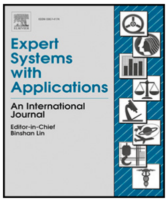
Expert Systems With Applications 252 (2024) 124220

Contents lists available at ScienceDirect

Expert Systems With Applications

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



Review

Advancements in preprocessing, detection and classification techniques for ecoacoustic data: A comprehensive review for large-scale Passive Acoustic Monitoring

Thomas Napier [*], Euijoon Ahn, Slade Allen-Ankins, Lin Schwarzkopf, Ickjai Lee

College of Science & Engineering, James Cook University, Townsville QLD 4811, Australia

ARTICLE INFO

Keywords:
Ecoacoustics
Passive acoustic monitoring
Deep learning
Machine learning
Acoustic signal processing
Bioacoustics
Species identification

ABSTRACT

Computational ecoacoustics has seen significant growth in recent decades, facilitated by the reduced costs of digital sound recording devices and data storage. This progress has enabled the continuous monitoring of vocal fauna through Passive Acoustic Monitoring (PAM), a technique used to record and analyse environmental sounds to study animal behaviours and their habitats. While the collection of ecoacoustic data has become more accessible, the effective analysis of this information to understand animal behaviours and monitor populations remains a major challenge. This survey paper presents the state-of-the-art ecoacoustics data analysis approaches, with a focus on their applicability to large-scale PAM. We emphasise the importance of large-scale PAM, as it enables extensive geographical coverage and continuous monitoring, crucial for comprehensive biodiversity assessment and understanding ecological dynamics over wide areas and diverse habitats. This large-scale approach is particularly vital in the face of rapid environmental changes, as it provides crucial insights into the effects of these changes on a broad array of species and ecosystems. As such, we outline the most challenging large-scale ecoacoustics data analysis tasks, including pre-processing, visualisation, data labelling, detection, and classification. Each is evaluated according to its strengths, weaknesses and overall suitability to large-scale PAM, and recommendations are made for future research directions.

1. Introduction

1.1. Background

Ecoacoustics is a rapidly emerging field that focuses on the study of environmental sounds to monitor and understand biodiversity (Farina & Gage, 2017). It encompasses the recording, analysis, and interpretation of sounds produced by wildlife, as well as those occurring in their natural habitats. Ecoacoustics plays a crucial role in understanding and monitoring biodiversity. By analysing environmental sounds, researchers can gain insights into the presence, behaviour, and interactions of various species. This approach is particularly important given the increasing challenges of biodiversity decline. Unfortunately, these challenges are global; species loss due to habitat destruction is a widespread issue, placing significant strain on ecosystems (Cardinale et al., 2012; Wilcove, Rothstein, Dubow, Phillips, & Losos, 1998). To support both humans and animals, ecosystem function must be maintained by reducing animal extinction rates. The primary aim of this study is to comprehensively review and assess various methods for analysing ecoacoustic data, with a particular focus on large-scale Passive Acoustic Monitoring (PAM). By evaluating these methodologies, we aim to identify their strengths, limitations, and areas for improvement, thus contributing to more effective biodiversity monitoring efforts. Given the comprehensive nature of our analysis and the depth of our evaluation, we believe that the findings and methodologies discussed herein could serve as a model for future ecoacoustic research efforts. This is particularly pertinent in the context of enhancing biodiversity monitoring.

Biodiversity monitoring is necessary to track species, understand declines, and evaluate management interventions and has thus been the focus of much recent research (Agranat, 2009; Enari, Enari, Okuda, Maruyama, & Okuda, 2019; Gibb, Browning, Glover-Kapfer, & Jones, 2018; Riede, 1993). Historically, ecological surveys of fauna have been manual. Typically, experienced zoologists are required to identify fauna in natural ecosystems to ascertain whether a species is present at a particular location at any point in time (Gregory, Gibbons, & Donald, 2004). Such surveys are limited in scale, expensive and

[*] Corresponding author.
E-mail addresses: thomas.napier@jcu.edu.au (T. Napier), euijoon.ahn@jcu.edu.au (E. Ahn), slade.allenankins@jcu.edu.au (S. Allen-Ankins), lin.schwarzkopf@jcu.edu.au (L. Schwarzkopf), ickjai.lee@jcu.edu.au (I. Lee).

https://doi.org/10.1016/j.eswa.2024.124220
Received 20 June 2023; Received in revised form 28 April 2024; Accepted 11 May 2024
Available online 16 May 2024

labour-intensive (Riede, 1993). Monitoring fauna using remote sensor technologies has become increasingly used to overcome these issues.

PAM is a technique that involves the use of Autonomous Recording Units (ARUs) equipped with microphones to capture environmental sounds. The core components of a PAM system include highly sensitive microphones for sound capture, data storage units for recording, and often include pre-amplifiers and filters to enhance sound quality and reduce noise (Sugai, Silva, Ribeiro, & Llusia, 2018). These systems are deployed in various habitats and can operate continuously over extended periods, capturing a wide range of biophonic (animal sounds), geophonic (non-biological natural sounds like wind or water), and anthrophonic (human-made noises) acoustic signals.

Until recently, technological costs and constraints were the primary barriers to deploying automated PAM systems for ecological studies. Due to this, research using PAM has primarily focused on only a handful of taxonomic groups (e.g., bats, cetaceans, birds), often resulting in relatively small-scale feasibility assessment studies (Gibb et al., 2018). In addition, researchers using PAM typically study an individual species or sub-sample data from the available recordings. Consequently, the focus is concentrated and limited to a specific time or location. Thus, while historically, analysis techniques were optimised for simpler tasks like human speech recognition in recorded conversation, advancements in technology have expanded their applicability to more complex ecoacoustic monitoring tasks.

In part, this has been facilitated by increasingly affordable digital sound recording devices and data storage solutions, which has enabled ecoacoustic recorders to be deployed at scale in terrestrial fauna surveys, both spatially and temporally (Roe et al., 2021). As such, there has been a recent trend towards very large-scale continental-wide projects such as the Australian Acoustic Observatory (A2O) (Roe et al., 2021), which has recorders deployed at the sites of the Terrestrial Ecosystem Research Network (TERN) and a range of other landholders, including National Parks, Australian Wildlife Conservancy, Bush Heritage Australia and other private individuals, the U.S. Northeast Passive Acoustic Sensing Network (NEPAN) (Van Parijs et al., 2015), the Okinawa Environmental Observation Network (OKEON) (Ross et al., 2018) as well as the National Oceanic and Atmospheric Administration (NOAA) / National Park Service (NPS) Ocean Noise Reference Station (NRS) (Haver et al., 2018; Ross et al., 2023). These systems offer an unprecedented opportunity for broad-scale monitoring of ecoacoustics patterns. However, alongside with capturing these broad patterns, it is equally important to identify the specific species present within a large soundscape. Biologists and ecologists conducting biodiversity surveys need detailed insights into the species present, as understanding the species composition is essential for conservation and ecological studies.

Large-scale recording using PAM can have distinct advantages over other forms of fauna monitoring, requiring fewer site visits and potentially providing continuous recordings as opposed to episodic samples associated with visits. PAM systems can also capture information on vocalising fauna for use with other monitoring techniques, including remote methods such as camera trapping or manual surveys (Enari et al., 2019). It is important to note, however, that PAM systems cannot detect the presence of non-vocal animals or species that produce sounds outside of the captured frequency range of deployed microphones, such as the ultrasonic echolocation sounds of bats (Roe et al., 2021), although specialised microphones can be deployed. Thus, while powerful, PAM must be supplemented with other monitoring methods to provide a complete picture of species presence if full species inventories are required. Realistically, no single approach can provide complete species inventories, and PAM has other advantages, as mentioned earlier.

An aspect of PAM systems that can be both advantageous and challenging is the generation of large volumes of sound information. While this is particularly true for continuous recording strategies such as those used by the A2O (Roe et al., 2021), real-time detection and episodic recording methods can mitigate this issue to some extent. However, it is important to recognise that the choice between continuous and episodic recording is often dictated by specific research needs and practical considerations. Episodic recording, where sounds are recorded at predetermined intervals, is frequently selected due to practical constraints such as limited battery life, data storage considerations, and the specific objectives of the research. For instance, studies focusing on particular temporal patterns or events may not require continuous data streams, making episodic recording a more suitable and resource-efficient approach. Despite this, the extensive temporal sampling available using PAM allows for the detection of interesting patterns that are not detectable from occasional visits typical of manual surveying. For example, PAM can capture diurnal and seasonal variations in animal vocalisations or detect rare vocal events that may be missed during manual surveys. In cases of large-scale projects which employ continuous recording strategies, the immense volumes of data generated are far in excess of what human experts can ever listen to and manually label. Consequently, this has raised a significant problem in analysing this data. Presently, most works conducted use a sub-sample of the full dataset containing the sounds of specific target species, and not all of the available information is utilised because it is too time-consuming to analyse.

However, rapid advancements in informatics, such as big data, Machine Learning (ML) and signal processing, have enabled large amounts of raw audio data to be effectively processed and transformed into useful data. Leveraging advancements in computational power, several emerging technology paradigms have been integrated into PAM systems. These include Deep Learning (DL), a specialised form of ML that mimics neural networks to analyse various forms of data; object recognition, which identifies distinct objects within digital images; and image segmentation, the technique of dividing a digital image into distinct segments to facilitate more precise image analysis. These techniques can provide species identification and, by their nature, will overcome some issues with previous approaches to species identification by sound. Though there have been some applications of ML and DL techniques in ecoacoustics (Fazekas, Schindler, Lidy, & Rauber, 2018; Kampichler, Wieland, Calmé, Weissenberger, & Arriaga-Weiss, 2010), their use has been more widespread in other domains for a variety of signal identification and recognition tasks, such as human speech recognition in large environments (Abdel-Hamid et al., 2014; Swamy & K.V, 2013) and orchestral music content analysis (Muller, Ellis, Klapuri, & Richard, 2011). While these approaches have seen increased adoption in ecoacoustics, even dating back to early applications in the 2000s, their impact has been somewhat fragmented. Specifically, the field still faces challenges when it comes to large-scale ecoacoustics applications due to a lack of standardisation in methodologies and datasets, which limits the broader applicability and comparability of these techniques (Stowell, 2022). Therefore, while acknowledging the growing role of ML and DL in ecoacoustics, several gaps and challenges still exist, particularly in the context of large-scale PAM applications.

1.2. Motivation

The primary motivation behind this study is that finding successful approaches to ecoacoustics, specifically for the analysis of large-scale PAM, constitutes a relatively new, cutting-edge, and promising branch of research. Developments and applications of DL in other fields, such as image recognition of weeds in agriculture and detection of fish species in marine biology, indicate that it has significant potential, yet it is underexplored in ecoacoustics. There are several key challenges halting progress. First, current approaches, typically manual identification, are too time-intensive and require expert knowledge of target animal vocalisations. Moreover, inadequate levels of labelled datasets are available for training supervised learning models and further labelled data is too expensive and challenging to acquire easily.

In the past few years, DL-based approaches to this problem have been State of The Art (SoTA). The most significant increase in accuracy for detection tasks has come from the advancement of image classification techniques, especially the use of Convolutional Neural Networks (CNN) in vision-based problems. However, two significant issues must be addressed for full applicability to ecoacoustics. First, many existing ML and DL approaches are evaluated on datasets with low variation amongst different taxonomic groups such as birds (Lasseck, 2019; Stowell, Wood, Pamuła, Stylianou, & Glotin, 2018), and frogs (Colonna et al., 2016; LeBien et al., 2020) which is not representative of the real world. Existing DL models are trained and tested on these limited, smaller, and often non-ecological datasets without consideration of the overarching problem. Thus, these models can generalise some aspects particularly well, e.g., a particular taxonomic group (Bardeli et al., 2010; Colonna et al., 2016; Lasseck, 2019; LeBien et al., 2020; Salamon et al., 2016; Stowell et al., 2018) or individual species (Frommolt & Tauchert, 2014; Willacy, Mahony, & Newell, 2015), but they cannot generalise beyond the circumscription of the dataset upon which they are trained (Kamilaris & Prenafeta-Boldú, 2018). DL models, especially CNNs, are known for their ability to excel in pattern recognition tasks such as image and sound classification. However, these models are trained to recognise the patterns that are abundant in the dataset on which they are trained. Therefore, if a model is trained on a narrow or non-representative dataset, its ability to generalise to new, unseen data may be poor. This is of particular concern in fields like ecoacoustics where the real-world data is incredibly diverse.

Secondly, many ML and DL approaches have historically been trained on manually pre-cleaned datasets devoid of environmental noise (Bardeli et al., 2010; Gasc, Sueur, Pavoine, Pellens, & Grandcolas, 2013; Phillips, Towsey, & Roe, 2018). Environmental sounds can have varying levels of impact on the results of ecoacoustics studies but is particularly influential in ones which use sensitive feature representations (Sánchez-Giraldo et al., 2020). There are, however, emerging studies that incorporate environmental noise in their training datasets (Grinfeder et al., 2022), but the extent to which it effects downstream tasks remains an understudied area. Traditional denoising techniques have been used like low-band and high-pass filtering (Brown, Garg, & Montgomery, 2018b; Neal, Briggs, Raich, & Fern, 2011; Pijanowski et al., 2011), however, they have limitations when it comes to recordings with overlapping calls or multiple species (Chen, Chen, Lin, Chen, & Lin, 2012; Huang et al., 2014). These nuances underscore the need for more comprehensive training datasets containing sufficient environmental noises for better performance on real-world datasets (Babaee, Anuar, Abdul Wahab, Shamshirband, & Chronopoulos, 2017), such as those generated by large sensor networks. Evidently, advancements in ecoacoustic analysis often prioritise increasing accuracy for specific tasks, such as those focused on particular taxonomic groups or geographic areas. While this approach may be sufficient for research questions with a narrow scope, it may not fully address the challenges and complexities inherent in large-scale, multi-taxonomic, and multi-regional datasets. Therefore, while current methods may be adequate for many specialised research questions, they may fall short in the context of comprehensive biodiversity inventories or large-scale ecological monitoring. As DL approaches rely on large quantities of labelled environmental recordings for supervised learning tasks, the problem of data scarcity remains a pivotal issue. Thus, progress towards a suitable solution remains inhibited by the lack of readily available, annotated, large-scale ecoacoustics datasets that adequately cover real-world natural soundscapes and the lack of clarity and consideration around the complexities and causes of variation present in large-scale systems (Gibb et al., 2018).

1.3. Unique contributions in comparison to other surveys

This paper contributes uniquely to the field by identifying all critical components of an end-to-end ecoacoustic analysis workflow for large-scale PAM systems. Further, a comprehensive survey of the SoTA technologies is conducted that falls within the proposed model. Focus is placed on the types of datasets, ML, DL, visualisation, and applications thereof through the unique lens of applicability to large-scale ecoacoustics data analysis. While large-scale ecoacoustics data analysis offers a broad view, it is essential to align this with the research questions being addressed. For example, large-scale datasets could be invaluable for studying the effects of urbanisation on wildlife, monitoring the migratory patterns of multiple vocal species across a continent, or investigating the spread of vocal invasive species across multiple habitats.

To date, and to the best of our knowledge, no other paper evaluates publicly available ecoacoustics datasets for their applicability to large-scale PAM. It will distinguish itself from previous surveys by considering all components individually and as a whole. Further emphasis is placed on an overall ecosystem view rather than investigating a specific species or taxonomic group. This is because, while focusing on specific species or taxonomic groups has its merits, an ecosystem-level approach offers a more holistic understanding of the acoustic environment. This broader perspective allows for the capture of complex interactions among multiple species and their responses to various environmental factors. For example, an ecosystem-level focus can reveal how different species' vocalisations overlap or interact, providing insights into community dynamics. Furthermore, this approach is particularly beneficial for identifying broader patterns and trends that may be missed when focusing solely on individual species. Such patterns could include shifts in community composition or changes in vocalisation timing across multiple species, which could be indicative of larger environmental changes. Focus is also placed on the effectiveness of unsupervised, self-supervised segmentation and labelling approaches, as the lack of labelled data is the most challenging issue to date, halting further progress for applications of ML and DL to large-scale PAM. As such, the primary objectives of this comprehensive review are to:

- Evaluate current methodologies in the pre-processing, detection, and classification of ecoacoustic data, particularly in the context of large-scale PAM.
- Identify efficient data annotation and segmentation techniques suitable for large-scale acoustic data.
- Assess the accuracy of various classification approaches in recognising species within extensive ecoacoustic datasets.
- Explore the challenges and limitations inherent in these methodologies when applied to large datasets and propose potential mitigation strategies.

Our investigation aims to address the gaps in the current literature, as highlighted in Table 1. This table juxtaposes our survey against existing studies, illustrating areas that previous reviews have not fully covered, particularly concerning large-scale PAM. By focusing on these gaps, our review endeavours to advance the field of ecoacoustics by providing insights into the current state of methodologies and suggesting avenues for future research.

1.4. Structure of survey

To better organise the structure of this survey, we present the sequence and interrelation of different steps in a typical PAM data analysis workflow as illustrated in Fig. 1. Based on this, the remainder of this paper is structured based on the major tasks presented as follows: Section 2 introduces key terms and covers the background of ecoacoustics for large-scale applications, including the characteristics and types of available datasets. Section 3 covers the pre-processing and denoising techniques observed for long-duration ecoacoustic applications, while Section 4 covers the task of visualising such data. In Section 5, we discuss and break down the current SoTA data labelling and segmentation techniques, focusing on scalability to large-scale PAM. Section 6 presents a further discussion on the detection and classification techniques and their applications. Finally, Section 7 presents an open set of problems and future challenges in the large-scale ecoacoustics area.

