# An Efficient Pipeline for the Unsupervised Segmentation of Heterogeneous Natural Soundscapes


## Abstract

Passive Acoustic Monitoring will transform biodiversity assessment and large-scale ecological surveys through continuous ecoacoustic data collection. Yet, dataset growth has outpaced existing segmentation methods, many of which are tuned to narrow taxonomic subsets lacking abiotic sounds, limiting ecological realism. Natural soundscapes contain overlapping biophony, geophony, anthrophony, and technophony, making resource-efficient and ecologically valid segmentation an ongoing challenge. As such, we introduce an unsupervised framework designed to reveal meaningful acoustic structures without predefined labels. The pipeline integrates systematic sampling, sound event detection, Mel-Frequency Cepstral Coefficient extraction, dimensionality reduction via Uniform Manifold Approximation and Projection (UMAP), and clustering with Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). We evaluate the framework across six biodiverse Australian soundscapes comprising of ($n$=19,230) 4.5-second non-overlapping segments. Results indicate that clusters are internally coherent, with Voronoi tessellations over the UMAP space showing distinct spatial boundaries. External validation with ($n$=2,000) manually annotated samples demonstrates strong alignment with ecologically meaningful sound types. F1-scores ranged from 83.7% to 98.3%, with precision and recall exceeding 91% across most sites. By combining unsupervised clustering with ecological validation, our framework offers a practical, generalisable solution for organising unlabelled ecoacoustic data while reducing manual effort and preserving ecological integrity.



*Keywords:* Clustering, Ecoacoustics, Large-scale data analysis, Passive acoustic monitoring, Unsupervised learning


## 1. Introduction

Ecoacoustics is an innovative and rapidly developing field dedicated to studying environmental sounds to monitor and understand biodiversity (Sueur and Farina, 2015). This field involves recording, analysing, and interpreting sounds produced by wildlife and their natural habitats. By analysing these environmental sounds, researchers can gain valuable insights into the presence, behaviour, and interactions of various soniferous species. This approach is increasingly vital due to the global decline in biodiversity (Butchart et al., 2010; Keil et al., 2015). Species loss from habitat destruction significantly strains ecosystems, emphasising the need to maintain ecosystem functions by reducing extinction rates (Chase et al., 2020; Haddad et al., 2015). Consequently, biodiversity monitoring is crucial for tracking species, understanding population declines, and evaluating the effectiveness of management interventions (Lindenmayer et al., 2012).

Historically, biodiversity monitoring relied on labour-intensive manual surveys (e.g., point counts) that were limited in spatial and temporal scope and susceptible to observer bias (Wimmer et al., 2013a). These methods often missed cryptic or nocturnal species and could not support continuous, long-term data collection across broad landscapes (Hoefer et al., 2023).

As a result, Passive Acoustic Monitoring (PAM) systems have emerged as a transformative solution to these challenges. By leveraging low-cost acoustic sensors to continuously capture natural soundscape data, PAM provides an efficient, non-invasive means of monitoring vocalising fauna and their responses to environmental changes (Sugai et al., 2019). As such, PAM systems can generate extensive datasets, offering great potential for remote, automated monitoring approaches for ecological studies. Yet, this shift from traditional methods to high-throughput sensing has introduced a new challenge: the sheer volume of data produced renders manual analysis impractical, creating a bottleneck in processing and interpretation (Gibb et al., 2019). Despite this, manual validation remains a widespread practice in ecoacoustic research, with experts routinely inspecting spectrograms and listening to recordings to ensure accuracy (Ross et al., 2023; Funosas et al., 2024). However, while necessary for verifying and refining results, these practices alone are becoming increasingly inadequate for addressing the volume and complexity of data produced by modern PAM systems.

Consequently, automation is essential to handle the volume of PAM data. Machine Learning (ML) and Deep Learning (DL) offer potential solutions, but face notable challenges. A primary issue is the high dimensionality and complexity of ecoacoustic data. Ecoacoustic recordings contain a rich mixture of biophony (animal vocalisations), geophony (natural abiotic sounds like wind and rain), anthrophony (human vocal and bodily sounds), and technophony (sounds generated by human technologies), creating highly variable, unstructured, multi-source datasets (Mullet et al., 2016; Farina et al., 2018; Vella et al., 2022). Recording conditions (e.g., microphone placement, background noise, species behaviour) add further variability (Happel and Happel, 2020). Such complexity can obscure meaningful patterns and relationships, posing significant challenges for

data structuring, segmentation, and interpretation in automated ecoacoustic analysis (Williams et al., 2022; Scarpelli et al., 2021).

A further limitation is the scarcity of labelled data, since supervised models require well-annotated training sets (Stowell, 2022). Manual annotation is time-consuming, so many ecoacoustic recordings lack comprehensive ground truth. Although some larger annotated datasets exist, particularly in birdsong analysis (Morfi et al., 2019; Rauch et al., 2025), broadly representative ecoacoustic datasets remain rare, hindering training for large-scale monitoring. As a result, studies often use small, manually labelled, species-specific datasets that, while useful in controlled experiments, do not capture the full variability of real-world soundscapes (Gibb et al., 2019; Scarpelli et al., 2021; Diaz et al., 2023). Consequently, while models trained on these datasets may perform well within their specific training domains, they often face challenges such as reduced accuracy, misclassification of unfamiliar sound types, or poor adaptation to new recording conditions when applied to diverse acoustic environments, which can limit their broader applicability in large-scale ecoacoustic analysis (Stowell, 2022; Vella et al., 2022; Wilkinghoff et al., 2025).

Scalability and computational constraints present additional hurdles, particularly for large-scale PAM datasets, which can generate terabytes of audio annually. This often far exceeds the processing capacity of many traditional ML pipelines (Wall et al., 2021; Stowell, 2022). Traditional ML methods, such as Support Vector Machines (SVMs) and Random Forests (RFs) have been applied to tasks like species identification and sound source clustering (Noda et al., 2016; Scarpelli et al., 2021; Nieto-Mora et al., 2023; Cominelli et al., 2024). However, these approaches often require extensive feature engineering and may not scale well with the high-dimensional, noisy, and heterogeneous data typically found in ecoacoustics (Genuer et al., 2017). Consequently, there is a pressing need for the development of optimised and resource-efficient ML architectures that can handle the volume of ecoacoustic data without compromising analytical performance.

Similarly, while DL approaches such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have demonstrated strong classification performance in bioacoustics (Quinn et al., 2022), they are computationally expensive, limiting their applicability for widespread deployment (Napier et al., 2024). As such, while some DL approaches have been successfully developed to tackle specific ecoacoustic analysis tasks (Zhong et al., 2020), these methods often underperform when applied to complex, real-world datasets, such as those from the Australian Acoustic Observatory (A2O) (Roe et al., 2021), or the U.S. Northeast Passive Acoustic Sensing Network (NEPAN) (Van Parijs et al., 2015). This is particularly true for those approaches evaluated on datasets including few species, which may only partially reflect the heterogeneity of real-world soundscapes (Dufourq et al., 2022).

The primary aim of this study will be to examine how unsupervised ML can help make sense of large-scale ecoacoustic recordings, with a focus on identifying meaningful acoustic patterns rather than trying to match known species labels. Unlike existing approaches that rely on species-level labels, our goal is to let the data speak for itself by grouping sounds purely based on acoustic similarity. In other words, we believe segmentation should be non-species-specific – not targetting any particular taxa, and thus allowing diverse sound sources (biotic or abiotic) to naturally form clusters. Another key design consideration in our approach is computational efficiency, in the narrower algorithmic sense of how the cost of a method grows as input size increases. With PAM producing millions of hours of recordings each year, there is a need for methods that remain practical under increasing data volumes in terms of processing time and deployment requirements. This framework aims to provide exactly this: a practical, resource-efficient way to organise unlabelled soundscape data so that experts can focus their time effectively. By automatically grouping similar sounds, the method can triage large audio collections, for example, highlighting prevalent sound types at a site or flagging unusual acoustic events, thereby allowing experts to focus their attention. In doing so, we aim to support more effective biodiversity monitoring of complex ecosystems and accelerate insights for management decisions. In summary, the main contributions of this work are as follows:

- We examine the challenges involved in analysing real-world soundscapes that contain a mix of biological, environmental, and human-made sounds, rather than focusing on narrow, taxonomically-filtered datasets;
- We present a modular pipeline that combines systematic data sampling, sound event detection, Mel-Frequency Cepstral Coefficients (MFCCs) feature extraction, dimensionality reduction, and unsupervised clustering;
- We evaluate a range of configurations across different sites to identify combinations that produce consistent, interpretable results under varied ecological and acoustic conditions;
- We assess and compare the computational efficiency of our framework, reporting on processing times to understand how the method performs compared to existing approaches, when applied to complex and multi-source datasets;
- We validate the ecological relevance of the discovered clusters using partially labelled data, demonstrating that meaningful acoustic patterns can be identified without requiring pre-labelled training sets;
- We provide a flexible structure that can support both targeted expert annotation and downstream automation, offering a practical tool for large-scale ecoacoustic analysis and environmental monitoring.

We emphasise that the present contribution does not lie in proposing a new standalone dimensionality reduction, clustering, or sound event detection algorithm. Rather, the novelty of this study lies in the methodological integration and evaluation of lightweight, established components within a unified unsupervised framework designed specifically for heterogeneous natural soundscapes. In contrast to prior approaches that focus

on narrower taxonomic targets, biophony-only corpora, or pre-trained feature representations, our framework is positioned at the soundscape level and is intended to organise mixed-source ecoacoustic data containing biophonic, geophonic, anthropophonic, and technophonic components. A further contribution is the evaluation strategy, which combines internal clustering metrics with post-hoc ecological validation to assess whether the resulting groupings are not only structurally coherent but also ecologically meaningful.

## 2. Related Works

### 2.1. Preliminaries

Large-scale ecoacoustics datasets differ from traditional bioacoustic datasets in their spatial, temporal, and taxonomic breadth, capturing continuous soundscapes that integrate biophony, geophony, and anthrophony (Vella et al., 2022; Turlington et al., 2024). Rather than focusing on individual species, these datasets aim to characterise entire acoustic environments, resulting in high-dimensional, temporally autocorrelated data with substantial source overlap and environmental variability (Truskinger et al., 2014; Napier et al., 2024). Consequently, their analysis requires scalable machine learning methods that remain robust under heterogeneous and noisy conditions.

A prominent example is the A2O, which comprises of approximately 250 monitoring stations across 64 sites, providing continuous, year-round recordings spanning diverse Australian ecosystems (Roe et al., 2021). This combination of geographic coverage and temporal depth enables the investigation of long-term ecological dynamics, phenological patterns, and responses to climatic and anthropogenic pressures that are difficult to capture using conventional survey methods (Vella et al., 2022).

However, the scale and complexity of datasets such as A2O impose substantial computational and methodological demands. Overlapping acoustic sources, site-specific variability, and high data volumes necessitate automated segmentation and structuring approaches capable of preserving ecological context. Unsupervised learning methods, including clustering and multi-index analysis, are therefore particularly relevant for organising large-scale soundscape data without reliance on extensive labelled training sets (Napier et al., 2023; Scarpelli et al., 2021).

### 2.2. Signal feature extraction

Selecting the optimal signal feature extraction in the analysis pipeline represents a difficult but necessary challenge. It involves transforming raw sound signals into structured representations suitable for ML and classification. The conversion of sound from the time domain into frequency-domain or time-frequency representations is often required as interpreting waveforms directly offers limited insight for sound event discrimination. Recent research has explored various signal feature extraction approaches for large-scale ecoacoustics and bioacoustics analysis, with a focus on spectral, cepstral, energy, and voicing-related features. Techniques such as the Short-Time Fourier Transform (STFT) and Discrete Wavelet Transform (DWT) have been employed for species-specific tasks (Thomas et al., 2020; Xie et al., 2018). These methods capture fine details in frequency modulations and are especially useful in distinguishing among species-specific vocalisations. For example, Stowell and Plumbley (2014) demonstrated that high-resolution signal processing can detect fine frequency modulations that complement more generalised frequency-domain features.

