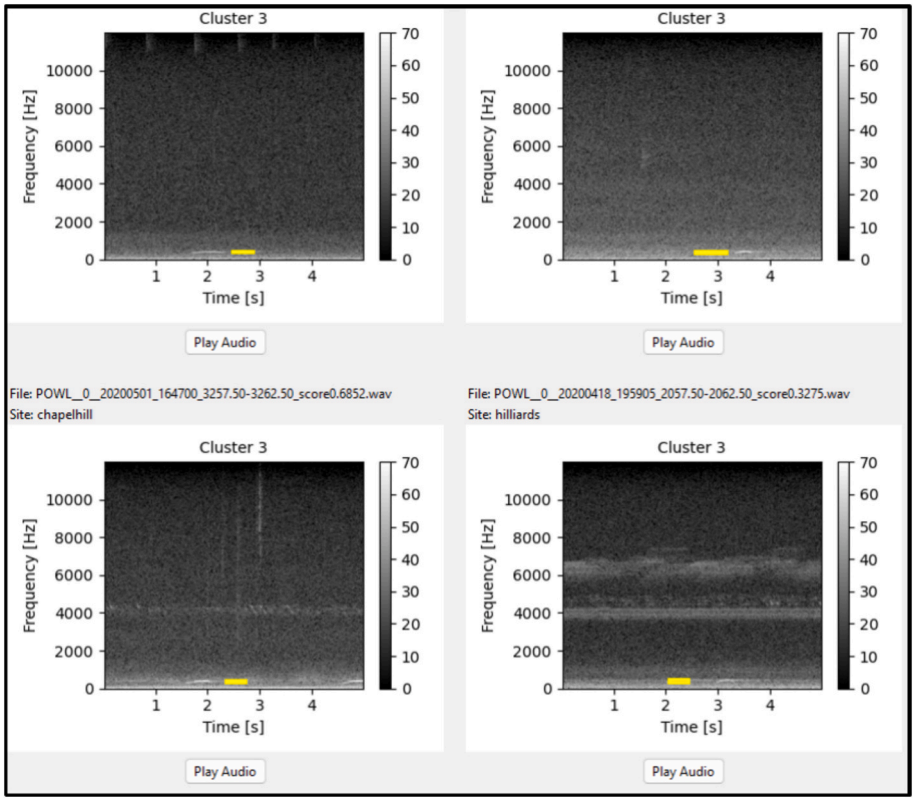


Fig. 3. Example of cluster validation interface: yellow annotation shows the boundary of a ROI. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

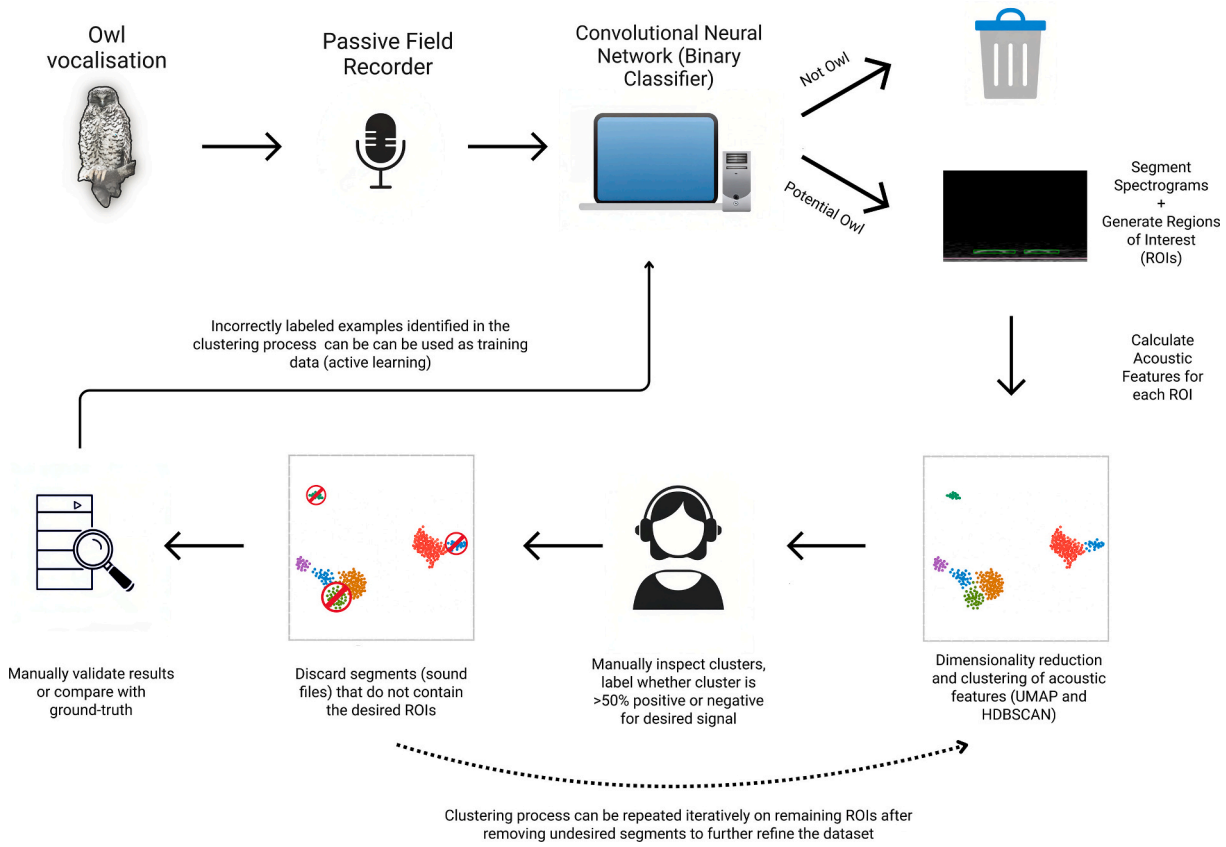


Fig. 4. Diagram depicting field recording iterative clustering process.

2.10. ROI ground truthing and iterative clustering

An 'ROI ground truth' dataset was developed after the segments were removed in the first round of clustering. The remaining 13,099 ROIs were manually verified by an expert labelling either as 'First Note', 'Second Note' 'Double Note' (both notes tagged together in one ROI instead of separately), 'FP' (false-positive) or 'Boobook' (*Ninox boobook*, a species with a similar vocalisation). As such, every ROI remaining in the dataset was validated with a ground truth to allow for comparison to additional clustering iterations. The UMAP and HDBSCAN clustering was then repeated and an additional manual inspection of clusters was undertaken. Segments without positive clusters (>50 % owl vocalisation ROIs) were again removed. The number of ground-truthed ROIs was then compared before and after clustering. UMAP and HDBSCAN clustering was then repeated for a third and final time on the remaining ROIs.

2.11. Semi-supervised learning and model testing

Improvements were then made to the initial model using a semi-supervised approach, using the clustered data to rapidly select low-confidence detections and high confidence false-positives and returning them to the training data. This was conducted by selecting all negative clusters (known to contain >50 % false positive files) and sorting by confidence score to select the most confident detections. The same approach was conducted for clusters containing >50 % true positive files and locating the files with the lowest confidence score. This was repeated until 1000 low-confidence (<0.3) positive and 1000 high confidence (>0.8) negative examples were selected. These examples were manually verified using the same approach as segment validation and sorted into TP or FP folders. All training parameters were kept the same and the model was retrained with the additional data included in the training dataset. The initial model and subsequent model were evaluated on two datasets.