Table 1
Comparison between our survey and existing surveys.

Reference	Components of ecoacoustics analysis framework	Focus on large-scale PAM	Evaluation of ecoacoustics dataset types	Coverage of signal pre-processing and noise removal	Ecoacoustics data labelling and segmentation	Long-duration data visualisation	Ecoacoustics detection and classification
Gibb et al. (2018)	◐	●	○	◐	○	○	◐
Babaee et al. (2017)	◐	○	◐	◐	◐	○	●
Xia, Togneri, Sohel, Zhao, and Huang (2019)	◐	○	◐	○	◐	○	◐
Kvsn, Montgomery, Garg, and Charleston (2020)	●	◐	○	●	◐	◐	●
Xie, Colonna, and Zhang (2020)	◐	◐	◐	●	○	○	○
Bonet-Solà and Alsina-Pagès (2021)	◐	○	◐	◐	◐	○	◐
Stowell (2022)	●	◐	◐	◐	●	◐	◐
Our Survey	●	●	●	●	●	●	●

●= Complete Information, ◐= Partial Information, ○= No Information.

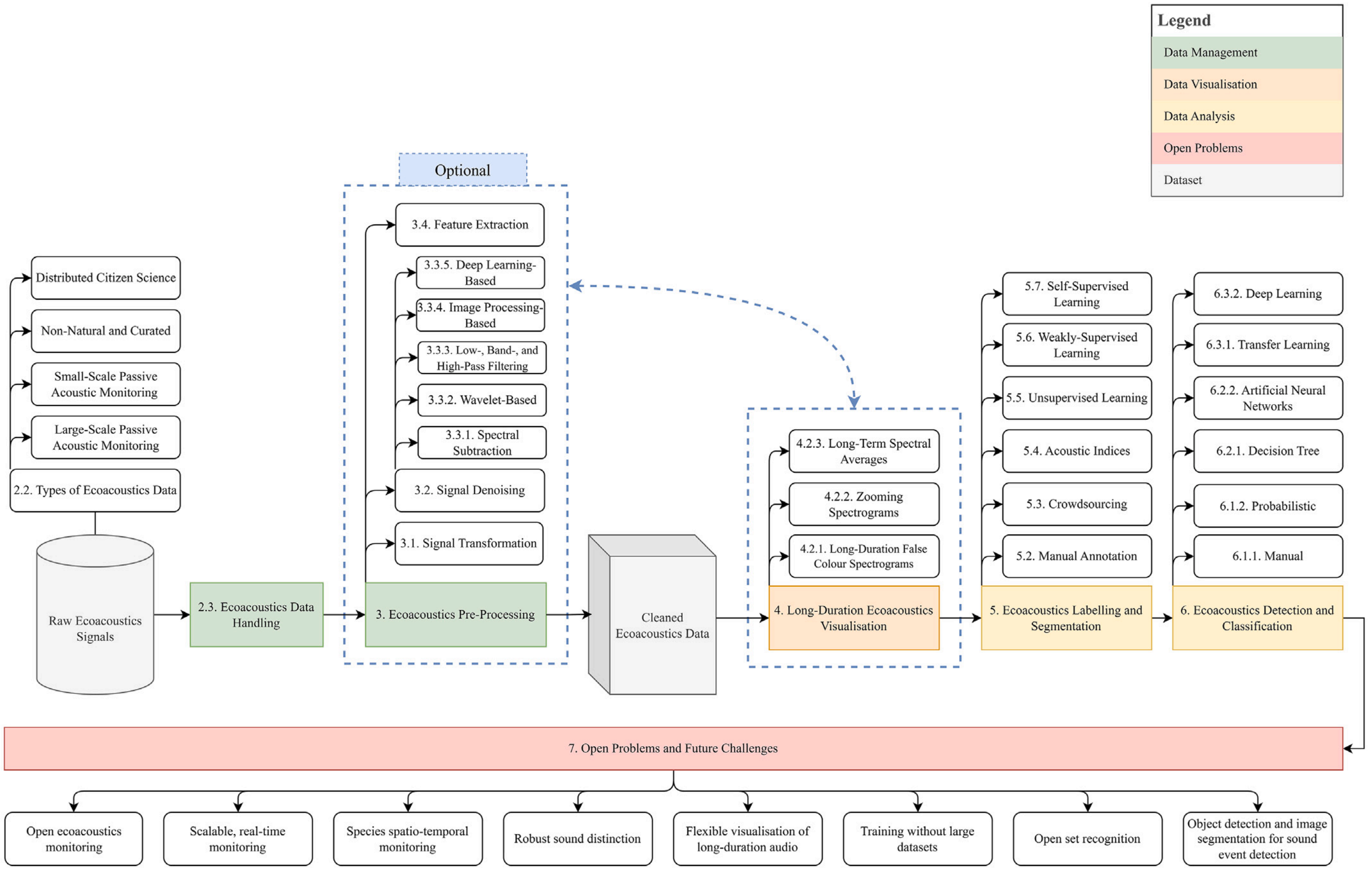

Legend
Data Management
Data Visualisation
Data Analysis
Open Problems
Dataset
Optional
Distributed Citizen Science
Non-Natural and Curated
Small-Scale Passive Acoustic Monitoring
Large-Scale Passive Acoustic Monitoring
2.2. Types of Ecoacoustics Data
Raw Ecoacoustics Signals
2.3. Ecoacoustics Data Handling
3.4. Feature Extraction
3.3.5. Deep Learning-Based
3.3.4. Image Processing-Based
3.3.3. Low-, Band-, and High-Pass Filtering
3.3.2. Wavelet-Based
3.3.1. Spectral Subtraction
3.2. Signal Denoising
3.1. Signal Transformation
3. Ecoacoustics Pre-Processing
Cleaned Ecoacoustics Data
4.2.3. Long-Term Spectral Averages
4.2.2. Zooming Spectrograms
4.2.1. Long-Duration False Colour Spectrograms
4. Long-Duration Ecoacoustics Visualisation
5.7. Self-Supervised Learning
5.6. Weakly-Supervised Learning
5.5. Unsupervised Learning
5.4. Acoustic Indices
5.3. Crowdsourcing
5.2. Manual Annotation
5. Ecoacoustics Labelling and Segmentation
6.3.2. Deep Learning
6.3.1. Transfer Learning
6.2.2. Artificial Neural Networks
6.2.1. Decision Tree
6.1.2. Probabilistic
6.1.1. Manual
6. Ecoacoustics Detection and Classification
7. Open Problems and Future Challenges
Open ecoacoustics monitoring
Scalable, real-time monitoring
Species spatio-temporal monitoring
Robust sound distinction
Flexible visualisation of long-duration audio
Training without large datasets
Open set recognition
Object detection and image segmentation for sound event detection

Fig. 1. A typical workflow model for large-scale ecoacoustics data analysis.

2. Ecoacoustics overview

2.1. Ecoacoustics vs. Bioacoustics

Ecoacoustics involves extracting ecological patterns from PAM systems on an ecosystem level for biodiversity assessments (Sueur & Farina, 2015). Although related, bioacoustics differs from this by examining the fabrication, transmission, and reception of animal sounds from vocal fauna. However, it is important to acknowledge that the use of PAM is not exclusive to ecoacoustics. Recent trends in bioacoustics also demonstrate a growing adoption of PAM techniques for species-specific studies. In contrast, ecoacoustics observations operate on a much broader scope and provide insights at an ecosystem or biome scale. They can offer a unique window into the habits and behaviours of populations and communities, which, once captured, can be used as a valuable means of monitoring vocal fauna (Kvsn et al., 2020). Such forms of monitoring are becoming of critical importance considering recent trends of species decline, yet this area, specifically on a large scale, remains understudied.

To date, much research has been conducted on human speech analysis and recognition (Abdel-Hamid et al., 2014; Swamy & K.V, 2013). However, there are several key differences between human speech and animal vocalisations. Human speech generally occupies a much smaller frequency range than animal vocalisations, and often the energy at frequencies above 5 kHz is ignored by speech recognition techniques (Monson, Hunter, Lotto, & Story, 2014). Human speech data

Table 2
Types of environmental ecoacoustics datasets.

–	Large-scale PAM datasets	Small-scale PAM datasets	Non-natural and curated datasets	Distributed citizen science datasets
Example Datasets	A2O (Roe et al., 2021), NEPAN (Van Parijs et al., 2015), NRS (Haver et al., 2018), OKEON (Ross et al., 2018)	BirdCLEF (Kahl, Denton, et al., 2021), CLO-43SD (Salamon et al., 2016)	AudioSet (Gemmeke et al., 2017), ESC (Piczak, 2015), VGG-Sound (Chen, Xie, Vedaldi, & Zisserman, 2020), UrbanSound (Salamon, Jacoby, & Bello, 2014)	FrogID (Rowley et al., 2019), BirdNET (Kahl, Wood, Eibl, & Klinck, 2021)
Scale	Large	Small	Medium-Large	Medium
Continuous	Yes	Mostly	No	No
Natural	Yes	Yes	No	Yes
Multi-Class	Yes	Sometimes	Yes	No
Multi-Species	Yes	Yes	Yes	Yes
Taxa Diversity Level	High	Medium	Low-Medium	Medium
Multi-Location	Yes	Sometimes	Yes	Yes

is also typically non-overlapping, whereas, in environmental recordings, several species may call simultaneously at variable distances from the receiver, with fluctuating directions and loudness. In addition, environmental factors such as geophony (wind, rain, etc.) or anthrophony (cars, planes, helicopters) can also obfuscate the signal, rendering identification of the original vocalisation difficult (Agranat, 2009). Signals in natural environments may also be reflected and scattered by objects, such as trees and rocks, further deforming the original signal in unknown ways. Thus, traditional techniques applied to human speech are suboptimal for animal vocalisation identification tasks.

Rich soundscapes, such as tropical rainforests or bushland, captured by large-scale systems can be remarkably complex, with many competing species seeking to communicate simultaneously. Vocalisations produced by terrestrial species serve multiple purposes, including promoting survival and facilitating reproduction. These vocalisations can exhibit variations due to a range of factors such as individual characteristics, environmental conditions, and social interactions among species. Such variations can manifest in frequency, timing, harmonics, and rhythm, adding layers of complexity to ecoacoustic data analysis (Happel & Happel, 2020). Further still, more types of non-biophonic sounds are captured, which are typically absent in smaller-scale studies.

2.2. *Types of ecoacoustics data*

Ecoacoustics data can be broadly categorised into four major groups. The first is large-scale PAM datasets such as the A2O (Roe et al., 2021). These datasets focus on an overall ecosystem or soundscape level rather than a specific site or species and generally possess a high ecoregion diversity and number of taxa monitored. The second is small-scale PAM datasets such as BirdCLEF (Kahl, Denton, et al., 2021) and CLO-43SD (Salamon et al., 2016). This dataset group still focuses on an overall environmental or soundscape level; usually, however, these datasets are gathered by individual researchers to develop methods around a single specific geographic area or target species. The high variance in the number of possible sounds captured distinguishes it from small-scale PAM. Other forms of related datasets can be grouped into either a non-natural or curated environmental dataset such as Environmental Sound Classification (ESC) (Piczak, 2015), which are manually extracted from public field recordings, or a distributed citizen science dataset such as FrogID (Rowley et al., 2019) which collects, and labels audio clips uploaded via crowdsourcing.

As illustrated in Table 2, large-scale PAM datasets present a unique challenge compared to the other identified dataset groupings. Specifically, large-scale PAM datasets feature considerably more variance in the possible number of sounds captured. Not only are sounds captured within natural environments and are thus subject to highly variable conditions on a day-to-day basis, but the addition of multiple locations and diverse, often overlapping, species calls distinguish these large-scale PAM datasets from any other. For this reason, additional considerations must be made for data handling and further still for the pre-processing of large-scale ecoacoustics datasets, as they must be crafted explicitly with these complications considered.

2.3. *Ecoacoustics data handling*

The task of handling large-scale PAM datasets presents several unique challenges, the foremost of which being computational bottlenecks. While advanced multi-species detectors do exist and are actively improving (Hao et al., 2022; Colin Quinn et al., 2022), the sheer volume of data, often in the terabytes, poses a significant challenge for timely processing, especially for researchers without access to high-performance computing resources. This issue is particularly acute for practitioners in resource-constrained regions.

Further challenges extend to data archiving and storage. The large volumes of data generated by PAM projects necessitate robust and scalable storage solutions (Roe et al., 2021). Traditional on-premises storage can be costly and difficult to scale, making cloud storage an increasingly popular alternative. However, the recurring costs and data transfer speeds can be significant barriers. Furthermore, the organisation of annotations and detections from ML and DL algorithms presents a logistical hurdle, a sentiment echoed by a recent study which found no standard approaches to annotation and a strong need for interoperability (Vella et al., 2022).

This lack of standardisation hampers the scaling of PAM projects and complicates data sharing and analysis. As such the need for open ecoacoustic monitoring is essential. This term refers to a collaborative, transparent, and accessible approach to ecoacoustic data collection, analysis, and sharing. Open ecoacoustics monitoring emphasises the use of standardised platforms and community-driven initiatives to collect, process, and share ecoacoustic data. This approach not only facilitates data accessibility and interoperability among researchers but also aims to democratise the process of ecoacoustic data analysis, allowing for a broader participation from the scientific community and stakeholders. Ensuring open access to this data is vital but it does come with its own set of challenges. This includes, but is not limited to, data privacy concerns, the need for standardised metadata, and the establishment of data-sharing agreements that respect the interests of all stakeholders.

3. Signal pre-processing and noise removal

In large-scale continental PAM sensor networks, such as the A2O (Roe et al., 2021) and NEPAN (Van Parijs et al., 2015), all omnidirectional recorders continuously collect data over many different sites and ecoregions, leading to high volumes of data. Audio recordings are often obtained under open-environment conditions, where a large variety of sounds near the microphone are captured. Ecoacoustic recordings can have multiple unknown sound sources with sometimes overlapping unknown mixing agents in both time and frequency (Agranat, 2009; Happel & Happel, 2020). Qualities such as sound reverberation can create distortions in the signal, are specific to each ecoregion, and are highly variable. Pre-processing must be employed before analysis to extract meaningful information from raw ecoacoustic data efficiently and accurately. Pre-processing can include noise filtering, downsampling, compression, conversion, and signal transformation. The previously applied pre-processing approaches are often unsuitable for large-scale studies due to computational processing time requirements and the complexity differences between small-scale datasets and large-scale acoustic scenes. This presents the need for more sophisticated data pre-processing techniques.

In addition to the necessity of pre-processing, it is also well recognised that there is a direct link between the results of signal denoising and the quality of output from feature extraction, segmentation, and classification (Xie et al., 2020). Without denoising, extracting meaningful information from the raw audio data can become difficult, or in some cases impossible, particularly for ML-based methods. This is not such an important consideration for DL-based methods as they are more robust to noise (Brown, Garg, & Montgomery, 2018a). Removing noise may benefit efficiency by reducing the total amount of data; however, over-removal may erase crucial information from the signal.

3.1. Signal transformation

While large-scale PAM systems often generate high volumes of data due to continuous recording across multiple sites and ecoregions, it is important to note that the length of these recordings is not solely a function of the system's scale. The length can also be determined by the recording schedule and can be obtained even with a single recorder, depending on its capabilities. However, audio sequences may be split into smaller segments before being used for feature extraction. Such techniques enable long-duration audio files to be more manageable and practical in this format while also allowing downstream methods to work in a more distributed and efficient way without requiring high RAM storage (Truskinger, Cottman-Fields, Johnson, & Roe, 2013). Across a broad range of studies, there are a few standard signals transforms that have been observed among them, including:

- Short-Time Fourier Transform (STFT) changes time-based information into frequency-based information. Depending on the application, several acoustic indices are used. Acoustic indices are quantitative measures designed to summarise a characteristic of the distribution of acoustic energy in an audio recording (Towsey, Wimmer, Williamson, & Roe, 2014). These indices can integrate frequency, time, and amplitude information, reflecting the multifaceted nature of sound recordings (Sueur, Farina, Gasc, Pieretti, & Pavoine, 2014). They range from simple summaries that provide an overview of sound intensity, to more sophisticated calculations that consider the spectral, temporal, and amplitude variations within a soundscape. Essential for ecoacoustic analysis, these indices can assist in the interpretation of ecoacoustic data, such as in some cicada and rain detection cases, where they analyse frequency information derived from STFT (Brown et al., 2018b).
- Downsampling is reducing the sample rates of an audio file to reduce file size. In some cases, such as purely bird-song applications, downsampling can be used to reduce sample rates to closely match the signals of interest, e.g., as birds do not normally vocalise below 11.025 kHz (Pijanowski et al., 2011), audio can be downsampled to match this more closely, while still retaining signals of interest;
- Conversion to Mono is converting stereophonic audio to monophonic-channel audio. It is ideal in most cases because one channel of audio preserves all sound in a single channel, which is all that is needed to detect significant audio signals; thus, the file size can be reduced for more efficient processing. It is important to note that decisions regarding the sample rate and whether to record in mono or stereo can be made at different stages. These choices can be determined during the study design stage to optimise data storage or can be adjusted post-recording based on specific research needs;
- Short-Term Windowing is the splitting of input signals into temporal segments. Typically, researchers take the approach of either splitting the signal into uniform fixed-length segments or splitting the signal into adjustable-length windows depending on the specific application requirements, as long files are typically non-viable for practical application due to high RAM requirements (Truskinger, Cottman-Fields, Eichinski, Towsey, & Roe, 2014);
- Event-Driven Windowing is the purposeful splitting of input signals through defining the beginning and end point of target sound events in time. Typically this also compresses the total amount of samples and enables the rapid retrieval of features of interest for further examination (Gage, Towsey, & Kasten, 2017; Qaisar, Simatic, & Fesquet, 2017).