In addition to traditional frequency-domain methods, spectrogram image-based features have gained popularity in bioacoustics and ecoacoustics analysis (Poutaraud et al., 2024). Spectrograms visually represent sound by plotting time on one axis and frequency on the other, with colour intensity indicating energy or amplitude. This image-based representation allows ML techniques, particularly CNNs, to process sound as visual patterns. Spectrogram image-based approaches are especially effective for detecting species-specific patterns that are difficult to capture using simpler spectral or temporal features. For example, Bisot et al. (2015) and Abidin et al. (2018) discussed the use of techniques such as Histograms of Oriented Gradients (HOG) and Local Binary Patterns (LBP) applied to spectrogram images, enabling the detection of time-frequency structures essential for sound event classification. This method has been particularly effective in complex soundscapes where overlapping sounds make feature extraction more challenging.

Among the various features, MFCCs have remained a recurrent technique for bioacoustic applications (Xie et al., 2020; Abdul and Al-Talabani, 2022; Soares et al., 2022; Ugarte and Arias-Arias, 2024). MFCCs, inspired by human auditory perception, provide a compact and perceptually relevant representation of sound. By focusing on the Mel scale, which emphasises frequencies to which the human ear is most sensitive, MFCCs capture the spectral envelope of sounds, making them particularly useful for species identification tasks where subtle differences in frequency are critical. While newer feature extraction methods like CNNs and matrix factorisation techniques offer powerful alternatives, we argue that MFCCs are still favourable due to their balance between efficiency and accuracy. MFCCs' human-inspired design aligns with the auditory processing capabilities of the human ear, making them highly effective for tasks involving complex soundscapes, such as species vocalisation classification (Noda et al., 2016; Serizel et al., 2018).

For instance, Akbal et al. (2022) introduced a novel ML model utilizing MFCCs to classify bird and anuran species automatically. Their approach achieved a remarkable 98.75% accuracy using a $k$-nearest neighbour classifier, demonstrating the robustness of MFCCs in multispecies bioacoustic sound classification. Similarly, Lakdari et al. (2024) evaluated various feature extraction methods for distinguishing individual gibbon calls under varying noise conditions. Their findings revealed that MFCCs outperformed embeddings from pre-trained CNNs, particularly in noisy environments, which highlight MFCCs' effectiveness under challenging acoustic scenarios.

### 2.3. Dimension reduction

The resulting outcomes of feature extraction often lead to high dimensionality, which can present significant challenges

for analysis, particularly when attempting to visualise or classify data accurately. To mitigate these issues, dimension reduction techniques are employed to compress the data into lower-dimensional spaces while preserving critical information. In this study, we apply and compare three prominent techniques: Principal Component Analysis (PCA), *t*-Distributed Stochastic Neighbour Embedding (*t*-SNE), and Uniform Manifold Approximation and Projection (UMAP). Each of these methods offers unique advantages and limitations in handling ecoacoustic data, which is characterised by a high degree of complexity and variability.

PCA is a linear technique that transforms data into new coordinates defined by the directions of maximum variance (Abdi and Williams, 2010). While PCA efficiently reduces dimensionality and is computationally light, it often fails to capture non-linear structures inherent in ecoacoustic datasets. The algorithm's reliance on global variance makes it less suited for datasets where local relationships and intricate ecological patterns are of interest (Yao et al., 2012). In contrast, *t*-SNE is a non-linear technique that excels at preserving local relationships, making it particularly effective for visualizing clusters in high-dimensional data (van der Maaten and Hinton, 2008). However, *t*-SNE is computationally intensive and primarily focuses on local neighbourhoods, often distorting global structures in the data (Zhou and Sharpee, 2022).

UMAP addresses many of the limitations inherent in PCA and *t*-SNE. UMAP not only preserves both local and global structures but also scales more efficiently to large datasets, offering faster computation times and more reproducible results (McInnes et al., 2018). As demonstrated by Milošević et al. (2022) and Becht et al. (2019), UMAP has outperformed other dimension reduction techniques in fields ranging from aquatic ecology to single-cell biology. Its ability to preserve the global topology of data while maintaining local structures makes UMAP particularly valuable for ecoacoustic datasets, where interactions between species, environmental conditions, and anthropogenic factors create complex, multi-dimensional patterns, and has thus seen a rise in use (Nieto-Mora et al., 2024; Cominelli et al., 2024).

In aquatic ecology, for example, UMAP has been combined with Louvain algorithms to improve the ordination and classification of biological indicators, offering superior performance over PCA in handling noisy, multi-scale data (Milošević et al., 2022). Similarly, in single-cell data analysis, UMAP has demonstrated its efficiency in organizing cell clusters meaningfully and reproducibly, outperforming alternatives like *t*-SNE in terms of runtime and interpretability (Becht et al., 2019). The theoretical underpinnings of UMAP, grounded in Riemannian geometry and algebraic topology, contribute to its effectiveness in handling complex, real-world data. By leveraging a mathematical approach that models the high-dimensional relationships of data points, UMAP achieves a balance between fidelity and distortion.

### 2.4. *Labelling and segmentation approaches*

Over the past decade, a variety of ecoacoustic segmentation approaches have emerged, each tackling the challenge of identifying structure in complex, multivariate soundscape data from distinct angles. These methods range from interpretable clustering using acoustic indices, to DL pipelines leveraging advanced embeddings. Among previously proposed approaches, Lin et al. (2017) introduced Periodicity-Coded Non-Negative Matrix Factorisation (PC-NMF) for the identification of periodic acoustic events within long-duration recordings. PC-NMF proved effective in removing environmental and anthropogenic noise from long-term recordings without needing training labels, making it valuable for habitats with periodic biotic activity. Although their approach shows potential for complex soundscapes, its success depends on chorus periodicity and temporal regularity, which may limit its applicability across different ecosystems or taxa.

Another class of methods uses acoustic indices for clustering. For example, Phillips et al. (2018) applied a combination of *k*-means and hierarchical clustering to twelve ecoacoustic indices, enabling broad acoustic partitioning across multi-taxa soundscapes. This approach demonstrated high computational scalability and included abiotic sounds but was limited in taxonomic resolution and generalisability. Similarly, Scarpelli et al. (2021) used motif discovery techniques on time-series data derived from selected indices (e.g., Acoustic Complexity Index (ACI), Temporal Entropy (ENT), and Event Count Index (EVN)) (Sueur et al., 2014; Towsey, 2017) and combined them with RF classifiers to segment diverse habitats. Their method included geophonic and anthropogenic components and performed well across multiple locations. However, while acoustic index-based approaches offer interpretable and scalable alternatives to feature-rich models, they often trade off taxonomic specificity or adaptability to novel acoustic conditions (Alcocer et al., 2022).

Recent studies have investigated novel approaches that integrate advanced feature extraction and dimensionality reduction with clustering or classification (Rendon et al., 2023). For instance, Cominelli et al. (2024) developed a pipeline using Visual Geometry Group (VGG)ish-derived acoustic embeddings, UMAP, and RF to classify cetacean vocalisations in marine soundscapes. Their approach achieved high classification accuracy ranging from 72–84% and demonstrated robustness across multiple locations. However, their approach was primarily restricted to a single taxonomic group and struggled to handle extreme low- and high-frequency calls outside the range of the VGGish model.

Several approaches have focused more directly on biophony and species-level separation. Guerrero et al. (2023) introduced the Learning Algorithm for Multivariate Data Analysis (LAMDA) $3\pi$ clustering algorithm using linearly spaced cepstral coefficients and frequency features. This approach achieved up to 96% detection accuracy across four datasets encompassing diverse taxa, including birds, frogs, and mammals. It was applied to geographically diverse locations and yielded strong generalisation within the biophony domain. However, the datasets used were manually segmented and limited to species-specific sounds, excluding geophony and anthropogenic noise. As a result, while the method is well-suited for targeted biodiversity assessments, its ability to scale to unstructured, multi-source

Table 1: Comparison of ecoacoustics segmentation approaches based on key properties and limitations.

| Ref. | Approach | Type | Features | Taxonomical Scope | Abiotic Included | Computational Efficiency | Generalisability |
|---|---|---|---|---|---|---|---|
| Lin et al. (2017) | PC-NMF | Clustering | Spectral Basis Matrix + Encoding Matrix | Multi-group | Yes | Moderate | High |
| Phillips et al. (2018) | $k$-means + HCA | Clustering | 12 Acoustic Indices | Multi-group, multi-species | Yes | High | Low |
| Scarpelli et al. (2021) | Time-Series Motif Discovery + RF | Classification | ACI, ENT, EVN | Multi-group, non-species specific | Yes | Moderate | Moderate |
| Guerrero et al. (2023) | LAMDA $3\pi$ | Clustering | Linearly-spaced cepstral coefficients + frequency information | Multi-group, multi-species | No | Moderate | High |
| Rendon et al. (2023) | Gaussian Mixture Models | Clustering | 15 Acoustic Indices | Multi-sound, non-species specific | Yes | Low-Moderate | Moderate |
| Cominelli et al. (2024) | Pre-trained acoustic models + UMAP + RF | Classification | VGGish acoustic features | Single group, multi-species | Yes | Low-Moderate | Moderate |
| Nieto-Mora et al. (2024) | Autoencoders + $k$-means + UMAP | Clustering | Autoencoder features | Multi-group, non-species-specific | Partially | Moderate | Moderate |
| Poutaraud et al. (2024) | MEC (Meta-Embedded Clustering) | Clustering | Spectrograms + CNN embeddings | Single group, Multi-species (birds) | No | Moderate | Moderate |

soundscapes remains uncertain.

Building on DL advances, (Poutaraud et al., 2024) presented Meta-Embedded Clustering (MEC), which combines CNN-derived spectrogram embeddings with meta-learning and HDBSCAN clustering. MEC achieved high cluster metrics and showed potential for unsupervised species detection. However, the method was tested exclusively on a single biophonic dataset and purposefully excluded abiotic sounds. Furthermore, the training and inference pipeline relied on high-end Graphics Processing Unit (GPU) infrastructure, indicating moderate scalability and limiting practical deployment in resource-constrained settings. DL representations have also seen use, such as in a study by Nieto-Mora et al. (2024) who proposed an autoencoder-based clustering pipeline that uses unsupervised feature learning combined with $k$-means and UMAP. Its capacity to operate without training labels is notable; however, validation was restricted to a single biome and recordings with high a signal-to-noise ratio were excluded, including heavy rainfall and insect choruses. Scalability and broader ecological generalisability thus remain open questions.

### 2.5. *Summary of literature gap*

As shown in Table 1, no single approach fully satisfies all dimensions required for robust and computationally tractable soundscape segmentation across heterogeneous ecological settings. While various methods have demonstrated utility in specific contexts, they only partially address computational efficiency, flexibility, or comprehensiveness when applied to large and multi-faceted datasets such as those associated with long-term PAM programs. Traditional approaches have shown potential in isolating biophonic elements or classifying species-specific sounds but are limited in handling overlapping sound sources or incorporating non-biophonic elements like geophony and anthrophony (Lin et al., 2017; Phillips et al., 2018).

