birds

MDPI

Review

Bird Sound Classification Using Machine Learning: A Bibliometric and Methodological Review for Ecological Monitoring

Murugaiya Ramashini [1,*], Liyanage Chandratilak De Silva [1] and Pg Emeroylariffion Abas [2,*]

[1] School of Digital Science, University Brunei Darussalam, Jalan Tungku Link, Bandar Seri Begawan BE1410, Brunei
[2] Faculty of Integrated Technologies, University Brunei Darussalam, Jalan Tungku Link, Bandar Seri Begawan BE1410, Brunei
* Correspondence: murugaiya.ramashini@ubd.edu.bn (M.R.); emeroylariffion.abas@ubd.edu.bn (P.E.A.)

Simple Summary

Bird sounds can help researchers identify species and monitor changes in the natural environment. Devices that record sound over long periods can collect far more audio than people can listen to, creating a need for reliable automatic identification. This review examined 583 publications from 2000 to 2026 to understand how bird sound research has developed and which methods are used. The studies show a shift from methods that rely on researchers selecting sound characteristics to computer models that learn patterns from large collections of recordings. Research has also increasingly connected bird identification with wildlife monitoring. However, systems that perform well on carefully prepared recordings can struggle outdoors, where insects, rain, and other birds may obscure a call. Recordings of rare species are also often scarce. Improving how these systems handle field conditions and identify less common birds could help conservation workers monitor wildlife across larger areas and respond to environmental change.

Abstract

Bird vocalizations provide valuable information for biodiversity assessment and ecological monitoring, while passive acoustic monitoring has greatly increased the volume of recordings available for analysis. This review examines research trends in bird sound classification and synthesizes the methods used to identify species from audio recordings. A Scopus search identified 604 publications; after applying the English-language and 2000–2026 publication-year criteria, 583 were retained for bibliometric analysis using VOSviewer V1.6.21. Keyword co-occurrence analysis identified four principal themes: deep learning-based classification, bioacoustics and passive acoustic monitoring, recognition and classification methods, and benchmark datasets. Bibliographic coupling analysis and a temporal overlay examined relationships among publications and the development of the field. The review also synthesizes approaches to preprocessing, segmentation, feature representation, classification and evaluation. The findings show a shift from handcrafted acoustic features and classical classifiers towards deep learning models trained on larger bioacoustic datasets. Despite these advances, environmental noise, overlapping vocalizations, limited labeled data, and hardware constraints continue to hinder reliable field deployment. By connecting research trends with methodological developments and deployment challenges, this review identifies priorities for developing bird sound classification systems that can support practical biodiversity monitoring.

Check for updates

Academic Editors: Jie Xie, Juan Gabriel Colonna and Shufei Duan

Received: 17 September 2026
Revised: 12 November 2026
Accepted: 16 November 2026
Published: 24 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: bird sound classification; bird species recognition; avian bioacoustics; machine learning; deep learning; bibliometric analysis; passive acoustic monitoring; biodiversity monitoring

1. Introduction

Birds are widely used as indicators in ecological research because their presence, distribution, and activity can provide information about habitats and environmental change. Many species are readily detected through their vocalizations, making bird sounds a valuable source of data for biodiversity assessment and ecological monitoring. Acoustic surveys are particularly useful in locations that are difficult to access or where repeated field visits would require substantial effort [1]. Bird watching has traditionally been based on the skills and experience of field scientists, identifying species by sight and sound and providing context to their observations based on the surrounding environment [2–4].

Passive acoustic monitoring has expanded upon these skills by allowing the sounds produced by the birds to be recorded repetitively or continuously in the natural environment of the species. Recording devices can be used to record audio in multiple locations and for extended periods of time with minimal human effort. This audio could be used to study species occurrence, habitat use, migrations, and changes in bird behaviour [5–7]. On the other hand, this process is also capable of producing huge quantities of audio data that cannot be processed manually, requiring automated methods to detect bird sounds and identify the species present.

Consequently, the bird sound classification process has evolved to become a multidisciplinary research field that combines bioacoustics with signal processing and artificial intelligence (AI). Machine learning, as a subset of AI, refers to computational methods that learn relationships from training data and use them to classify or predict new observations without relying solely on explicitly programmed rules. On the other hand, Deep learning (DL), also a subset of ML, uses neural networks with multiple layers to learn increasingly complex representations, often directly from audio recordings or their time–frequency representations. Earlier approaches typically involved pre-processing of audio recordings followed by detection of vocalization and extraction of hand-crafted acoustic features to be used in statistical or conventional machine learning classifiers. More advanced approaches have involved the utilization of deep learning networks to learn representations directly from spectrograms and other audio inputs. The availability of open databases of recordings has supported these advancements, allowing researchers to evaluate multiple approaches on a wider range of species.

Advances in classification techniques have not removed the difficulties in deploying such algorithms on field recordings. The presence of background noise, overlap in vocalizations and different characteristics of recording devices may limit the practical applications of these techniques, especially when they are applied to audio recorded outside of curated datasets. Moreover, training data are, more often than not, unbalanced and insufficient, as there are too few examples for labeling for the vast majority of rare or geographically limited birds. Additionally, methods designed for continuous monitoring should work under resource constraints in terms of computing power, memory capacity, and energy consumption on field devices. Thus, evaluation of the methods' practicality requires not only investigation of their classification capabilities but also examination of the circumstances of use.

Despite the number of previous reviews dedicated to bird sound classification techniques, it is still necessary to provide an up-to-date review, especially considering the fast

development of deep learning, as well as the growing use of passive acoustic monitoring and the resulting development of large bioacoustic datasets. Consequently, this current review attempts to fill this gap by providing bibliometric analysis of publications together with a synthesis of technological development in bird sound classification. Keyword co-occurrence analysis is applied to identify key research themes, whilst bibliographic coupling network analysis is performed to analyze relationships among publications and their evolution. The methodological review comprises preprocessing, segmentation, feature representation, and classification steps, as well as the datasets and benchmarks utilised to design and assess methods developed in relation to bird sound classification. Additionally, the review discusses the key challenges that need to be addressed before automated bird sound classification can be applied reliably and effectively in ecological monitoring.

Previous reviews have provided important perspectives on the development of automated bioacoustics analysis. Stowell [8] reviewed the application of deep learning across computational bioacoustics, discussing neural network architectures, acoustic representations, data limitations, and research directions for analyzing animal vocalizations. This review expanded the scope of the use of deep learning in bioacoustics by covering other types of sounds in addition to birds. More recently, Maithripala et al. [9] have reviewed bird sound recognition using deep learning in connection with artificial intelligence advancements, concentrating on data issues, training procedures, deep learning models, and open problems. The authors pointed out the advantages of convolutional and recurrent neural networks for the purpose of classifying bird species, while acknowledging the significance of data, robustness, and monitoring systems.

In comparison to these works, the current review takes an additional perspective through integrating bibliometric analysis and methodological review devoted to the topic of bird sound classification. In addition to analyzing classification techniques, the study also focuses on the evolution of research themes, inter-relations between publications, and trends in the literature. At the same time, datasets, labeling and benchmarking practices, along with ecological application-related problems, are taken into consideration. This approach offers a wider view of the links between the research activity, methodological innovations, and ecological requirements while providing insights into possible areas for further studies.

Accordingly, this review addresses the following research questions:

- RQ1: What are the main research themes in bird sound classification, and how have they evolved over time?
- RQ2: How have bird sound classification methods developed from hand-crafted acoustic features and traditional machine learning approaches to deep learning architectures?
- RQ3: How do datasets, labeling practices, and benchmarking contribute to the development of bird sound classification systems?
- RQ4: What are the main challenges that restrict the deployment of automated bird sound classification systems in ecology, and what future research is needed to overcome these challenges?

2. Review Methodology

This review combined a structured literature search with bibliometric analysis and a methodological synthesis of bird sound classification research. The review process comprised three stages: retrieval and filtering of relevant publications, bibliometric analysis of the resulting dataset, and synthesis of the principal methods, datasets and research themes identified in the literature. The overall process is summarized in Figure 1.

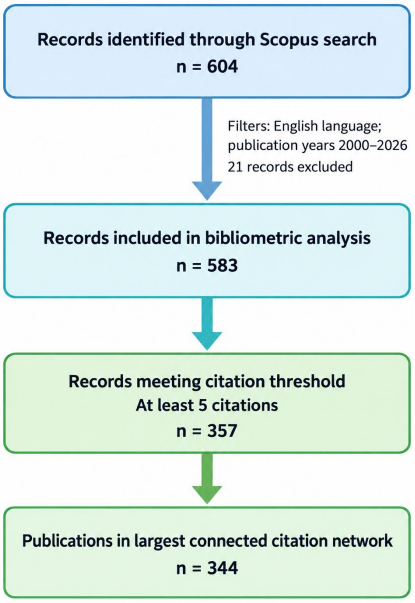

Figure 1. Literature retrieval and filtering process used in this review. Different colours distinguish the successive stages of literature selection and bibliometric analysis.

2.1. Information Source and Literature Search Strategy

The literature search was conducted using the Scopus database on 16 July 2026 and covered publications published between 1 January 2000 and 16 July 2026. Accordingly, the records for 2026 include only publications indexed in Scopus by the search date. Scopus was selected because of its multidisciplinary abstract and citation database covering computer science, engineering and ecological research, which are the principal disciplines represented in bird sound classification. Scopus indexes bibliographic records from numerous publishers and professional organizations, including many journals and conference proceedings available through ScienceDirect, SpringerLink and IEEE Xplore. Using a single indexing database also provided a consistently structured source of publication metadata for keyword co-occurrence, bibliographic coupling and temporal analyses. Nevertheless, the restriction to Scopus may have excluded relevant publications that were not indexed in the database.

The search strategy was designed to retrieve publications addressing bird or avian sounds in combination with classification, recognition, identification or detection tasks. Terms relating to machine learning, deep learning, bioacoustics and signal processing were included to cover both conventional and contemporary methodological approaches. Accordingly, the search terms were organized into three concept groups relating to the target acoustic domain, analytical task, and methodological or disciplinary context, as summarized in Table 1. Terms within each concept group were combined using the Boolean operator OR, while the three concept groups were combined using AND.