3.2. Signal denoising

In an ecoacoustics context, signal noise relates to the unwanted modifications of a signal that may have been captured during the original audio recording, storage, or transmission. In many cases, noise represents an error or negative quality in sound recordings, which can often be detrimental when recovering the original audio sources. Importantly, background noises from an anthropogenic or geophony source are often discarded in areas such as bioacoustics (Cai, Ee, Pham, Roe, & Zhang, 2007). However, in some cases, noise can be beneficial, such as in the case of ecoacoustics for large-scale PAM. Many non-biophonic sounds are important to keep intact, as this can further assist ecological studies into these types of noise's effects on animal behaviour (Kok et al., 2023; Potvin et al., 2023). However, some denoising is still required as some types of noise, such as white noise, can render downstream tasks difficult or, in some cases, impossible. This is particularly true when extracting meaningful information audio data, specifically with ML-based methods, which are highly noise-sensitive during feature extraction stages (Nettleton, Orriols-Puig, & Fornells, 2010; Xie et al., 2020).

3.3. Denoising methods

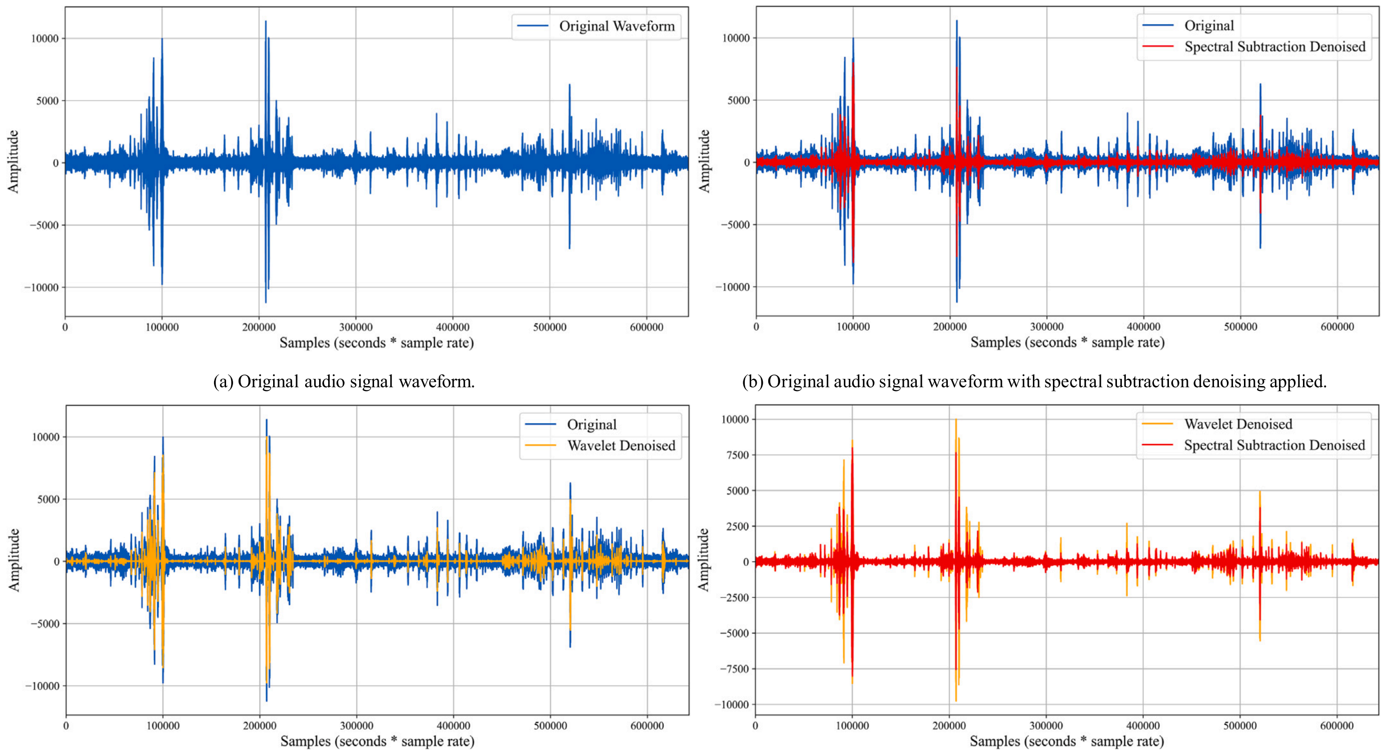
3.3.1. Spectral subtraction

Spectral subtraction is a common approach to audio noise reduction consisting of subtracting the frequency components from noisy audio portions to obtain a cleaned and enhanced recording. An example of its effect can be observed in Fig. 2(b). The primary working principle is based on generating a noise profile for a given recording, which is suitable for short-duration segments where the noise stays relatively consistent within the same recording. An example of where spectral subtraction worked effectively was a study conducted in 2016 for frog call classification (Xie, Towsey, Zhang, & Roe, 2016). Here, spectral subtraction was successfully implemented to improve the segmentation result for short 44-second environmental recordings of 26 different anuran species.

However, for longer, continuous recordings, the sound variation becomes far more significant, and accurately selecting an approximate noise profile is vital for accurate results. Large-scale PAM systems typically possess a highly diverse range of noises, so it becomes increasingly difficult for one noise profile to cover the possible variance adequately. Thus, spectral subtraction may be suboptimal for applications using long-duration recordings with rapidly changing background noise and is better suited for applications employing shorter, non-continuous event-driven recordings.

3.3.2. Wavelet-based

Wavelet-based denoising uses the time–frequency domain created by wavelet transforms to localise features for generating sparse signal representations. An underlying assumption is that undesired noise is decoupled from the signal of interest by their frequency ranges. As such, wavelet-based denoising must be carefully applied, as noise can occur at any frequency. Consequently, its effectiveness will significantly depend on the type of noise and the signal of interest frequency ranges. Wavelet-based denoising has been used in ecoacoustic applications, across a range of species, particularly among those with vocalisations containing considerable low-frequency energy. The most common occurrence is cases where wavelet-based techniques have been applied, including humpback whales (Ren, Johnson, & Tao, 2008), anurans (Huang et al., 2014), and various bird species (Alonso et al., 2017).



(a) Original audio signal waveform.

(b) Original audio signal waveform with spectral subtraction denoising applied.

(c) Original audio signal waveform with wavelet-based denoising applied.

(d) A comparison between spectral subtraction and wavelet-based denoising.

Fig. 2. A comparison of two different denoising effects on a waveform for a 30-second-long recording taken from the A2O containing several overlapping bird calls. The first (a) waveform represents the original, noisy audio recording. The second (b) and third (c) signal waveform plots represent the effects of spectral subtraction and wavelet-based denoising, as red and orange, respectively, compared to the original signal (a). The fourth panel (d) shows a comparison between spectral subtraction (b) and wavelet-based denoising (c). These denoising effects were achieved using a combination of SciPy and PyWavelets (Lee, Gommers, Waselewski, Wohlfahrt, & O'Leary, 2019; Virtanen et al., 2020).

In general, wavelet-based denoising can outperform other methods by measuring Signal-to-Noise Ratio (SNR) and Segmental SNR (SSNR) across a range of noise conditions. Still, it is also less efficient than techniques such as low-, band- and high-pass filters, meaning that it is not as suitable or scalable to large-scale systems (Xie et al., 2020). Wavelet-based approaches represent a trade-off between flexibility and efficiency. They are not as efficient as filter-based approaches; however, as seen in Fig. 2(c) and (d), they can still achieve superior noise reduction and are not restricted to recordings of similar-frequency species as the underlying wavelet transform algorithm has high adaptation potential (Chen, Xie, & Zhao, 2013).

3.3.3. Low-, band- and high-pass filtering

Low-band- and high-pass filtering is the process of removing unwanted sounds within a particular frequency range. For example, if birds only call between 1 kHz and 12 kHz, then all other frequency ranges can be ignored (Pijanowski et al., 2011). Removal of such unwanted sounds outside of the specified frequency range is typically achieved through attenuation. Such approaches to denoising are conceptually simple and take comparatively less computational power than wavelet-based denoising and spectral subtraction. As such, it has seen wide usage as a pre-processing step for acoustic recordings such as birdsong (Brown et al., 2018b; Neal et al., 2011).

However, as illustrated in Fig. 3(b), (c) and (d) a caveat with such techniques is that they are unfit for recordings with overlapping calls due to the potential of mis-attenuation and removal of useful information outside of the specified ranges. In addition, accurate denoising is generally restricted to a single species or taxonomic group with vocalisations in a known, similar frequency range, thus limiting the effectiveness to typically singular or closely related species with low inter-individual call variation such as anurans (Chen et al., 2012; Huang et al., 2014).

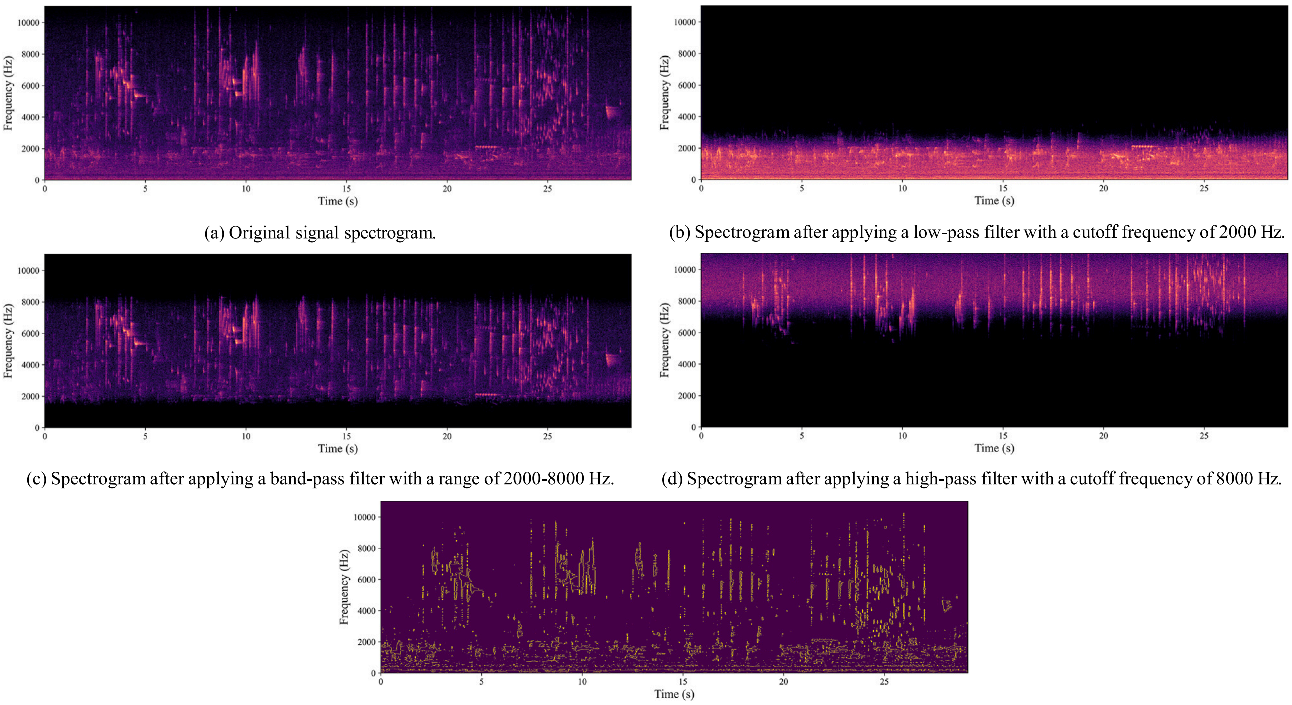
3.3.4. Image-processing based noise reduction

Image-processing-based approaches in ecoacoustics are applied to spectrograms, which are the transformed 2-dimensional representation of an ecoacoustic waveform. Once in this form, there are a variety of image-based techniques that can be used to reduce noise, such as edge detection (Hussein, Hussein, & Becker, 2012), smoothing (Lin, Chou, Akamatsu, Chan, & Chen, 2013), and other enhancement processes (Esfahanian, Erdol, Gerstein, & Zhuang, 2017) that can improve the spectrogram quality for downstream detection and classification tasks. An example illustration of edge detection as applied to a noisy ecoacoustics spectrogram sample can be observed in Fig. 3(e).

Such techniques have varying performance and flexibility, which is primarily determined by the species of interest. For example, image-based denoising has been shown to improve the classification accuracy for identifying bat vocalisations but has varied performance for other species (Heim et al., 2019). In the literature, the primary usage of image-based denoising is in tandem with other denoising methods for improving the SNR of ecoacoustic recordings. However, this can lead to increased computation requirements and thus may not be as suitable for large-scale systems in terms of scalability.

3.3.5. Deep learning-based noise reduction

Deep learning is a well-used method for localising the source of sound in acoustic environments, and reducing the overall noise levels, particularly when using deep convolutional networks trained to estimate the Direction of Arrival (DOA) for speech sources (Grumiaux, Kitić, Girin, & Guérin, 2022). Much work has gone into localising speech sources due to their importance in speech recognition tasks. For example, in a paper from 2019 (Chakrabarty & Habets, 2019), the researchers used a supervised learning-based CNN for multi-speaker environmental source localisation and accurately separated speakers in a dynamic environment.



(a) Original signal spectrogram.

(b) Spectrogram after applying a low-pass filter with a cutoff frequency of 2000 Hz.

(c) Spectrogram after applying a band-pass filter with a range of 2000-8000 Hz.

(d) Spectrogram after applying a high-pass filter with a cutoff frequency of 8000 Hz.

(e) This plot shows the results of image-processing techniques applied to the spectrogram, specifically thresholding and edge detection, to identify and enhance the contours of bird vocalisations.

Fig. 3. A comparison of spectrograms with the low-, band-, high-pass filter, and image-processing-based noise reduction (edge detection) applied for a 30-second-long recording taken from the A2O containing several overlapping bird calls. Librosa, a Python package for audio analysis was used for the low-, band- and high-pass filters, while OpenCV was used for finding the contours for edge detection for image-processing-based noise reduction (McFee et al., 2023).

Outside of speech, other methods of sound event localisation extend to various types of sound classes, such as in the DCASE 2019 challenge (Politis, Mesaros, Adavanne, Heittola, & Virtanen, 2021). In these cases, and depending on the challenge and the given dataset, several other types of sound, including anthrophony such as knocking on doors, phone ringing, etc., and some limited biophony such as barking dogs are present in the recordings. An advantage of source localisation is that it is just as capable of localising sources in clear single-source cases as with noisy, multi-source acoustic scenes, irrespective of time overlaps. Owing to the uniqueness of spectral characteristics, neural networks can exploit source localisation to great effect.

However, despite the advantages of sound event localisation, applying such techniques in ecoacoustics remains challenging due to interference by simultaneous unknown sound sources, typically also with an unknown mixing agent in the form of noise, rendering source isolation difficult (Lin & Tsao, 2019). It also remains inhibited by the lack of data completeness with accompanying labels, particularly for the full range of animal species present in natural ecoacoustics soundscapes.

3.3.6. Performance measures

The primary purpose of denoising is to increase the effectiveness of downstream tasks, including segmentation and classification, where evaluation is measured by the output of the final task rather than on the choice of denoising. However, two primary metrics are proposed in the literature for directly evaluating denoising algorithms, namely SNR and SSNR (Xie et al., 2020). SNR can be characterised as the ratio of signal power to noise power and is often expressed in decibels. It describes the level of the wanted signal relative to the quantity of background noise (Farina, Eldridge, & Li, 2021).

Moreover, SSNR is computed by calculating the SNR on a frame-by-frame basis over the signal and averaging those values, thereby factoring equal weighting for loud and soft recording portions. In addition to quantitative metrics which directly measure the signal-to-noise ratio, efficiency is another significant consideration, particularly as the need for scalability grows due to the implementation of large PAM networks.

3.3.7. Discussion of ecoacoustics denoising approaches

In ecoacoustics, noise can be detrimental to recovering original audio sources, or it can be used to provide additional information, such as in the case of large-scale PAM for studying the effects of non-biophonic sounds on animal behaviour. As illustrated in Fig. 3, low-, band-, and high-pass filtering approaches are simple and commonly used for attenuating unwanted sounds, but they lack the flexibility required for recordings with overlapping calls. Spectral subtraction can be practical for short, episodic recordings but not for long, continuous sequences with rapidly changing background noise. Wavelet-based denoising can outperform other methods in terms of noise reduction, but it is computationally expensive and lacks scalability. Image-processing-based approaches are applied to spectrograms to reduce noise, but they have varying performance and flexibility depending on the species being investigated. Deep learning is a well-used method for localising the source of sound and reducing noise levels, particularly for speech sources, but it remains inhibited by a lack of labelled natural soundscape data. Table 3 below displays the representative denoising approaches according to their target animal species. Image-processing-based approaches are the most widely used across all target species, with wavelet-based approaches in a close second.