Another observed limitation is narrowness in taxonomic scope or focus on biophony, leaving the interactions between biophony, geophony, and anthrophony underexplored. We argue that this is key to understanding complex community dynamics, such as the effect of anthrophony on biophony, which is a developing area of study (Grinfeder et al., 2022; Kok et al., 2023). Furthermore, the integration of abiotic sound sources and environmental context into these analyses remains inconsistent, which is required for capturing the full detail of soundscapes. Additionally, although recent efforts utilising dimensionality reduction methods like UMAP and clustering have demonstrated improved performance in handling complex ecoacoustic datasets, some existing studies are constrained by the computational scalability or generalisability demands of large-scale applications (Nieto-Mora et al., 2024; Cominelli et al., 2024). Together, these studies provide important methodological contributions to ecoacoustic segmentation, each addressing specific challenges such as clustering fidelity or taxonomic identification. However, we note that no existing approach to date concurrently fulfills all core requirements for large-scale, generalisable ecoacoustic analysis.

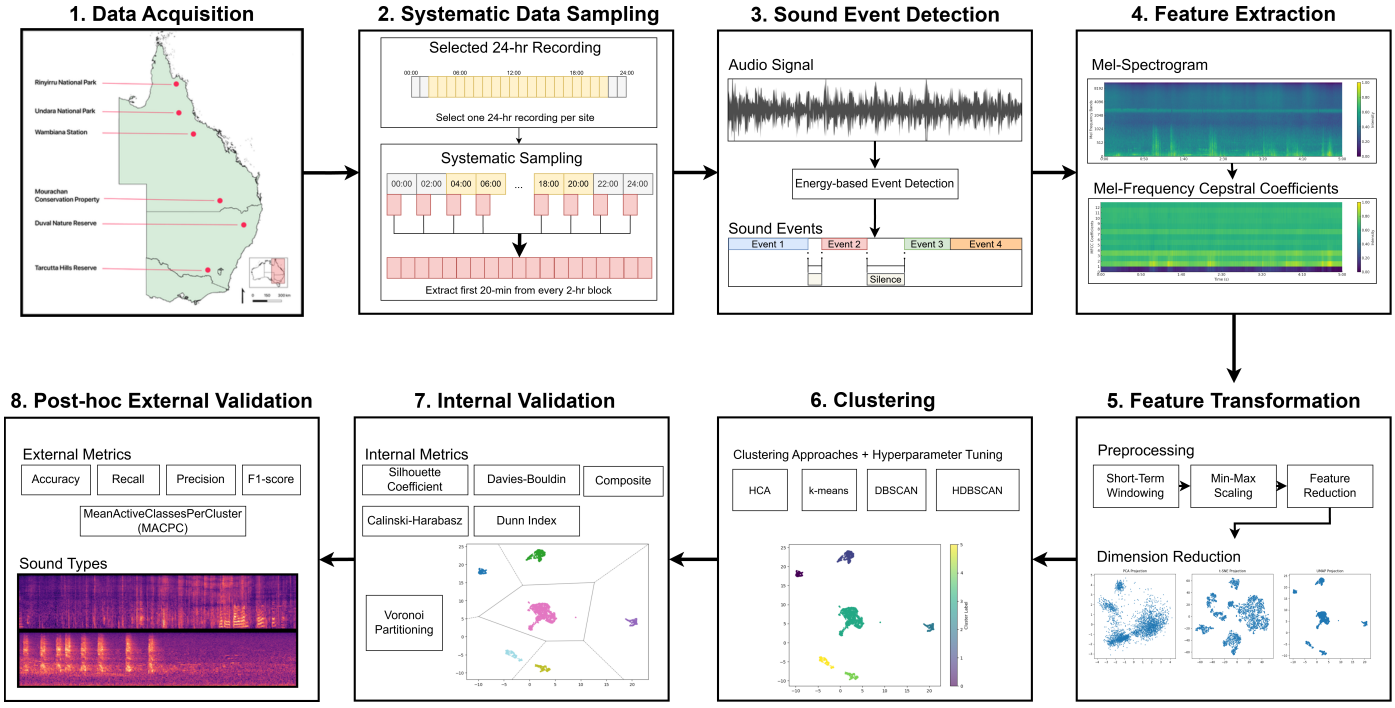



Figure 1: Overview of the proposed unsupervised segmentation method.

## 3. Proposed Unsupervised Segmentation Method

The proposed framework begins by collecting data from the A2O, which provides a richly varied ecoacoustic dataset spanning multiple sites and ecological zones. To ensure representative yet computationally feasible coverage, we apply systematic subsampling (see Section 4.1 for details of this sampling strategy). After sampling, we apply Sound Event Detection (SED) to isolate meaningful events from background noise. We compute short-term signal energy and mark intervals that exceed a threshold; a peak-picking algorithm then pinpoints the start and end of each high-energy event. These high-energy intervals (e.g., animal calls, wind gusts, rain) are extracted while skipping low-energy (silent) periods. Each detected event is further split into non-overlapping 4.5-second segments.

The decision to adopt a 4.5-second window length derives from balancing the need to capture sufficient acoustic information for species call analysis against the practical constraints of computational efficiency. Similar durations (e.g., 3-5 seconds) are widely adopted in ecoacoustic literature as they are short enough to assume local stationarity of sound while long enough to encompass individual vocalisations or sound events (Bradfer-Lawrence et al., 2019; Kahl et al., 2021; Williams et al., 2025).

After segmentation, each 4.5-second audio snippet undergoes preprocessing to standardise the signal prior to feature extraction. Normalisation follows using min-max scaling, which rescales each feature to a common range and helps prevent attributes with larger numeric magnitudes from dominating the downstream dimensionality reduction and clustering analysis. These steps improve consistency prior to dimensionality reduction and clustering.

After preprocessing, each 4.5-second segment is represented by MFCCs, which approximate human auditory perception and serve as a strong baseline for wildlife sound classification even in noisy conditions (Mcloughlin et al., 2019; Bonet-Solà and Alsina-Pagès, 2021; Lakdari et al., 2024). We use static MFCCs only, excluding delta and delta-delta coefficients, to keep the feature space compact and reduce redundancy and noise amplification associated with derivative features (Li et al., 2017; Clink et al., 2019; Napier et al., 2023). Although classical hybrid Gaussian Mixture Model-Hidden Markov Model pipelines (GMM-HMM) often included these derivatives (Brown and Smaragdis, 2009), a compact static representation was more appropriate for the present clustering objective.

To mitigate the curse of dimensionality and improve clustering efficiency, we applied three dimensionality reduction methods: PCA, *t*-SNE, and UMAP. PCA is a linear method that preserves as much variance as possible, whereas *t*-SNE and UMAP are non-linear techniques that emphasise local structure in the data. For each method, we evaluated both 2D and 3D output spaces, allowing us to obtain compact latent spaces that improved tractability for clustering and enabled direct visual comparison across approaches. These settings are consistent with prior use of *t*-SNE and UMAP for low-dimensional representation and exploratory visualisation (Thomas et al., 2022; Cominelli et al., 2024). For consistency of interpretation and presentation, we retained the 2D embeddings in the final reporting, as these preserved the principal separation patterns observed during testing while providing the clearest basis for comparative figures and partition-based visualisation. We then judged these methods by how well their reduced feature spaces preserved the important relationships among segments, that is, their

ability to group similar acoustic patterns. In the resulting low-dimensional feature space, we clustered the segments using several unsupervised algorithms to automatically discover sound-type groupings. For example, we used *k*-means, a centroid-based method requiring a preset number of clusters, and Hierarchical Clustering Analysis (HCA), which produces a dendrogram and allows clusters to be defined at various linkage thresholds. We used criteria such as the silhouette coefficient and elbow method to help choose the number of clusters for these methods. Using multiple clustering algorithms allowed the framework to accommodate the heterogeneity of the data, as some sound types formed tight, well-defined clusters, whereas others overlapped in time or frequency and were harder to segregate.

To assess the ecological relevance of the clustering results, we conducted an expert-guided post-hoc validation process using binary manual annotations available for each dataset. After clustering, a random subset of ten percent of points from each cluster was manually labelled according to audible sound types. The most frequently occurring label among these samples was used to describe the dominant class within each cluster (Napier et al., 2025). This procedure provides a qualitative and quantitative check on whether the clusters correspond to ecologically interpretable acoustic units, rather than being only mathematically well separated in feature space.

We then compared the resulting cluster-level assignments to the original binary annotations using five metrics: accuracy, precision, recall, F1-score and Mean Active Classes Per Cluster (MACPC). These scores quantify how well the clusters correspond to consistent, semantically meaningful sound types that are ecologically grounded.

## 4. Experiment Settings

### *4.1. Datasets*

We selected six ecologically distinct sites from the A2O network, spanning savannahs and dry sclerophyll woodland habitats at low to moderate altitude along the eastern side of Australia (Allen-Ankins et al., 2023). These sites were chosen for their wide variety of acoustic signatures representative of different habitats. Data collection followed a systematic sampling strategy. Recordings contained a mix of biophonic, geophonic, and occasional anthropogenic sounds, reflecting real-world ecoacoustic conditions where human-generated noise is unevenly distributed.

Each site as part of the A2O network, includes four autonomous acoustic recorders at fixed locations (two near a water source and two further away) (Roe et al., 2021). For this study, we selected one recorder per site to keep the dataset size tractable while still capturing the range of microhabitats. From the chosen recorder at each site, a single 24-hour period was selected. To ensure that this period was representative of the site's acoustic diversity, we reviewed long-duration false-colour spectrograms (Towsey et al., 2018) and chose days with varied and active soundscapes. This step aimed to maximise ecological coverage without favouring any specific taxa or acoustic condition.

Within each selected day, we systematically extracted the first 20 minutes from every two-hour block. This approach, consistent with established ecoacoustic sampling practices (Linke and Deretic, 2020; Wimmer et al., 2013b; Cifuentes et al., 2021), balances computational efficiency with ecological representativeness. This sampling schedule maximised temporal diversity by capturing diurnal, nocturnal, and crepuscular activity. All recordings were segmented into 4.5-second non-overlapping windows without denoising, preserving the natural complexity of overlapping sound events in both time and frequency domains.

The resulting dataset comprised $n$=19,230 high-energy segments (3,205 per site), each 4.5s in length for a total of 4 hours of audio per location over a complete 24-hour cycle. By incorporating spatial heterogeneity and multi-habitat sampling, this dataset provides a strong testbed for evaluating our framework's generalisability. It ensures that our approach is not confined to a single environment or species group but is instead broadly applied across multiple realistic ecological settings.

### *4.2. Comparison algorithms and hyperparameter settings*

Four clustering algorithms were applied: HDBSCAN, *k*-means, DBSCAN, and HCA. These algorithms represent a range of clustering paradigms: HDBSCAN identifies clusters of varying density and determines the number of clusters automatically; *k*-means partitions data by minimising intra-cluster variance but requires a predefined number of clusters *k*; DBSCAN forms clusters based on density connectivity and neighbourhood criteria; and HCA constructs a dendrogram of pairwise distances, allowing clusters to be extracted using adjustable linkage thresholds.

We systematically tuned each clustering algorithm's key parameters across ranges chosen to balance sensitivity to small recurring acoustic structures, robustness to noise and fragmentation, and computational tractability. For HDBSCAN, we varied the minimum cluster size from 5 to 50 and the minimum samples from 5 to 20 to assess a continuum from more permissive settings, which retain smaller and potentially rare sound groupings, to more conservative settings, which suppress noise-driven clusters and enforce denser local structure, consistent with the hierarchical density-based formulation of the method (Campello et al., 2013). For *k*-means, we tested values of $k$ from 2 to 50 so that both coarse and finer acoustic partitions could be explored without imposing an implausibly large number of clusters a priori; selection was guided by established cluster-number criteria including elbow-style inspection and the silhouette coefficient (Rousseeuw, 1987). For DBSCAN, we varied radius $\varepsilon$ from 0.1 to 1.0 and `min_samples` from 3 to 10 to examine stricter versus looser neighbourhood-density definitions in the reduced feature space, following standard parameterisations of density-based clustering (Ester et al., 1996; Schubert et al., 2017). For HCA, we compared single, complete, and average linkage because these reflect different assumptions about cluster geometry, with single linkage favouring chaining structures, complete linkage favouring compact clusters, and average linkage providing an intermediate compromise (Murtagh and Contreras, 2012). Candidate configurations were then assessed using the internal validation metrics described in Sec-

tion 4.3.1 to identify settings that produced the most coherent and well-separated cluster structures.