The search was conducted across the title, abstract, and keyword fields using the following search expression:

TITLE-ABS-KEY((“bird*” OR “avian”) W/2 (“sound” OR “song” OR “call” OR “vocal*” OR “acoustic*” OR “audio”)) AND TITLE-ABS-KEY(“classification” OR “recognition” OR “identification” OR “detection”) AND TITLE-ABS-KEY(“machine learning” OR “deep learning” OR “bioacoustic*” OR “signal processing”)

The W/2 proximity operator required the bird-related and sound-related terms to occur within two words of each other, irrespective of their order. The asterisk wildcard was used to capture variations of the relevant terms, such as “bird” and “birds” or terms

beginning with "vocal". The search returned 604 records before the application of the language and publication-year filters.

Table 1. Concept groups and search terms used in the Scopus search strategy.

Concept Group	Search Terms	Purpose
Target acoustic domain	("bird*" OR "avian") W/2 ("sound" OR "song" OR "call" OR "vocal*" OR "acoustic*" OR "audio")	Identifies publications concerning bird or avian sounds
Analytical task	"classification" OR "recognition" OR "identification" OR "detection"	Restricts the results to relevant analytical tasks
Methodological or disciplinary context	"machine learning" OR "deep learning" OR "bioacoustic*" OR "signal processing"	Identifies studies involving computational methods or bioacoustics

2.2. Eligibility Criteria and Dataset Preparation

The 604 records retrieved from Scopus were subsequently screened against predefined eligibility criteria to obtain the final dataset for bibliometric analysis. Publications were retained if they:

- were written in English;
- were published between 2000 and 2026; and
- were classified in Scopus as journal articles or conference papers.

Application of these criteria excluded 21 records that did not satisfy one or more of the eligibility criteria, leaving 583 publications in the final bibliometric dataset. As the search was conducted before the end of 2026, the records for that year represent publications indexed in Scopus only up to the search date. Bibliographic information for the 583 publications, including titles, authors, abstracts, author keywords, citation counts, and reference lists, was exported from Scopus in CSV format for analysis using VOSviewer V1.6.21. The literature retrieval and filtering process is presented in Figure 1.

Screening was conducted by a single researcher using the predefined eligibility criteria; therefore, independent duplicate screening and formal inter-reviewer agreement assessment were not performed.

2.3. Bibliometric Analysis

Bibliometric analysis was conducted using VOSviewer version 1.6.21 to examine the structure and development of research on bird sound classification. Three complementary analyses were performed: keyword co-occurrence analysis, bibliographic coupling, and overlay visualization.

Keyword co-occurrence analysis was used to identify the principal research topics represented in the 583 publications. Author keywords were analyzed according to how frequently they occurred together within the same publications. Keywords with strong co-occurrence relationships were then clustered to visualize related research topics and the connections among them.

Bibliographic coupling was used to examine relationships among publications based on the references they shared. Publications citing similar sources were considered more closely related in terms of their research topics or methodological foundations. A minimum threshold of five citations per publication was applied. Of the 583 publications in

the bibliometric dataset, 357 met this threshold, and 344 formed the largest connected bibliographic-coupling network used for the subsequent analysis.

Overlay visualization was used to examine the temporal development of research topics. In this visualization, keywords were coloured according to their average year of publication, allowing established topics to be distinguished from more recent areas of research. In the network visualizations, nodes represented either keywords or publications, depending on the analysis. Node size represented the frequency of keyword occurrence or the citation count of a publication, while links represented co-occurrence or bibliographic-coupling relationships. Colours distinguished clusters of closely related keywords or publications. The resulting networks were interpreted according to the prominence, clustering and interconnections of the represented topics and publications.

The five-citation threshold was applied only to the bibliographic coupling analysis and its temporal overlay and was not used to restrict the literature considered in the methodological synthesis presented in Sections 4–6. The broader synthesis considered the full eligible dataset, including recent publications that had not yet accumulated five citations. The threshold was introduced to improve the interpretability of the network visualization. Nevertheless, because citation counts accumulate over time, recent publications may be underrepresented in the bibliographic coupling network.

3. Bibliometric Results and Analysis

Bibliometric analysis revealed the principal research themes, relationships among publications, and changes in research interests over time. These findings were examined through keyword co-occurrence analysis, bibliographic coupling, and temporal analysis of keywords, respectively.

3.1. Keyword Co-Occurrence Analysis and Research Themes

The keyword co-occurrence network generated from the 583 publications is presented in Figure 2. Larger nodes represent keywords occurring more frequently in the literature, while links indicate relationships between keywords appearing in the same publications. The colour-coded groups represent clusters of closely related research topics. Four principal thematic areas were identified: deep learning-based classification; bioacoustics and passive acoustic monitoring; bird sound recognition and classification methods; and benchmark datasets and evaluation frameworks.

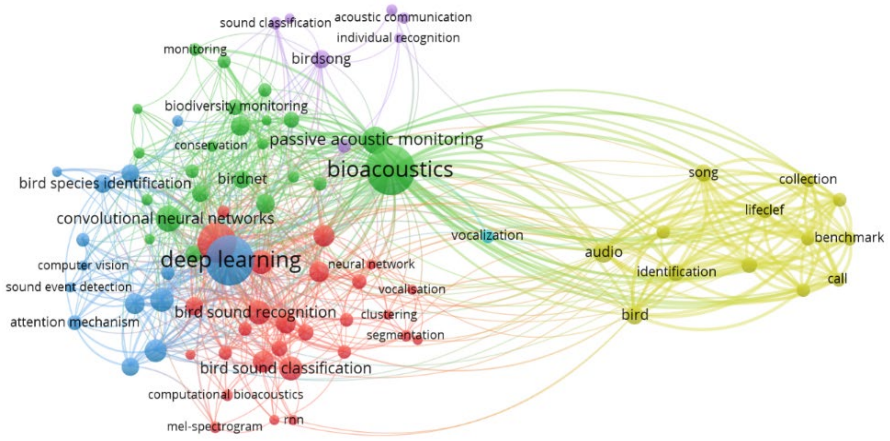

Figure 2. Keyword co-occurrence network of bird sound classification research generated using VOSviewer V1.6.21. Node size represents keyword frequency, and different colours distinguish clusters of closely related research keywords.

To complement the network visualization, Table 2 summarizes the representative keywords and principal focus of each thematic cluster identified in Figure 2.

Table 2. Summary of the principal research themes identified through keyword co-occurrence analysis.

Research Theme	Representative Keywords	Principal Focus
Deep learning-based classification	Deep learning, convolutional neural networks, attention, computer vision, sound event detection and BirdNET	Data-driven learning of acoustic representations and automated bird sound classification
Bioacoustics and passive acoustic monitoring	Bioacoustics, passive acoustic monitoring, biodiversity monitoring, conservation and BirdNET	Application of automated acoustic analysis to ecological monitoring and conservation
Recognition and classification methods	Bird sound classification, Mel-spectrograms, segmentation and clustering	Signal processing, acoustic representation and species recognition methods
Benchmark datasets and evaluation frameworks	LifeCLEF, benchmark, dataset, collection and identification	Shared datasets, evaluation protocols and comparison of classification methods

3.1.1. Deep Learning–Based Classification

Deep learning formed the largest cluster in the keyword co-occurrence network, highlighting its central role in the development of automated bird sound classification. Keywords such as convolutional neural networks (CNNs), attention, computer vision, and sound event detection appeared in close association, suggesting that these topics are often investigated together. The prominence of this cluster also reflects the increasing use of data-driven approaches alongside those based on manually designed acoustic features.

This shift is driven by the ability of deep neural networks to learn discriminative representations directly from spectrograms and other time–frequency representations, reducing the need for manually crafted acoustic descriptors. CNNs have become one of the most widely used architectures for bird sound classification due to their effectiveness in capturing local spectral and temporal patterns that characterize species-specific vocalizations. More recently, researchers have extended these models by incorporating attention mechanisms, transfer learning, and other sophisticated architectures.

The presence of BirdNET in this cluster demonstrates the application of deep learning to bioacoustics research. By combining large-scale annotated datasets with deep neural networks, BirdNET has shown that automated bird species recognition can achieve high accuracy and support ecological monitoring [10]. Its presence in the keyword network also reflects the broader shift towards large-scale, data-driven approaches to biodiversity monitoring.

3.1.2. Bioacoustics and Passive Acoustic Monitoring

The bioacoustics and passive acoustic monitoring cluster indicates a growing interaction between artificial intelligence, ecological monitoring, and conservation research. The close relationship among keywords such as bioacoustics, passive acoustic monitoring, biodiversity monitoring, conservation, and BirdNET suggests that, in addition to improving algorithms, researchers are increasingly examining their applications in ecological monitoring and conservation.

This growing emphasis is closely linked to the widespread adoption of autonomous recording units (ARUs), which enable continuous acoustic data collection across large

geographical areas with minimal human intervention. Compared with conventional field surveys, passive acoustic monitoring provides a more scalable and less invasive approach for tracking bird populations, seasonal migration, habitat use, and changes in biodiversity over extended periods. These advantages have established passive acoustic monitoring as an important component of modern ecological research.

The growing use of passive acoustic monitoring has also created a need for automated techniques to process the large volumes of recordings collected through long-term monitoring programmes. Machine learning, especially deep learning, is thus an important part of transforming acoustic recordings into usable ecological data. The strong connections observed within this bibliometric cluster reflect the increasing convergence of bioacoustics, ecology, and artificial intelligence.

3.1.3. Recognition and Classification Methods

Another cluster identified in the keyword co-occurrence network focuses on bird sound recognition and classification methods, with terms such as bird sound classification, mel-spectrograms, segmentation, and clustering appearing closely related. These connections highlight the continued importance of signal processing and pattern recognition techniques for transforming raw audio into meaningful representations that can be used for species classification.

Studies associated with this cluster reflect the evolution of methods in this field. Early studies mainly combined manually extracted acoustic features from recordings of bird vocalizations with traditional statistical and machine learning methods. In this way, researchers laid the foundations for recognizing different bird species based on their sounds with an acceptable level of accuracy.

With further development of the field, researchers have moved to more advanced techniques focused on data-driven feature extraction from audio or spectrogram representations. Nevertheless, keywords related to segmentation and feature extraction remain prominent in the bibliometric network; this shows that both continue to be important topics in bird sound classification research.

3.1.4. Benchmark Datasets and Evaluation Frameworks

Benchmark datasets and evaluation frameworks constitute another cluster of research topics identified through the bibliometric analysis. Keywords such as LifeCLEF, benchmark, dataset, collection, and identification highlight the increasing importance of shared evaluation platforms for testing bird sound classification methods.