3.4. Feature extraction

Feature extraction relates to obtaining a set of values representative of the original properties of the signal data to reduce the input data to facilitate subsequent learning and generalisation. Every audio signal consists of many features. However, the method of extracting the

Table 3
A summary of signal denoising approaches used for different animal species.

Author	Denoising approach	Target species
Xie et al. (2016)	Spectral Subtraction	Anurans
Huang et al. (2014)	Wavelet-Based	Anurans
Alonso et al. (2017)	Spectral Subtraction	Anurans
Heim et al. (2019)	Image-Processing Based	Bats
Hussein et al. (2012)	Image-Processing Based	Bats
Brown et al. (2018b)	High-Pass Filter	Birds
Ren et al. (2008)	Wavelet-Based	Ortolan Bunting (Birds)
Neal et al. (2011)	Band-Pass Filter	Birds
Hussein et al. (2012)	Image-Processing Based	Birds
Ren et al. (2008)	Wavelet-Based	Humpback Whale
Lin et al. (2013)	Image-Processing Based	Cetaceans
Esfahanian et al. (2017)	Image-Processing Based	North Atlantic Right Whale
Ren et al. (2008)	Wavelet-Based	Rhesus Monkey

characteristics of the dataset that are relevant is highly dependent on the problem to be solved. The process of extracting the right features for subsequent analysis is critical for ensuring downstream tasks are efficient and, likely; no single feature set will consistently outperform another across all possible scenarios.

Mel-Frequency Cepstrum Coefficients (MFCCs) are one of the most common representations used, which signifies the short-term power spectrum of a sound using quasi-logarithmic spacing to roughly resemble the resolution of the human auditory system. MFCCs have demonstrated several advantages, particularly in environmental acoustics, due to their simplistic nature and inherent robustness (Davis & Mermelstein, 1980; Mcloughlin, Stewart, & McElligott, 2019; Trawicki, Johnson, & Osiejuk, 2005). MFCCs are particularly useful in speech recognition and birdsong studies, meaning that the algorithms developed are well-optimised and robust. However, they have also been shown to be susceptible to interference from background noise due to an underlying dependence on the spectral form (Wu & Cao, 2005). Moreover, using MFCC features can introduce redundant information, and excluding this whilst maintaining a precise representation of the original signal can become challenging during algorithm optimisation.

Furthermore, a study in the 2016 DCASE challenge revealed that MFCCs ranked among the topmost popular feature representation, with the best result of 89.7% classification accuracy of a multi-class acoustic scene using a combination of MFCCs and spectrograms for input. MFCCs were also among the lowest in terms of mean Equal Error Rate (EER), with a score of 0.174 for a domestic audio tagging task (Mesaros, Heittola, Benetos, Foster, Lagrange, Virtanen, & Plumbley, 2018). MFCCs were commended for accurately representing a signal's spectral properties, allowing for high inter-class variability for class discrimination by classical ML approaches. Although MFCCs achieved good results, the TUT database (Mesaros, Heittola, & Virtanen, 2016) on which these approaches were trained was obtained from real-life environmental conditions and thus had overlapping sounds, consisting mainly of anthrophony, with comparatively smaller quantities of biophony and geophony.

Another standard feature set used in ML applications is Acoustic Indices. Acoustic indices are quantitative measures that describe various acoustic properties, such as frequency spectrum, temporal structure or amplitude fluctuations (Sueur et al., 2014). In the context of ML, a diverse array of acoustic indices are commonly extracted from raw audio recordings, serving as feature sets to describe various aspects of the input sound. These indices have proliferated significantly in recent years, reflecting the growing complexity and scope of ecoacoustic analysis. Prominent examples include the Acoustic Entropy Index (AEI) (Sueur, Aubin, & Simonis, 2008) and the Acoustic Complexity Index (ACI) (Pieretti, Farina, & Morri, 2011) which are used to broadly characterise the diversity of sounds in a given audio recording. They have been successfully employed in automated species classification applications, such as the identification of 12 different species of frog choruses in environmental recordings (Brodie, Allen-Ankins, Towsey, Roe, & Schwarzkopf, 2020). However, a notable caveat is that acoustic indices can be sensitive to recording conditions. Environmental factors such as background noise interference and overlapping calls can influence the acoustic properties of the sound and render models unable to generalise to other datasets apart from those upon which they were trained (Alcocer, Lima, Sugai, & Llusia, 2022).

Spectrograms and mel-spectrograms are also common feature sets used in deep learning (Stowell et al., 2018). Spectrograms provide a detailed, and visually interpretable representation of the frequency content of an audio signal over time, which encapsulates both spectral and temporal features. Due to this, however, spectrograms can also have a high-dimensional feature space. In contrast, mel-spectrograms offer reduced dimensionality by aggregating spectral information into mel-frequency bands, which can impact the ability of models to accurately capture certain spectral characteristics of the audio signal. Despite this, mel-spectrograms have seen use in deep learning species classification tasks based on natural soundscape recordings (LeBien et al., 2020; Mcloughlin et al., 2019).

Several other acoustic features exist, including spectral and temporal features. While an exhaustive description of all features is beyond this survey's scope, a few notable characteristics have merit for large-scale ecoacoustics. First is zero-cross rating, the rate at which the signal changes from positive to negative and helps detect sounds in noisy environments. Spectral flatness is also helpful for detecting how noisy a signal is, and the spectral centroid can be used to describe the timbre of a signal. Typically, these individual features are not often the only parameter measure for a signal but are often combined to enable better characterisation of a target signal.

3.5. Discussion of signal pre-processing techniques

As seen in Table 4, careful application of signal pre-processing techniques to large-scale ecoacoustics datasets must be considered to counterbalance signal degradation effects, decrease processing times, and obtain accurate downstream results. This process involves signal transformation, signal denoising, and feature extraction, each serving a unique purpose in enhancing the quality and interpretability of ecoacoustic recordings. The implications of these pre-processing techniques extend beyond data preparation; they influence the efficacy of subsequent analyses and the ecological interpretations derived from them.

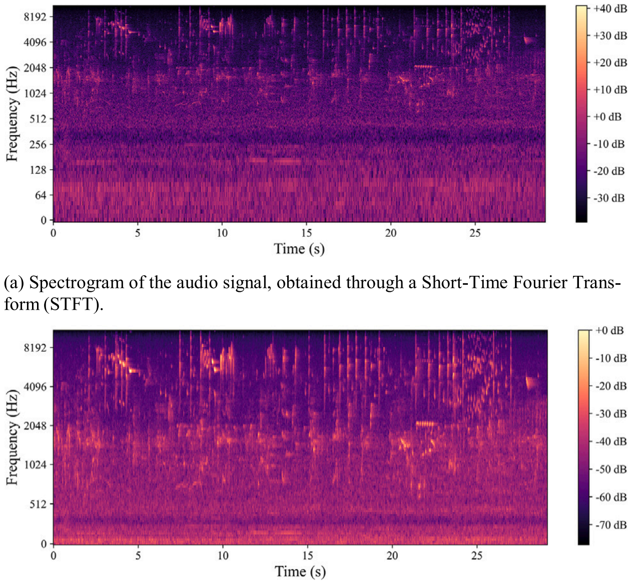
Signal transformation is instrumental for making raw ecological data amenable to detection and classification algorithms. Similarly, denoising techniques remove extraneous noises that may mask critical bioacoustic signals, thereby clarifying the desired vocalisations. Moreover, feature extraction is an essential step for distilling complex audio data into representative features that encapsulate the key characteristics of a soundscape. These processes not only simplify the dataset, making it more manageable for analysis, but also aid in reducing the risk of model overfitting, thus ensuring more robust and generalisable findings. This is essential in studies aimed at understanding animal behaviour and communication within natural habitats.

However, several challenges remain across these signal pre-processing areas. For signal denoising, a key difficulty is the balancing the removal of unwanted noise while preserving ecologically significant sounds, particularly in diverse and dynamic soundscapes typical of large-scale PAM. For feature extraction, the selection of meaningful features is crucial, as it directly impacts the accuracy and relevance of ecological interpretations derived from the data.

Future improvements in signal pre-processing could focus on enhancing algorithmic efficiency, reducing computational demands, and developing more adaptive methods capable of handling the inherent variability and complexity of natural soundscapes. Such advancements would not only bolster the accuracy and reliability of ecoacoustic analyses but also expand the scope and scale of biodiversity monitoring efforts globally.

Table 4
A summary of the advantages and disadvantages of signal pre-processing techniques.

Approach	Summary	Advantages	Disadvantages	Example references
Signal Transformation	Modifies the original raw signal into a smaller, more practical format that is more meaningful and compatible with detection and classification algorithms	Organises the signal data and improves the efficiency and quality of any subsequent analysis techniques	For large volumes of data, it can become a time-consuming and resource intensive process	Brown et al. (2018b), Qaisar et al. (2017), Truskinger et al. (2014)
Signal Denoising	Removes unwanted and disruptive sounds from the input signal to clarify the underlying desirable vocalisations and other information.	Can enhance the quality of the audio recording and attenuate interference and distortions while retaining the underlying signals.	Can potentially remove additional helpful information from the signal such as overlapping vocalisations or sounds with a distant source	Brown et al. (2018a, 2018b), Gibb et al. (2018);
Feature Extraction	Relates to the extraction of a set of values that are representative of the original properties of the signal data	Reduces the complexity and volume of the original data, which increases the efficiency of subsequent analysis and reduces the risk of model overfitting	Accurate results are reliant on the selection of meaningful features extracted from the original data	Babaee et al. (2017), Bonet-Solà and Alsina-Pagès (2021), Rama Rao, Garg, and Montgomery (2018)



(a) Spectrogram of the audio signal, obtained through a Short-Time Fourier Transform (STFT).

(b) Mel-frequency spectrogram of the same audio signal, visualised with 128 Mel-bands, highlighting the perceptual relevance of the frequency bands scaled to the Mel-scale, which more closely approximates to the human auditory system.

Fig. 4. A standard spectrogram (a) and a Mel-frequency spectrogram (b) illustrating the frequency components of a 30-second-long recording taken from the A2O containing several overlapping bird calls. Intensity of colours represents the magnitude of the frequency components in decibels (dB). The horizontal axis shows time in seconds, and the vertical axis represents frequency in hertz (Hz)

4. Visualisation of large ecoacoustic data

4.1. Background and challenges

With the introduction and technological advances of large PAM systems, it is now possible to record and store years' worth of audio at multiple locations across a continent—far in excess of what human experts can manually examine. As such, audio recordings of such length must be reduced or compressed without excess loss of information or detail. Visualisation of sound is one approach, as human visual interpretation has the most significant capacity to synthesise and integrate large amounts of information. However, visualising large-scale audio datasets remains a largely underexplored yet crucial area.

Traditionally, ecologists have used spectrograms, which are two-dimensional representations of sound. As illustrated in Fig. 4(a), time is generally expressed on the x-axis and frequency (Hertz or kilohertz) increases up the y-axis, with the sound's amplitude illustrated with colour intensity in decibels (dB). Typical spectrograms are a few seconds long, but they can be extended as long as required to capture the target animal's vocalisation. However, recordings longer than a few seconds must be first temporally split into fixed-length segments for visualisation with spectrogram representations. This is because the traditional representations cannot compress longer time sections meaningfully. In a typical spectrogram whose pixel rows and columns comprise the frequency bins and spectra respectively, a 24-hour long recording, if using a standard temporal scale of 0.02 spectra per frame and 35.7 pixels per centimetre display monitor, would be shown as a 1.2 km wide image (Towsey, Truskinger, & Roe, 2015).

To enhance the ecological utility of these visualisations, Fig. 4(b) introduces the Mel-frequency spectrogram. This approach modifies the frequency scale to align more closely with human auditory perception. This makes Mel-frequency spectrograms particularly effective for distinguishing subtle differences in sound that are often crucial in ecological monitoring, such as differentiating between similar vocalisations of species or discerning between animal sounds and environmental noise.

4.2. Visualisation approaches

4.2.1. Long duration false colour spectrograms

Long Duration False Colour (LDFC) spectrograms are an evolution of the standard spectrogram in that they incorporate combinations of acoustic indices to increase visualisation scalability (Towsey et al., 2018). Typically LDFC spectrograms are created by mapping three uncorrelated acoustic indices to red, green and blue (RGB) colour channels. Acoustic indices are calculated at a 1-minute resolution, allowing a full 24 h of recording to be represented on a standard monitor. They are designed to assist in gaining a comprehensive insight into a day's acoustic activity, which can enable the rapid detection of periods containing low ecoacoustic activity. In addition, they can also be useful for identifying species that produce consistent sounds over long periods such as chorusing frogs or insects (Brodie et al., 2020). Some example LDFC spectrograms taken from the day-long recordings from the A2O are illustrated in Fig. 5.

Meaningful results in the case of LDFC spectrograms depend heavily on the specific hand-picked acoustic indices (Bradfer-Lawrence et al., 2019). Consequently, information on species identity, specific geophonic activity, and anthropogenic events is lost, although it can be potentially changed by constructing the LDFC using specifically crafted acoustic indices. Another drawback is that they are largely ineffective in identifying species that call during morning chorus sequences or in tandem with many other species at similar times due to the 1-minute temporal resolution at which they are typically calculated. Accurate interpretation of LDFC spectrograms requires some specialist knowledge to fully understand their meaning, indicating that they are not non-expert friendly and thus may be difficult to standardise. Furthermore,

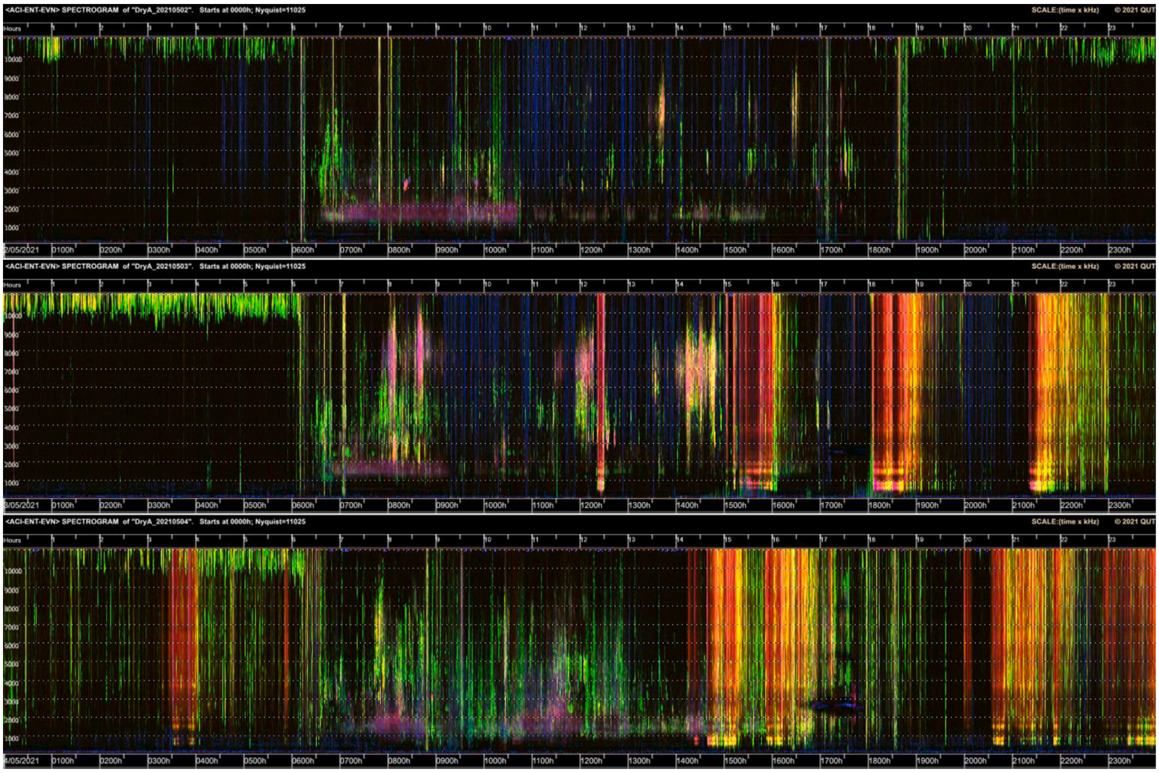


Fig. 5. An example of three LDFC spectrograms generated from consecutive days in May-April 2021 in the Tarcutta Hills, New South Wales, Australia region from the A2O sensor network showcasing the variation in acoustic activity on a day-to-day basis. The horizontal grid lines are at 1 kHz intervals. Each false colour was obtained by combining the ACI, Temporal Entropy (H[t]), and Acoustic Cover (ACR) indices in red, green, and blue colours respectively, as per Towsey, Zhang, et al. (2014).

they still suffer from the limitations of visualising more than 24 h of recording simultaneously for the same reason as standard spectrograms; they cannot be effectively scaled for multi-day visualisation without loss of information meaning.

4.2.2. Zooming spectrograms

A promising technical contribution that attempts to overcome the challenges of multiple day-long recordings is using zooming spectrograms (Towsey et al., 2015). This technique's visualisation enables the multi-scale viewing of long-duration recordings at minute, hour, day, or year intervals. Zooming spectrograms enable researchers to focus on specific portions of the recording containing acoustic events of interest, such as bird songs or insect calls. Through this, the structure and timing of acoustic signals can be more easily examined in relation to environmental factors such as weather, time of day, or location.