### 4.3. Performance measures

In this study, segmentation quality is evaluated in two complementary ways. First, we assess whether the discovered groupings are structurally coherent in feature space, that is, whether acoustically similar segments form compact and well-separated clusters under unsupervised conditions. Second, we assess whether these groupings are ecologically interpretable by examining their correspondence to manually annotated sound types and their within-cluster acoustic complexity. We therefore define high-quality segmentation not as recovery of a dense event-level ground truth, which is unavailable at this scale, but as the consistent formation of acoustically coherent and ecologically meaningful acoustic units.

#### 4.3.1. Internal validation

Internal validation metrics assess the structure and quality of clusters based on compactness, separation, and definition. These metrics guide hyperparameter optimisation for algorithms such as $k$-means and DBSCAN, particularly in selecting the optimal number of clusters in an unsupervised setting.

We use four widely applied clustering validity indices: the Calinski–Harabasz (CH) Index (Caliński and Harabasz, 1974), which measures the ratio of between-cluster to within-cluster dispersion (higher values indicate more compact and well-separated clusters, suggesting clearer acoustic boundaries); the Dunn Index (DI) (Dunn†, 1974), which captures the ratio between minimum inter-cluster distance and maximum intra-cluster diameter (higher values indicate stronger separation between distinct sound types); the Silhouette Coefficient (SC) (Rousseeuw, 1987), which quantifies how well each point lies within its assigned cluster relative to others (higher scores suggest that ecoacoustic segments are acoustically cohesive and distinct from neighbouring clusters); and the Davies–Bouldin (DB) Index (Davies and Bouldin, 1979), which evaluates average cluster similarity (lower values indicate better separation and reduced overlap between sound categories). Together, these metrics provide a robust, label-free evaluation framework for quantifying cluster definition and separability in ecoacoustic feature space. However, because structural compactness alone does not guarantee ecological relevance, these internal measures are complemented by post-hoc external validation against manually annotated sound types.

#### 4.3.2. External validation

External validation compares the clustering output against partially ground truthed labels obtained through manual annotation. Following a majority-vote label propagation from a randomly selected subset of points in each cluster, we compute accuracy, precision, recall, and F1-score to quantify the agreement between predicted cluster membership and annotated classes. In this study, these labels do not represent fine-grained event boundaries for every sound in the corpus, but instead correspond to ecologically recognisable sound types audible within each segment. Accordingly, external validation is used to assess whether the discovered clusters align with biologically meaningful acoustic units, such as birds, insects, frogs, mammals, wind, rain, vehicles, or mixed-source categories, rather than to imply recovery of exhaustive species-level ground truth. In the ecoacoustic context, high precision indicates that clusters are acoustically specific (few false inclusions of unrelated sound types), high recall reflects that clusters capture most relevant events of a given type, and a high F1-score demonstrates a balanced trade-off between these properties.

To complement traditional metrics, we introduce MACPC. MACPC quantifies the average number of active classes per sample within each cluster, averaged across all clusters:

$$\text{MACPC} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{i\in C_k}\sum_{j=1}^{L}1[y_{ij} > \tau] \quad (1)$$

where $K$ is the number of clusters, $C_k$ is the set of sample indices in cluster $k$, $L$ is the total number of possible classes, $y_{ij}$ is a binary indicator for the presence of class $j$ in sample $i$, and $\tau$ is a detection threshold (typically 0.5).

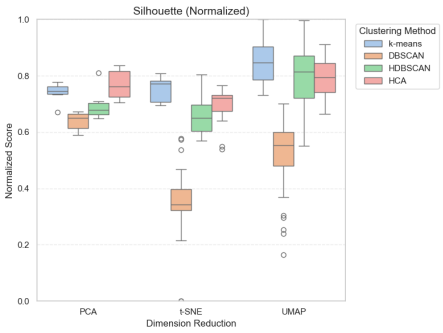
MACPC is a cluster-conditional adaptation of Label Cardinality (LC) from multi-label learning (Zhang and Zhou, 2013). Classical LC is the average number of active labels per instance in a dataset (Tsoumakas and Katakis, 2008). This links an established multi-label property to ecoacoustic clustering quality. In ecological terms, MACPC serves as a proxy for within-cluster acoustic complexity, where lower MACPC values suggest greater acoustic homogeneity, meaning clusters are dominated by a single sound type. Higher MACPC values indicate increased acoustic heterogeneity, often due to overlapping calls, mixed-species choruses, or the co-occurrence of biophony and abiotic noise. While this reflects richer ecological scenes, it typically reduces cluster purity and may complicate downstream segmentation. By comparing MACPC values with external validation scores, we can assess whether greater within-cluster diversity corresponds to reduced alignment with ground-truthed labels, as observed in sites with high MACPC and lower F1-scores.

To negate label–model circularity, all clustering and dimensionality reduction were performed without labels. Labels were used only post-hoc for (i) sparse majority-vote propagation from a random per-cluster subset and (ii) external validation metrics. No labels were used to select UMAP/HDBSCAN hyperparameters or to train representations. This procedure preserves the unsupervised nature of the workflow while permitting independent assessment against human annotations.

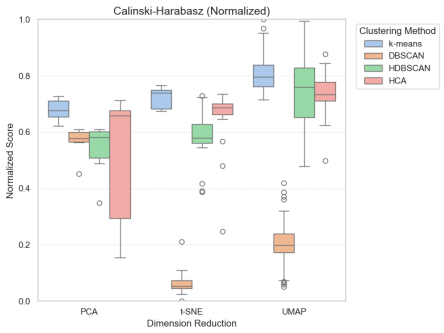
## 5. Results and Discussion

### 5.1. Clustering performance comparison

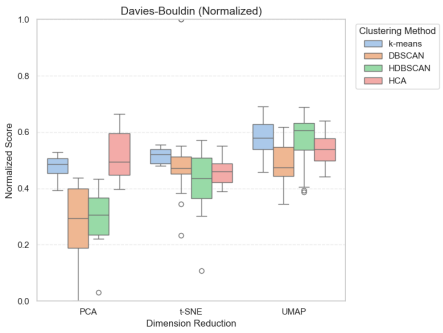
UMAP combined with HDBSCAN outperformed all other methods on the five internal validation metrics, achieving the highest normalized scores and smallest interquartile ranges (IQRs). Specifically, as seen in the SC (Figure 2a) and CH Indices (Figure 2b), HDBSCAN under UMAP achieved high normalised



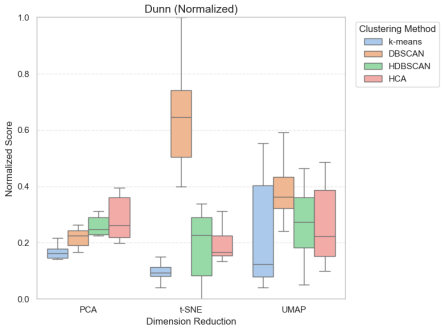

(a) Silhouette score, reflecting intra-cluster cohesion versus inter-cluster separation.



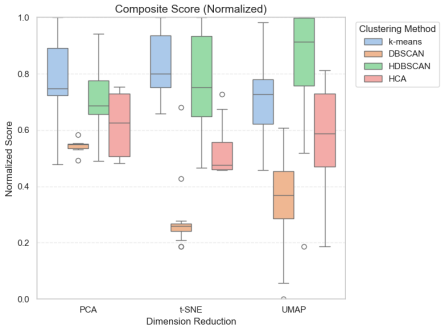

(b) Calinski-Harabasz index, rewarding compact and well-separated clusters.




(c) Davies-Bouldin index (inverted), where higher scores indicate better clustering.




(d) Dunn index, capturing the ratio between smallest inter-cluster and largest intra-cluster distances.




(e) Composite Score (CS), a normalised aggregation of all internal metrics.

Figure 2: Comparison of internal clustering validation metrics across different clustering algorithms (*k*-means, DBSCAN, HDBSCAN, HCA) and dimension reduction methods (PCA, *t*-SNE, UMAP). Each score is normalized to the [0, 1] range for comparability. These results support the selection of UMAP with HDBSCAN as a leading configuration for ecoacoustic clustering based on internal consistency.

values with tight IQRs, reflecting compact and well-separated clusters. Its performance on the inverted DB Index (Figure 2c) further supports this with the highest mean, indicating reduced overlap and high inter-cluster contrast. The DI (Figure 2d) values, though generally lower across all methods, remained competitive for UMAP-HDBSCAN, suggesting meaningful separation even in difficult conditions. The CS (Figure 2e) confirmed this balanced performance, with HDBSCAN under UMAP yielding the highest aggregate scores across datasets.

The strength of this pairing likely stems from UMAP's ability to preserve both local and global data structures while flattening complex manifolds, a property particularly useful in ecoacoustic feature spaces where different sound types may form nested or curved clusters (McInnes et al., 2018). HDBSCAN, which excels in identifying clusters with variable density and robustness to noise, complements this by not assuming globular boundaries and allowing for the detection of subtle acoustic groupings (Campello et al., 2013).

*k*-means also performed strongly, particularly under PCA and UMAP embeddings. However, its performance dropped on the DI (Figure 2d), suggesting weaker inter-cluster separation when clusters are closely packed or non-spherical. Its CS remained high (Figure 2e), especially under UMAP, indicating reliable performance when clusters are well-separated and the data structure aligns with its assumptions. Nevertheless, its sensitivity to initialisation and cluster count makes it less adaptive to the dynamic and overlapping nature of ecoacoustic data.

DBSCAN exhibited the most variability across metrics and dimensionality reduction techniques. Its scores for the SC and CH Indices were relatively low under t-SNE and UMAP (Figures 2a and 2b), suggesting poor cohesion and dispersion of clusters. The method did perform relatively well on the DI under t-SNE (Figure 2d), but this was inconsistent across other embeddings. DBSCAN's CS (Figure 2e) were the lowest of all algorithms, highlighting its difficulty in reliably extracting meaningful structure from the data. This is likely due to its reliance on fixed density thresholds, which are ill-suited to the high variability in density and temporal overlap found in ecoacoustic recordings.

HCA demonstrated moderate, stable performance across all metrics. It achieved mid-range scores in the SC and CH Index (Figures 2a and 2b), and maintained competitive CS (Figure 2e), though it was never the top performer. HCA's DI values were slightly higher than those of *k*-means and DBSCAN in some configurations (Figure 2d), indicating better relative separation in certain contexts. However, its overall performance lacked standout qualities, suggesting it may be more useful for interpretability or hierarchical insights rather than maximising cluster quality.