Large, publicly accessible datasets have also become essential for the development and evaluation of machine learning approaches to bird sound classification. Standardized datasets used for model training and evaluation make it easier to compare different classification approaches and to assess how well models generalize across datasets. They also help improve the reproducibility of research findings.

The LifeCLEF challenges are among the initiatives that not only provide annotated datasets but also establish evaluation criteria. In addition to benchmark classification performance, these challenges encourage the development of new features, classification methods, and species identification techniques. The prominence of this cluster in the bibliometric network reflects the significant contribution of benchmark datasets to methodological progress in bird sound classification research.

3.2. Bibliographic Coupling Analysis and Research Clusters

Bibliographic coupling analysis was performed to examine relationships among publications on bird sound classification and the intellectual structure of the field. In contrast

to keyword co-occurrence analysis, which identifies relationships among research topics, bibliographic coupling links publications that cite common references. Only publications with at least five citations were considered in this analysis. Of the 583 publications in the dataset, 357 met this threshold, and 344 formed the largest connected network shown in Figure 3.

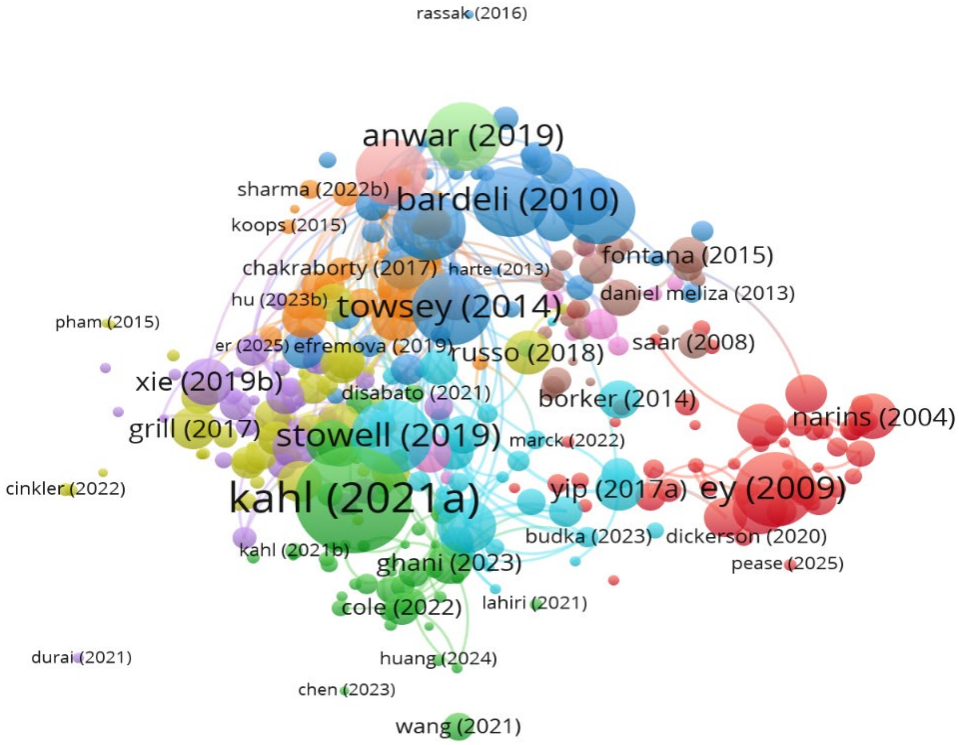

Figure 3. Bibliographic coupling network of publications on bird sound classification generated using VOSviewer V1.6.21. Links indicate shared cited references, node size represents the citation count of each publication, and colours represent clusters of related studies.

Several key publications connecting research directions can be identified in the bibliographic coupling network [11–14]. An early study developed methods for detecting the vocalizations of two endangered bird species in complex acoustic environments and applied them to habitat mapping [11]. Another compared manual surveys with autonomous recording and automated call detection, demonstrating the potential efficiency of acoustic monitoring while identifying differences in detection sensitivity and weather-related bias [12]. A later study evaluated methods submitted to the first Bird Audio Detection challenge, showing how machine learning could detect bird sounds across varied recordings [13]. More recently, BirdNET was developed to identify hundreds of bird species by sound and was evaluated using both individual recordings and field soundscapes [14]. The presence and connectivity of these publications in the network indicate their influence on research into automated bird sound analysis, passive acoustic monitoring and bioacoustics.

In addition, the bibliographic coupling network demonstrates the gradual diversification of research themes. Earlier work focused on acoustic signal processing and the detection of particular vocalizations [11], while later studies incorporated deep learning approaches and larger annotated datasets [13,14]. The importance of evaluating automated methods against field surveys was also established [12] and remains relevant as classification systems become more sophisticated. These publications illustrate the cumulative development of bird sound classification research, in which advances in signal processing, machine learning and bioacoustics have increasingly been applied to ecological monitoring.

3.3. Emerging Research Directions

The temporal overlay of the bibliographic coupling network in Figure 4 provides another perspective on the development of bird sound classification research. By colouring publications according to their year of publication, the visualization shows how earlier and more recent studies are positioned within the network and how they are connected.

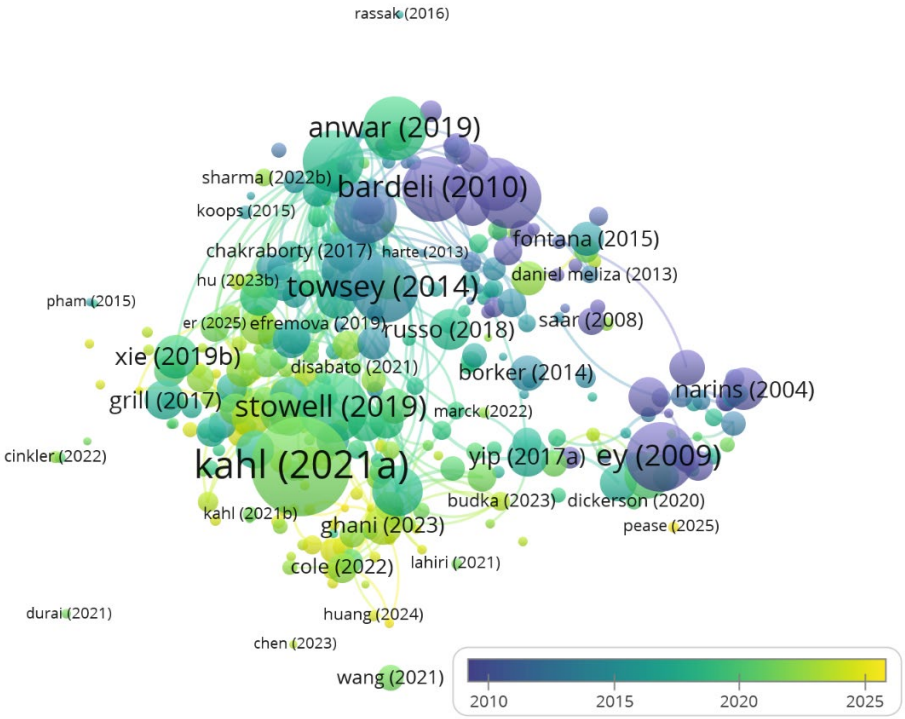

Figure 4. Temporal overlay of the bibliographic coupling network for bird sound classification research generated using VOSviewer V1.6.21. Nodes represent publications, links indicate bibliographic coupling relationships based on shared cited references, and node colours indicate publication year, from earlier studies in purple to more recent studies in yellow.

Earlier publications include work on detecting bird vocalizations in complex acoustic environments [11] and comparing autonomous acoustic monitoring with manual field surveys [12]. More recent publications examine the use of deep learning for general bird sound detection [13] and the identification of large numbers of species in recordings collected for ecological monitoring [14]. Their connections within the network illustrate how subsequent research builds on earlier work in signal processing and acoustic monitoring.

The overlay therefore suggests a broadening of research from methods developed for particular detection and monitoring tasks towards systems designed to analyze larger and more diverse acoustic datasets. The publications highlighted in Figure 4 provide examples of this development; the figure alone does not establish which specific topics will become future research directions.

4. Methodological Synthesis of Bird Sound Classification

Figure 5 shows the general pipeline used by automated systems for bird sound classification. This section will consider the stages of this process, including pre-processing, segmentation, feature extraction, and classification.

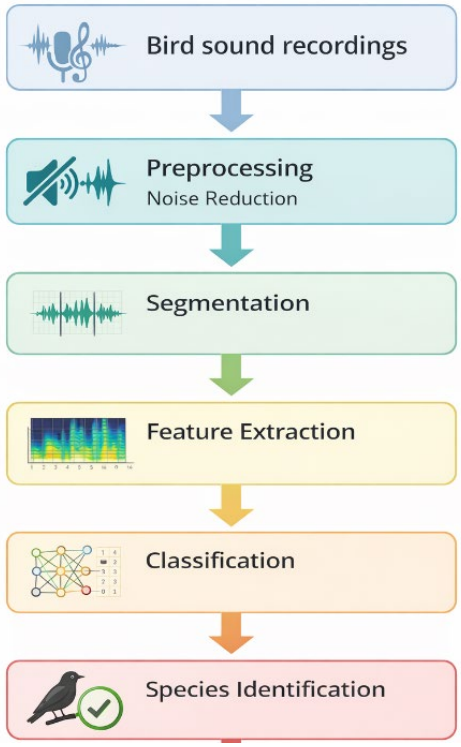

Figure 5. General pipeline of automated bird sound classification systems.

4.1. Preprocessing and Noise Reduction

Collecting good bird sound recordings is one of the important elements of creating an automated classification system for bird sounds. Some earlier research collected their data in controlled environments where sounds were recorded from caged birds. In these types of environments, microphones were installed either near the cages or in sound-attenuation rooms to record good-quality sounds of different types, such as songs and calls. This helped in making the first datasets for the development of classification systems [15,16].

Bird vocalizations are also recorded in natural environments using devices such as conventional microphones, circular microphone arrays, and autonomous recording units (ARUs). These approaches enable long-term observation of bird populations in their natural habitats [5,11,17]. ARUs have become particularly important in ecological research because they can continuously collect large volumes of environmental audio and record multiple species at the same time. More specialized approaches have also been investigated, including lightweight sensing systems and microphones attached to birds, with these methods allowing researchers to study vocal behaviour alongside movement patterns and environmental conditions [18].