Zooming spectrograms can overcome the issue of visualising multi-day acoustic activity. However, due to the underlying reliance on a combination of false-colour spectrogram images, it is still ineffective at accurately representing species identities during overlapping vocalisations. Furthermore, the surrounding context may be lost when zooming in on a specific portion of the sound, making it more difficult to understand the overall acoustic environment and the relationship between independent events.

4.2.3. Long-term spectral averages

Long-Term Spectral Averages (LTSAs) offer another approach to visualising large-scale acoustic datasets. In LTSAs, the signal is divided into smaller segments, and the Fourier Transform is applied to each segment to obtain its frequency spectrum. These individual spectra are then averaged over time to produce a single spectral profile that represents the average frequency content of the entire signal. By averaging the frequency content over these longer durations, LTSAs provide a condensed yet informative view of the acoustic environment, making them particularly useful for viewing long-term trends and patterns in recordings with sparse acoustic activity such as deep-sea ecosystems (Lin & Tsao, 2018; Ryan et al., 2016).

The strength of LTSAs lies in their ability to reduce the data dimensionality while preserving essential features of the soundscapes. This makes them highly applicable for monitoring long-term changes in ecoacoustic environments, such as seasonal variations in animal vocalisations. However, while LTSAs excel at capturing broader trends, the averaging process can obscure finer details, such as short-term vocalisations or transient noise events. This makes them less suitable for studies requiring precise identification of individual species or specific acoustic events. Additionally, LTSAs are constrained by the resolution of the display medium, which can limit their effectiveness for visualising extremely large datasets.

4.3. Discussion of large ecoacoustic data visualisation

As depicted in Table 5, selecting an appropriate visualisation approaches to expansive ecoacoustics datasets is crucial for enhancing data interpretability and facilitating insightful ecological analyses. These approaches, encompassing LDFC Spectrograms, Zooming Spectrograms, and LTSAs, serve distinct roles in rendering acoustic data more accessible and informative. The impact of these visualisation strategies extends beyond mere data representation; they significantly influence the depth and breadth of ecological insights that can be gleaned from large datasets. As such, their utility in large-scale PAM is contingent upon continuous methodological refinement to enhance detail resolution, computational efficiency, and adaptability to diverse ecological contexts.

Future advancements in visualisation approaches should aim to address these challenges by developing more sophisticated algorithms capable of providing clearer, more detailed representations of ecoacoustic data. Enhancements in computational strategies and the integration of ML techniques may also offer promising avenues for improving the scalability and effectiveness of these visualisation methods in large-scale

Table 5
A summary of visualisation approaches for large-scale ecoacoustic data.

Visualisation approach	Key characteristics	Advantages	Disadvantages	References
Long Duration False Colour (LDFC) Spectrograms	LDFC Spectrograms enhance the scalability of visualising ecoacoustic data by integrating various acoustic indices into the red, green, and blue (RGB) channels.	This method allows for the observation of daily variations in acoustic activity, providing a scalable solution for large datasets and enhancing the visual interpretability of complex ecoacoustic data.	However, this approach may struggle with accurately representing specific species identities during periods of overlapping vocalisations, and the interpretation of these spectrograms generally requires a certain level of expertise.	Towsey, Zhang, et al. (2014), Towsey et al. (2018)
Zooming Spectrograms	Zooming Spectrograms offer a multi-scale perspective on long-duration recordings, allowing for detailed inspection at intervals ranging from minutes to years, thus focusing on particular acoustic events of interest.	They enable a thorough examination of the structure and timing of acoustic signals, providing insights into the relationship between acoustic events and environmental factors, which is invaluable for focused ecological studies.	The main drawback of this approach is that surrounding context may be lost when examining specific portions of time, which may limit understanding of the of the acoustic environment and its dynamics.	Towsey et al. (2015)
Long-Term Spectral Averages (LTSAs)	LTSAs simplify the visualisation of large-scale acoustic datasets by dividing the signal into smaller segments, applying the Fourier Transform to each, and then averaging these to represent the overall frequency content.	LTSAs are particularly effective for providing a condensed yet informative overview of the acoustic environment, making them ideal for observing long-term trends and patterns in areas with sparse acoustic activity.	However, the averaging process inherent in LTSAs can obscure important details such as short-term vocalisations and transient noise events, and their effectiveness is constrained by the resolution of the display medium, limiting their utility for very large datasets.	Lin and Tsao (2018), Ryan et al. (2016)

ecoacoustic studies. This extends to the improvement of the resolution and clarity of visual representations, developing automated tools for pattern recognition and anomaly detection, and creating interactive platforms that allow users to explore and analyse acoustic datasets intuitively.

5. Ecoacoustics data labelling and segmentation

5.1. Background

Audio segmentation involves isolating a signal of interest from a mixture of signals to be used for further processing. Several segmentation methods observed throughout the literature utilise other datasets to pre-train before training on the target dataset because there are insufficient levels of labelled data (Dufourq, Batist, Foquet, & Durbach, 2022; Tan et al., 2018; Zhong et al., 2020). This represents the most pressing issue halting further progress. As large-scale PAM sensor networks generate thousands of hours of audio daily, accurate labelling remains a significant challenge. The need for accurately labelled data is great because downstream DL classification relies on large quantities of data for generalisation. More is needed to adequately train models to detect a majority of the species vocalising, and perhaps the geophony occurring in a large ecoacoustics soundscape.

Thus, a more scalable and efficient solution for labelling species and non-biophonic sounds from large amounts of long-duration recordings is required. Current signal processing methods on smaller and curated ecoacoustics datasets do not address inherent complexities with large-scale datasets such as overlapping target sounds, environmental and background noises, reverberation, distance of target species from the recorder, and the variability of sound inter- and intra-species (Bravo Sanchez, Hossain, English, & Moore, 2021). PAM systems record continuously over many differing ecoregions. As exemplified in Figure 6, each consecutive day recorded, even within the same area, can have a completely different acoustic makeup from the previous day, further demonstrating the need for automation.

5.2. Manual annotation

Before the roll-out of large-scale sensor networks, traditional approaches generally involved manual annotation by ornithologists and ecologists using their expertise. However, with the volume of data captured by large-scale PAM sensor networks, manual labelling of ground truths is now too resource intensive. Now, other approaches such as data augmentation (Lasseck, 2019) and simulated synthetic data generation (Glotin, Ricard, & Balestriero, 2022) has been used to increase training data artificially (Stowell, 2022).

Manual approaches to labelling ecoacoustics data are highly time demanding and unscalable. Typically, experienced operators are needed, and it can take an average of 120 s to listen and annotate per 30-second sample (Linke & Deretic, 2019; Stowell & Sueur, 2020). The scalability challenges become particularly pronounced in large-scale PAM operations, which may involve extensive geographic coverage, high sensor density, or long-duration monitoring (Truskinger et al., 2014). Such operations generate vast datasets that far exceed the capacity for manual annotation, thus necessitating a more scalable solution.

5.3. Crowdsourcing approaches

A recent approach to overcome the lack of labelled data has been to use crowdsourcing to annotate audio scenes; however, progress is slow, and data still requires expert verification, and hence they are limited in scale and have varying performance (Cartwright et al., 2017; Shamir et al., 2014). It has seen wide adoption in ML areas for other domains such as computer vision (Krishna et al., 2016; Parent & Eskenazi, 2011). However, technique improvement has been slow for non-speech applications, such as ecoacoustics, due to the need for more data compared to other image-based tasks and practical audio labelling tools (Cartwright et al., 2017; Shamir et al., 2014). Another major issue is that citizen scientists have varying levels of skill and often lack expertise, and training them is a substantial undertaking. Furthermore, another reason for the lack of adoption is that crowdsourcing techniques require an ongoing engagement from the community of non-experts, which can sometimes be challenging to attract and not all questions can be answered using citizen science methods. There is also some concern regarding the scalability when applied to strong multi-label annotation. It may require considerably more effort due to the need for multiple passes over the data (Cartwright et al., 2017).

Despite its drawbacks, some tasks, such as weak labelling (Truskinger et al., 2013), can benefit from crowdsourcing. It can be an efficient and effective method for recruiting volunteers and providing reliable data to experts if the system is well-designed. Three major characteristics define such a system: (1) anyone can participate, (2) all participants use the same protocol and tools so data can be combined effectively and be of high quality, and (3) the data collected assists experts with deriving conclusions efficiently. However, such approaches still need to be underpinned by human involvement and are not fully scalable to large-scale annotation tasks.

5.4. Acoustic indices

An acoustic index is a statistic used to summarise a particular aspect of the distribution of acoustic energy within a recording. Primarily they have been increasingly employed to save time during the labelling process by providing a meaningful summary of the acoustic events within a recording. As acoustic indices are algorithmically simple, they can be rapidly developed and are highly scalable, making them ideal for large-scale PAM applications (Stowell & Sueur, 2020). There are many different acoustic indices, such as the acoustic richness index (Sueur et al., 2014) or ACI (Pieretti et al., 2011), and often a select few are used in tandem to describe an acoustic scene and can be specifically crafted to filter the acoustic features of a particular species. One such example is from a study conducted in 2021 which was able to achieve around 70% accuracy for broad label assignment (insect, birds, geophony, etc.) using a semi-automated, multi-index approach (Scarpelli, Liquet, Tucker, Fuller, & Roe, 2021).

However, it has been observed that, while increasing the efficiency of data labelling compared to traditional methods, the efficacy of acoustic indices may be hindered by several key challenges. While acoustic indices can be used to summarise a set of acoustic energy, this comes with the risk that target sounds may be obscured by non-target noise, thus masking the signal of interest and rendering their summaries ineffective (Metcalf et al., 2020). Due to its continuous nature, this can be particularly problematic in the context of large-scale environmental PAM, which can have higher proportions of non-biophonic sounds, such as geophony and anthrophony. Consequently, signals of interest can become partially or fully masked, leading to reduced index effectiveness and misleading correlations. Additionally, a recent study has shown that acoustic indices and ML models may not be universally reliable for predicting species richness across diverse ecosystems (Sethi et al., 2023). Here, the authors indicate that while changes in the soundscapes could indicate shifts in ecological communities, the acoustic features themselves were not consistently predictive of species richness across different datasets. This raises questions about the unreliability of acoustic indices, particularly when applied to varied ecological contexts, and as such, that acoustic monitoring should be used cautiously and in tandem with traditional in-person surveys for more reliable biodiversity assessments. This sentiment is shared by another study (Alcocer et al., 2022), which also shows that acoustic indices are not always good approximations of biodiversity. However, a recent study (Allen-Ankins et al., 2023) has successfully shown that acoustic indices were reasonable for predicting frog and bird diversity, but varied with habitat, and that combinations of indices were better than one index alone.

5.5. Unsupervised learning approaches

Unsupervised segmentation or clustering, in the context of ecoacoustics, is the process of grouping signals which are the most similar to each other and as dissimilar from signals in other groups as possible. It enables rapid automatic separation of input audio signals with similar features, without the requirement for labelled data. Using unsupervised learning to drive the use of DL remains understudied. However, unsupervised learning can provide high degrees of scalability to the labelling of long-duration recordings, which is particularly useful because the primary bottleneck for downstream DL classification tasks is typically the quantity of labelled data. However, it is worth noting that, despite the efficiency, segmentation results from unsupervised learning typically have lower accuracy than supervised or semi-supervised techniques and method validation generally requires expert intervention for the interpretation of discovered patterns (Babaee et al., 2017; Rama Rao et al., 2018); however, evaluation can be performed on the formed clusters and the model used to create them.

There are two main approaches when evaluating unsupervised learning clustering results, depending on the availability of ground-truth labels. Since the most likely use case of unsupervised learning is when ground-truth labels are unavailable, the evaluation must be performed on the model itself. Cluster evaluation can be achieved for k-means and Gaussian Mixture Models (GMM) by looking at the external measure of purity, reaffirmed by the internal measure of the silhouette index. Furthermore, Class-average Mean Average Precision (CMAP) is the mean of the per-class precision scores, and label-weighted label-ranking average precision (lwlrap) is the mean of the per-example precision's scores (Denton, Wisdom, & Hershey, 2022). Both are closely related metrics useful for multi-class multi-label contexts and have been the primary target metric for many BirdCLEF competitions (Kahl, Denton, et al., 2021) and DCASE audio challenges (Politis et al., 2021).

An approach such as Ozanich, Thode, Gerstoft, Freeman, and Freeman (2021) applies deep-embedded clustering, a form of unsupervised learning that combines an element of deep learning with clustering algorithms to automatically detect unlabelled signals in a coral reef soundscape. The feature vector used by the authors included hand-picked spectral and temporal features. It resulted in an accuracy level of 77.5% and could be achieved for a combination of fish and whale signals. For other marine species, unsupervised clustering has a near-perfect allocation of whistles for individual dolphins (Kershenbaum, Sayigh, & Janik, 2013). Here, the authors compared the results from k-means clustering, hierarchical clustering, and an Adaptive Resonance Theory (ART) neural network. In a similar area, but in a terrestrial application, unsupervised learning has also been used for the syllable clustering of ultrasonic rodent vocalisations, such as in the DeepSqueak analysis software (Coffey, Marx, & Neumaier, 2019) which implements two unsupervised clustering algorithms, both k-means and ART. Additionally, a recent study introduces a noteworthy advancement in the field of unsupervised learning for ecological soundscapes, particularly in its ability to handle multi-species detection without requiring labelled data (Guerrero, Bedoya, López, Daza, & Isaza, 2023). This is a significant step forward for large-scale PAM, especially in biodiverse regions where labelled data is scarce. However, the paper does have limitations that could impact its scalability and accuracy in real-world applications. For instance, the algorithm's performance can be impacted by background noise, a common issue in large-scale PAM. Additionally, the methodology involves complex pre-processing and segmentation steps, which could be computationally intensive for large datasets.

Thus, existing clustering approaches have demonstrated reasonable accuracy and reliability in most cases. Still, existing approaches typically focus on a select few target species or a particular taxonomic group. Thus, they vary in flexibility to different environments and scalability to large sensor networks.

5.6. Weakly-supervised approaches

Weakly-supervised approaches allow for labelling to take place that is imprecise or lacks detail. It represents a middle-ground between data label quality and efficiency, as weak-labelled data can be more rapidly generated than most other techniques. In the literature, some studies such as Kong, Xu, and Plumbley (2017) can only determine if there is the presence of a birdcall within an audio recording, and thus it does not consider other potential targets like other taxonomic groups such as anurans or other groups of sound like anthrophony or geophony often found in environmental recordings. Despite this, the authors achieved an accuracy of around 81% on unseen data, with the trade-off being that such an approach loses granularity and transferability to novel applications.

Another example can be observed in Coban, Syed, Pir, and Mandel (2021), but like the previous study, achieved a single-label true and false-positive detection for each species. Despite this, however, the approach does include other call varieties outside of birdsong, but the training data only covers a select number of bird and amphibian species in great conservation need, and of that, only a maximum of two common variations of call types are used. The authors here used a semi-automated template-based sound detection approach with a graphical user interface for post-validation and achieved a mean-average precision equal to 89.3% and total average precision of 97.5

5.7. Self-supervised learning approaches

Self-Supervised Learning (SSL) is a technique to generate a labelled dataset from unlabelled data. SSL eliminates the need for data labelling by taking unstructured data as input and generating its labels. The SSL model decides whether the labels generated are reliable and accordingly uses them in the next iteration to adjust their weights. However, SSL is computationally expensive because it needs to make sense of the unlabelled data and generate the corresponding labels; thus, generally, SSL has lower accuracy than traditional supervised learning models because they generate their labels without any external support for determining whether its computations are correct. Despite this, SSL approaches consistently outperform semi-supervised methods while being conceptually simpler.

An example can be seen in Saeed, Grangier, and Zeghidour (2021) where the authors use a contrastive learning approach pre-trained on AudioSet to achieve an accuracy equal to 80.2% on test data taken from the Bird Song Detection dataset. However, it is important to note that the authors use a binary classification approach that only considers the presence or absence of bird sound within the given recording. Several approaches, such as Denton et al. (2022), use an unsupervised approach based on acoustic indices as its input feature vector. The authors here use a combination of k-means and hierarchical clustering to achieve high quality separation of birdsong. A similar method is applied in Coban et al. (2021); however, the authors still had to apply data augmentation due to small quantities of labelled data. Data was manually listened to and annotated with multiple labels, further exemplifying the strong need for automation; however, the dataset used did contain a baseline level of background noise to be expected in natural environments.

The use of SSL methods in ecoacoustics has the potential to allow for increased accuracy gain from unlabelled data. Some of the more notable and applicable approaches include Swapping Assignments between multiple Views (SwAV) (Caron et al., 2021), SimCLR v2 (Chen, Kornblith, Swersky, Norouzi, & Hinton, 2020), SimSiam (Chen & He, 2020) or DeepCluster (Caron, Bojanowski, Joulin, & Douze, 2019). These represent some recent audio SSL computer vision approaches with the most applicability and potential for achieving accurate labels for large-scale natural soundscapes (Liu et al., 2022).