3. Results

3.1. Manual validation and segmentation

After inference using the 2764 h dataset, the binary classifier labelled 11,475 segments (0.28 %) as potential owl vocalisations with a confidence >0.1. These segments were manually validated to form the

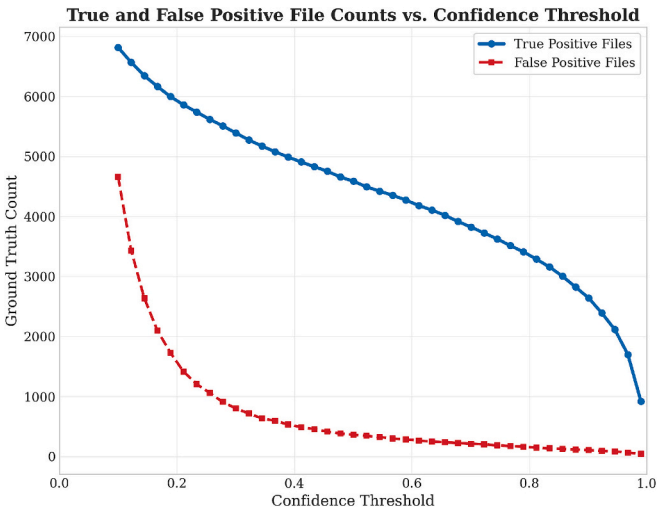


Fig. 5. Number of verified true vs false positive owl call detections at each confidence score.

'segment ground-truth'. 6817 of these segments were identified as true-positives and 4658 were false-positives. Fig. 5 details the number of true vs false-positive ground truths at each confidence score. After application of threshold segmentation 28,215 ROIs were extracted from the classified segments.

3.2. Dimensionality reduction and unsupervised classification

Manual validation of the cluster analysis using UMAP and HDBSCAN on the 28,215 ROIs revealed three clusters (9, 10 and 11) that consisted of >50 % owl vocalisations, and the remaining clusters were primarily false-positive detections (see Fig. 6). This manual inspection took an expert observer approximately 1.5 h. Fig. 7 demonstrates how ROIs appear in a spectrogram and are assigned to different clusters.

All clusters apart from 9, 10 and 11 were removed from the dataset. The remaining segments (those containing at least one ROI from a positive cluster) were then compared with the 'segment ground-truth' (see Table 2). This clustering step reduced false positives by 88 % (4658 to 550) while retaining 6479 of the 6817 validated true positives. This method retrieved 775 more owl vocalisations than the classifier operating at its optimal confidence threshold with less false-positive detections (F1 = 0.94 vs 0.83, Table 2).

The clustering process (Fig. 6) separated vocalisations detected by the Audiomoths which contained a subtle background noise artifact, and these were primarily found in cluster 10. Cluster 3 largely contained a mixture of false-positives that were very similar shape to owl vocalisations, but also a high number of Australian Boobook vocalisations (a species which makes a similar call to the Powerful Owl). The remaining clusters were primarily background noise and other non-owl noises, with SM2 and Audiomoth noise floors also largely being separated (SM2 background noise in cluster 1, Audiomoth in cluster 0).

After removing all negative clusters (all except 9,10 and 11) the number of ROIs was reduced from 28,215 to 13,099. The remaining ROIs were then manually verified (see Table 3). A second iteration of the clustering was conducted on these ROIs (see Fig. 8). Cluster 0 identified instances where the segmentation had annotated two notes together instead of separately, cluster 2 consisted of mostly non-owl ROIs. Cluster 3 and 4 consist primarily of owl vocalisations, with cluster 3 only featuring calls recorded on Audiomoth devices. Label noise and inaccurate labels were reduced from 3568 to 755, a reduction of approximately 78.84 % by clustering (Table 3) with a loss of less than 2 % of the TP labels. The actual reduction in label noise is likely much higher as this value only includes the ground-truthed ROIS and does not factor in the reduction from 27,771 to 13,099 in the initial cluster.

The third iteration of clustering was conducted on HDBSCAN cluster 3 and 4 from Fig. 8. reveals one of the main utilities of segmentation for feature extraction as part of the automated detection pipeline. Individual notes are annotated and can be inspected for site differences or even vocal individuality. Fig. 9 indicates that unsupervised clustering can distinguish certain sites and suggests that there may be geographic or individual variation for this species. UMAP dimensionality reduction scores indicate that the most important features in separating notes from the 'tingalpa' site differ most are min_f_shape and min_f_centroid (see Appendix 1). Vocalisations from the 'slaughter' and 'venman' sites also cluster together, although not sufficiently enough to form distinct clusters. Clusters 0 and 2 are potentially splitting male and female vocalisations, but this requires further investigation.