Thanks to improvements in modern recording techniques, acquiring vast amounts of data for studying bird sounds has become cheaper and easier. This has led to the development of some bird sound databases and audio libraries for conducting research on bird sound analysis and classification [5,19]. Popular datasets include those that are freely accessible online, such as the Xeno-Canto database and the Cornell Macaulay Library. These are rich in recordings of various bird species from different geographical locations. Subsequently, many recent studies have used these online datasets for their training and testing.

Nonetheless, bird recordings from field devices and online resources can contain substantial environmental noise and overlapping acoustic events, including sounds from insects, wind, rain, and other animals. Bird vocalizations may therefore be only partially audible or embedded within complex soundscapes [11,20]. Modern large-scale datasets such as BirdSet, which contain recordings of thousands of species, highlight both the

strengths and challenges of large bioacoustics collections. In these recordings, bird songs and calls may occur alongside background noise. This often makes effective signal preprocessing and segmentation important before feature extraction and classification.

Noise reduction is a crucial preprocessing step in bird sound classification, especially for recordings gathered from natural environments containing wind, insects, rain, and other animal noise. Different signal-filtering methods have been used to reduce noise before feature extraction, with the choice of technique usually depending on the characteristics of the bird vocalizations [3,21].

Early approaches used frequency-based filtering methods, namely high-pass and low-pass filters, to reduce noise in bird recordings. High-pass filters are generally used to suppress low-frequency noise, whereas low-pass filters can remove high-frequency disturbances outside the frequency range of the vocalizations being studied [11,17]. Some researchers limited their analyses to frequency bands containing most of the energy of the target vocalizations, thereby reducing the influence of background sounds.

Apart from filtering, additional preprocessing methods have been proposed to improve robustness when working with noisy signals. In particular, whitening filters can be used to reduce components of background noise [22,23], and wavelet methods can be used to improve signal quality through wavelet packet decomposition combined with band-pass or low-pass filtering [24,25]. Other studies have considered methods such as Wiener filtering, spectrogram-based noise cancellation, and correlation filter banks [26,27]. More recent work has used spectral subtraction to improve noise robustness [28].

Despite the importance of noise reduction in preprocessing, some recordings in public repositories have already undergone selection or curation. Consequently, some studies simply omit a separate noise-reduction step and instead rely on the robustness of their feature extraction or classification methods [17,29]. For recordings collected directly from bird habitats, however, noise reduction may remain an important part of processing. Following preprocessing, continuous recordings can be segmented into acoustic units for further analysis.

4.2. Segmentation of Bird Vocalizations

After the preprocessing, individual bird vocalization events may be separated from a continuous recording. Segmentation is particularly useful for classification because recordings from natural environments can contain multiple calls, multiple species, and intervals dominated by background noise. Separating vocal events can facilitate subsequent feature extraction and improve the efficiency of an automated classification system.

Segmentation methods have evolved alongside bird sound classification methods. In earlier studies, vocal units were usually identified manually through visual inspection of waveforms or spectrograms [30]. More recent research has used automated algorithms capable of analyzing large acoustic datasets. Bird songs and calls often contain distinct vocal units, such as syllables, that can serve as target segments [30]. Detecting these units can support subsequent feature extraction and classification.

Recent advances in deep learning have also shown that explicit segmentation of individual vocal units is not always required. Models can classify audio segments using their spectrograms without first separating individual syllables.

4.2.1. Manual Segmentation

In earlier studies, manual segmentation was widely used during preprocessing. Researchers visually inspected waveforms and spectrograms to identify the beginning and end of individual syllables based on the characteristics of bird vocalizations [17,31]. This generally involved examining the signal in the time domain and selecting intervals contain-

ing vocal events [20,32]. While manual segmentation can be effective for small datasets, it is labor-intensive and difficult to apply to large-scale datasets. It is also unsuitable for real-time applications such as automated bird monitoring and systems designed to reduce aircraft-bird collisions.

4.2.2. Energy-Based Time-Domain Segmentation

Energy-based time-domain techniques have been widely employed for automatic segmentation. The input signal is split into small frames, and the energy level of each frame is computed. Frames whose energy exceeds a predefined threshold are considered to be probable vocalization frames and classified as syllable segments [19,33]. Several variations of the technique have been reported that employ adaptive energy thresholds as well as energy- and zero-crossing point-based endpoint detection [34,35]. While it is an easy and efficient algorithm, energy-based segmentation can be quite sensitive to noise interference and hence may result in false detections, especially in complex acoustic environments.

4.2.3. Time–Frequency Segmentation Methods

Time–frequency methods analyze spectrograms, allowing temporal and frequency characteristics to be considered together. Examples include the use of a random forest classifier for segmentation [36] and two-dimensional spectrogram analysis for syllable extraction [23]. Other approaches employ frequency-band thresholding or sinusoidal modelling to identify vocal elements in bird recordings [37,38]. Time–frequency methods can offer greater robustness to noise than purely time-domain approaches, although they may introduce additional computational complexity.

4.2.4. Model-Based and Specialized Segmentation Approaches

More specialized segmentation methods have been proposed for complex acoustic environments. These include spectrogram-based morphological filtering to detect bird activity [39], sinusoidal detection methods to identify tonal components [40] and unsupervized clustering techniques, such as k-means, to identify acoustic units automatically [41]. Speaker diarization methods have also been explored in biodiversity-monitoring applications to divide long recordings containing several sound sources into meaningful segments [42].

Some studies have used segmentation tools integrated into bioacoustic analysis software. For example, template-matching functions in Avisoft SASLab Pro have been applied to detect particular bird calls within long recordings [11,43]. Signal-envelope tracking methods, including the Hilbert follower, have also been used to identify syllable boundaries from characteristic vocalization patterns [5].

4.2.5. Segmentation-Free Approaches

Although segmentation is commonly used, separating individual vocalizations is not always necessary. Several studies have successfully classified bird species from longer audio segments without first identifying syllable boundaries, particularly when machine learning models are applied directly to spectrogram representations [44,45]. This approach can reduce preprocessing requirements and may be especially useful for large datasets and real-time monitoring applications. Whether or not individual vocalizations are segmented, the following stage usually involves extracting or learning acoustic features that represent their characteristics.

4.3. Feature Representation for Bird Sound Classification

After preprocessing and, where necessary, segmentation, the acoustic content of the recordings must be represented in a form suitable for species classification. This is particularly important in long-term monitoring, where large amounts of audio may contain both informative vocalizations and irrelevant or noisy signals. An effective representation should preserve features that help distinguish species while reducing the influence of background noise and recording conditions.

Earlier studies mainly relied on manually designed acoustic descriptors, whereas more recent approaches increasingly use deep learning to learn representations directly from spectrograms and other time–frequency inputs. Handcrafted descriptors nevertheless remain useful, especially when training data or computational resources are limited. The following subsections review the temporal, spectral, cepstral, spectrogram-based and other specialized representations used in bird sound classification.

4.3.1. Temporal Features

Temporal characteristics provide some of the earliest and most intuitive information that can be extracted from bird vocalizations. Features describing the timing, duration and amplitude of acoustic events have long been used to characterize bird songs and calls because many species exhibit distinctive rhythmic patterns. Although temporal descriptors alone may not always provide sufficient discrimination between acoustically similar species, they remain valuable when combined with complementary spectral or cepstral information, particularly in lightweight classification systems.

Temporal features are extracted directly from the time-domain waveform and capture the amplitude variations and temporal structure of bird vocalizations. These features are among the earliest descriptors used in audio signal analysis and remain useful for identifying rhythmic and temporal patterns in bird calls [30,46].

Common temporal descriptors include measures of signal energy, zero-crossing rate, autocorrelation, and syllable duration. These features describe characteristics such as intensity, periodicity, and temporal variation within the audio signal. Although temporal features are computationally simple and efficient, they can be sensitive to background noise and are therefore often combined with frequency-domain features for improved classification performance. The temporal features used in bird sound classification studies are summarized in Table 3.

Table 3. Common temporal features used in bird sound classification studies.

Feature	Description	Representative Studies
Zero Crossing Rate (ZCR)	Measures the rate of sign changes in the waveform	[47,48]
Signal Energy	Measures the intensity of the signal	[17,49,50]
Entropy of Energy	Represents energy distribution across frames	[29]
Autocorrelation	Detects periodic structure	[11,22]
Syllable Duration	Temporal length of vocal units	[47,51]
Temporal Statistics	Mean, variance, skewness, kurtosis	[23]

4.3.2. Spectral Features

Spectral features describe how energy is distributed across the frequency components of bird vocalizations. Because many bird species produce distinctive frequency patterns, these features are widely used to capture species-specific acoustic characteristics. Common

examples include spectral centroid, spectral bandwidth, spectral roll-off, and spectral flux, which describe different aspects of the shape and spread of the frequency spectrum. These features are especially useful when species differ mainly in the frequency structure of their vocalizations. More recent studies have also examined multiscale spectral representations, including log-wavelet, log-Mel and log-power spectrograms, to capture spectral information at different frequency resolutions [52]. Table 4 summarizes common spectral features and their applications in bird sound classification.

Table 4. Common spectral features used in bird sound classification studies.

Feature	Description	Representative Studies
Spectral Centroid	Indicates center of spectral energy	[47,53]
Spectral Bandwidth	Measures spread of frequency spectrum	[23,54]
Spectral Roll-off	Frequency below which energy threshold is reached	[30,55]
Spectral Flux	Measures spectral change between frames	[48]
Spectral Entropy	Indicates spectral randomness	[56,57]
Spectral Flatness	Distinguishes tonal vs. noise signals	[53]

4.3.3. Cepstral Features

Cepstral features are among the most widely used acoustic representations for bird sound classification. These features capture the spectral envelope of audio signals and are derived from their frequency-domain representations.

Mel-Frequency Cepstral Coefficients (MFCCs) are the most commonly used cepstral features for bird sound analysis. MFCCs use a frequency scale based on human auditory perception and provide a compact representation of the spectral envelope [17,19]. Variants such as Linear Predictive Cepstral Coefficients (LPCC), Perceptual Linear Prediction Cepstral Coefficients (PLP), and Gammatone Frequency Cepstral Coefficients (GTCC) have also been used to capture different spectral characteristics. Recent studies have explored combinations of cepstral features with deep learning architectures to improve classification performance. For example, MFCC features have been used as inputs to convolutional neural networks or combined with other descriptors, such as log-Mel filter banks or Power-Normalized Cepstral Coefficients (PNCCs). Table 5 summarizes common cepstral features and their applications in bird sound classification.