Table 2: Post-hoc external validation metrics for unsupervised clustering across six ecoacoustic datasets. Each site underwent cluster-based annotation via majority-vote label propagation from a manually labelled subset. Accuracy, precision, recall, and F1-score assess the alignment between cluster assignments and the propagated classes. MACPC quantifies the average number of active sound types per sample within each cluster (based only on sampled points), serving as a proxy for within-cluster ecological complexity. Also shown are the number of unique sound types and clusters identified.

| Dataset | Accuracy | Precision | Recall | F1-score | MACPC | # Sound Types | # Clusters | Sound Types Present |
|---|---|---|---|---|---|---|---|---|
| Undara | 0.848 | 0.839 | 0.848 | 0.837 | 1.525 | 8 | 18 | birds, frogs, human_speech, insects, misc/uncertain, rain_light, vehicles, wind_light |
| Duval | 0.898 | 0.849 | 0.898 | 0.870 | 1.430 | 9 | 15 | birds, human_speech, insects, mammals, misc/uncertain, rain_heavy, rain_light, vehicles, wind_light |
| Rinyirru | 0.940 | 0.959 | 0.940 | 0.945 | 1.227 | 6 | 12 | birds, insects, mammals, misc/uncertain, vehicles, wind_light |
| Mourachan | 0.954 | 0.914 | 0.954 | 0.933 | 1.723 | 10 | 19 | birds, frogs, insects, mammals, misc/uncertain, rain_heavy, rain_light, vehicles, wind_strong, wind_light |
| Wambiana | 0.985 | 0.982 | 0.985 | 0.983 | 1.100 | 4 | 6 | insects, birds, wind_light, wind_strong |
| Tarcutta | 0.930 | 0.920 | 0.930 | 0.925 | 1.350 | 5 | 10 | insects, birds, rain_heavy, wind_light, wind_strong |

### 5.2. *Post-hoc external validation of cluster assignments*

Table 2 includes the number of clusters identified at each site and the number and type of sound sources present. Across sites, configurations with lower MACPC tended to achieve higher external F1 (e.g., Wambiana: MACPC = 1.10, F1 = 0.983), while acoustically complex sites (e.g., Undara, Duval) exhibited higher MACPC and lower F1. This pattern suggests that cluster purity, and therefore downstream interpretability, declines as polyphony increases. In practice, MACPC provides a quantitative handle on where automated segmentation may require human review or higher temporal–spectral resolution. Because MACPC is label-agnostic, it generalises to biotic and abiotic mixtures, offering a practical screen for soundscapes in which geophony/anthrophony confound biophony indices.

The results demonstrate a clear relationship between clustering effectiveness and acoustic complexity. Wambiana, a site with relatively few sound types and distinct acoustic boundaries, showed extremely high alignment with F1 = 0.983 and the lowest MACPC = 1.100. Across all datasets, clustering consistently grouped acoustically similar events, supporting the ecological validity of the segmentation approach in varied environments.

### 5.3. *Computational efficiency and runtime growth*

Although the present evaluation was conducted on a systematic subset rather than a full observatory-scale corpus, runtime behaviour remains important for judging practical deployment. Figure 3 therefore reports empirical processing times for HDBSCAN paired with PCA, t-SNE, and UMAP within the evaluated data regime, while the dotted extensions indicate extrapolated trends rather than directly demonstrated end-to-end performance at larger scales. Please note that HDBSCAN exhibits subquadratic complexity, approximately $O(n \log n)$ where $n$ is the number of objects. This supports its applicability to large datasets (Campello et al., 2013). The three paired dimension reduction approaches require slightly different complexity. PCA is the most efficient approach requiring linear time complexity $O(n \times d)$ where $d$ is the number of dimensions, and typically $d$ is much smaller than $n$, thus ending up with $\approx O(n)$ (Abdi and Williams, 2010). $t$-SNE has subquadratic time complexity, $O(n \log n)$ when implemented with tree-based approximation (van der Maaten, 2014), making it feasible for large scale applications. UMAP demonstrates near-linear scalability with an overall complexity of approximately $O(n \log n)$ primarily enabled by approximate nearest neighbour search (McInnes et al., 2018). This makes it more scalable than $t$-SNE while generally less computationally efficient than PCA. This is evidenced by Figure 3.

The results indicate a marked difference in algorithmic efficiency across the three methods. PCA exhibits the most favourable efficiency properties, with processing times remaining consistently low even as dataset sizes increase. This is expected, given PCA's linear nature and relatively low computational complexity. UMAP demonstrates competitive scalability, outperforming $t$-SNE across all dataset sizes while still capturing essential data structures. Although UMAP's computation time increases with dataset size, it remains within practical limits, making it a suitable choice for large-scale ecoacoustic analysis. This is further corroborated by examining the extrapolated trends (dotted lines) which suggest that processing times for $t$-SNE become prohibitive for sample sizes beyond $10^5$, making it unsuitable for large-scale applications.

To contextualise our approach, we approximated the processing times of two external benchmark methods from the literature: MEC and LAMDA $3\pi$. Please note that both MEC and LAMDA $3\pi$ do not have a single closed-form complexity, as they are composite pipelines. Their computation time is determined by the underlying embedding, dimensionality reduction, and clustering components. For MEC, reported run-

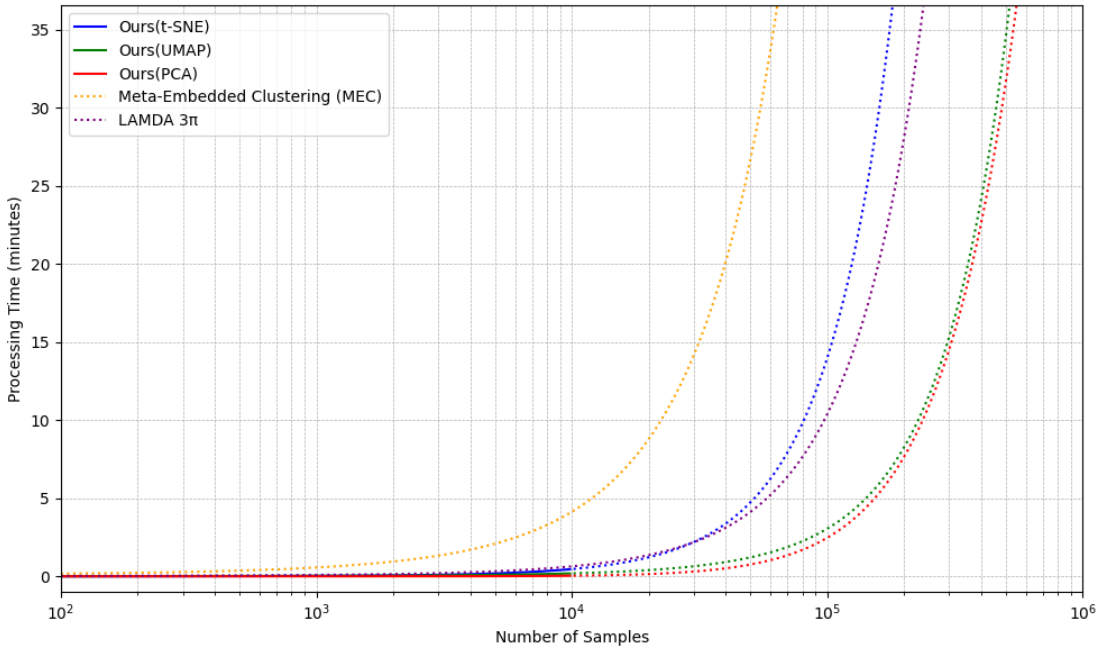



Figure 3: Runtime analysis of proposed clustering configurations versus external benchmarks. Processing time (minutes, $y$-axis) is plotted against the number of 4.5-second audio samples (log-scaled, $x$-axis). Solid lines show empirically measured durations for our proposed combinations ($t$-SNE, UMAP, PCA) with HDBSCAN, while dotted segments beyond $10^4$ samples represent extrapolated estimates. LAMDA $3\pi$ (Guerrero et al., 2023) and Meta-Embedded Clustering (MEC) (Poutaraud et al., 2024) are shown as dotted curves based on approximated runtimes from published implementations. The results indicate that UMAP provides the most favourable balance between runtime and representation quality among the evaluated methods.

times include an extensive GPU-based fine-tuning stage involving 50,000 region-of-interest images over 10 hours, along with hierarchical patch decomposition and multi-view CNN encoding (Poutaraud et al., 2024). Based on this, MEC was estimated to be approximately 22× slower than our UMAP-based approach. For LAMDA $3\pi$, the workflow incorporates custom noise reduction, Otsu-based segmentation, spectral entropy calculations, and sequential LAMDA clustering (Guerrero et al., 2023), which collectively yield a runtime of an estimated 20 seconds per 1-minute recording. This was benchmarked against our UMAP + HDBSCAN pipeline, which processes the same file length in approximately 6 seconds, resulting in a relative approximation of 3.4× slower for LAMDA $3\pi$.

The choice of dimension reduction method thus has substantial implications for ecoacoustic workflows. While $t$-SNE remains useful for visual exploration of small to mid-sized datasets, UMAP provides a more scalable alternative without significant loss of cluster structure fidelity. PCA, while computationally efficient, lacks the ability to preserve non-linear relationships and may not be optimal for all ecoacoustic segmentation tasks. Under the present experimental conditions, these results support the use of lightweight components for resource-constrained ecoacoustic workflows.

Species-level recognisers excel in targeted contexts lack scalability to continental observatories due to the long tail of taxa, annotation scarcity, and domain shift across habitats and recording conditions (Stowell, 2022). Networks like the A2O collect millions of hours of audio (Roe et al., 2021), rendering species-specific models difficult to maintain and validate across space and time. Our ecosystem or soundscape-level framing preserves information from overlapping sound sources that otherwise confound species indices, and it generalises without assuming a closed label set.

Relative to recent unsupervised ecoacoustic pipelines, our framework emphasises open-set, soundscape-level segmentation with lightweight components, trading maximal species resolution for robustness across heterogeneous sites. MEC (Poutaraud et al., 2024) improves clustering quality in unlabelled bird datasets via pseudo-labelling and meta-learning on CNN embeddings, showing gains in information-theoretic criteria but focusing on biophony and requiring pretrained backbones. LAMDA $3\pi$ (Guerrero et al., 2023) infers sonotypes and the number of clusters directly, demonstrating multi-dataset generality, yet primarily targets animal vocalisations and presumes cleaner segmentation. In contrast, we explicitly include abiotic sources, report overlap-aware diversity via MACPC and scale analysis to 24 hours with modest compute. Performance-wise, our external F1 scores (Table 2) sit alongside strong internal validity on simpler sites, with degraded mapping quality at high MACPC.

While prior studies have applied acoustic indices (Phillips et al., 2018; Scarpelli et al., 2021; Rendon et al., 2023) or CNN-based embeddings (Nieto-Mora et al., 2023; Poutaraud et al., 2024) for soundscape analysis, our approach differs by operating entirely unsupervised across biomes and incorporating abiotic sources without exclusion filters or pre-trained feature dependence. Relative to acoustic-index pipelines (Phillips et al., 2018; Scarpelli et al., 2021), our UMAP–HDBSCAN consistently achieved superior internal separation and stability across

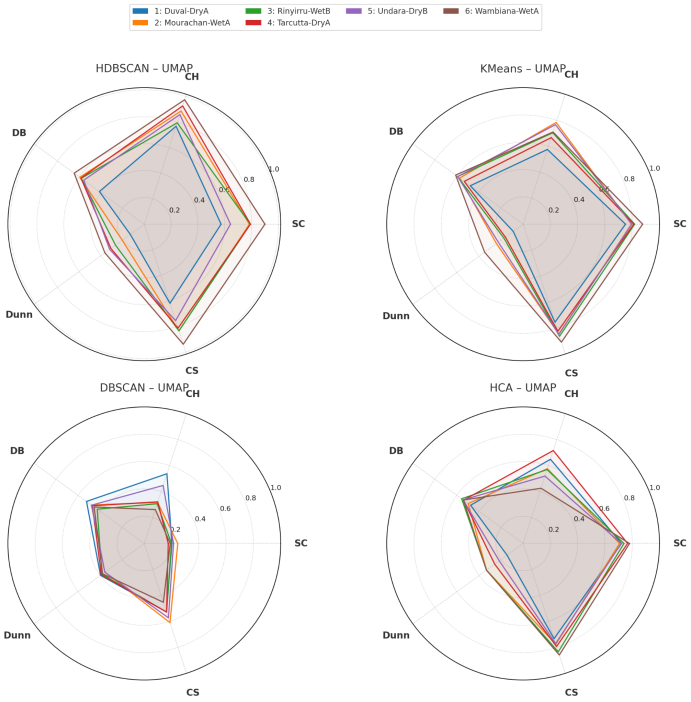



Figure 4: Radar plots comparing internal clustering validation metrics across six ecoacoustic sites for each clustering technique (HDBSCAN, *k*-means, DBSCAN, HCA), all using UMAP as the dimensionality reduction method. Each plot shows the normalised values (0–1 scale) for Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index (inverted), Dunn Index, and the Composite Score. Higher values indicate better clustering quality. Results show that HDBSCAN and *k*-means consistently outperform DBSCAN and HCA across most sites and metrics. Ecologically, stronger and more balanced scores suggest that a site contains more acoustically coherent and separable soundscape structure, whereas lower or uneven scores indicate greater polyphony, overlapping sound sources, or stronger abiotic interference.