5.8. Discussion of ecoacoustics labelling and segmentation approaches

Table 6 shows several labelling and segmentation approaches in the literature; however, few can handle the full unique breadth of sounds observed in large-scale PAM datasets. While traditional and citizen science proved to be useful earlier in history, they lack the scalability requirements to handle long-duration data. Automated methods such as acoustic indices have been used to save time by providing a biologically meaningful summary of sound; however, accurate results still heavily depend on specific hand-picked methods (Bradfer-Lawrence et al., 2019). Furthermore, these types of methods lack consistency and may not be transferable to different environments, and in addition, are susceptible to poor performance because background sounds in recordings may contribute more to indices than sounds of interest, rendering their summaries ineffective (Scarpelli et al., 2021).

Results from unsupervised model-based approaches often possess lower accuracy than supervised, semi-supervised or self-supervised techniques and methods validation generally requires expert intervention for the interpretation of discovered patterns (Rama Rao et al., 2018; Thakur & Rajan, 2016). Existing clustering approaches have demonstrated relatively high levels of accuracy in some cases, but they typically only focus on a select few target species or a particular taxonomic group and are thus varied in terms of flexibility.

Weakly-supervised learning in large-scale ecoacoustics is beneficial in some select cases, but evidently, it will only be of significant benefit when the training dataset used is large enough to represent the full distribution of possible sounds. While it does represent a more efficient way of obtaining labelled data over traditional methods, it suffers from a lack of flexibility. Thus, further testing must be conducted using a dataset from multiple sources, ideally including Google's Audioset, VGG-Sound and other bioacoustics-specific datasets such as BirdCLEF, depending on the application.

In light of these observations, it is clear that there are persistent challenges that currently exist. While current methods have made strides in data labelling and segmentation, they often fall short in addressing issues like data imbalance, computational efficiency, and the adaptability to diverse acoustic environments. These limitations not only restrict the scalability of these approaches but also raise questions about their real-world applicability, and highlights the need for interdisciplinary research collaborations and more comprehensive datasets to advance the field effectively.

6. Ecoacoustics detection and classification

In ecoacoustics, classification predicts one or more categorical labels, such as species or call type. Although often used interchangeably, the detection task differs from classification in that it generally provides temporal detail of when a sound event occurs. In general, detection in ecoacoustics can be broken into three main approaches. The first is detection as binary classification, which returns a binary (yes/no) decision as to whether the presence of a target signal is found within a recording, which is commonly named occupancy detection (Stowell, 2022). The primary advantage of such a technique is that it allows for large amounts of data to be quickly sifted through by ignoring sections of the recording containing no significant sound events.

The second involves defining both the start and end times and the types of sound events, otherwise known as Sound Event Detection (SED) or Acoustic Event Detection (AED) (Morfi et al., 2021). While this adds a degree of complexity and annotation time, it allows for superior segmentation of significant acoustic events. The third approach is image object detection, which consists of localising a vocalisation by surrounding it in a 2-Dimensional (2D) bounding box within a spectrogram image, where each bounding box represents a single sound event. Without adding considerable complexity to the annotation process, using bounding boxes permits downstream classification tasks to leverage developments and optimisations in computer vision techniques. Many studies have successfully evaluated the classification performance of models based on prior signal detection; however, techniques that integrate both a component of detection and classification for multi-species identification in noisy soundscapes are notably rarer (LeBien et al., 2020).

6.1. Traditional approaches

6.1.1. Manual approaches

In recent history, the identification of animals has been a task reserved only for expert ecologists and ornithologists. Typically, the identification of species would take place mainly aurally and visually

Table 6
A summary of ecoacoustics data labelling and segmentation approaches.

Approach	Scalability	Flexibility to different ecoregions and taxa	Accuracy	Efficiency	Noise handling	Long-duration handling	Example references
Manual Annotation	Low	High	90%	Slow	No	No	Linke and Deretic (2019)
Crowdsourcing	Low-Medium	High	90%	Slow	No	Sometimes	Cartwright et al. (2017), Shamir et al. (2014)
Acoustic Indices	High	Low-Medium	70%	Fast	Yes	Sometimes	Metcalf et al. (2020), Scarpelli et al. (2021)
Unsupervised Learning/Clustering	High	Varied	77%–90%	Varied	No	Yes	Coffey et al. (2019), Guerrero et al. (2023), Kershenbaum et al. (2013), Ozanich et al. (2021)
Weakly-Supervised Learning	High	Medium	90%	Fast	No	Yes	Coban et al. (2021), Kong et al. (2017)
Self-Supervised Learning	High	Varied	80%	Slow	No	Yes	Denton et al. (2022), Saeed et al. (2021)

using the playback of recordings alongside the associated spectrogram (Swiston & Mennill, 2009). However, identification in this form relies on the availability of individuals with the expertise to identify diverse animal sounds, which presents potential concerns with observer bias. Despite this, it can still take an expert some time to make an accurate assessment, limiting manual observations to small volumes (Joshi, Mulder, & Rowe, 2017). Even if an expert can rapidly identify animals from recordings without bias, there is simply too much sound to examine and label in long-term environmental recordings.

6.1.2. Probabilistic approaches

Probabilistic methods have been developed to assist in identifying species. One of the primary approaches is Hidden Markov Models (HMMs), which probabilistically infers whether a signal of interest is present based on an underlying multi-state model. The main advantages of these approaches are that they incorporate temporal detail on signals. However, such models are typically complex for non-experts to develop and understand. In addition to requiring large amounts of training data, many existing approaches only consider a particular species.

For instance, in Trawicki et al. (2005), the authors used a combination of HMMs with MFCC feature vectors as input for the classification of a type of bird, the Norwegian Ortolan Bunting (*Emberiza hortulana*), with relative success, equal to 63% to 92% accuracy depending on the number and similarity of songs used. The dataset used contained approximately 8500 manually labelled songs averaging a length of 1.5 s, which is not insignificant. Moreover, this manual labelling would need to be repeated if the technique was applied to another species.

Another example of probabilistic methods' performance was in a study conducted in 2014 (Zilli, Parson, Merrett, & Rogers, 2014) for detecting Cicada activity using HMMs applied to crowdsourced smartphone recordings. Here, the authors used data captured by visitors to New Forest, a national park on the south coast of England, via a smartphone app. F1-scores of 67% to 82% were reached depending on the feature vector used, reflecting the balance between the model's precision and recall. Precision, in this context, measures the accuracy of the system in identifying true Cicada events among all detected events, while recall assesses the system's ability to capture all actual Cicada events within the noisy, crowd-sourced data. An F1-score of 67% indicates a moderate balance between these two metrics, suggesting room for improvement in either precision, recall, or both. In contrast, a score of 82% denotes a stronger alignment between the system's ability to accurately identify true Cicada events and to minimise missed actual events. Despite these results, the technique lacked robustness to noise, leading to noticeable decreases in performance. Additionally, the calls used lasted long periods without interruption and were clearly distinct from background noises, which may hold true when applied to cicadas, but may not always be the case in other contexts within a large soundscape.

Another emerging probabilistic approach is the use of Gaussian Mixture Models (GMMs), particularly in the context of assessing acoustic heterogeneity in transformed landscapes. A recent study introduced a novel Acoustic Heterogeneity Index (AHI) that employs GMMs to model the acoustic dissimilarity between sites experiencing varying levels of landscape transformation (Rendon, Rodríguez-Buritica, Sanchez-Giraldo, Daza, & Isaza, 2022). This methodology involves a comprehensive five-step process, which includes noise filtering, acoustic index estimation, temporal pattern inclusion, and GMM-based classification. Tested in tropical dry forests, the approach achieved F1-scores of 92% and 90% in different regions. However, it is worth noting that the method was specifically tailored for landscape transformation and may not be directly applicable for species identification. Additionally, the study did not address how well the GMMs would perform in acoustically complex and noisy environments, leaving room for further investigation.

6.2. Machine learning approaches

6.2.1. Decision tree

Decision trees are supervised learning algorithms that classify unknown signals based on their similarity to previously learned features from training data. In 2010 there was a study that experimented with Ocellated Turkey (*Meleagris ocellata*) acoustics and used several classification algorithms, including decision trees, Artificial Neural Networks (ANN), Support Vector Machines (SVM), random forest and fuzzy classifiers (Kampichler et al., 2010). They found that neural networks and SVMs were not performant enough. Instead, the use of both decision trees and random forests was capable of achieving a high level of accuracy while also having discernible transparency.

Furthermore, a more recent study demonstrated the efficacy of using time-series motif discovery and random forest classification for categorising broad sources of sound in ecoacoustic data (Scarpelli et al., 2021). Specifically, the study's approach focuses on broad categories like birds, insects, and geophony. Their approach was able to achieve 70% accuracy in the label assignment across two datasets. While this type of approach has its merits, it may not be sufficient for researchers interested in identifying specific species or understanding nuanced environmental interactions.

6.2.2. Artificial neural network

Artificial Neural Networks (ANN) have been extensively used in various classification and detection tasks due to their inherent robustness against fuzzy data and their ability to carry out non-linear discrimination. ANNs were used for insect detection in a study over two decades ago in 2001 (Chesmore & Nellenbach, 2001). Here, the authors used ANN to identify 25 different species of British grasshoppers and crickets accurately. Combining time domain signal processing with an ANN, the authors' preliminary findings demonstrated high classification accuracy approaching 100%. Furthermore, they proposed that the technique was applicable to other insect varieties and other taxa, including birds. However, as more noise was introduced, the system's performance gradually deteriorated, and the identification accuracy was non-uniform across all species, with two species, in particular, showing a significant decline.

A more recent example of ANN was used to classify call sequences emitted by bats (Preatoni et al., 2005). The authors aimed to evaluate four different classification methods, with ANN among them, by collecting 126 3-second sound samples. The authors purposefully eliminated noise through digital filtering to ensure only the cleanest ultrasonic clicks remained, consequently bringing into question its efficacy in real-world test cases. Following a comparison of each of the four methods, it was concluded that ANN performed accurately, but to warrant its further use it needed to demonstrate more benefits over other techniques, such as multiple discriminate function analysis. In another similar study, the authors developed a technique for classifying independent bird calls, yielding mixed results ranging from 67% to 97% accuracy (Zilli et al., 2014). As such, ANNs have shown some promise in various classification tasks, however, their performance can be impacted by the presence of noise and their superiority over alternative techniques needs further validation.

6.3. Advanced neural approaches

6.3.1. Transfer learning

Transfer learning is another approach, which can bypass the need for random initialisation of a neural network, instead allowing for the transference of knowledge from one task, to another closely related one, through the reuse of learned information (Tan et al., 2018). The most common pretraining is now conducted using Google's AudioSet – an immense dataset of audio recordings taken from YouTube videos – enabling the inheritance of knowledge to be applied to new domains, such as predicting calls of interest, without needing to annotate large amounts of data.

Transfer learning has shown mixed results. A recent approach in 2022 saw the use of transfer learning as applied to PAM (Dufourq et al., 2022). Here, the authors attempted to conduct a large-scale investigation on four passive acoustic datasets containing the calls of a select few endangered species. Contrary to most preconceptions that large amounts of data are needed for training deep learning models, the authors showed that transfer learning performed well, achieving an F1-score of 82% on a small dataset of 25 samples. However, it was revealed that such results could result from low inter-species call variation, and high signal-to-noise ratios, indicating that such a technique might not be as applicable to continent-scale datasets.

Another example can be seen in Zhong et al. (2020) where they use transfer learning combined with pseudo-labelling to train the model, initially trained on ImageNet, on the pseudo-labelled data to predict the labels on unlabelled data. The authors use a dataset taken from Puerto Rico, in which they apply a template matching process to segment sound events of interest, which are then labelled manually as either positive or negative matches. They concluded that transfer learning and pseudo-labelling could classify the presence or absence of 24 species better than using a pre-trained ResNet50 CNN and training using a VGG16 architecture. However, while these results are promising, the data used to train the model is only partially representative of a whole soundscape and may not cover the diversity of inter-species call variation and external sounds such as geophony and anthrophony. This is particularly evident because their model does not perform equally for each of the 24 species in their dataset.

As shown, transfer learning approaches can achieve relatively accurate results. However, they lack adaptability because the quality of the output labels still relies on sufficient data that covers the range of possible sounds. Despite the mostly positive relationship between the size of training data and the classification accuracy, the output of transfer learning is only as good as the initial annotations. Due to the broad spectrum of sounds found in large-scale PAM datasets, ensuring adequate coverage may still take considerable effort.

6.3.2. Deep learning

For DL-based approaches, signals are often detected and classified based on similarity to a learned, labelled training dataset. The primary advantage of these techniques is that they overcome issues of noise-sensitivity feature extraction stages present in other approaches by learning features directly from the input data. However, to avoid overfitting, DL approaches require extensive datasets; thus, significant progress is halted by the need for publicly available labelled datasets which adequately cover the full range of inter and intra-sound variations found in large-scale ecosystems.

In the literature, two major approaches to DL classification exist. The first is to use spectrograms, or Mel-spectrograms, as the most common input (LeBien et al., 2020; Mcloughlin et al., 2019; Stowell et al., 2018). This is because techniques developed around image-based input data are well-optimised; however, some information is lost due to the transformation from raw input to spectrogram representation. The other recent approach directly uses raw audio data as input for minimised information loss. As developments continue into raw audio models, with developments such as WaveNet (van den Oord et al., 2016), progress towards automated species classification may move in this direction once more standardised neural network architectures are proven to work well for a variety of tasks. Thus, this research may benefit, soon, from a hybrid approach that incorporates a component of raw audio waveforms in addition to spectrogram images.

DL has been used in some species classification tasks; for example, in a study in 2018 (Fazekas et al., 2018), a CNN was used to identify bird songs. Input data used by the authors consisted of field recordings collected from habitats which were subsequently cleaned and separated into acoustic events, noise, and irrelevant segments. Using the cleaned data, the authors concluded that frequency features are more easily distinguishable than time features. While the result indicated that noise within the recording improved performance, from this study, it is inconclusive if the recordings contained sufficient environmental noise and variation that could be applied to different ecoregions. More recently, the scope and complexity of DL applications has expanded. For instance, a study conducted in Sonoma County, California, used a pre-trained CNN fine-tuned with a custom-labelled dataset to classify a broad range of soundscape components, achieving an overall F0.75-score of 0.88 (Colin Quinn et al., 2022). Another study in Guangzhou, China, used CNNs and Target Sound Area Ratios (TSAR) to quantify the dominance of seven types of acoustic scenes in urban forests, achieving an F1-score of 0.97 (Hao et al., 2022). These studies not only underscore the growing role of DL in ecoacoustics analysis, but also highlight the potential for these methods to capture complex interactions between human activities and biodiversity, thereby contributing to ongoing conservation efforts. Thus, DL approaches offer advantages in feature learning directly from input data but require extensive datasets. The choice between image-based and raw audio-based approaches is a topic of exploration, and future advancements may involve hybrid models combining both approaches.

6.4. Performance measures

For sound event detection, two primary metrics are used for evaluation: segment-based and event-based metrics, which are typically drawn from comparing the system output and ground truthed labels (Xia et al., 2019). In the context of this study, TP (True Positives) refers to the number of correctly identified positive instances, TN (True Negatives) to the number of correctly identified negative instances, FP (False Positives) to the number of negative instances incorrectly

Table 7
A summary of ecoacoustics detection and classification approaches.

Category	Approach	Summary	Advantages	Disadvantages	Accuracy	Efficiency	Example references
Traditional Approaches	Manual Methods	Species identification would typically take place by experts based on examination of physical characteristics.	Often required for initial verification (ground-truthing) to ensure that the species captured by PAM are actually visually seen at the given location.	Reliant on the availability of human experts and can be susceptible to observer bias.	High	Low	Joshi et al. (2017), Swiston and Mennill (2009)
	Probabilistic Methods	Probabilistically infers whether a signal of interest is present, based on an underlying multistate model.	Incorporates temporal detail on the signal.	Complex to develop for non-experts. Requires large quantities of training data. Many existing methods only consider a particular type of species or sound.	Moderate-High	Moderate-High	Trawicki et al. (2005), Zilli et al. (2014)
Machine Learning Approaches	Decision Trees	Supervised algorithms classify unknown signals based on their similarity to previously learned features from training data.	Uses a while-box model to provide transparent and comprehensible Boolean logic decisions via visualisation.	Requires a high quantity of expert-verified data. Feature extraction techniques are often highly noise-sensitive, and learners can create complicated trees that overfit the data.	Moderate	Moderate	Kampichler et al. (2010), Scarpelli et al. (2021)
	Artificial Neural Networks	Artificial neural networks model the human brain with building blocks designed to mimic the neurons in the human brain.	Can be flexible and be used for both regression and classification problems. Any data which can be made numeric can be used in the model.	Uses a black-box model where results can be more difficult to interpret.	Moderate-High	Moderate	Fox et al. (2008), Zilli et al. (2014)
Advanced Neural Approaches	Transfer Learning	Transfer learning allows a neural network to leverage pre-existing knowledge from one task to improve performance on a closely related, but different, task, reducing the need for extensive data annotation.	Has demonstrated promising classification accuracy, even with small datasets, by utilising pre-trained models on large, diverse audio datasets like Google's AudioSet.	May struggle with adaptability across diverse soundscapes, as the quality of its output is heavily dependent on the initial annotations and may not account for the full range of inter-species call variations or external sounds like geophony and anthrophony.	High	Moderate-High	Dufourq et al. (2022), Zhong et al. (2020)
	Deep Learning	Signals are detected and classified based on similarity to a learned training dataset.	Superior at distinguishing latent features in a dataset and highly suitable for large ecoacoustics datasets.	Accurate results require data to undergo several stages of cleaning and pre-processing. Training can also be computationally expensive and time intensive.	High	Low-Moderate	Fazekas et al. (2018), Hao et al. (2022), Colin Quinn et al. (2022)

identified as positive, and FN (False Negatives) to the number of positive instances incorrectly identified as negative.