Table 5. Common cepstral features used in bird sound classification studies.

Feature	Description	Representative Studies
MFCC	Mel-scale cepstral representation	[17,58]
LPCC	Linear prediction cepstral coefficients	[20]
PLP-CC	Perceptual linear prediction cepstral coefficients	[5]
GTCC	Gammatone cepstral coefficients	[59]
PNCC	Power-normalized cepstral coefficients	[60]

4.3.4. Spectrogram-Based Features

Spectrogram-based representations are now widely used in bird sound classification, particularly in deep learning applications. By showing how the frequency content of a signal changes over time, spectrograms provide a two-dimensional representation that can be processed as an image by convolutional neural networks. Common forms include Mel, harmonic, and log-Mel spectrograms. These representations enable models to learn distinctive acoustic patterns directly from the recorded audio [61,62]. More recently, architectures

such as EfficientNet and ConvNeXt have been applied to spectrogram images generated from large bird sound datasets [63,64].

4.3.5. Other Acoustic Descriptors and Feature Learning

In addition to the feature representations described above, several specialized acoustic descriptors have been developed for bird sound classification. These include pitch-based features, harmonic ratios, tonal descriptors and frequency modulation patterns, all of which capture particular characteristics of bird vocalizations [65].

Feature extraction tools such as openSMILE, MARSYAS and Sound Ruler have also been used to generate large sets of acoustic descriptors [3,66]. Dimensionality reduction methods, including Principal Component Analysis (PCA), are often applied before classification to reduce the number of features [58]. More recent studies have also explored unsupervized feature-learning approaches, in which models learn acoustic representations directly from large datasets without relying on manually defined feature descriptors [67].

4.4. Classification Methods for Bird Sound Recognition

Once feature extraction is done, the obtained features are then classified so that a bird species is assigned to each audio sample. The classification techniques employed in the recognition of bird sounds have evolved from template-based and statistical techniques to machine learning and deep learning techniques.

4.4.1. Template-Based and Sequential Models

Early sound classification techniques utilized template matching approaches. Among the popular algorithms is Dynamic Time Warping (DTW), which compares sequences of acoustic parameters to gauge the similarity of bird calls [16,20]. While DTW is capable of handling the variation in time between sequences, it is likely to perform poorly in cases where there is significant noise within the recording or when there are considerable variations in the acoustic structure of the call.

Another technique that has been utilized for the recognition of bird sounds is Hidden Markov Models (HMMs). HMMs model sequences of acoustic observations and can represent the temporal structure of bird song syllables [5,17].

4.4.2. Classical Machine Learning Classifiers

Conventional machine learning methods generally classify bird sounds using acoustic features that are extracted from the recordings before model training. Different classical algorithms for machine learning have been tested on bird sound classification. SVM is one of them, which is widely popular due to its capability to work with high-dimensional vectors and use non-linear decision boundaries via a kernel function. It has shown good results in bird species recognition [47,68].

There are some other algorithms like k-Nearest Neighbor, Decision Trees, Naïve Bayes classifiers, and Random Forest. Such classifiers can give good performance while using appropriate acoustic features. Also, ensemble algorithms like Random Forest can be applied to bird sound classification [36,44].

4.4.3. Deep Learning Methods

In recent years, deep learning techniques have become increasingly prominent in bird sound classification research. Deep learning methods use multilayer neural networks to learn acoustic representations and classification patterns directly from spectrograms or other audio inputs, reducing the need for separately designed features. Convolutional Neural Networks (CNNs) can learn hierarchical representations from spectrogram images. Several

studies have applied CNN-based models to bird sound recognition using spectrograms as inputs [58,61]. These models typically use convolutional and pooling layers to capture patterns in time–frequency representations.

Deep learning methods have also been evaluated in large-scale competitions such as BirdCLEF, which has encouraged the development of bird sound recognition systems [13]. Hybrid architectures that combine CNNs with recurrent networks, such as Long Short-Term Memory (LSTM) models, have been introduced to capture both spectral and temporal information from audio signals. More recent studies have explored architectures such as EfficientNet, ConvNeXt, and transformer-based models trained on large bioacoustic datasets [64,69]. These models have shown strong classification performance when applied to large datasets containing recordings from multiple bird species.

4.4.4. Other Classification Strategies

Among the other techniques considered for identification of bird sounds are the Gaussian Mixture Model (GMM), which models probability distributions of the acoustic features, and Vector Quantization and self-organizing maps, which cluster the acoustic features. Among distance-based methods used are the Euclidean or Mahalanobis distance, and among template matching techniques, acoustic feature matching is common.

4.4.5. Comparative Studies

Several studies have compared classification strategies to identify suitable approaches for automated bird identification. These comparisons have examined classifiers such as SVMs, neural networks, k-NN, and decision trees using the same sets of acoustic features.

The results show that model performance depends strongly on the feature representation and the characteristics of the dataset. Classical machine learning classifiers, particularly SVMs, have demonstrated good performance in bird sound recognition tasks. Deep learning models have also achieved strong results, especially when large datasets are available for training. Table 6 summarizes classification methods used in bird sound recognition research.

The comparison in Table 6 highlights that the suitability of bird sound classification methods depends on dataset size, acoustic complexity, computational resources, and monitoring objectives. Traditional feature-based classifiers remain useful for smaller datasets and applications requiring relatively low computational cost, whereas CNN-based approaches offer greater flexibility in learning complex acoustic representations when sufficient training data are available. Transformer-based and self-supervized approaches can use large collections of unlabeled recordings to learn transferable representations, although their pretraining and computational requirements may limit deployment on resource-constrained devices. Transfer learning and weakly supervized approaches provide additional opportunities for addressing limited annotations, although their effectiveness may depend on the similarity between training and deployment conditions and the quality and precision of the available labels.

Several review papers have covered bird sound classification methods. The review in reference [8] provides an overview of deep learning applications in bioacoustics, while reference [70] examines the use of deep learning to recognize bird vocalizations. A more recent review [9] discusses machine learning methods for identifying bird species through their vocalizations.

Table 6. Comparative summary of bird sound recognition methods, learning strategies, and their practical applications.

Method	Input Data and Task	Advantages	Limitations	Suitable Use Cases	Rep. Studies
Dynamic Time Warping (DTW)	LPCC sequences; vocalization matching	Aligns temporally varying acoustic sequences	Reduced recognition performance in noisy conditions	Template-based bird-call matching	[20]
Hidden Markov Model (HMM)	MFCC/PLP-CC sequences; species recognition and detection	Models temporal patterns and variable-length vocalizations	Sensitive to recording quality and model configuration	Species recognition and call detection	[5,17]
Support Vector Machine (SVM)	MFCC, GTCC and entropy-based features; species classification	Effective with informative handcrafted features	Depends on feature quality and parameter tuning	Feature-based bird species identification	[47,71]
K-Nearest Neighbor (k-NN)	Syllable-pair histograms; species recognition	Simple classification using fixed-dimensional song features	Limited species coverage in the evaluated study	Syllable-based song recognition	[33]
Tree-based ensemble learning	Engineered acoustic features; species classification	Feature selection and ensemble predictions	Depends on feature extraction and model design	Large-scale feature-based classification	[56]
Random Forest	Time–frequency features; bird-song segmentation	Effective segmentation in noisy recordings	Evaluated for segmentation, not species classification	Bird-call preprocessing and segmentation	[36]
Artificial Neural Networks (ANN/AR-TDNN)	LPC and LDA-transformed features; classification	Learns nonlinear patterns and temporal context	Computational and training requirements vary	Feature-based and sequential vocalization recognition	[72,73]
Convolutional Neural Network (CNN)	STFT spectrogram segments; species classification	Learns time–frequency representations; supports large-scale recognition	Requires substantial training resources; may miss background species	Large-scale acoustic species identification	[61]
CNN–LSTM	MFCC and chroma features; species classification	Combines learned features with temporal modelling	Increased architectural complexity	Temporal bird vocalization classification	[64,74]
Transformer-based self-supervized learning	Mel-filterbank representations; species recognition	Learns from unlabeled recordings; benefits underrepresented species	Pretraining requirements and overlapping vocalizations	Long-term passive acoustic monitoring	[75]
Deep transfer learning	Mel-spectrogram images; multi-species classification	Reuses pretrained representations and supports fine-tuning	Requires adaptation to target acoustic recordings	Multi-species classification using pretrained deep learning models	[61]
Multi-instance multi-label learning (MIML; weakly supervized)	Time–frequency acoustic instances; simultaneous species classification	Supports recognition of multiple species within recordings	Depends on bag-level labels and suitable instance representations	Monitoring recordings containing overlapping bird vocalizations	[21]
Interpretable CNN/prototype learning	Log-mel spectrograms; multi-label bird classification	Provides interpretable predictions using learned acoustic prototypes	Prototypes may incompletely represent vocal repertoires or capture background-noise patterns	Explainable bird sound classification for ecological monitoring	[62]

5. Datasets, Labels, and Benchmarks

The existence of well-annotated data sets is key for the creation and validation of systems that automatically classify birds' sounds. Such bio-acoustics data sets enable scientists to train their machine learning algorithms and evaluate their effectiveness and performance. During the last two decades, there have been several large-scale data sets created which accelerated the process of research into bird sounds classification.

5.1. Publicly Available Bird Sound Datasets

Many publicly available databases contain large collections of bird sounds recorded in different regions of the world. One of the most widely used is Xeno-Canto, a community-contributed database containing recordings submitted by birdwatchers. The database provides information on species, recording location, recording quality, and vocalization type, making it an important source of training data for bird sound classification research.

Another major source of bioacoustic recordings is the Macaulay Library, maintained by the Cornell Lab of Ornithology. It contains a large collection of wildlife recordings, including bird vocalizations, and is widely used for ecological research and species identification.

In addition to these general databases, several curated datasets have been developed for machine learning tasks. For instance, Warblr contains bird sound recordings collected in urban areas of the United Kingdom, while BirdVox contains recordings associated with nocturnal bird migration.