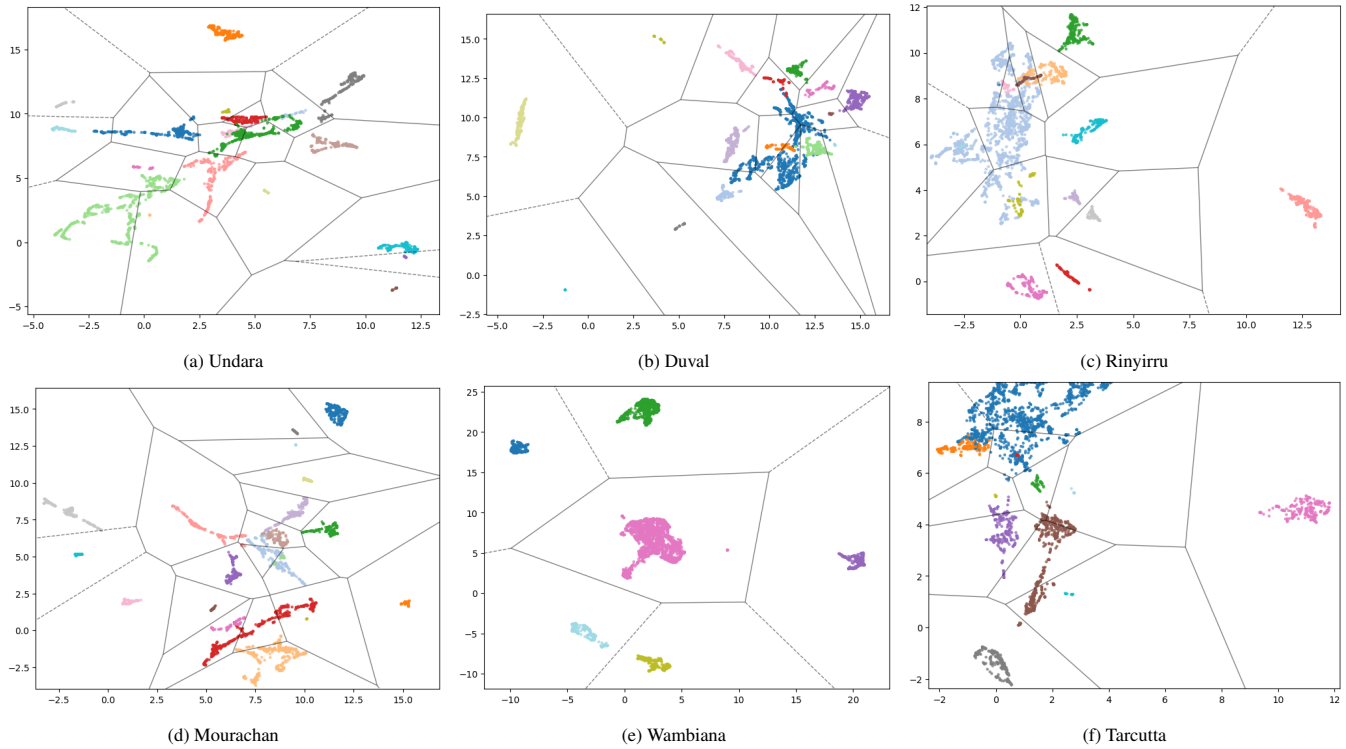
a greater number of ecoregions (Figures 2a–2e).

### *5.4. Per-site patterns and ecological interpretability*

Per-site patterns in Figure 4 align with external validation in Table 2, indicating that acoustic heterogeneity drives differences in separability and downstream labelling quality. Wambiana shows the strongest agreement with propagated ground truthing (F1 = 0.983) alongside the lowest within-cluster diversity (MACPC = 1.10) and a small set of sound types (4) and clusters (6). Tarcutta also performs well (F1 = 0.925; MACPC = 1.35; 5 sound types; 10 clusters). In contrast, Mourachan exhibits higher within-cluster diversity (MACPC = 1.723) and more clusters (19) across a larger palette of sound types (10), with correspondingly lower F1 than Wambiana (F1 = 0.933). Duval shows a similar pattern of elevated diversity (MACPC = 1.430; 9 sound types; 15 clusters) and lower F1 (0.870) relative to Wambiana and Tarcutta. Rinyirru achieves high external validity (F1 = 0.945) with moderate diversity (MACPC = 1.227), and Undara is comparatively harder (F1 = 0.837; MACPC = 1.525).

These external trends are consistent with the internal validity patterns in Figure 4 and Table 3: sites dominated by a small number of coherent sources yield tighter, better separated clusters and higher external agreement, whereas sites with overlapping biophony and variable abiotic noise show higher MACPC. Practically, this suggests that UMAP paired with HDBSCAN remains a strong default across sites, but parameter selection and post-labelling effort should be adjusted by site according to observed within-cluster diversity and the mix of sound types.

From an ecological perspective, the value of the proposed framework lies in its ability to organise heterogeneous soundscape archives into coherent acoustic units that remain interpretable at the level of ecological sound types. Soundscape ecology is concerned with how biological, geophysical, and human-produced sounds structure landscapes across space and time, and these components can therefore all contribute useful ecological information rather than serving merely as nuisance variation (Pijanowski et al., 2011; Fuller et al., 2015). In this context, "biologically meaningful acoustic units" should be interpreted at the soundscape-event level rather than as a guarantee of species-level syllable or call-boundary recovery. In heterogeneous natural recordings, meaningful units may correspond to dominant bird vocalisations, frog calls, insect choruses, rainfall, wind events, vehicle noise, or mixed acoustic scenes that remain ecologically informative in relation to habitat condition, disturbance, and temporal activity patterns (Teixeira et al., 2024). The purpose of the framework is therefore to recover ecologically coherent acoustic structure from complex soundscapes, not to assert perfect recovery of fine-grained behavioural or taxonomic units in all cases. Within long-term passive acoustic monitoring workflows, such structuring can reduce manual burden, help prioritise expert review toward the most informative segments, and provide a practical intermediary step between raw recordings and finer-grained ecological analyses or species-specific follow-up models (Gibb et al.,



(a) Undara (b) Duval (c) Rinyirru

(d) Mourachan (e) Wambiana (f) Tarcutta

Figure 5: Voronoi partitioning of clustered ecoacoustic data projected onto a 2D space using the best method. Each region represents the spatial domain closest to a given cluster centroid, illustrating the degree of separation between clusters. Larger, well-defined cells suggest relatively discrete and ecologically coherent soundscape units, whereas smaller, densely packed, or irregular regions indicate acoustically complex scenes with stronger source overlap, boundary ambiguity, or mixed biotic and abiotic content. Ecologically, these patterns help identify which sites contain clearer dominant sound structures, and where greater post-labelling effort or expert review may be needed.

2019).

### 5.5. *Spatial interpretation of clustering via Voronoi diagrams*

Voronoi tessellations are a standard device for partitioning embedding spaces and diagnosing local separation and overlap structure without labels (Okabe et al., 2009). Larger Voronoi regions suggest well-separated clusters, while smaller, densely packed cells indicate overlapping or ambiguous groupings. In our setting, they complement internal metrics by exposing boundary fragility and mixed neighbourhoods in UMAP space, which has become a common manifold-learning projection for high-dimensional features (McInnes et al., 2018). As seen in Figure 5, the observed variability across sites suggests that clustering performance is not uniform, aligning with the mixed SC (mean 0.484) and high variance in the CH index (std 4991.11), which indicate inconsistencies in cluster compactness.

The differences between sites are notable. As seen in Table 3, Tarcutta (SC = 0.63) and Wambiana (0.84) showed large, well-defined Voronoi cells, indicative of low acoustic overlap. Conversely, Mourachan, with a lower SC (0.23), showed fragmented Voronoi partitions, reflecting higher within-cluster ambiguity due to overlapping species calls and abiotic noise. Sites like Mourachan and Duval display intermediate patterns, reinforcing the dataset's heterogeneity and the difficulty of achieving universally well-separated clusters. In the context of the study's aims, the Voronoi diagrams validate the method's ability to process large-scale ecoacoustic data. These findings suggest that while the method is scalable, its robustness can be uneven, warranting further investigation into factors affecting cluster quality.

### 5.6. *Limitations*

We acknowledge that our segmentation approach does not address all challenges of ecoacoustic data. In particular, overlapping sounds during high-activity periods (e.g., dawn or dusk choruses) and complex background noise remain beyond our current scope. Such high polyphony can limit the effectiveness of both SED and clustering by reducing cluster separability and complicating label assignment even in advanced pipelines (Parrilla and Stowell, 2022). Instead, we have prioritised the establishment of a scalable and adaptable approach for organising acoustic data into coherent broad-scale groups, laying the groundwork for downstream tasks that require further refinement or expert interpretation.

However, while the framework successfully segments large-scale ecoacoustic data into broad sound types, several limitations may affect its consistency and interpretability. First, the taxonomic granularity of our clustering is coarse. Clusters reflect broad sound categories rather than single species, so in cases where multiple species produce very similar calls, our method may group them together. This means that species-level differences can be lost – for example, two bird species with similar spectro-temporal calls may end up in one cluster.

Table 3: Overall best internal clustering results per site.

| Site | Algorithm | Silhouette | Calinski-Harabasz | Davies-Bouldin | Dunn-Index | Composite |
|---|---|---|---|---|---|---|
| **Undara** | UMAP-HDBSCAN | 0.606 | 1852.615 | 0.388 | 0.592 | 3.999 |
| **Duval** | UMAP-HDBSCAN | 0.238 | 317.915 | 0.871 | 0.111 | 3.999 |
| **Rinyirru** | UMAP-*k*-means | 0.472 | 2980.872 | 0.785 | 0.021 | 4.000 |
| **Mourachan** | UMAP-HDBSCAN | 0.629 | 4039.231 | 0.439 | 0.067 | 3.999 |
| **Wambiana** | UMAP-HDBSCAN | 0.846 | 27522.670 | 0.211 | 0.314 | 3.999 |
| **Tarcutta** | UMAP-HDBSCAN | 0.6309 | 7740.708 | 0.426 | 0.333 | 3.999 |

As a result, if one attempted species-level labelling of the clusters, precision and recall would likely drop due to such mixed-species groupings. Our validation at the sound type level avoids this issue by using broader labels, but it also highlights that fine taxonomic resolution is not achieved by the current pipeline.