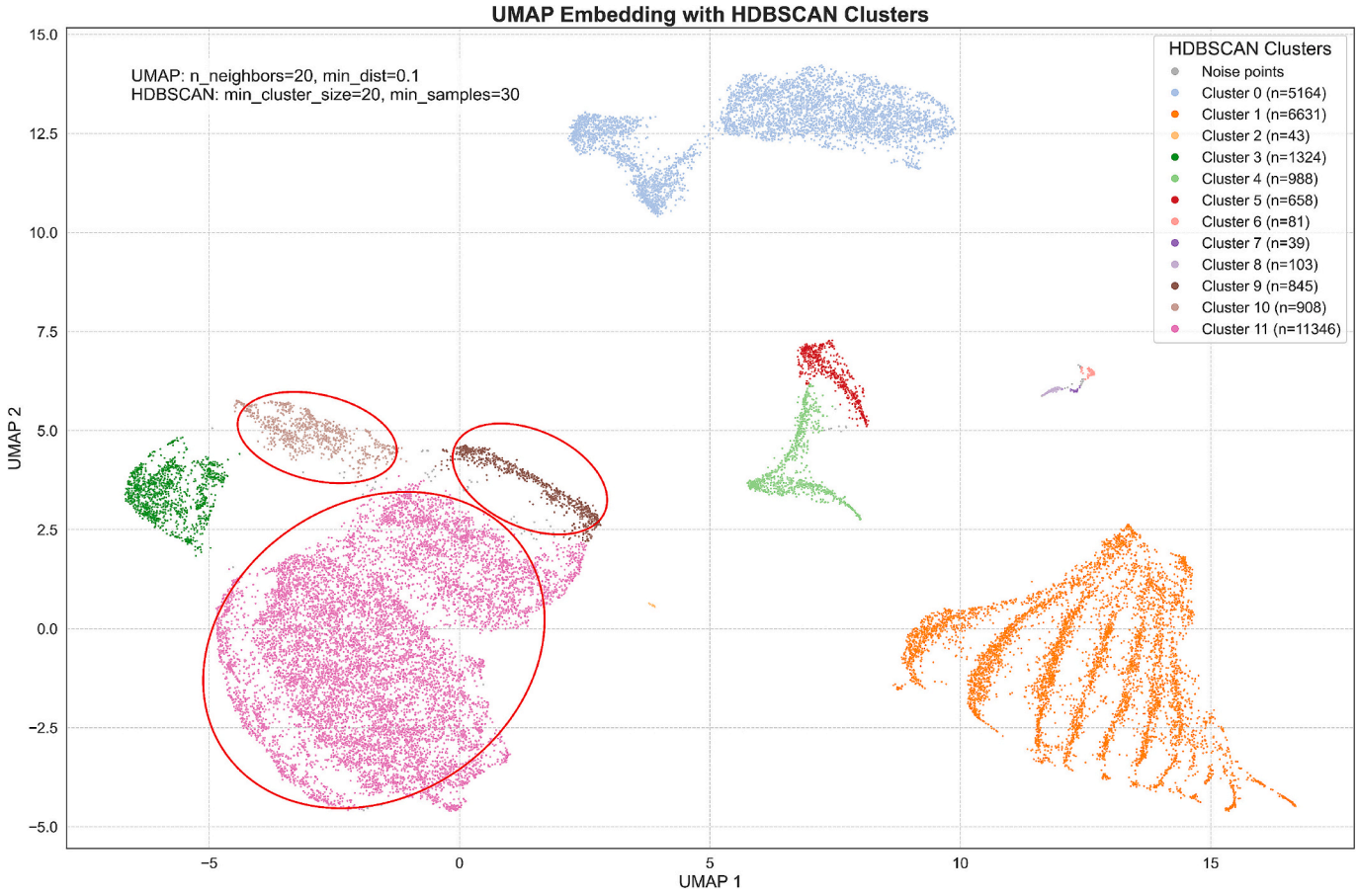


Fig. 6. UMAP projection of acoustic features for all regions of interest. Clusters and example spectrograms are coloured according to HDBSCAN classification. HDBSCAN clusters 9, 10 and 11(in red) consist of predominantly owl vocalisation ROIs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

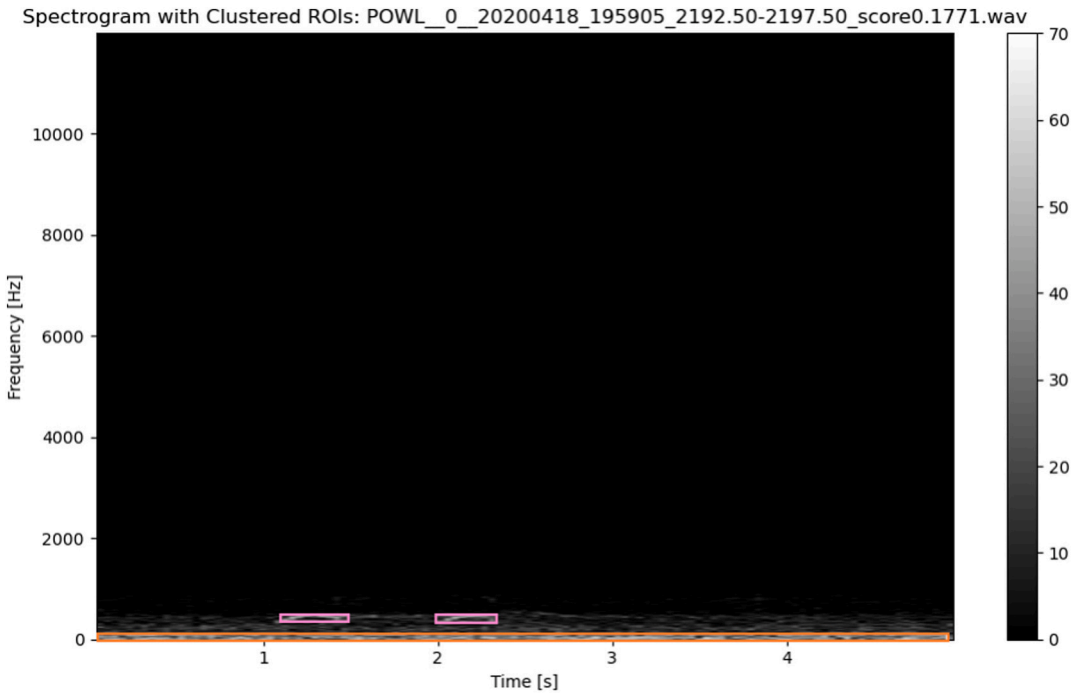


Fig. 7. Example of ROI annotation resulting from the segmentation process, colouring corresponds with the assigned cluster in Fig. 6 (cluster 11 - pink for owl vocalisations, cluster 1 orange for the low frequency noise floor. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Number of ground-truthed owl call detections before and after cluster removal.

Scenario	TP	FP	Precision	Recall	F1
Full validated dataset (baseline)	6,817	4,658			
Post-cluster (only HDBSCAN clusters 9–11)	**6,479**	**550**	**0.922**	**0.951**	**0.936**
Threshold >0.2	5,940	1,571	0.791	0.871	0.829
Threshold >0.24 (optimal)	**5704**	**1148**	**0.832**	**0.837**	**0.834**
Threshold >0.3	5,396	806	0.870	0.792	0.829
Threshold >0.4	4,953	513	0.906	0.727	0.806
Threshold >0.5	4,587	363	0.927	0.673	0.780

Table 3
Ground-truthed counts of ROIs before and after the second clustering iteration.

Label	Before Clustering	After Clustering
TP	9561	9399
FP	1382	430
Double Note	1303	231
Faulty Recorder	800	5
Boobook	53	51
Total	13,099	10,116

3.3. Binary classifier performance

The initial model (v1) exhibited similar ROC-AUC values in both test datasets (0.956 vs 0.961). The second model performance improved from 0.961 to 0.975 on the field data and 0.961 to 0.999 on the randomly selected held-out validation set (see Fig. 10).

From a practical detection standpoint, the v2 model outperformed v1 on the field dataset (see Fig. 11b). At the optimal threshold levels model v2 detected 219/263 vocalisations in the dataset with 18 false-positives (F1 = 0.876) compared to 202/263 with 43 false-positives (F1 = 0.795). v1 demonstrates a higher recall across thresholds, but with a higher number of false positives (see Fig. 11a, c & d).

4. Discussion

This study underscores the potential of combining acoustic feature extraction and unsupervised clustering with neural network-based bioacoustic detection. Our approach, which integrates transfer-learning classification with segmentation and iterative UMAP and HDBSCAN clustering has demonstrated the capacity to rapidly extract individual note annotations from noisy field recordings. This ultimately facilitates large-scale data collection from soundscape recordings, providing the potential to up-scale bioacoustic studies. We tested our pipeline on a substantial dataset containing 2764 h of recordings, focusing on territorial calls of the threatened Powerful Owl (*Ninox strenua*). The process rapidly reduced the dataset into 10,116 annotations, with 9399 (93 %) of these being correctly annotated individual notes of the target species. Our methodology follows a similar approach to the one applied by Michaud et al. (2023) using segmentation and clustering to reduce label noise in Xeno Canto recordings. In our case, we apply segmentation and clustering to field data using a neural network as an initial filtering step. This method requires minimal manual validation and mitigates some of the challenges posed by noisy field recordings, as highlighted in previous studies (Priyadarshani et al., 2018; Sainburg et al., 2020; Teixeira et al., 2024).