More recently, large-scale datasets such as BirdSet have been introduced to support deep learning research in bioacoustics. BirdSet brings together recordings from multiple sources and represents thousands of bird species. The scale and diversity of such datasets are valuable for developing and evaluating deep learning models, which generally require large volumes of labeled data. The characteristics and typical uses of the bird sound datasets discussed above are summarized in Table 7.

Table 7. Summary of several widely used bird sound datasets employed in machine learning studies.

Dataset	Characteristics	Typical Use
Xeno-Canto [76]	Large community-contributed collection of bird recordings with global coverage	Model training and bird species recognition
Macaulay Library [77]	Extensive curated archive of wildlife audio archive	Biodiversity research and ecological studies
Warblr [78]	Field recordings collected mainly from urban environments	Bird sound detection
BirdVox [79]	Recordings focused on nocturnal bird migration	Acoustic monitoring and migration research
BirdSet [80]	Large-scale dataset combining recording from multiple sources	Deep learning-based bird sound classification

5.2. Annotation and Labeling Challenges

Although large datasets are now available, labeling bird vocalizations remains a major challenge. Manual annotation requires expert knowledge of bird species and can be highly time-consuming. The task is further complicated by overlapping vocalizations, environmental noise, and differences in recording conditions.

Another difficulty arises when several bird species are present in the same audio segment. In these cases, assigning a single label may not accurately represent the acoustic scene. Multi-label classification approaches have therefore been explored to account for recordings that contain more than one species.

Annotation quality can also differ considerably between datasets. Community-contributed recordings may contain inconsistencies or errors in species identification, while curated datasets often provide more reliable labels but may be smaller or have more limited coverage. Balancing dataset size with annotation quality therefore remains an important issue in bird sound classification research.

5.3. Dataset Limitations and Generalization Challenges

Although bird sound datasets are becoming more widely available, several limitations can affect how well they represent real-world conditions and support automated species classification. One major issue is class imbalance. Common and frequently recorded species may be represented by many examples, while rare or geographically restricted species often have far fewer labeled recordings. This imbalance can affect both model training and evaluation, as classifiers may perform well for well-represented species but less reliably for those with limited data. Rauch et al. [81] identify class imbalance as an important challenge in avian bioacoustics and account for species-distribution characteristics in their evaluation of the BirdSet benchmark.

Geographic and temporal sampling biases are also important concerns. Recordings may be concentrated in certain locations, habitats, seasons, or periods of vocal activity, which can limit the range of acoustic variation represented for a species. Borror [82] documented seasonal and geographic variation in bird vocalizations, including differences associated with breeding and migration. Differences in microphones, autonomous recording units and recording configurations can also affect signal quality, frequency response and background noise. Rauch et al. [81] note that variation in recording environments and equipment can introduce domain shifts that reduce model generalization.

Label quality is another important concern, particularly in community-contributed datasets and recordings that contain overlapping vocalizations from multiple species. Incorrect species identification, incomplete annotations, weak recording-level labels, and inconsistent labeling practices may introduce uncertainty into model training and evaluation. Rauch et al. [81] identify label uncertainty as a major challenge in bird sound classification, particularly when recordings lack precise annotations of individual vocalization events.

Consequently, strong classification performance obtained from conventional benchmark datasets may not accurately reflect model effectiveness under real-world ecological monitoring conditions, especially when training and evaluation sets contain recordings from the same locations, devices, contributors or time periods. Under these conditions, models may learn dataset-specific acoustic characteristics that do not generalize to independent field recordings.

To improve the ecological relevance of future evaluations, bird sound classification studies should consider geographically independent evaluation sets and cross-device evaluation. Greater attention to species-level performance, annotation reliability, and underrepresented populations would provide a more realistic assessment of model robustness and generalization. The BirdSet study by Rauch et al. [81], which evaluates classification models across multiple geographically distinct soundscape datasets, provides a relevant example of assessing model performance under diverse recording conditions.

5.4. Benchmark Challenges and Evaluation Platforms

Benchmark challenges have been very useful in the development of bird sound classification because they provide datasets and evaluation methodologies that enable comparison of approaches under the same conditions. The main benchmark challenge in this area is the BirdCLEF, which is held as part of the LifeCLEF challenges.

The BirdCLEF dataset consists of annotated bird sound recordings, and participants in the competition need to design systems that can automatically classify birds species. Different kinds of environments and bird species are used to evaluate different approaches under different circumstances. In order to evaluate performance, metrics like precision, recall, and mean average precision (MAP) are used.

Participation in benchmark challenges has encouraged the development of more sophisticated models, particularly deep learning architectures for large-scale, multi-species classification. By allowing methods to be compared under consistent evaluation conditions, these challenges have also contributed to progress in automated bioacoustic analysis.

6. Deployment Challenges and Future Research Directions

Good performance on a benchmark dataset does not always translate into reliable performance under field conditions. Benchmark recordings may contain clear vocalizations, consistent labels, and recording conditions similar to those used during model training. Field recordings, however, are often much more variable. Wind, rain, insects, other animals and human activity can mask bird calls, while several species may vocalize at the same time. The same species may also sound different across locations, populations and recording devices. These sources of variation can reduce classification accuracy, particularly for rare species with limited labeled data.

Field deployment also introduces additional practical constraints. Autonomous recording units may need to operate for several months in remote locations with limited battery power, storage capacity, processing resources, and connectivity. Classification models must therefore be computationally efficient enough to process large volumes of audio while still producing detections that are sufficiently reliable for the intended ecological application. Identifying a possible species presence, estimating changes in bird activity, and issuing a conservation alert may each require different levels of accuracy and timeliness. For automated classification to support ecological monitoring, its outputs must also be linked to recording locations and times and be available for interpretation or verification by researchers.

6.1. Robustness in Complex Acoustic Environments

A significant drop in performance is observed for models that show good performance on carefully selected benchmark datasets when deployed on field recordings. This is due to the disparity in training and testing environments. Common factors that influence field recordings include wind, rain, insect chorusing, distant machinery sounds, and artifacts from equipment, which may affect the clarity of bird songs, with the volume and sound characteristics of the above-mentioned factors varying during the recording process. Consequently, models that have not been trained on similar conditions may therefore suffer in terms of classification performance.

Overlapping vocalizations present an additional challenge. Dawn chorus recordings, which are particularly valuable for biodiversity monitoring, may contain several species vocalizing simultaneously within similar frequency bands. A loud or nearby call may mask a quieter species, while calls from different birds may appear to form a single acoustic event. Conventional preprocessing methods can struggle to separate these sounds, increasing the difficulty of detecting every species present. Robustness to overlapping vocalizations is therefore important when a system is intended to describe an entire bird community rather than recognize an isolated call.

Differences between recording devices and deployment sites can also affect performance. Microphone sensitivity, recording distance, placement and compression can change the apparent characteristics of a vocalization. A model trained primarily on close, high-quality recordings may consequently perform less well on sounds captured by a distant ARU. Evaluation on recordings collected from different habitats and devices is needed to determine how reliably a model will perform outside its original training conditions.

Addressing these difficulties will require more than improvements to classification architectures. Sound source separation may help disentangle mixed signals, while noise-reduction methods must preserve quieter bird vocalizations as they suppress unwanted sounds. Training with realistic acoustic variation may improve robustness, and multi-instance learning could help when a recording is known to contain a species, but the exact timing of its call is uncertain. The effectiveness of these approaches across diverse field conditions remains only partially resolved.

6.2. Biological and Ornithological Sources of Vocal Variation

Bird vocalizations are shaped by biological and behavioural factors that create variation both within and between species. Birdsong is closely linked to reproductive and territorial behaviour, including mate attraction and territorial defence. As a result, vocal activity and repertoire can vary with season and behavioural context. Territorial and courtship songs may be more common during the breeding season, whereas contact calls, alarm calls and flight calls may occur more frequently during non-breeding or migration periods. Vocal characteristics can also differ with age, season, behavioural context and individual variation. Borror [82] documented variation in passerine songs across individuals, seasons and geographic populations, including changes in song patterns during migration and the breeding period. Konishi [83] further examined the behavioural and developmental processes involved in birdsong production and learning.

Geographic and population-level differences can also affect the acoustic characteristics of bird vocalizations. Regional song patterns and individual repertoires may create considerable variation within the same species. Migration can introduce further variability because recordings collected along migratory routes may include populations from different breeding regions. Closely related species may also share similar frequency ranges, temporal patterns, and syllable structures because of their common evolutionary history. Body size, evolutionary history, and habitat acoustics can further influence song frequency and structure [84]. Acoustic convergence and vocal mimicry may create additional difficulties for automated species identification [82].

These biological sources of variation have important implications for machine learning-based bird sound classification. Models trained on recordings from a limited number of geographic regions, seasons, populations or individuals may not capture the full vocal diversity of a species. Their performance may therefore decline when applied to unfamiliar populations, behavioural contexts or recording periods. To improve ecological applicability, future studies should include recordings from different seasons, populations and geographic regions and evaluate model performance across independent monitoring sites and time periods.

6.3. Limited Labeled Data, Class Imbalance and Species Diversity

Models that employ supervized learning for bird sound classification rely strongly on properly labeled training data. Despite an increasing availability of recordings from publicly available databases such as Xeno-Canto and the Macaulay Library, as well as datasets created for BirdCLEF, annotation process remains resource-intensive and often requires expert knowledge. It becomes especially challenging in cases when recorded audio

contains overlapping sounds of birds, presence of environmental noises, or vocalizations of similar acoustic species. The annotation process becomes quite laborious when applied to long recordings collected via passive acoustic monitoring.

Furthermore, the existing recordings are not equally distributed among various species. Some widespread and abundant bird species may be well-represented in datasets, featuring thousands of examples of their vocalizations, while rare or elusive species will provide only a few annotated recordings. In such cases, models trained on an imbalanced dataset will show better results in recognition of common and widespread species than underrepresented ones. It is especially dangerous for the purposes of biodiversity monitoring, which may include rare or endangered species, and might be concealed by overall accuracy if dominated by the presence of numerous test samples of some species.

Besides, natural variations of vocalization within species add to the difficulties with model generalization. Vocalizations of the same species may vary depending on the individuality of the bird, its age, subspecies, or geographic origin. Song types may change according to the behaviour of the bird at the moment. Thus, the classifier trained on audio recordings from one particular area may fail to recognize the same bird sounds in another location. To reliably deploy the model, one should make sure it takes into account the variability of sounds rather than adapts to certain acoustic patterns of the training set.