Standard metrics can be used for classification to evaluate results (Bravo Sanchez et al., 2021). Here, metrics such as accuracy, precision, recall, F1-score, and the Receiver Operator Characteristic (ROC) - Area Under the Curve (AUC), or (ROC-AUC) can be calculated. Accuracy is the ratio of correctly predicted instances to the total instances and is calculated using the equation:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)}.$$

Precision is the ratio of true positive instances to the total positive instances, as high precision relates to a low false positive rate. Precision is calculated using the equation:

$$Precision = \frac{TP}{(TP + FP)}.$$

Recall, also known as sensitivity, is the ratio of true positives to all instances in a class. Recall is calculated using the equation:

$$Recall = \frac{TP}{(TP + FN)}.$$

F1-score is the weighted average of precision and recall. The F1-score considers precision and recall, which usually makes it a better metric for evaluating a model than accuracy as long as false positives and false negatives have a similar cost. The F1-score is calculated using the equation:

$$F1 - score = 2 * \frac{(Recall * Precision)}{(Recall + Precision)}.$$

In addition, ROC-AUC metric are the measures of the ability of a classifier to distinguish between classes, where generally the more significant the AUC, the better the performance of the model at different threshold points between positive and negative classes.

6.5. Discussion of ecoacoustics detection and classification approaches

As summarised in Table 7, no single detection and classification approach is ideal in all scenarios. The application spectrum of these approaches is vast, ranging from conservation biology and environmental monitoring to landscape ecology and behavioural studies. For instance, probabilistic methods, such as HMMs, offer a robust framework for detecting specific vocalisations against a backdrop of ambient noise, proving invaluable in studies focusing on species with well-defined call structures. On the other hand, ML techniques, including Decision Trees and Artificial Neural Networks, provide a powerful tool for classifying

Table 8
A comprehensive overview of the open problems and future challenges in ecoacoustics data analysis.

Open problem	Current status	Future directions
Open Ecoacoustics Monitoring	• The transition to open ecoacoustics monitoring is hampered by complex issues, including the allocation of computational resources, efficient data handling, and maintaining data privacy. • These challenges hinder the progression to a more collaborative and accessible ecoacoustics research field.	• Priority should be placed on the development of open ecoacoustics platforms to encourage community-driven initiatives. • These platforms must support standardised data sharing and annotation formats to unify and expedite advancements in ecoacoustics research
Scalable, Real-Time Monitoring	• The current state of algorithms in PAM studies present a significant challenge when it comes to their application in real-time monitoring of species on a large scale.	• Emphasis should be placed on optimising algorithmic design to reduce computational requirements without compromising the quality of analysis. • It is also imperative that these algorithms are tested in diverse and realistic field conditions to ensure their robustness and adaptability in various ecological settings.
Species Spatiotemporal Monitoring	• The analysis of species' movement and behaviour through spatiotemporal monitoring is impeded by the problem of non-trivial sound propagation in diverse outdoor environments. • This challenge is compounded by environmental noise and the complexity of overlapping vocalisations, which obscure target signals and hinder accurate analysis.	• Advancements in spatiotemporal monitoring techniques, particularly in direction-of-arrival and spatial analysis, are crucial. • Future developments should focus on refining these techniques to accurately capture the directionality and spatial distribution of vocalising species, even in acoustically complex environments.
Robust Sound Distinction	• Identifying and isolating overlapping sounds, especially during periods of high acoustic activity such as dawn choruses, remains a significant challenge. • This difficulty extends to differentiating between biophonic, geophonic, and anthrophonic sounds, which are crucial for understanding environmental interactions and species behaviour.	• Future research should aim to enhance the ability to distinguish overlapping sounds accurately and retain crucial information, particularly for non-biophonic sounds that play a key role in ecological studies.
Flexible Visualisation of Extended Duration Data	• Current visualisation methods struggle to effectively represent the diversity of ecoregions and the continuous nature of recordings, limiting their utility in broad-scale ecological analysis.	• Investigating and developing innovative interactive visualisation methods is essential. These methods should be capable of effectively managing and compressing long-duration data, providing clear and comprehensive visual representations that cater to the needs of large-scale PAM.
Training Without Large Datasets	• SSL methods show promise for training models with limited, unbalanced data. However, their effectiveness for large-scale, multi-class applications in ecoacoustics remains under-explored and uncertain.	• Intensive research into the efficacy of SSL for ecoacoustics is needed. Additionally, exploring the use of simulated sounds to augment existing datasets could provide a pathway to improve model training and accuracy.
Open Set Recognition	• There is a growing need for the capture and creation of a model that is generalised such that it can detect new vocalisations that were not accounted for in the original training set. • The ability to extend the boundaries of the known set of target classes, also known as open set recognition is a growing necessity (Stowell, 2022).	• Future research should focus on developing hierarchical unsupervised learning approaches. These approaches would be instrumental in efficiently detecting novel occurrences within continuous datasets.
Object Detection and Image Segmentation for Sound Event Detection	• Leveraging object detection and image segmentation techniques in sound event detection, such as using architectures like You Only Look Once (YOLO) and Region-Based Convolutional Neural Networks (R-CNN), is a relatively unexplored area.	• Advancing the use of object detection and image segmentation techniques in ecoacoustics would enable more nuanced and detailed analysis of soundscapes, leading to higher-resolution recognition of acoustic events and more informative downstream analysis for ecological research.

complex acoustic data, facilitating the identification of multiple species within a recording.

The wide variance in detection and classification approaches reflects the dynamic complexity of natural soundscapes, where factors such as species diversity, habitat characteristics, and ambient noise levels interplay. This diversity necessitates a tailored approach to ecoacoustic analysis, where the choice of method aligns closely with the specific research objectives, whether it be broad-scale biodiversity assessment or targeted species monitoring. Thus, corresponding approaches will depend greatly on the associated dataset, species of interest and required detection granularity, with binary presence detection entailing the least annotation work and bounding boxes the most.

However, it is crucial to note that these methods also come with their own set of limitations and challenges. For instance, adaptability to diverse soundscapes remains a significant hurdle, as many techniques are tailored for specific environments or species and may not perform well when applied universally. Another challenge lies in the handling of background noise, which can significantly impact the accuracy of both traditional and ML-based methods. Moreover, the dependency on extensive, well-annotated datasets for training ML models poses a barrier to the scalability and applicability of these techniques. The future of ecoacoustics detection and classification lies in addressing these challenges through the development of more adaptable, efficient, and universally applicable methods. Additionally, strengthening interdisciplinary collaborations and open ecoacoustics platforms will be crucial for advancing the field, enabling the sharing of data, tools, and methodologies across the global ecoacoustics community.

7. Open problems and future challenges

As the field of ecoacoustics continues to evolve and expand, it becomes increasingly apparent that several key challenges need to be addressed to fully realise the potential of automated, large-scale ecoacoustics data analysis. Despite notable progress in the field, these challenges represent significant barriers to the development and implementation of effective and comprehensive ecoacoustic monitoring systems.

The challenges highlighted in this section stem from a variety of factors including technological limitations, data complexity, and the need for advanced analytical methodologies. Addressing these challenges requires a multifaceted approach, involving innovations in ML and DL, improvements in data processing and management techniques, and the development of more sophisticated tools for data visualisation and interpretation. The successful navigation of these challenges will pave the way for more accurate, efficient, and scalable ecoacoustics data analysis, thereby enhancing our ability to monitor and understand the natural world.

Table 8 provides a comprehensive overview of the most pressing open problems and future challenges identified in the field of ecoacoustics data analysis. For each challenge, we discuss the current status and outline potential future directions, offering insights into the advancements needed to address these critical issues.

8. Conclusion

With the recent developments in PAM hardware and processing, there are now greater volumes of raw ecoacoustics data than ever before, far surpassing the capabilities of current techniques to analyse it fully. Analysis methods must be refined to achieve truly useful biodiversity monitoring, including multi-species detection, classification, and the processes surrounding data labelling and visualisation.

This comprehensive review has delineated the current state and advancements in pre-processing, detection, and classification techniques within the field of ecoacoustics, particularly emphasising their application in large-scale PAM. We have critically examined various methodologies, highlighting their strengths, limitations, and suitability for large-scale ecoacoustic data analysis. Our discussion underscores the imperative need for more nuanced, robust, and reproducible approaches that can handle the complexities inherent in large-scale ecoacoustic datasets.

Current techniques lack the flexibility or accuracy required to translate the unique breadth of sounds captured by large-scale PAM sensor networks. This can mainly be attributed to the need for labelled datasets that adequately cover the complexities. Such datasets are difficult to create due to the challenges surrounding the labelling of long-duration ecoacoustics data. As such, this survey has identified several open challenges and future directions for ecoacoustics analysis research as applied to large-scale PAM.

Looking forward, several avenues for future research emerge from our review. Firstly, the development of more sophisticated ML and DL models that can effectively handle the high variability and volume of ecoacoustic data is paramount. These models must be adept at distinguishing between overlapping biophonic, geophonic, and anthrophonic sounds, ensuring accurate species identification and ecological monitoring. Secondly, priority should be placed on the development of robust, optimised algorithm design to reduce computational requirements and ensure that large-scale PAM datasets can be effectively analysed. Lastly, fostering transparent, accessible datasets and collaborative platforms will enable for more rigorous validation of techniques, and accelerate innovation.

By addressing these gaps, we can significantly enhance our understanding of biodiversity and ecosystem health, contributing to more informed conservation and management decisions. The potential of ecoacoustics in providing insights into environmental changes, species diversity, and ecosystem dynamics is immense, and with continued research and technological advancements, its impact can only grow.

CRediT authorship contribution statement

Thomas Napier: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Euijoon Ahn:** Writing – review & editing, Supervision. **Slade Allen-Ankins:** Writing – review & editing, Supervision. **Lin Schwarzkopf:** Writing – review & editing, Supervision. **Ickjai Lee:** Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

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

Data availability

Data will be made available on request.

References

Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*, 1533–1545.

Agranat, I. (2009). Automatically identifying animal species from their vocalizations.

Alcocer, I., Lima, H., Sugai, L. S. M., & 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., et al. (2023). Effectiveness of acoustic indices as indicators of vertebrate biodiversity. *Ecological Indicators*, *147*, Article 109937.

Alonso, J. B., Cabrera, J., Shyamnani, R., Travieso, C. M., Bolaños, F., García, A., et al. (2017). Automatic anuran identification using noise removal and audio activity detection. *Expert Systems with Applications*, *72*, 83–92.

Babaee, E., Anuar, N. B., Abdul Wahab, A. W., Shamshirband, S., & Chronopoulos, A. T. (2017). An overview of audio event detection methods from feature extraction to classification. *Applied Artificial Intelligence*, *31*, 661–714.

Bardeli, R., Wolff, D., Kurth, F., Koch, M., Tauchert, K.-H., & Frommolt, K.-H. (2010). Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. *Pattern Recognition Letters*, *31*, 1524–1534.

Bonet-Solà, D., & Alsina-Pagès, R. M. (2021). A comparative survey of feature extraction and machine learning methods in diverse acoustic environments. *Sensors*, *21*, 1274.

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

Bravo Sanchez, F. J., Hossain, M. R., English, N. B., & Moore, S. T. (2021). Bioacoustic classification of avian calls from raw sound waveforms with an open-source deep learning architecture. *Scientific Reports*, *11*.

Brodie, S., Allen-Ankins, S., Towsey, M., Roe, P., & Schwarzkopf, L. (2020). Automated species identification of frog choruses in environmental recordings using acoustic indices. *Ecological Indicators*, *119*, Article 106852.

Brown, A., Garg, S., & Montgomery, J. (2018a). Automatic and efficient denoising of bioacoustics recordings using MMSE STSA. *IEEE Access*, *6*, 5010–5022.

Brown, A., Garg, S., & Montgomery, J. (2018b). Scalable preprocessing of high volume environmental acoustic data for bioacoustic monitoring. *PLoS One*, *13*, Article e0201542.

Cai, J., Ee, D., Pham, B., Roe, P., & Zhang, J. (2007). Sensor network for the monitoring of ecosystem: Bird species recognition. (pp. 293–298).

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., et al. (2012). Biodiversity loss and its impact on humanity. *Nature*, *486*, 59–67.

Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2019). Deep clustering for unsupervised learning of visual features. arXiv:1807.05520 [cs].

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2021). Unsupervised learning of visual features by contrasting cluster assignments. arXiv:2006.09882 [cs].

Cartwright, M., Seals, A., Salamon, J., Williams, A., Mikloska, S., Macconnell, D., et al. (2017). Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations. *Proceedings of the ACM on Human-Computer Interaction*, *29*.

Chakrabarty, S., & Habets, E. A. P. (2019). Multi-speaker DOA estimation using deep convolutional networks trained with noise signals. *IEEE Journal of Selected Topics in Signal Processing*, *13*, 8–21.

Chen, W.-P., Chen, S.-S., Lin, C.-C., Chen, Y.-Z., & Lin, W.-C. (2012). Automatic recognition of frog calls using a multi-stage average spectrum. *Computers & Mathematics with Applications*, *64*, 1270–1281.

Chen, X., & He, K. (2020). Exploring simple siamese representation learning. arXiv:2011.10566 [cs].

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., & Hinton, G. (2020). Big self-supervised models are strong semi-supervised learners. arXiv:2006.10029 [cs, stat].

Chen, H., Xie, W., Vedaldi, A., & Zisserman, A. (2020). Vggsound: A large-scale audio-visual dataset. (pp. 721–725).

Chen, G., Xie, W., & Zhao, Y. (2013). Wavelet-based denoising: A brief review. In *2013 Fourth international conference on intelligent control and information processing* (pp. 570–574).

Chesmore, E., & Nellenbach, C. (2001). Acoustic methods for the automated detection and identification of insects. *Acta Horticulturae*, 223–231.

Coban, E. B., Syed, A. R., Pir, D., & Mandel, M. I. (2021). Towards large scale ecoacoustic monitoring with small amounts of labeled data. In *2021 IEEE workshop on applications of signal processing to audio and acoustics* (pp. 181–185).

Coffey, K. R., Marx, R. G., & Neumaier, J. F. (2019). DeepSqueak: A deep learning-based system for detection and analysis of ultrasonic vocalizations. *Neuropsychopharmacology*, *44*, 859–868.

Colonna, J., Peet, T., Ferreira, C. A., Jorge, A. M., Gomes, E. F., & Gama, J. (2016). Automatic classification of anuran sounds using convolutional neural networks. In *Proceedings of the ninth international C* conference on computer science & software engineering*.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *28*, 357–366.

Denton, T., Wisdom, S., & Hershey, J. R. (2022). Improving bird classification with unsupervised sound separation. In *ICASSP 2022 - 2022 IEEE international conference on acoustics, speech and signal processing* (pp. 636–640).

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

Enari, H., Enari, H. S., Okuda, K., Maruyama, T., & Okuda, K. N. (2019). An evaluation of the efficiency of passive acoustic monitoring in detecting deer and primates in comparison with camera traps. *Ecological Indicators*, *98*, 753–762.

Esfahanian, M., Erdol, N., Gerstein, E., & Zhuang, H. (2017). Two-stage detection of north Atlantic right whale upcalls using local binary patterns and machine learning algorithms. *Applied Acoustics*, *120*, 158–166.

Farina, A., Eldridge, A., & Li, P. (2021). Ecoacoustics and multispecies semiosis: Naming, semantics, semiotic characteristics, and competencies. *Biosemiotics*, *14*, 141–165.

Farina, A., & Gage, S. H. (2017). Ecoacoustics: A new science. In *Ecoacoustics* (pp. 1–11). John Wiley & Sons, Ltd, chapter 1.

Fazekas, B., Schindler, A., Lidy, T., & Rauber, A. (2018). A multi-modal deep neural network approach to bird-song identification. arXiv:1811.04448 [cs, eess].

Fox, E. J., et al. (2008). Call-independent individual identification in birds. *Bioacoustics*, *18*, 51–67.

Frommolt, K.-H., & Tauchert, K.-H. (2014). Applying bioacoustic methods for long-term monitoring of a nocturnal wetland bird. *Ecological Informatics*, *21*, 4–12.

Gage, S. H., Towsey, M., & Kasten, E. P. (2017). Analytical methods in ecoacoustics. *Ecoacoustics*, 273–296.

Gasc, A., Sueur, J., Pavoine, S., Pellens, R., & Grandcolas, P. (2013). Biodiversity sampling using a global acoustic approach: Contrasting sites with microendemics in New Caledonia. *PLoS One*, *8*(5), 1–10.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., et al. (2017). Audio set: An ontology and human-labeled dataset for audio events. (pp. 776–780).

Gibb, R., Browning, E., Glover-Kapfer, P., & Jones, K. E. (2018). Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution*, *10*, 169–185.

Glotin, H., Ricard, J., & Balestriero, R. (2022). Fast chirplet transform injects priors in deep learning of animal calls and speech. openreview.net.

Gregory, R. D., Gibbons, D. W., & Donald, P. F. (2004). Bird census and survey techniques. *Bird Ecology and Conservation*, 17–56.