4.1. Key benefits

Threshold-agnostic validation: The clustering of extracted acoustic features acts as a form of 'threshold-agnostic' validation. Navine et al. (2024) note that threshold choices may produce biased vocalisation counts which can vary across subsets of the data. Score thresholds applied to classifiers have been found affect the meaning and utility of processed data (Knight and Bayne, 2019). Clustering the features as a post-processing step was able to separate vocalisations of interest (across all thresholds) from other noise with high accuracy (Table 2; Fig. 6). In our dataset, the clustering process removed 88 % of false positive detections while retaining 95 % of the true positives (F1 = 0.94), outperforming the v1 model at the optimal threshold (F1 = 0.834) by

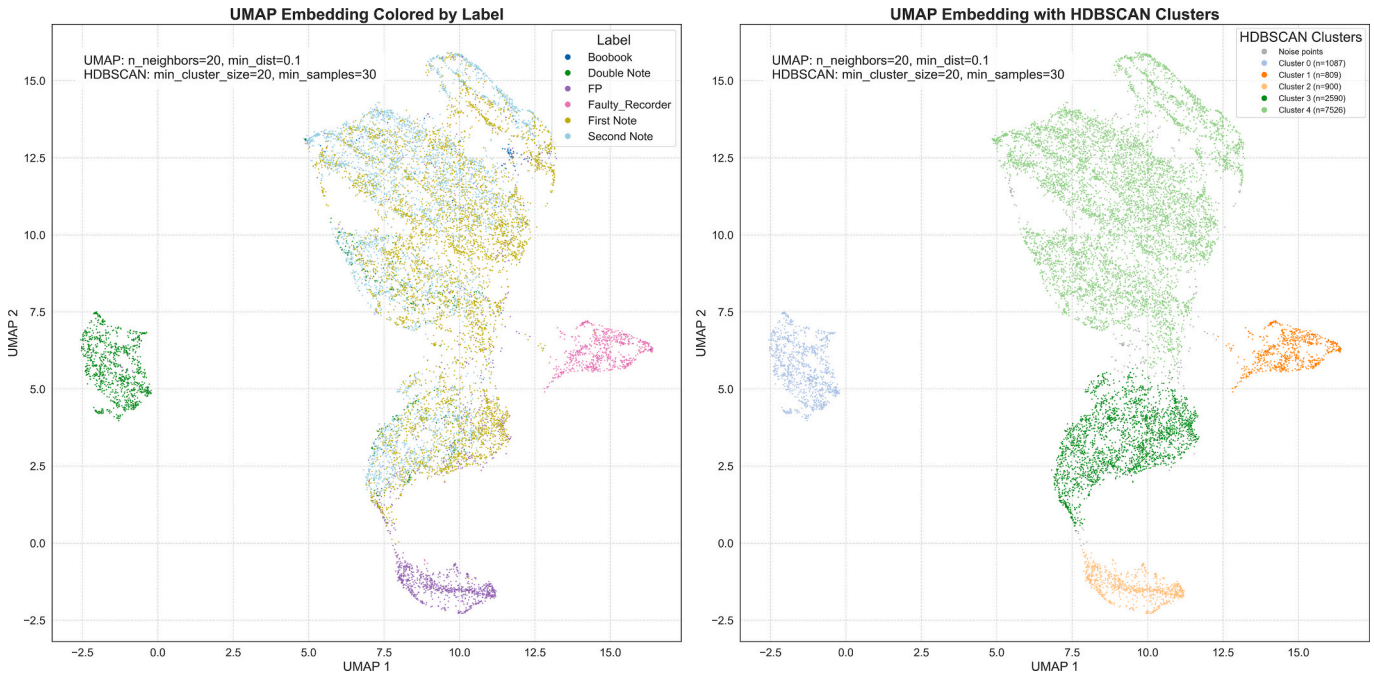


Fig. 8. UMAP projections of the second iteration of the clustering process consisting of clusters 9,10 and 11 from Fig. 6. Coloured by ground-truthed ROI labels (left) and HDBSCAN cluster labels (right).

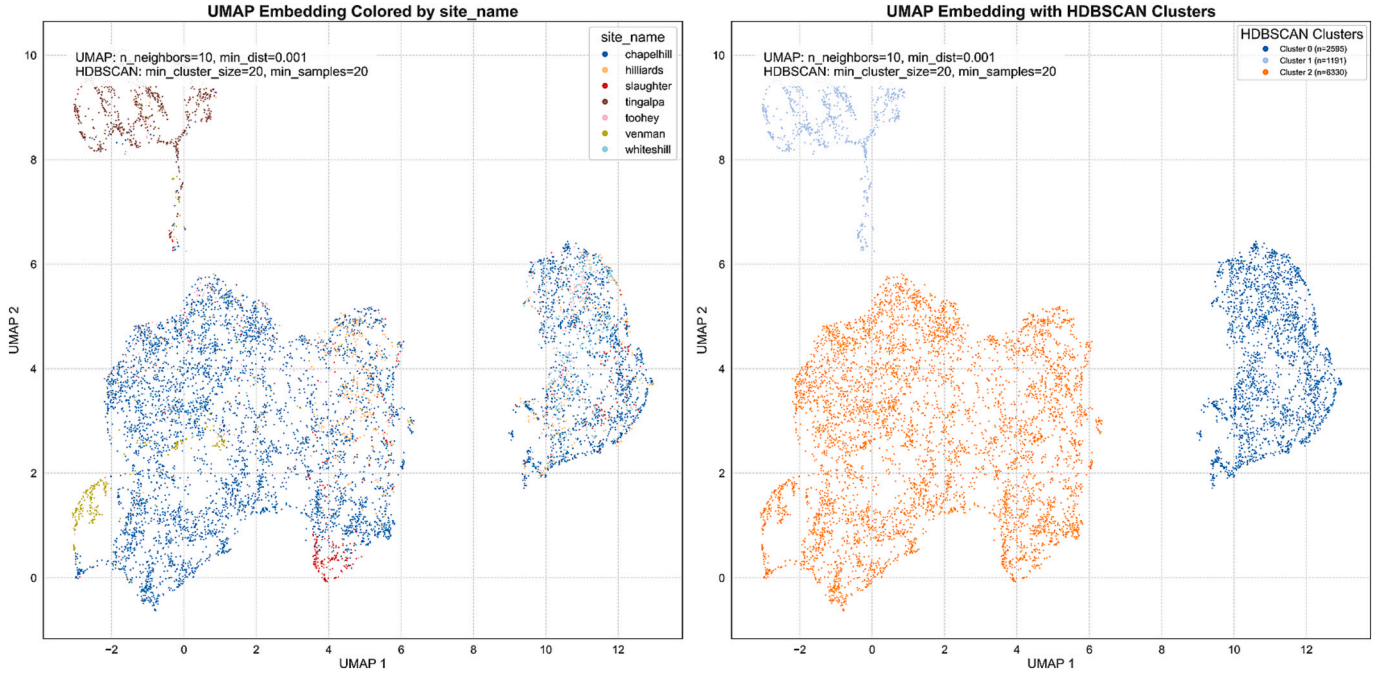


Fig. 9. UMAP projection of acoustic features comprising of clusters 3 and 4 from Fig. 8. Shown are ground-truthed labels (top left), HDBSCAN cluster labels (top right) and site labels (bottom).

detecting more vocalisations with fewer false positives (Table 2).