As a result, reducing dependency on large, manually annotated, and balanced datasets remains one of the primary goals in this area of study. Self-supervized learning makes use of unlabeled recordings for learning general acoustic representation, while semi-supervized learning combines labeled and unlabeled samples in the course of training. Active learning helps to focus the efforts of experts on labeling the recordings whose annotation may be especially valuable, and few-shot learning becomes helpful when only few samples of the species are present. Such techniques as data augmentation and synthetic audio generation may diversify training samples or reduce imbalance between classes. Evaluation of such approaches across different species, populations, ecosystems, recording devices, and environmental conditions is required.

From an ecological perspective, addressing limited labeled data for rare species requires combining computational approaches with targeted data collection and expert involvement. Autonomous recording units could be deployed in habitats, geographic regions and periods when conservation-priority species are most likely to vocalize, thereby improving the collection of relevant recordings. Collaboration among field ornithologists, conservation organizations and community recording initiatives may also help expand geographically diverse reference collections. Recordings used for model development should undergo appropriate species identification and quality verification, particularly where acoustically similar species are present. Stowell et al. [13] demonstrated the practical value of automated bird detection in remote acoustic monitoring and highlighted the importance of recording conditions, annotation quality and generalization when developing systems for ecological applications.

Active learning offers a practical approach to reducing annotation effort by selecting uncertain or informative recordings for expert review rather than requiring manual examination of entire acoustic datasets, as discussed by Settles [85]. Likewise, transfer learning and few-shot learning can assist models in better utilization of sparse verified recordings. For instance, Snell et al. [86] illustrated the utility of prototypical networks in few-shot classification problems and hence provided a framework to explore similar approaches in bird sound classification. Data augmentation can increase the acoustic variability of training data, though artificially created samples cannot completely substitute the natural acoustic variation captured through recordings.

In the case of rare or endangered birds, accurate detection is a priority over maximized classification accuracy. Therefore, in the evaluation process, species-specific precision and recall values should be calculated, with special emphasis on false negatives and false positives. Moreover, automatic detection methods need to provide room for expert evaluation to allow for their use in making conservation-related decisions. If classification results are used for the estimation of the species distribution, then non-detection does not necessarily mean absence, and repeated sampling along with modelling techniques accounting for imperfect detection might be necessary. False positives, false negatives and variations in performance among different acoustic habitats were emphasized by Stowell et al. [13]. The combination of targeted recording, efficient annotation, expert validation and thorough evaluation at the level of species presents a promizing way forward in constructing classification systems.

6.4. Hardware, Edge Computing and Real-Time Monitoring

Large-scale bird sound monitoring often involves deploying ARUs in remote locations for long periods. These devices may have limited storage, processing power, memory, and battery capacity. Continuous high-quality recording can quickly consume both storage space and energy, while retrieving devices to replace batteries or collect data can be costly. Transmitting large audio files from remote locations may place additional demands on communication systems and power supply. As a result, a model that performs well on powerful computing infrastructure may not be practical for continuous operation on a field device.

Edge computing offers a way to address some of these limitations by processing recordings locally. An edge-based system could identify potentially relevant audio segments, discard recordings with little useful information, and transmit detections or selected audio for further analysis. Local processing may also reduce the delay between a vocalization and its detection. However, the models that achieve the strongest classification results can be computationally demanding, and reducing their size may affect their ability to recognize quieter calls or less common species.

The need for real-time processing depends on the monitoring application. Delayed analysis may be sufficient for retrospective biodiversity surveys or studies of long-term trends. Applications involving migration tracking, habitat disturbance or conservation alerts may instead require detections soon after a sound is recorded. A real-time system must process incoming audio continuously without accumulating delays, while also managing intervals containing dense or overlapping vocalizations. BirdNET demonstrates the potential of deep learning to analyze field recordings, but adapting a research model for reliable operation on a resource-constrained device remains a separate engineering challenge.

Lightweight neural network architectures, knowledge distillation and model pruning provide possible ways to reduce memory and computational requirements. Systems may also use an initial, inexpensive detection stage so that a more demanding classifier is run only when bird vocalizations are likely to be present. For practical monitoring, evaluation should consider classification performance alongside processing speed, memory usage, energy consumption and reliability over extended deployments. A system that performs slightly less well in a benchmark but operates continuously in the field may be more useful than one that requires substantially greater computing resources.

6.5. Multi-Species and Multi-Label Classification

Many bird sound classification methods assign a single species label to each audio segment. Field recordings, particularly those captured during dawn choruses, frequently contain vocalizations from multiple species at the same time. Treating these recordings as single-label problems may omit species that are present or assign a label only to the

loudest or most prominent call. The resulting output provides an incomplete account of the acoustic scene and may underestimate species presence in biodiversity surveys.

The multi-label classification strategy would probably be more appropriate, as it would enable a model to identify a variety of species in the same audio recording. The model could be trained to provide individual scores for all possible species, making it possible to identify several species at once when needed. However, the chances of obtaining the desired results will become increasingly difficult with the growing number of species and overlap of their sounds. There will be no possibility of hearing the quiet vocalization due to the loud ones, and vocalizations of different species can be mistaken for each other.

It is essential to remember that multi-label classification requires appropriate training data as well. If the recording contains the voice of one particular species which is easily identifiable, there can be voices of other species recorded as well. The training of the model in the absence of these vocalizations can result in errors in predictions. The use of approaches that allow for handling missing labels, and source separation, and acoustic scene analysis might help here. It is crucial to conduct further research in order to investigate how effective such approaches will be in species identification.

6.6. Integration with Ecological Monitoring Systems

Bird sound classification systems have limited practical value if their outputs remain disconnected from the databases, mapping platforms and workflows used by ecologists and conservation practitioners. Integrating acoustic detections with geographic information systems, sensor networks and biodiversity databases could support monitoring across sites and over time. Recording locations, timestamps, and associated environmental measurements can provide context for interpreting a detection and examining changes in bird activity.

The form of the output is as important as the species prediction itself. Ecologists may need access to detection scores, the time within a recording when a call occurred, and the underlying audio so that uncertain results can be checked. A possible detection may be useful for directing further investigation but insufficient on its own to establish species presence for a conservation decision. Monitoring systems should therefore make it possible to review detections and understand the limitations of the classifications they provide.

Integration requirements also depend on the intended use. A retrospective survey may prioritize the ability to analyze and verify large collections of recordings, whereas a conservation alert may require rapid delivery of a smaller number of highly reliable detections. A system designed to study changes in bird populations must also produce results that can be compared consistently across recording sites, devices and monitoring periods. Connecting a classifier to an ecological workflow therefore requires careful consideration of data management, interpretation and verification.

Achieving this integration requires sustained collaboration among machine learning researchers, signal-processing specialists, field ecologists and conservation practitioners. Their combined input is needed to determine which species and ecological questions should be prioritized, what level of error is acceptable and how automated detections should contribute to monitoring decisions.

Acoustic monitoring has increasingly been applied to ornithological investigations beyond automated species identification, including assessments of species richness, population abundance, occupancy and temporal changes in bird communities. Previous studies have demonstrated the potential of acoustic indices and computer-assisted sampling for species richness assessment, while autonomous recording units combined with statistical models have supported population density and abundance estimation. These applications are particularly valuable for nocturnal and elusive species, such as owls, nightjars, and rallids, which can be difficult to survey using conventional observational methods. Recent

longitudinal investigations have further demonstrated that passive acoustic monitoring can detect changes in species occupancy and responses to habitat disturbance. Table 8 summarizes representative studies illustrating these applications. Acoustic detection frequency, occupancy, abundance and population density represent different ecological measures and should therefore be interpreted according to the sampling design and analytical methods employed.

Table 8. Representative applications of acoustic monitoring for avian species diversity, population assessment, temporal monitoring and nocturnal bird conservation.

Study	Target Species/Group	Monitoring Objective	Acoustic Approach and Principal Finding
[87]	Temperate and tropical bird communities	Species richness	Acoustic indices were more strongly associated with avian species richness in temperate than tropical habitats.
[88]	Bird communities	Species richness	Computer-assisted acoustic sampling improved the efficiency of expert-assisted species identification.
[89]	New Zealand forest birds	Population density	Paired acoustic recordings and point counts supported density estimation after correcting detection bias.
[90]	Nightjars, Dartford warblers and woodlarks	Occupancy and detectability	Automated clustering, manual verification and occupancy modelling supported monitoring of rare breeding birds.
[91]	Príncipe scops-owl	Nocturnal species presence	Automated call detection produced site-level presence results comparable to field surveys.
[92]	Eastern black rail	Occupancy and abundance	Call-playback point counts and detection-corrected models supported population assessment.
[93]	Nocturnally migrating songbirds	Migration timing and magnitude	Acoustic detections corresponded with mist-net observations of migration.
[94]	Pennsylvania forest birds	Abundance estimation	Machine-learning-derived acoustic detections supported abundance estimates comparable to conventional point counts.
[95]	California spotted owl	Multi-year occupancy trends	Passive acoustic monitoring identified declining occupancy trajectories following severe wildfires.

Collectively, these studies demonstrate that acoustic monitoring can contribute to biodiversity assessment and population-level ecological investigations when combined with appropriate analytical methods. Bombaci and Pejchar [89] demonstrated the importance of correcting detection bias in acoustic-based density estimation, while Fiss et al. [94] showed that machine-learning-derived acoustic detections can support abundance modelling. For nocturnal species, automated monitoring of the Príncipe Scops-Owl demonstrated the feasibility of detecting elusive birds in difficult terrain, whereas Winiarski et al. [95] used multi-year acoustic surveys to identify declining occupancy trajectories in California spotted owls following severe wildfire disturbance. These findings highlight the value of integrating automated bird sound classification with ecological models and long-term survey designs. Nevertheless, reliable assessment of population trends requires repeated sampling and careful consideration of imperfect detection, changes in vocal activity, and variation in environmental conditions.

6.7. Research Opportunities

Pre-training by self-supervized learning through a large set of unlabeled sounds is one of the promising approaches for classifying bird sounds. Passive acoustic monitoring produces many times more recordings than can be labeled manually by experts. It would be much more efficient to learn general audio representations through these recordings before training on specific species. The effectiveness of the approach should be demonstrated on underrepresented species and in cases where sounds are recorded in a different habitat compared to pre-training.