Another limitation is the problem of simultaneous overlapping from multiple species and other sound types. The clustering method does not explicitly separate overlapping sources; we only quantify it with MACPC, meaning that clusters may yet still contain only partially disentangled sounds rather than discrete taxonomic groups. A further consideration is domain shift, as configurations that perform well under one set of habitats or acoustic conditions may not transfer perfectly to other ecological settings without additional validation. Broader variation in recording conditions may therefore still influence clustering consistency. This reduces the interpretability of results and somewhat limits the ability to attribute clusters to specific taxa or sound events without additional post-processing or expert intervention.

Furthermore, while our evaluation spanned 24 hours, scaling to larger datasets (e.g., a month of continuous audio) may pose additional challenges. In terms of computation, the pipeline's runtime and memory use grow with the number of segments. Certain steps, like dimensionality reduction and clustering, may become bottlenecks at larger scales. Using more scalable techniques such as distributed computing would be necessary to handle a larger dataset. In terms of clustering outcomes, a larger dataset might produce more clusters and finer-grained distinctions, but it could also introduce more variability. A full month of recordings will capture rarer events and a greater variety of conditions, which could lead to cluster fragmentation and require the need for higher clustering parameter adjustments. This reliance on tuning would increase the complexity of applying the method effectively, particularly for users without extensive ML expertise. Thus, although our framework is designed with efficiency in mind, its performance and output would need careful assessment at higher data volumes.

## 6. Conclusion

We present a broad-scale exploratory study focused on the unsupervised segmentation of large-scale, high-dimensional ecoacoustic datasets. Our evaluation across multiple geographic sites supports the robustness and computational practicality of the proposed framework within the evaluated data regime. We leverage computational intelligence to reveal latent acoustic structures without predefined taxonomic labels. Our findings demonstrate the potential of unsupervised clustering to extract meaningful ecoacoustic patterns, offering an adaptive methodology for guiding expert annotation efforts.

However, analysis of clustering performance, including Voronoi diagram interpretations, reveals some inconsistencies in robustness across different sites. The method remains sensitive to hyperparameters, and assumes continuous temporal autocorrelation, limiting its effectiveness in rapidly shifting soundscapes. These findings emphasise the need for further refinements, including adaptive clustering strategies, improved feature representations, and automated hyperparameter tuning.

Beyond segmentation, this framework could serve as a foundation for automated biodiversity monitoring by identifying changes in acoustic communities over time or detecting anomalous events (e.g., absence of expected calls, emergence of new anthropogenic noise). Unsupervised segmentation offers a complementary approach in scenarios where species-specific classifiers are impractical such as remote or under-annotated regions. Rather than outputting species identities, this framework groups audio segments by sound similarity into broad acoustic categories that can support subsequent species-focused monitoring. This accomplishes a different but valuable goal: it preserves ecological information from overlapping sources and unknown calls without requiring any training labels, thus providing an organised view of the complete soundscape that species-level methods cannot easily give.

Future work should explore hybrid approaches that incorporate weak supervision and techniques for handling dense acoustic overlaps in both frequency and time need to be refined for more accurate segmentation. Larger-scale empirical benchmarking is also needed to assess runtime and memory behaviour under month-scale and observatory-scale deployments, rather than within the systematic subset evaluated here. This includes profiling peak memory usage, storage overheads, and batching or distributed-processing strategies as dataset size increases. Further, more comprehensive assessment of the framework's generalisability across global-scale ecoacoustic datasets is needed for true real-world application. Addressing these challenges will enhance the framework's reliability, making it a genuinely more effective tool for automated ecoacoustic analysis and large-scale passive acoustic monitoring applications.

## Data Availability

The code and processed feature data used in this study are publicly available at: https://github.com/thomasnapier/unsupervised-soundscape-segmentation. The original raw audio recordings are publicly available from the Australian Acoustic Observatory: https://data.acousticobservatory.org/. Because the analyses were conducted on derived 4.5-second segments extracted from the original recordings, the repository additionally provides scripts for retrieving the relevant source recordings and for recreating the segmentation procedure. This enables reconstruction of the derived analysis dataset from the public source recordings.

## References


Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

Abdul, Z. K. and Al-Talabani, A. K. (2022). Mel frequency cepstral coefficient and its applications: A review. *IEEE Access*, 10:122136–122158.

Abidin, S., Togneri, R., and Sohel, F. (2018). Spectrotemporal analysis using local binary pattern variants for acoustic scene classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(11):2112–2121.

Akbal, E., Dogan, S., and Tuncer, T. (2022). An automated multispecies bioacoustics sound classification method based on a nonlinear pattern: Twine-pat. *Ecological Informatics*, 68:101529.

Alcocer, I., Lima, H., Sugai, L. S. M., and Llusia, D. (2022). Acoustic indices as proxies for biodiversity: a meta-analysis. *Biological Reviews*, 97(6):2209–2236.

Allen-Ankins, S., McKnight, D. T., Nordberg, E. J., Hoefer, S., Roe, P., Watson, D. M., McDonald, P. G., Fuller, R. A., and Schwarzkopf, L. (2023). Effectiveness of acoustic indices as indicators of vertebrate biodiversity. *Ecological Indicators*, 147:109937.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., Ginhoux, F., and Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38–44.

Bisot, V., Essid, S., and Richard, G. (2015). Hog and subband power distribution image features for acoustic scene classification. In *2015 23rd European signal processing conference (EUSIPCO)*, pages 719–723. IEEE.

Bonet-Solà, D. and Alsina-Pagès, R. M. (2021). A comparative survey of feature extraction and machine learning methods in diverse acoustic environments. *Sensors*, 21(4).

Bradfer-Lawrence, T., Gardner, N., Bunnefeld, L., Bunnefeld, N., Willis, S. G., and Dent, D. H. (2019). Guidelines for the use of acoustic indices in environmental research. *Methods in Ecology and Evolution*, 10(10):1796–1807.

Brown, J. C. and Smaragdis, P. (2009). Hidden markov and gaussian mixture models for automatic call classification. *The Journal of the Acoustical Society of America*, 125(6):EL221–EL224.

Butchart, S. H., Walpole, M., Collen, B., Van Strien, A., Scharlemann, J. P., Almond, R. E., Baillie, J. E., Bomhard, B., Brown, C., Bruno, J., et al. (2010). Global biodiversity: indicators of recent declines. *Science*, 328(5982):1164–1168.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27.

Campello, R. J. G. B., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Pei, J., Tseng, V. S., Cao, L., Motoda, H., and Xu, G., editors, *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Chase, J. M., Blowes, S. A., Knight, T. M., Gerstner, K., and May, F. (2020). Ecosystem decay exacerbates biodiversity loss with habitat loss. *Nature*, 584(7820):238–243.

Cifuentes, E., Vélez Gómez, J., and Butler, S. J. (2021). Relationship between acoustic indices, length of recordings and processing time: a methodological test. *Biota colombiana*, 22(1):26–35.

Clink, D. J., Crofoot, M. C., and Marshall, A. J. (2019). Application of a semi-automated vocal fingerprinting approach to monitor bornean gibbon females in an experimentally fragmented landscape in sabah, malaysia. *Bioacoustics*, 28(3):193–209.

Cominelli, S., Bellin, N., Brown, C., Rossi, V., and Lawson, J. (2024). Acoustic features as a tool to visualize and explore marine soundscapes: Applications illustrated using marine mammal passive acoustic monitoring datasets. *Ecology and Evolution*, 14.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

Diaz, S. D. U., Gan, J. L., and Tapang, G. A. (2023). Acoustic indices as proxies for bird species richness in an urban green space in metro manila. *PLoS One*, 18(7):e0289001.

Dufourq, E., Batist, C., Foquet, R., and Durbach, I. (2022). Passive acoustic monitoring of animal populations with transfer learning. *Ecological Informatics*, 70:101688.

Dunn†, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4(1):95–104.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Farina, A. et al. (2018). Perspectives in ecoacoustics: A contribution to defining a discipline. journal of ecoacoustics. *Journal of Ecoacoustics*.

Fuller, S., Axel, A. C., Tucker, D., and Gage, S. H. (2015). Connecting soundscape to landscape: Which acoustic index best describes landscape configuration? *Ecological indicators*, 58:207–215.

Funosas, D., Barbaro, L., Schillé, L., Elger, A., Castagneyrol,

B., and Cauchoix, M. (2024). Assessing the potential of birdnet to infer european bird communities from large-scale ecoacoustic data. *Ecological Indicators*, 164:112146.

Genuer, R., Poggi, J.-M., Tuleau-Malot, C., and Villa-Vialaneix, N. (2017). Random forests for big data. *Big Data Research*, 9:28–46.

Gibb, R., Browning, E., Glover-Kapfer, P., and Jones, K. E. (2019). Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution*, 10(2):169–185.

Grinfeder, E., Haupert, S., Ducrettet, M., Barlet, J., Reynet, M.-P., Sèbe, F., and Sueur, J. (2022). Soundscape dynamics of a cold protected forest: dominance of aircraft noise. *Landscape Ecology*, pages 1–16.

Guerrero, M. J., Bedoya, C. L., López, J. D., Daza, J. M., and Isaza, C. (2023). Acoustic animal identification using unsupervised learning. *Methods in Ecology and Evolution*, 14(6):1500–1514.

Haddad, N. M., Brudvig, L. A., Clobert, J., Davies, K. F., Gonzalez, A., Holt, R. D., Lovejoy, T. E., Sexton, J. O., Austin, M. P., Collins, C. D., et al. (2015). Habitat fragmentation and its lasting impact on earth's ecosystems. *Science advances*, 1(2):e1500052.

Happel, R. E. and Happel, R. J. (2020). Soundscape ecology.

Hoefer, S., McKnight, D. T., Allen-Ankins, S., Nordberg, E. J., and and, L. S. (2023). Passive acoustic monitoring in terrestrial vertebrates: a review. *Bioacoustics*, 32(5):506–531.

Kahl, S., Wood, C. M., Eibl, M., and Klinck, H. (2021). Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236.

Keil, P., Storch, D., and Jetz, W. (2015). On the decline of biodiversity due to area loss. *Nature communications*, 6(1):8837.

Kok, A. C., Berkhout, B. W., Carlson, N. V., Evans, N. P., Khan, N., Potvin, D. A., Radford, A. N., Sebire, M., Shafiei Sabet, S., Shannon, G., et al. (2023). How chronic anthropogenic noise can affect wildlife communities. *Frontiers in Ecology and Evolution*, 11:1130075.

Lakdari, M. W., Ahmad, A. H., Sethi, S., Bohn, G. A., and Clink, D. J. (2024). Mel-frequency cepstral coefficients outperform embeddings from pre-trained convolutional neural networks under noisy conditions for discrimination tasks of individual gibbons. *Ecological Informatics*, 80:102457.

Li, J., Dai, W., Metze, F., Qu, S., and Das, S. (2017). A comparison of deep learning methods for environmental sound.

Lin, T.-H., Fang, S.-H., and Tsao, Y. (2017). Improving biodiversity assessment via unsupervised separation of biological sounds from long-duration recordings. *Scientific reports*, 7(1):4547.

Lindenmayer, D. B., Gibbons, P., Bourke, M., Burgman, M., Dickman, C. R., Ferrier, S., Fitzsimons, J., Freudenberger, D., Garnett, S. T., Groves, C., et al. (2012). Improving biodiversity monitoring. *Austral ecology*, 37(3):285–294.

Linke, S. and Deretic, J.-A. (2020). Ecoacoustics can detect ecosystem responses to environmental water allocations. *Freshwater Biology*, 65(1):133–141.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction.