Individual note annotation: Unlike clustering of deep learning embeddings, our approach provides the added benefit of producing individually annotated notes. This approach allows for large amounts of fine-scale annotations to be 'harvested' quickly from field data. In this dataset, the methodology only required a few hours of expert validation time (1–2 h per clustering iteration) and is likely to annotate more consistently than a human observer. The process can easily be applied to larger datasets, enabling the collection of many annotations without the need for manual labelling. The process works by clustering ROIs based on their acoustic feature similarity. Undesired clusters are then removed, reducing label noise. Multiple iterations of clustering proved beneficial, as some label noise remained after the initial round. The second iteration reduced the remaining label noise by 78.84 % (Table 3, Fig. 8).

Semi-supervised learning: The clustering process allowed uncertain segment examples to be easily selected for a semi-supervised learning approach, noticeably improving model performance from an

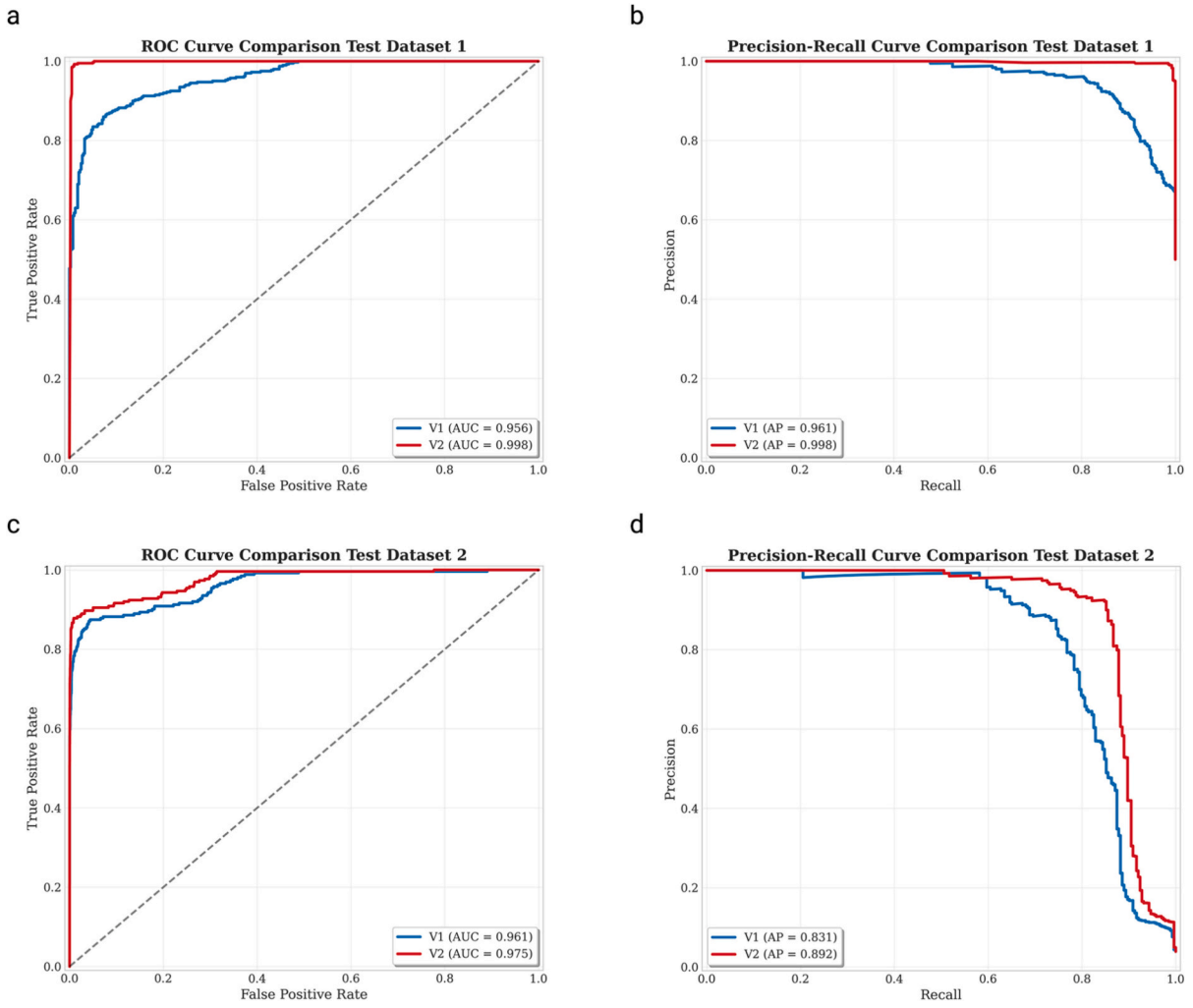


Fig. 10. Validation plots comparing performance of both models against two test datasets.

ROC-AUC of 0.961 to 0.975 on a noisy field dataset (Figs. 10 and 11). The second model improved detection at the optimal threshold (improving F1 from 0.795 to 0.876) and substantially reduced false-positives across thresholds, albeit with a slight reduction in recall (Fig. 10). It is worth noting that the performance improvement of the model here is not the focus of this study, rather an additional utility produced as a by-product of the approach. State-of-the-art models trained on global birdsong datasets provided with similar training data would likely outperform both models (Ghani et al., 2023). The MobileNet binary classifier was largely selected in this instance as a rapid way to demonstrate the utility of clustering the acoustic features as a post-processing step. As with the reduction of false positives, clustering of embeddings produced by deep learning species classifiers would likely provide similar utility for the purpose of a self-supervised approach, however the additional step of segmentation and clustering acoustic features could assist in reducing the influence of background sounds and has the additional potential benefit of producing annotations for each additional note.

4.2. Recorder differences

An unexpected result from this study was the noticeable difference in output from different passive recorders. Our study revealed variations in noise profiles between the two different devices, and the clustering process was able to distinguish between vocalisations and noise profiles recorded on different recorders (see Fig. 8). We suggest that care should be taken when interpreting results when different recorder types are used in acoustic studies, and a similar UMAP and HDBSCAN approach on outputs could provide useful information regarding the way in which recorder types may be influencing detection. Potenza et al. propose an equalisation method for soundscape recordings from multiple recorder sources (Potenza et al., 2024).

4.3. Limitations and future direction

While this methodology proved effective for Powerful Owl monitoring, future studies could expand its applicability to other species with more complex vocalisations or explore regional and individual variations in owl calls, as regional vocal dialects have been observed in many avian species (Baker and Cunningham, 1985). As owls tend to call at night, they avoid acoustic competition with diurnal birds, and therefore owl vocalisations tend to be more obvious in the spectrogram as there is less interference from simultaneously calling species. However owl vocalisations provide a different set of challenges as their low frequency vocalisations can also make their calls prone to being obscured by anthropogenic noise, as well as wind and rain in passive recordings (Shonfield and Bayne, 2017b).