Grinfeder, E., Haupert, S., Ducrettet, M., Barlet, J., Reynet, M.-P., Sèbe, F., et al. (2022). Soundscape dynamics of a cold protected forest: Dominance of aircraft noise. *Landscape Ecology*, *37*(2), 567–582.

Grumiaux, P.-A., Kitić, S., Girin, L., & Guérin, A. (2022). A survey of sound source localization with deep learning methods. *Journal of the Acoustical Society of America*, *152*, 107–151.

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

Hao, Z., Zhan, H., Zhang, C., Pei, N., Sun, B., He, J., et al. (2022). Assessing the effect of human activities on biophony in urban forests using an automated acoustic scene classification model. *Ecological Indicators*, *144*, Article 109437.

Happel, R. E., & Happel, R. J. (2020). Soundscape ecology. In *Encyclopedia of the world's biomes*: *vol. 5*, (pp. 195–202).

Haver, S. M., Gedamke, J., Hatch, L. T., Dziak, R. P., Van Parijs, S., McKenna, M. F., et al. (2018). Monitoring long-term soundscape trends in U.S. waters: The NOAA/NPS ocean noise reference station network. *Marine Policy*, *90*, 6–13.

Heim, O., Heim, D. M., Marggraf, L., Voigt, C. C., Zhang, X., Luo, Y., et al. (2019). Variant maps for bat echolocation call identification algorithms. *Bioacoustics*, *29*, 557–571.

Huang, C.-J., Chen, Y.-J., Chen, H.-M., Jian, J.-J., Tseng, S.-C., Yang, Y.-J., et al. (2014). Intelligent feature extraction and classification of anuran vocalizations. *Applied Soft Computing*, *19*, 1–7.

Hussein, W., Hussein, M., & Becker, T. (2012). Spectrogram enhancement by edge detection approach applied to bioacoustics calls classification. *International Journal of Signal and Image Processing*, *3*, 1–20.

Joshi, K. A., Mulder, R. A., & Rowe, K. M. C. (2017). Comparing manual and automated species recognition in the detection of four common south-east Australian forest birds from digital field recordings. *Emu - Austral Ornithology*, *117*(3), 233–246.

Kahl, S., Denton, T., Klinck, H., Glotin, H., Goëau, H., Vellinga, W.-P., et al. (2021). Overview of BirdCLEF 2021: Bird call identification in soundscape recordings. In *Proceedings of the working notes of CLEF 2021*: *vol. 2936*, (p. 14). CEUR-WS.org.

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

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, *147*, 70–90.

Kampichler, C., Wieland, R., Calmé, S., Weissenberger, H., & Arriaga-Weiss, S. (2010). Classification in conservation biology: A comparison of five machine-learning methods. *Ecological Informatics*, *5*, 441–450.

Kershenbaum, A., Sayigh, L. S., & Janik, V. M. (2013). The encoding of individual identity in dolphin signature whistles: How much information is needed? *PLoS One*, *8*, Article e77671.

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

Kong, Q., Xu, Y., & Plumbley, M. D. (2017). Joint detection and classification convolutional neural network on weakly labelled bird audio detection. In *2017 25th European signal processing conference* (pp. 1749–1753).

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., et al. (2016). Visual genome: Connecting language and vision using crowdsourced dense image annotations. arXiv:1602.07332 [cs].

Kvsn, R. R., Montgomery, J., Garg, S., & Charleston, M. (2020). Bioacoustics data analysis – A taxonomy, survey and open challenges. *IEEE Access*, *8*, 57684–57708.

Lasseck, M. (2019). Audio-based bird species identification with deep convolutional neural networks. In *Proceedings of the working notes of CLEF 2021* (p. 11).

LeBien, J., Zhong, M., Campos-Cerqueira, M., Velev, J. P., Dodhia, R., Ferres, J. L., et al. (2020). A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. *Ecological Informatics*, *59*, Article 101113.

Lee, G., Gommers, R., Waselewski, F., Wohlfahrt, K., & O'Leary, A. (2019). PyWavelets: A Python package for wavelet analysis. *The Journal of Open Source Software*, *4*(36), 1237.

Lin, T.-H., Chou, L.-S., Akamatsu, T., Chan, H.-C., & Chen, C.-F. (2013). An automatic detection algorithm for extracting the representative frequency of cetacean tonal sounds. *Journal of the Acoustical Society of America*, *134*, 2477–2485.

Lin, T.-H., & Tsao, Y. (2018). Listening to the deep: Exploring marine soundscape variability by information retrieval techniques. In *2018 OCEANS - MTS/IEEE kobe techno-oceans* (pp. 1–6).

Lin, T., & Tsao, Y. (2019). Source separation in ecoacoustics: A roadmap towards versatile soundscape information retrieval. *Remote Sensing in Ecology and Conservation*, *6*, 236–247.

Linke, S., & Deretic, J. (2019). Ecoacoustics can detect ecosystem responses to environmental water allocations. *Freshwater Biology*, *65*, 133–141.

Liu, S., Mallol-Ragolta, A., Parada-Cabeleiro, E., Qian, K., Jing, X., Kathan, A., et al. (2022). Audio self-supervised learning: A survey. arXiv:2203.01205 [cs, eess].

McFee, B., et al. (2023). Librosa/librosa: 0.10.1.

Mcloughlin, M. P., Stewart, R., & 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*, Article 20190225.

Mesaros, A., Heittola, T., Benetos, E., Foster, P., Lagrange, M., Virtanen, T., et al. (2018). Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*, 379–393.

Mesaros, A., Heittola, T., & Virtanen, T. (2016). TUT database for acoustic scene classification and sound event detection. In *2016 24th European signal processing conference* (pp. 1128–1132).

Metcalf, O. C., Barlow, J., Devenish, C., Marsden, S., Berenguer, E., & Lees, A. C. (2020). Acoustic indices perform better when applied at ecologically meaningful time and frequency scales. *Methods in Ecology and Evolution*, *12*, 421–431.

Monson, B. B., Hunter, E. J., Lotto, A. J., & Story, B. H. (2014). The perceptual significance of high-frequency energy in the human voice. *Frontiers in Psychology*, *5*.

Morfi, V., Nolasco, I., Lostanlen, V., Singh, S., Strandburg-Peshkin, A., Gill, L., et al. (2021). Few-shot bioacoustic event detection: A new task at the DCASE 2021 challenge. In *Proceedings of the detection and classification of acoustic scenes and events 2021 workshop*. Detection and Classification of Acoustic Scenes and Events 2021 Workshop, DCASE2021 ; Conference date: 15-11-2021 Through 19-11-2021.

Muller, M., Ellis, D. P. W., Klapuri, A., & Richard, G. (2011). Signal processing for music analysis. *IEEE Journal of Selected Topics in Signal Processing*, *5*, 1088–1110.

Neal, L., Briggs, F., Raich, R., & Fern, X. Z. (2011). Time-frequency segmentation of bird song in noisy acoustic environments. In *2011 IEEE international conference on acoustics, speech and signal processing* (pp. 2012–2015).

Nettleton, D. F., Orriols-Puig, A., & Fornells, A. (2010). A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review*, *33*, 275–306.

Ozanich, E., Thode, A., Gerstoft, P., Freeman, L. A., & Freeman, S. (2021). Deep embedded clustering of coral reef bioacoustics. *Journal of the Acoustical Society of America*, *149*, 2587–2601.

Parent, G., & Eskenazi, M. (2011). Speaking to the crowd: Looking at past achievements in using crowdsourcing for speech and predicting future challenges. In *Interspeech 2011*.

Phillips, Y. F., Towsey, M., & 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.

Piczak, K. J. (2015). ESC. In *Proceedings of the 23rd ACM international conference on multimedia* (pp. 721–725).

Pieretti, N., Farina, A., & Morri, D. (2011). A new methodology to infer the singing activity of an avian community: The acoustic complexity index (ACI). *Ecological Indicators*, *11*(3), 868–873, Cited by: 350.

Pijanowski, B. C., Villanueva-Rivera, L. J., Dumyahn, S. L., Farina, A., Krause, B. L., Napoletano, B. M., et al. (2011). Soundscape ecology: The science of sound in the landscape. *Source: BioScience BioScience*, *61*, 203–216.

Politis, A., Mesaros, A., Adavanne, S., Heittola, T., & Virtanen, T. (2021). Overview and evaluation of sound event localization and detection in DCASE 2019. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 684–698.

Potvin, D. A., et al. (2023). Effects of temporal variations in ecotourist noise on an avian community: A case study from a UNESCO world heritage site. *Journal of Ecotourism*, 1–19.

Preatoni, D. G., et al. (2005). Identifying bats from time-expanded recordings of search calls: Comparing classification methods. *Journal of Wildlife Management*, *69*, 1601–1614.

Qaisar, S. M., Simatic, J., & Fesquet, L. (2017). High-level synthesis of an event-driven windowing process. (pp. 1–8).

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

Rama Rao, K., Garg, S., & Montgomery, J. (2018). Investigation of unsupervised models for biodiversity assessment. *AI 2018: Advances in Artificial Intelligence*, *11320*, 160–171.

Ren, Y., Johnson, M. T., & Tao, J. (2008). Perceptually motivated wavelet packet transform for bioacoustic signal enhancement. *Journal of the Acoustical Society of America*, *124*, 316–327.

Rendon, N., Rodríguez-Buritica, S., Sanchez-Giraldo, C., Daza, J. M., & Isaza, C. (2022). Automatic acoustic heterogeneity identification in transformed landscapes from Colombian tropical dry forests. *Ecological Indicators*, *140*, Article 109017.

Riede, K. (1993). Monitoring biodiversity: Analysis of Amazonian rainforest sounds. *Ambio*, *22*, 546–548.

Roe, P., Eichinski, P., Fuller, R. A., McDonald, P. G., Schwarzkopf, L., Towsey, M., et al. (2021). The Australian acoustic observatory. *Methods in Ecology and Evolution*, *12*, 1802–1808.

Ross, S. R. J., Friedman, N. R., Dudley, K. L., Yoshimura, M., Yoshida, T., & Economo, E. P. (2018). Listening to ecosystems: Data-rich acoustic monitoring through landscape-scale sensor networks. *Ecological Research*, *33*, 135–147.

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

Rowley, J., Callaghan, C., Cutajar, T., Portway, C., Potter, K., Mahony, S., et al. (2019). FrogID: Citizen scientists provide validated biodiversity data on frogs of Australia. *Herpetological Conservation and Biology*, *14*, 155–170.

Ryan, J., Cline, D., Dawe, C., McGill, P., Zhang, Y., Joseph, J., et al. (2016). New passive acoustic monitoring in Monterey bay national marine sanctuary. In *OCEANS 2016 MTS/IEEE Monterey* (pp. 1–8).

Saeed, A., Grangier, D., & Zeghidour, N. (2021). Contrastive learning of general-purpose audio representations. In *ICASSP 2021 - 2021 IEEE international conference on acoustics, speech and signal processing*.

Salamon, J., Bello, J. P., Farnsworth, A., Robbins, M., Keen, S., Klinck, H., et al. (2016). Towards the automatic classification of avian flight calls for bioacoustic monitoring. *PLoS One*, *11*, Article e0166866.

Salamon, J., Jacoby, C., & Bello, J. P. (2014). A dataset and taxonomy for urban sound research. In *Proceedings of the ACM international conference on multimedia* (pp. 1041–1044).

Sánchez-Giraldo, C., Bedoya, C. L., Morán-Vásquez, R. A., Isaza, C. V., & Daza, J. M. (2020). Ecoacoustics in the rain: Understanding acoustic indices under the most common geophonic source in tropical rainforests. *Remote Sensing in Ecology and Conservation*, *6*(3), 248–261.

Scarpelli, M. D. A., Liquet, B., Tucker, D., Fuller, S., & 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*.

Sethi, S. S., Bick, A., Ewers, R. M., Klinck, H., Ramesh, V., Tuanmu, M.-N., et al. (2023). Limits to the accurate and generalizable use of soundscapes to monitor biodiversity. *Nature Ecology & Evolution*, *7*(9), 1373–1378.

Shamir, L., Yerby, C., Simpson, R., von Benda-Beckmann, A. M., Tyack, P., Samarra, F., et al. (2014). Classification of large acoustic datasets using machine learning and crowdsourcing: Application to whale calls. *Journal of the Acoustical Society of America*, *135*, 953–962.

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

Stowell, D., & Sueur, J. (2020). Ecoacoustics: Acoustic sensing for biodiversity monitoring at scale. *Remote Sensing in Ecology and Conservation*, *6*.

Stowell, D., Wood, M. D., Pamuła, H., Stylianou, Y., & Glotin, H. (2018). Automatic acoustic detection of birds through deep learning: The first bird audio detection challenge. *Methods in Ecology and Evolution*, *10*, 368–380.

Sueur, J., Aubin, T., & Simonis, C. (2008). Seewave, a free modular tool for sound analysis and synthesis. *Bioacoustics*, *18*(2), 213–226.

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

Sueur, J., Farina, A., Gasc, A., Pieretti, N., & 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, J., & Llusia, D. (2018). Terrestrial passive acoustic monitoring: Review and perspectives. *BioScience*, *69*(1), 15–25.

Swamy, S., & K.V, R. (2013). An efficient speech recognition system. *Computer Science & Engineering: An International Journal*, *3*, 21–27.

Swiston, K. A., & Mennill, D. J. (2009). Comparison of manual and automated methods for identifying target sounds in audio recordings of pileated, pale-billed, and putative ivory-billed woodpeckers. *Journal of Field Ornithology*, *80*(1), 42–50.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). A survey on deep transfer learning. *Artificial Neural Networks and Machine Learning – ICANN 2018*, 270–279.

Thakur, A., & Rajan, P. (2016). Model-based unsupervised segmentation of birdcalls from field recordings. In *2016 10th international conference on signal processing and communication systems* (pp. 1–6).

Towsey, M. W., Truskinger, A. M., & Roe, P. (2015). The navigation and visualisation of environmental audio using zooming spectrograms. In *2015 IEEE international conference on data mining workshop* (pp. 788–797).

Towsey, M., Wimmer, J., Williamson, I., & Roe, P. (2014). The use of acoustic indices to determine avian species richness in audio-recordings of the environment. *Ecological Informatics*, *21*, 110–119, Ecological Acoustics.

Towsey, M., Zhang, L., Cottman-Fields, M., Wimmer, J., Zhang, J., & Roe, P. (2014). Visualization of long-duration acoustic recordings of the environment. *Procedia Computer Science*, *29*, 703–712, 2014 International Conference on Computational Science.

Towsey, M., Znidersic, E., Broken-Brow, J., Indraswari, K., Watson, D. M., Phillips, Y., et al. (2018). Long-duration, false-colour spectrograms for detecting species in large audio data-sets. *Journal of Ecoacoustics*, *2*, 1.

Trawicki, M., Johnson, M., & Osiejuk, T. (2005). Automatic song-type classification and speaker identification of norwegian ortolan bunting (emberiza hortulana) vocalizations. In *2005 IEEE workshop on machine learning for signal processing*.

Truskinger, A., Cottman-Fields, M., Eichinski, P., Towsey, M., & 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*.

Truskinger, A., Cottman-Fields, M., Johnson, D., & Roe, P. (2013). Rapid scanning of spectrograms for efficient identification of bioacoustic events in big data. In *2013 IEEE 9th international conference on e-science*.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., et al. (2016). WaveNet: A generative model for raw audio. CoRR abs/1609.03499.

Van Parijs, S. M., et al. (2015). NEPAN: A U.S. 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., & Roe, P. (2022). Key issues for realizing open ecoacoustic monitoring in Australia. *Frontiers in Ecology and Evolution*, *9*.

Virtanen, P., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*, 261–272.

Wilcove, D. S., Rothstein, D., Dubow, J., Phillips, A., & Losos, E. (1998). Quantifying threats to imperiled species in the United States. *BioScience*, *48*, 607–615.

Willacy, R. J., Mahony, M., & Newell, D. A. (2015). If a frog calls in the forest: Bioacoustic monitoring reveals the breeding phenology of the endangered Richmond Range mountain frog (Philoria richmondensis). *Austral Ecology*, *40*, 625–633.

Wu, Z., & Cao, Z. (2005). Improved MFCC-based feature for robust speaker identification. *Tsinghua Science and Technology*, *10*, 158–161.

Xia, X., Togneri, R., Sohel, F., Zhao, Y., & Huang, D. (2019). A survey: Neural network-based deep learning for acoustic event detection. *Circuits, Systems, and Signal Processing*, *38*, 3433–3453.

Xie, J., Colonna, J. G., & Zhang, J. (2020). Bioacoustic signal denoising: A review. *Artificial Intelligence Review*, *54*, 3575–3597.

Xie, J., Towsey, M., Zhang, J., & Roe, P. (2016). Adaptive frequency scaled wavelet packet decomposition for frog call classification. *Ecological Informatics*, *32*, 134–144.

Zhong, M., LeBien, J., Campos-Cerqueira, M., Dodhia, R., Lavista Ferres, J., Velev, J. P., et al. (2020). Multispecies bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling. *Applied Acoustics*, *166*, Article 107375.

Zilli, D., Parson, O., Merrett, G. V., & Rogers, A. (2014). A hidden Markov model-based acoustic cicada detector for crowdsourced smartphone biodiversity monitoring. *Journal of Artificial Intelligence Research*, *51*, 805–827.