Mcloughlin, M. P., Stewart, R., and McElligott, A. G. (2019). Automated bioacoustics: methods in ecology and conservation and their potential for animal welfare monitoring. *Journal of the Royal Society Interface*, 16(155):20190225.

Milošević, D., Medeiros, A. S., Stojković Piperac, M., Cvijanović, D., Soininen, J., Milosavljević, A., and Predić, B. (2022). The application of uniform manifold approximation and projection (umap) for unconstrained ordination and classification of biological indicators in aquatic ecology. *Science of The Total Environment*, 815:152365.

Morfi, V., Bas, Y., Pamuła, H., Glotin, H., and Stowell, D. (2019). Nips4bplus: a richly annotated birdsong audio dataset. *PeerJ Computer Science*, 5:e223.

Mullet, T. C., Gage, S. H., Morton, J. M., and Huettmann, F. (2016). Temporal and spatial variation of a winter soundscape in south-central alaska. *Landscape Ecology*, 31(5):1117–1137.

Murtagh, F. and Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 2(1):86–97.

Napier, T., Ahn, E., Allen-Ankins, S., and Lee, I. (2023). An optimised grid search based framework for robust large-scale natural soundscape classification. In *Australasian Joint Conference on Artificial Intelligence*, pages 468–479. Springer.

Napier, T., Ahn, E., Allen-Ankins, S., Schwarzkopf, L., and Lee, I. (2024). Advancements in preprocessing, detection and classification techniques for ecoacoustic data: A comprehensive review for large-scale passive acoustic monitoring. *Expert Systems with Applications*, page 124220.

Napier, T., Ahn, E., Allen-Ankins, S., Schwarzkopf, L., and Lee, I. (2025). Leaves: An open-source web-based tool for the scalable annotation and visualisation of large-scale ecoacoustic datasets using cluster analysis. *Ecological Informatics*, 87:103026.

Nieto-Mora, D., Rodríguez-Buritica, S., Rodríguez-Marín, P., Martínez-Vargaz, J., and Isaza-Narváez, C. (2023). Systematic review of machine learning methods applied to ecoacoustics and soundscape monitoring. *Heliyon*.

Nieto-Mora, D. A., Ferreira de Oliveira, M. C., Sanchez-Giraldo, C., Duque-Muñoz, L., Isaza-Narváez, C., and Martínez-Vargas, J. D. (2024). Soundscape characterization using autoencoders and unsupervised learning. *Sensors*, 24(8).

Noda, J. J., Travieso, C. M., and Sánchez-Rodríguez, D. (2016). Methodology for automatic bioacoustic classification of anurans based on feature fusion. *Expert Systems with Applications*, 50:100–106.

Okabe, A., Boots, B., Sugihara, K., and Chiu, S. N. (2009). Spatial tessellations: concepts and applications of voronoi diagrams.

Parrilla, A. G. A. and Stowell, D. (2022). Polyphonic sound

event detection for highly dense birdsong scenes.

Phillips, Y. F., Towsey, M., and Roe, P. (2018). Revealing the ecological content of long-duration audio-recordings of the environment through clustering and visualisation. *PLOS ONE*, 13(3):1–27.

Pijanowski, B. C., Farina, A., Gage, S. H., Dumyahn, S. L., and Krause, B. L. (2011). What is soundscape ecology? an introduction and overview of an emerging new science. *Landscape ecology*, 26(9):1213–1232.

Poutaraud, J., Sueur, J., Thébaud, C., and Haupert, S. (2024). Meta-embedded clustering (mec): A new method for improving clustering quality in unlabeled bird sound datasets. *Ecological Informatics*, 82:102687.

Quinn, C. A., Burns, P., Gill, G., Baligar, S., Snyder, R. L., Salas, L., Goetz, S. J., and Clark, M. L. (2022). Soundscape classification with convolutional neural networks reveals temporal and geographic patterns in ecoacoustic data. *Ecological Indicators*, 138:108831.

Rauch, L., Schwinger, R., Wirth, M., Heinrich, R., Huseljic, D., Herde, M., Lange, J., Kahl, S., Sick, B., Tomforde, S., and Scholz, C. (2025). Birdset: A large-scale dataset for audio classification in avian bioacoustics.

Rendon, N., Giraldo, J. H., Bouwmans, T., Rodríguez-Buritica, S., Ramirez, E., and Isaza, C. (2023). Uncertainty clustering internal validity assessment using fréchet distance for unsupervised learning. *Engineering Applications of Artificial Intelligence*, 124:106635.

Roe, P., Eichinski, P., Fuller, R. A., McDonald, P. G., Schwarzkopf, L., Towsey, M., Truskinger, A., Tucker, D., and Watson, D. M. (2021). The australian acoustic observatory. *Methods in Ecology and Evolution*, 12(10):1802–1808.

Ross, S. R.-J., O'Connell, D. P., Deichmann, J. L., Desjonquères, C., Gasc, A., Phillips, J. N., Sethi, S. S., Wood, C. M., and Burivalova, Z. (2023). Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions. *Functional Ecology*, 37(4):959–975.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Scarpelli, M. D. A., Liquet, B., Tucker, D., Fuller, S., and Roe, P. (2021). Multi-index ecoacoustics analysis for terrestrial soundscapes: A new semi-automated approach using time-series motif discovery and random forest classification. *Frontiers in Ecology and Evolution*, 9.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., and Xu, X. (2017). Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM transactions on database systems (tods)*, 42(3):1–21.

Serizel, R., Bisot, V., Essid, S., and Richard, G. (2018). Acoustic features for environmental sound analysis. *Computational analysis of sound scenes and events*, pages 71–101.

Soares, B. S., Luz, J. S., de Macêdo, V. F., e Silva, R. R. V., de Araújo, F. H. D., and Magalhães, D. M. V. (2022). Mfcc-based descriptor for bee queen presence detection. *Expert Systems with Applications*, 201:117104.

Stowell, D. (2022). Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10:e13152.

Stowell, D. and Plumbley, M. D. (2014). Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. *PeerJ*, 2:e488.

Sueur, J. and Farina, A. (2015). Ecoacoustics: the ecological investigation and interpretation of environmental sound. *Biosemiotics*, 8(3):493–502.

Sueur, J., Farina, A., Gasc, A., Pieretti, N., and Pavoine, S. (2014). Acoustic indices for biodiversity assessment and landscape investigation. *Acta Acustica united with Acustica*, 100(4):772–781.

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

Teixeira, D., Roe, P., van Rensburg, B. J., Linke, S., McDonald, P. G., Tucker, D., and Fuller, S. (2024). Effective ecological monitoring using passive acoustic sensors: Recommendations for conservation practitioners. *Conservation Science and Practice*, 6(6):e13132.

Thomas, M., Jensen, F. H., Averly, B., Demartsev, V., Manser, M. B., Sainburg, T., Roch, M. A., and Strandburg-Peshkin, A. (2022). A practical guide for generating unsupervised, spectrogram-based latent space representations of animal vocalizations. *Journal of Animal Ecology*, 91(8):1567–1581.

Thomas, M., Martin, B., Kowarski, K., Gaudet, B., and Matwin, S. (2020). Marine mammal species classification using convolutional neural networks and a novel acoustic representation. In *Machine learning and knowledge discovery in databases: European conference, ECML PKDD 2019, würzburg, Germany, september 16–20, 2019, proceedings, Part III*, pages 290–305. Springer.

Towsey, M., Znidersic, E., Broken-Brow, J., Indraswari, K., Watson, D. M., Phillips, Y., Truskinger, A., and Roe, P. (2018). Long-duration, false-colour spectrograms for detecting species in large audio data-sets. *Journal of Ecoacoustics*, 2:1–13.

Towsey, M. W. (2017). The calculation of acoustic indices derived from long-duration recordings of the natural environment.

Truskinger, A., Cottman-Fields, M., Eichinski, P., Towsey, M., and Roe, P. (2014). Practical analysis of big acoustic sensor data for environmental monitoring. In *2014 IEEE fourth international conference on big data and cloud computing*, pages 91–98. IEEE.

Tsoumakas, G. and Katakis, I. (2008). Multi-label classification: An overview. *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications*, pages 64–74.

Turlington, K., Suárez-Castro, A. F., Teixeira, D., Linke, S., and Sheldon, F. (2024). Exploring the relationship between the soundscape and the environment: A systematic review. *Ecological Indicators*, 166:112388.

Ugarte, J. P. and Arias-Arias, J. (2024). Unveiling relevant acoustic features for bird species automatic classification. *Expert Systems with Applications*, 257:125046.

van der Maaten, L. (2014). Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 15(93):3221–3245.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Van Parijs, S. M., Baumgartner, M., Cholewiak, D., Davis, G., Gedamke, J., Gerlach, D., Haver, S., Hatch, J., Hatch, L., Hotchkin, C., et al. (2015). Nepan: A us northeast passive acoustic sensing network for monitoring, reducing threats and the conservation of marine animals. *Marine Technology Society Journal*, 49(2):70–86.

Vella, K., Capel, T., Gonzalez, A., Truskinger, A., Fuller, S., and Roe, P. (2022). Key issues for realizing open ecoacoustic monitoring in australia. *Frontiers in Ecology and Evolution*, 9:809576.

Wall, C. C., Haver, S. M., Hatch, L. T., Miksis-Olds, J., Bochenek, R., Dziak, R. P., and Gedamke, J. (2021). The next wave of passive acoustic data management: How centralized access can enhance science. *Frontiers in Marine Science*, 8:703682.

Wilkinghoff, K., Fujimura, T., Imoto, K., Roux, J. L., Tan, Z.-H., and Toda, T. (2025). Handling domain shifts for anomalous sound detection: A review of dcase-related work.

Williams, B., Balvanera, S. M., Sethi, S. S., Lamont, T. A., Jompa, J., Prasetya, M., Richardson, L., Chapuis, L., Weschke, E., Hoey, A., et al. (2025). Unlocking the soundscape of coral reefs with artificial intelligence: pretrained networks and unsupervised learning win out. *PLOS Computational Biology*, 21(4):e1013029.

Williams, B., Lamont, T. A., Chapuis, L., Harding, H. R., May, E. B., Prasetya, M. E., Seraphim, M. J., Jompa, J., Smith, D. J., Janetski, N., et al. (2022). Enhancing automated analysis of marine soundscapes using ecoacoustic indices and machine learning. *Ecological Indicators*, 140:108986.

Wimmer, J., Towsey, M., Planitz, B., Williamson, I., and Roe, P. (2013a). Analysing environmental acoustic data through collaboration and automation. *Future Generation Computer Systems*, 29(2):560–568.

Wimmer, J., Towsey, M., Roe, P., and Williamson, I. (2013b). Sampling environmental acoustic recordings to determine bird species richness. *Ecological Applications*, 23(6):1419–1428.

Xie, J., Hu, K., Zhu, M., and Guo, Y. (2020). Bioacoustic signal classification in continuous recordings: Syllable-segmentation vs sliding-window. *Expert Systems with Applications*, 152:113390.

Xie, J., Towsey, M., Zhang, J., and Roe, P. (2018). Frog call classification: a survey. *Artificial Intelligence Review*, 49:375–391.

Yao, F., Coquery, J., and Lê Cao, K.-A. (2012). Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. *BMC bioinformatics*, 13:1–15.

Zhang, M.-L. and Zhou, Z.-H. (2013). A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837.

Zhong, M., LeBien, J., Campos-Cerqueira, M., Dodhia, R., Ferres, J. L., Velev, J. P., and Aide, T. M. (2020). Multi-species bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling. *Applied Acoustics*, 166:107375.

Zhou, Y. and Sharpee, T. O. (2022). Using global t-sne to preserve intercluster data structure. *Neural computation*, 34(8):1637–1651.