This methodology has only been applied to one vocalisation from the Powerful Owl repertoire, and future work should investigate whether this approach translates effectively to other vocalisations and species. This methodology can be used with multi-class classifiers, and there is potential to apply it to the outputs of existing models such as BirdNET or

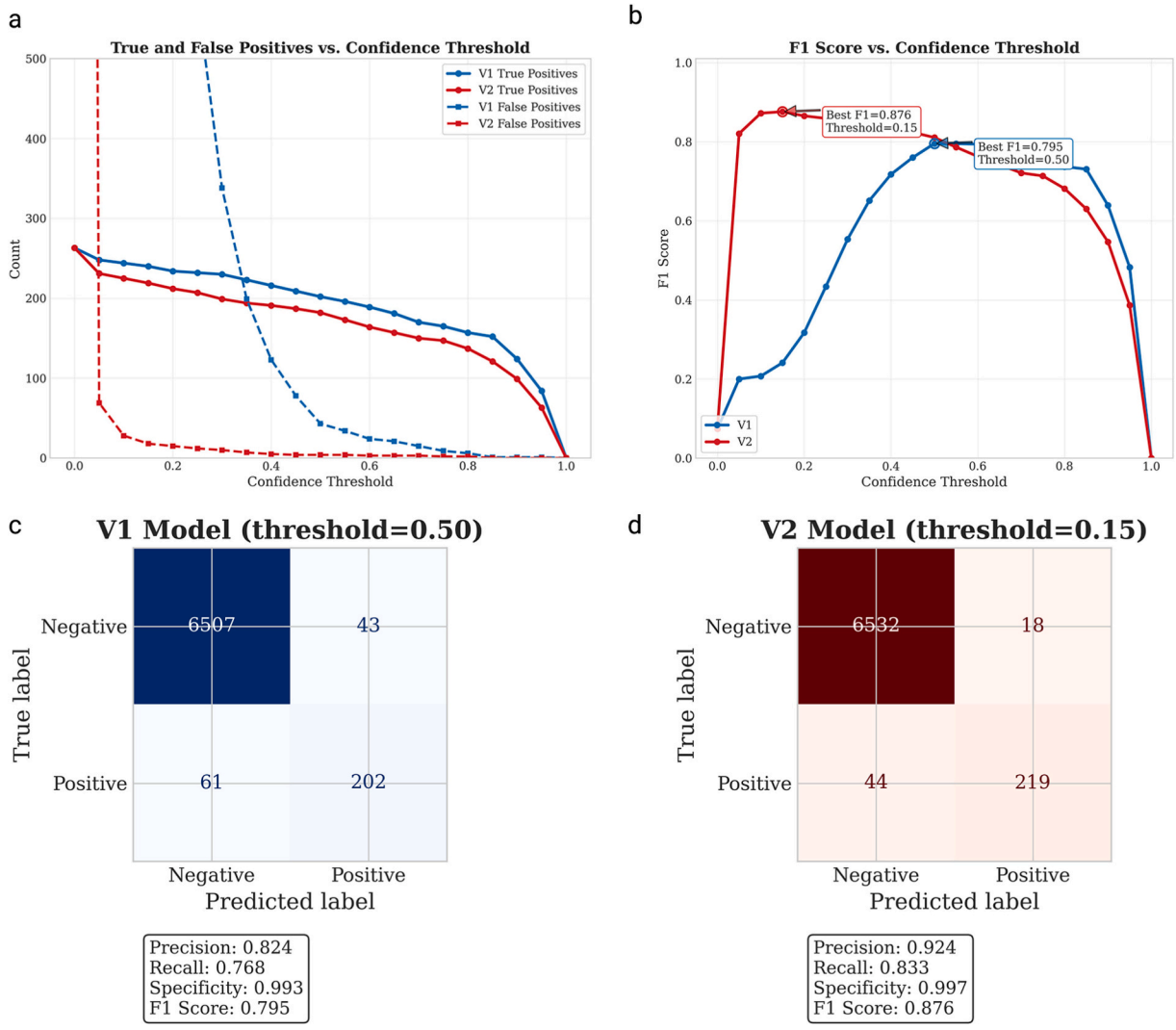


Fig. 11. Validation plots and confusion matrices comparing performance of both models on test dataset 2.

Perch. In the future, this approach should be applied to larger Powerful Owl datasets spanning broad geographic areas in order to investigate regional variation in calling behaviour, repertoire classification, sexual dimorphism and vocal individuality. It will enable rapid collection of data, significantly reducing the large amount of manual annotation typically required for a study of this nature. Future research could focus on refining these methods for multi-species detection or for analysing individual vocal repertoires, thereby further advancing the field of bioacoustics and its conservation applications. It is plausible that clustering approaches could be used to identify when species from certain geographic locations are expanding into new areas, and could be used in correlation with biological surveillance methodologies similar to those described in Wood et al. (2024).

5. Conclusion

This study found that unsupervised clustering of extracted acoustic features was a highly effective post-processing step following neural network classification. UMAP and HDBSCAN clustering contributed to a significant reduction in false-positive detections and label noise. Using this approach, 9399 individual notes of Powerful Owl vocalisations obtained from noisy field data were automatically annotated with minimal manual input required. This approach not only enhances detection accuracy but also provides a scalable solution for processing large datasets, reducing the labour-intensive task of manual annotation. The resulting clustering outputs can also support semi-supervised learning workflows and improve model performance. Overall, this study contributes to the growing body of work highlighting the importance and utility of passive acoustic monitoring and machine learning approaches for conservation research (Duchac et al., 2020; Kahl et al., 2021). By enabling the rapid collection of annotations, particularly in data-deficient regions, this approach can open the door to exploring geographic call variation, vocal individuality, and behavioural ecology.

CRediT authorship contribution statement

Callan Alexander: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Robert Clemens:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Paul Roe:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Susan Fuller:** Writing – review & editing, Supervision, Methodology, 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.

Acknowledgements

This project was supported and funded by the Queensland University of Technology through an Australian Government Research Training Program Stipend. The research was conducted in collaboration with BirdLife Australia's Powerful Owl Project. We thank Finn Roff-Marsh, Matt Wright, Jasmine Zeleny and BirdLife Southern Queensland for their assistance. We acknowledge the Turrbal and Yugara, as the First Nations owners of the lands where the research was conducted.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2025.103222.

Data availability

The code used in this study is available along with a tutorial and test dataset are available at https://github.com/callanalexander/call-collector. Full library of field recordings available upon request.

References

Alexander, C., 2022. Passive Acoustic Monitoring of Australia's Largest Owl: Using Automatic Species Recognition to Detect the Powerful Owl (*Ninox strenua*) [PhD Thesis].. Queensland University of Technology.

Allaoui, M., Kherfi, M.L., Cheriet, A., 2020. Considerably improving clustering algorithms using UMAP dimensionality reduction technique: A comparative study. In: El Moataz, A., Mammass, D., Mansouri, A., Nouboud, F. (Eds.), Image and Signal Processing. Springer International Publishing, pp. 317–325. https://doi.org/10.1007/978-3-030-51935-3_34.

Backhouse, F., Dalziell, A.H., Magrath, R.D., Rice, A.N., Crisologo, T.L., Welbergen, J.A., 2021. Differential geographic patterns in song components of male Albert's lyrebirds. Ecol. Evol. 11 (6), 2701–2716. https://doi.org/10.1002/ece3.7225.

Baker, M.C., Cunningham, M.A., 1985. The biology of bird-song dialects. Behav. Brain Sci. 8 (1), 85–100.

Best, P., Paris, S., Glotin, H., Marxer, R., 2023. Deep audio embeddings for vocalisation clustering. PLoS One 18 (7), e0283396. https://doi.org/10.1371/journal.pone.0283396.

Blanco-Portals, J., Peiró, F., Estradé, S., 2022. Strategies for EELS data analysis. Introducing UMAP and HDBSCAN for dimensionality reduction and clustering. Microsc. Microanal. 28 (1), 109–122.

Bradsworth, N., White, J.G., Isaac, B., Cooke, R., 2017. Species distribution models derived from citizen science data predict the fine scale movements of owls in an urbanizing landscape. Biol. Conserv. 213, 27–35.

Bravo Sanchez, F.J., English, N.B., Hossain, M.R., Moore, S.T., 2024. Improved analysis of deep bioacoustic embeddings through dimensionality reduction and interactive visualisation. Eco. Inform. 81, 102593. https://doi.org/10.1016/j.ecoinf.2024.102593.

Campello, R.J.G.B., Moulavi, D., Zimek, A., Sander, J., 2015. Hierarchical density estimates for data clustering, visualization, and outlier detection. ACM Trans. Knowl. Discov. Data 10 (1), 1–51. https://doi.org/10.1145/2733381.

Charif, R.A., Waack, A.M., Strickman, L.M., 2010. Raven Pro 1.4 User's Manual. Cornell Lab of Ornithology, Ithaca, NY, 25506974.

Coffey, K.R., Marx, R.E., Neumaier, J.F., 2019. DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations. Neuropsychopharmacology 44 (5), 859–868. https://doi.org/10.1038/s41386-018-0303-6.

Cohen, Y., Nicholson, D.A., Sanchioni, A., Mallaber, E.K., Skidanova, V., Gardner, T.J., 2022. Automated annotation of birdsong with a neural network that segments spectrograms. Elife 11, e63853.

Debus, S.J.S., 1995. Surveys of large forest owls in northern New South Wales: methodology, calling behaviour and owl responses. Corella 19, 38–50.

Denton, T., Wisdom, S., Hershey, J.R., 2022. Improving Bird Classification with Unsupervised Sound Separation. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 636–640. https://doi.org/10.1109/ICASSP43922.2022.9747202.

Duchac, L.S., Lesmeister, D.B., Dugger, K.M., Ruff, Z.J., Davis, R.J., 2020. Passive acoustic monitoring effectively detects Northern Spotted Owls and Barred Owls over a range of forest conditions. Condor 122 (3), duaa017.

Fauna and Flora Guarantee Act 1988.

Eichinski, P., Alexander, C., Roe, P., Parsons, S., Fuller, S., 2022. A convolutional neural network bird species recognizer built from little data by iteratively training, detecting, and labeling. Front. Ecol. Evol. 10, 810330.

Ghani, B., Denton, T., Kahl, S., Klinck, H., 2023. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. Sci. Rep. 13 (1), 22876. https://doi.org/10.1038/s41598-023-49989-z.

Grava, T., Mathevon, N., Place, E., Balluet, P., 2008. Individual acoustic monitoring of the European eagle owl Bubo bubo. Ibis 150 (2), 279–287.

Henkel, C., Pfeiffer, P., Singer, P., 2021. Recognizing bird species in diverse soundscapes under weak supervision (arXiv:2107.07728). arXiv. https://doi.org/10.48550/arXiv.2107.07728.

Hill, A.P., Prince, P., Piña Covarrubias, E., Doncaster, C.P., Snaddon, J.L., Rogers, A., 2018. AudioMoth: Evaluation of a smart open acoustic device for monitoring biodiversity and the environment. Methods Ecol. Evol. 9 (5), 1199–1211.

Johnsgard, P.A., 1988. North American Owls: Biology and Natural History.

Kahl, S., Wood, C.M., Eibl, M., Klinck, H., 2021. BirdNET: a deep learning solution for avian diversity monitoring. Eco. Inform. 61, 101236.

Kershenbaum, A., Blumstein, D.T., Roch, M.A., Akçay, Ç., Backus, G., Bee, M.A., Bohn, K., Cao, Y., Carter, G., Cäsar, C., Coen, M., DeRuiter, S.L., Doyle, L., Edelman, S., Ferrer-i-Cancho, R., Freeberg, T.M., Garland, E.C., Gustison, M., Harley, H.E., Zamora-Gutierrez, V., 2016. Acoustic sequences in non-human animals: a tutorial review and prospectus. Biol. Rev. 91 (1), 13–52. https://doi.org/10.1111/brv.12160.

Knight, E.C., Bayne, E.M., 2019. Classification threshold and training data affect the quality and utility of focal species data processed with automated audio-recognition software. Bioacoustics 28 (6), 539–554. https://doi.org/10.1080/09524622.2018.1503971.

Koch, T.M., Marks, E.S., Roberts, T.F., 2024. AVN: A Deep Learning Approach for the Analysis of Birdsong. bioRxiv. https://pmc.ncbi.nlm.nih.gov/articles/PMC11370480/.

Lostanlen, V., Salamon, J., Farnsworth, A., Kelling, S., Bello, J.P., 2019. Robust sound event detection in bioacoustic sensor networks. PLoS One 14 (10), e0214168. https://doi.org/10.1371/journal.pone.0214168.

Loyn, R.H., McNabb, E.G., Volodina, L., Willig, R., 2001. Modelling landscape distributions of large forest owls as applied to managing forests in north-East Victoria, Australia. Biol. Conserv. 97 (3), 361–376.

Madhavan, M., Linhart, P., 2024. Vocal individuality in owls: a taxon-wide review in the context of Tinbergen's four questions. J. Ornithol. https://doi.org/10.1007/s10336-024-02230-8.

Manzano-Rubio, R., Bota, G., Brotons, L., Soto-Largo, E., Pérez-Granados, C., 2022. Low-cost open-source recorders and ready-to-use machine learning approaches provide effective monitoring of threatened species. Eco. Inform. 72, 101910. https://doi.org/10.1016/j.ecoinf.2022.101910.

McFee, B., Raffel, C., Liang, D., Ellis, D., McVicar, M., Battenberg, E., Nieto, O., 2015. Librosa: Audio and Music Signal Analysis in Python, pp. 18–24. https://doi.org/10.25080/Majora-7b98e3ed-003.

McGinn, K., Kahl, S., Peery, M.Z., Klinck, H., Wood, C.M., 2023. Feature embeddings from the BirdNET algorithm provide insights into avian ecology. Eco. Inform. 74, 101995. https://doi.org/10.1016/j.ecoinf.2023.101995.

McInnes, L., Healy, J., Astels, S., 2017. hdbscan: hierarchical density based clustering. J. Open Source Softw. 2 (11), 205. https://doi.org/10.21105/joss.00205.

McInnes, L., Healy, J., Melville, J., 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (arXiv:1802.03426). arXiv. https://doi.org/10.48550/arXiv.1802.03426.

Michaud, F., Sueur, J., Le Cesne, M., Haupert, S., 2023. Unsupervised classification to improve the quality of a bird song recording dataset. Eco. Inform. 74, 101952.

Navine, A.K., Denton, T., Weldy, M.J., Hart, P.J., 2024. All thresholds barred: direct estimation of call density in bioacoustic data. Front. Bird Sci. 3. https://doi.org/10.3389/fbirs.2024.1380636.

Neal, L., Briggs, F., Raich, R., Fern, X.Z., 2011. Time-frequency segmentation of bird song in noisy acoustic environments. In: 2011 IEEE International Conference on Acoustics, speech and signal processing (ICASSP). IEEE, pp. 2012–2015.

Pérez-Granados, C., 2023. BirdNET: applications, performance, pitfalls and future opportunities. Ibis 165 (3), 1068–1075. https://doi.org/10.1111/ibi.13193.

Picciulin, M., Kéver, L., Parmentier, E., Bolgan, M., 2019. Listening to the unseen: passive acoustic monitoring reveals the presence of a cryptic fish species. Aquat. Conserv. Mar. Freshwat. Ecosyst. 29 (2), 202–210.

Potenza, A., Zaffaroni-Caorsi, V., Benocci, R., Guagliumi, G., Fouani, J.M., Bisceglie, A., Zambon, G., 2024. Biases in ecoacoustics analysis: a protocol to equalize audio recorders. Sensors (Basel, Switzerland) 24 (14), 4642. https://doi.org/10.3390/s24144642.

Priyadarshani, N., Marsland, S., Castro, I., 2018. Automated birdsong recognition in complex acoustic environments: a review. J. Avian Biol. 49 (5) jav-01447.

Recalde, N.M., 2023. pykanto: A python library to accelerate research on wild bird song arXiv Preprint arXiv:2302.10340.

Rognan, C.B., Szewczak, J.M., Morrison, M.L., 2012. Autonomous recording of great Gray owls in the Sierra Nevada. Northwest. Nat. 93 (2), 138–144.

Sainburg, T., Thielk, M., Gentner, T.Q., 2020. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. PLoS Comput. Biol. 16 (10), e1008228.

Shaw, T., Schönamsgruber, S.-R., Cordeiro Pereira, J.M., Mikusiński, G., 2022. Refining manual annotation effort of acoustic data to estimate bird species richness and composition: the role of duration, intensity, and time. Ecol. Evol. 12 (11), e9491. https://doi.org/10.1002/ece3.9491.

Shonfield, J., Bayne, E., 2017a. Autonomous recording units in avian ecological research: current use and future applications. Avian Conserv. Ecol. 12 (1).

Shonfield, J., Bayne, E., 2017b. The effect of industrial noise on owl occupancy in the boreal forest at multiple spatial scales. Avian Conserv. Ecol. 12 (2).

Shonfield, J., Heemskerk, S., Bayne, E.M., 2018. Utility of automated species recognition for acoustic monitoring of owls. J. Raptor Res. 52 (1), 42–55.
Steinfath, E., Palacios-Muñoz, A., Rottschäfer, J.R., Yuezak, D., Clemens, J., 2021. Fast and accurate annotation of acoustic signals with deep neural networks. Elife 10, e68837.
Stowell, D., 2022. Computational bioacoustics with deep learning: a review and roadmap. PeerJ 10, e13152. https://doi.org/10.7717/peerj.13152.
Teixeira, D., Maron, M., van Rensburg, B.J., 2019. Bioacoustic monitoring of animal vocal behavior for conservation. Conserv. Sci. Pract. 1 (8), e72.
Teixeira, D., Roe, P., Van Rensburg, B.J., Linke, S., McDonald, P.G., Tucker, D., Fuller, S., 2024. Effective ecological monitoring using passive acoustic sensors: recommendations for conservation practitioners. Conserv. Sci. Pract. 6 (6), e13132. https://doi.org/10.1111/csp2.13132.
Terry, A.M., Peake, T.M., McGregor, P.K., 2005. The role of vocal individuality in conservation. Front. Zool. 2 (1), 10. https://doi.org/10.1186/1742-9994-2-10.
Tolkova, I., Chu, B., Hedman, M., Kahl, S., Klinck, H., 2021. Parsing Birdsong with Deep Audio Embeddings (arXiv:2108.09203). arXiv. https://doi.org/10.48550/arXiv.2108.09203.
Tseng, S., Hodder, D.P., Otter, K.A., 2024. Using autonomous recording units for vocal individuality: insights from Barred Owl identification. Avian Conserv. Ecol. 19 (1). https://doi.org/10.5751/ACE-02680-190123.
Ulloa, J.S., Haupert, S., Latorre, J.F., Aubin, T., Sueur, J., 2021. scikit-maad: an open-source and modular toolbox for quantitative soundscape analysis in Python. Methods Ecol. Evol. 12 (12), 2334–2340.
Wood, C.M., Kahl, S., 2024. Guidelines for appropriate use of BirdNET scores and other detector outputs. J. Ornithol. 165 (3), 777–782. https://doi.org/10.1007/s10336-024-02144-5.
Wood, C.M., Gutiérrez, R.J., Peery, M.Z., 2019. Acoustic monitoring reveals a diverse forest owl community, illustrating its potential for basic and applied ecology. Ecology 100 (9), 1–3.
Wood, C.M., Champion, J., Brown, C., Brommelsiek, W., Laredo, I., Rogers, R., Chaopricha, P., 2023. Challenges and opportunities for bioacoustics in the study of rare species in remote environments. Conserv. Sci. Pract. 5 (6), e12941.
Wood, C.M., Günther, F., Rex, A., Hofstadter, D.F., Reers, H., Kahl, S., Peery, M.Z., Klinck, H., 2024. Real-time acoustic monitoring facilitates the proactive management of biological invasions. Biol. Invasions 26 (12), 3989–3996. https://doi.org/10.1007/s10530-024-03426-y.
Yan, X., Zhang, H., Li, D., Wu, D., Zhou, S., Sun, M., Hu, H., Liu, X., Mou, S., He, S., 2019. Acoustic recordings provide detailed information regarding the behavior of cryptic wildlife to support conservation translocations. Sci. Rep. 9 (1), 1–11.
Zhao, S., Xie, J., Ding, C., 2023. Automatic individual recognition of wild crested Ibis based on hybrid method of self-supervised learning and clustering. Eco. Inform. 75, 102089. https://doi.org/10.1016/j.ecoinf.2023.102089.
Znidersic, E., Towsey, M., Roy, W.K., Darling, S.E., Truskinger, A., Roe, P., Watson, D.M., 2020. Using visualization and machine learning methods to monitor low detectability species—the least bittern as a case study. Eco. Inform. 55, 101014.