Multimodal learning provides another promising direction. Acoustic recordings can be combined with contextual information such as geographic location, recording time, weather conditions, and habitat characteristics. Such information may help resolve ambiguous predictions, particularly when acoustically similar species occur in different regions or environmental settings. At the same time, models should be evaluated to ensure that they recognize acoustic evidence rather than relying too heavily on location or other contextual information. This is especially important when species distributions change or a system is deployed in a new area.

Investing in open datasets and benchmark challenges remains crucial as well. Datasets and benchmark procedures help researchers compare their methods, replicate results, and find out whether improvements are generalizable to a different set of recordings. It will be more meaningful if future benchmarks include not only recordings made on different devices and in different environments, but also recordings made on rare species, overlapping sounds, and difficult background noise. The reporting of results under these circumstances separately may uncover problems which are hidden behind the overall accuracy number.

Finally, future research should assess complete monitoring systems under realistic deployment conditions. Improvements in classification algorithms matter most when they lead to more reliable detections, more efficient use of recording hardware, or more useful ecological information. Field evaluation, together with input from the people who will interpret and use the results, is needed to determine whether automated bird sound classification can support routine biodiversity monitoring and conservation decisions.

Further research needs to focus on finding solutions that will enhance the reliability and applicability of automatic bird vocalization classification. It is possible that data augmentation through noise mixing and time-shifting, among other relevant modifications, may be helpful to deal with the lack of labeled data and ensure robustness in various conditions [96]. Nevertheless, any data augmentation technique must preserve biologically valid vocal traits in order not to change the acoustic properties of the species in question. In addition, domain adaptation techniques can improve generalization of the classifier by making the difference between training and application conditions smaller [97]. Few-shot learning seems to be an interesting solution that can help recognize rare species which have few labeled recordings [86]. Active learning can decrease labeling costs by allowing researchers to focus only on ambiguous samples [85].

Another promising research area is open set recognition since the ecological monitoring system may face some species that were not present in the training data. Instead of trying to classify every recorded sample into a species class, it will be necessary to separate unknown sounds for further analysis [98]. Noise-robust learning is another promising area of investigation for the sake of improving classification in difficult field conditions such as overlap of vocalizations, noise from wind and rain, and other disturbances. Such techniques as multi-condition training, noise-augmented data generation, sound-source separation, and consistency learning can be tested for preserving classification performance under those circumstances. Light-weight architectures such as MobileNet might be useful for deployment on field acoustic monitors [99]. Model-pruning, quantization and knowledge-distillation

techniques may further reduce memory use, processing time and energy consumption. Studies of edge deployment should therefore report latency, memory requirements and energy consumption alongside classification accuracy to determine whether models are suitable for continuous field operation.

Future studies should evaluate these approaches using independent field recordings collected across different habitats, geographic regions, seasons, and recording devices. Evaluation should consider not only classification accuracy but also generalization to unfamiliar conditions, annotation requirements, computational efficiency, and suitability for long-term ecological monitoring.

As a methodological extension of the present study, future bibliometric reviews could compare Scopus with additional databases and literature platforms, including Web of Science, IEEE Xplore, ScienceDirect and SpringerLink, to determine whether broader literature coverage changes the identified research themes and methodological trends.

7. Conclusions

Bird vocalizations provide a valuable source of information about species presence and activity, while autonomous recording units make it possible to collect audio across large areas and extended periods. The resulting volume of recordings has increased the need for automated analysis. This review examined how bird sound classification research has developed, identified its principal research themes, and synthesized the methods used to turn recordings into species classifications.

The study combined bibliometric analysis with a structured review of methodological developments. A Scopus search initially retrieved 604 publications; after applying the English-language and 2000–2026 publication-year criteria, 583 publications were retained. Keyword co-occurrence analysis identified the main research themes, while bibliographic coupling analysis examined connections among publications. A temporal overlay of the bibliographic coupling network provided a view of how earlier and more recent studies are positioned within the field. The methodological synthesis considered the full classification pipeline, from data acquisition and preprocessing to segmentation, feature representation, classification, and evaluation.

Based on the keyword co-occurrence analysis, four key topics were identified as the following: deep learning-based classification; bioacoustics and passive acoustic monitoring; bird sound recognition and classification approaches; and benchmark datasets and frameworks for model evaluation. This set of topics demonstrates that there is a dual focus of the field on developing better recognition approaches and applying them for ecological monitoring. The methodological synthesis revealed a transition from approaches based on signal processing, hand-designed acoustic features, and classic classifiers to approaches that learn representations from spectrograms and large bioacoustic datasets. However, the former can be useful in certain circumstances, especially when there is a lack of annotated samples and computational resources. Public datasets and benchmark challenges allowed for the consistent comparison of various recognition approaches and enabled the creation of species recognition systems based on large datasets. The bibliographic coupling network showed connections between studies related to acoustic detection, autonomous monitoring, and deep learning-based classification. Temporal overlay of the network confirmed the presence of early methodological studies as well as later studies with larger datasets and deep learning-based models. These results place current classification approaches into the broader research field related to signal processing, machine learning, bioacoustics, and ecology.

However, there are numerous challenges that arise when the developed system is used beyond benchmark datasets. Environmental noise and overlapping sounds can make the recognition of particular species more difficult; lack of annotations and imbalanced classes can negatively impact the recognition of rare species; differences between habitats, populations, and recorders can decrease the ability of models to generalize. The long-term application in field conditions requires consideration of processing time, memory consumption, battery power, and methods of verifying detection of species. Future evaluations of approaches need to include testing of models on diverse field recordings with reporting of the recognition of underrepresented species and challenging acoustic conditions and of computational and energy requirements.

The contribution of this paper lies in linking the research topics of the field with the methods and constraints involved in the development of automatic monitoring systems. Advances in classification accuracy are valuable, but the importance of such systems in ecological contexts depends on the possibility of getting accurate, interpretable results in real-life conditions. To solve this problem, it is necessary to continue work in machine learning, signal processing, ecology, and conservation practice.

There are three principal insights that can be drawn from the methodological synthesis. Firstly, deep learning approaches, specifically CNNs and more advanced attention-based models, are more likely to outperform conventional approaches when there is a large and representative training dataset, and the task is complex and involves multiple species. They allow learning of discriminative representations directly from spectrograms or audio samples without designing hand-engineered acoustic features. However, they are advantageous only in the case of correct model selection, sufficient computational power, and similarity between the training and deployment conditions. It is not possible to conclude that deep learning models outperform conventional approaches in every case.

Secondly, conventional machine learning algorithms, like SVMs and RF in combination with hand-designed acoustic features, are preferred in the case of applications involving small datasets and limited sets of species or limited computational resources due to lower complexity and less intensive training. Selection of the classification approach needs to take into account the amount of data and its representativeness, complexity of the recognition task, and deployment conditions rather than the model complexity.

Thirdly, accuracy on curated benchmark datasets cannot be considered as the measure of the reliability of the ecological monitoring. Environmental noise, overlapping vocalizations, geographic and seasonal variation, label uncertainty, and differences between recording devices can affect performance under field conditions. Consequently, meaningful evaluation should include independent recordings from different locations and periods, species-level performance assessment, and consideration of computational efficiency. These factors are essential for determining whether automated classification systems can provide reliable information for biodiversity monitoring and conservation applications.

Author Contributions: Conceptualization, P.E.A. and L.C.D.S.; methodology, P.E.A. and M.R.; software, M.R.; validation, P.E.A.; formal analysis, M.R.; investigation, M.R.; resources, P.E.A. and L.C.D.S.; data curation, M.R.; writing—original draft preparation, M.R.; writing—review and editing, P.E.A.; visualization, M.R.; supervision, P.E.A. and L.C.D.S.; project administration, P.E.A. and L.C.D.S.; funding acquisition, P.E.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Universiti Brunei Darussalam, grant number UBD/RSCH/1.3/FICBF/2026/007.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in the article.

Acknowledgments: During the preparation of this manuscript/study, the authors used Grammarly Premium v1.2.214.1792 for the purposes of English correction. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Suthers, R.A. How Birds Sing and Why It Matters. In *Nature's Music: The Science of Birdsong*; Elsevier Academic Press: San Diego, CA, USA, 2004; ISBN 9780124730700.
2. Alvarsson, J.J.; Wiens, S.; Nilsson, M.E. Stress Recovery during Exposure to Nature Sound and Environmental Noise. *Int. J. Environ. Res. Public Health* **2010**, *7*, 1036–1046. [CrossRef]
3. Vilches, E.; Escobar, I.A.; Vallejo, E.E.; Taylor, C.E. Data Mining Applied to Acoustic Bird Species Recognition. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August 2006; Volume 3, pp. 400–403. [CrossRef]
4. Frommolt, K.-H.; Bardeli, R.; Clausen, M. (Eds.) Computational Bioacoustics for Assessing Biodiversity. In *Proceedings of the International Expert Meeting on IT-Based Detection of Bioacoustical Patterns, Isle of Vilm, Germany, 7–10 December 2007*; BfN-Skripten 234; Bundesamt für Naturschutz: Bonn, Germany, 2008.
5. Potamitis, I.; Ntalampiras, S.; Jahn, O.; Riede, K. Automatic Bird Sound Detection in Long Real-Field Recordings: Applications and Tools. *Appl. Acoust.* **2014**, *80*, 1–9. [CrossRef]
6. Gregory, R.D.; Willis, S.G.; Jiguet, F.; Voříšek, P.; Klvaňová, A.; van Strien, A.; Huntley, B.; Collingham, Y.C.; Couvet, D.; Green, R.E. An Indicator of the Impact of Climatic Change on European Bird Populations. *PLoS ONE* **2009**, *4*, 4678. [CrossRef]
7. Marques, T.A.; Thomas, L.; Martin, S.W.; Mellinger, D.K.; Ward, J.A.; Moretti, D.J.; Harris, D.; Tyack, P.L. Estimating Animal Population Density Using Passive Acoustics. *Biol. Rev.* **2013**, *88*, 287–309. [CrossRef]
8. Stowell, D. Computational Bioacoustics with Deep Learning: A Review and Roadmap. *PeerJ* **2022**, *10*, e13152. [CrossRef]
9. Maithripala, A.; Arachchi, S.M.; Karunanayaka, K.; Perera, R.; Pallewatta, P. A Review of Automated Bird Sound Recognition and Analysis in the New AI Era. In Proceedings of the 2024 8th SLAAI—International Conference on Artificial Intelligence (SLAAI-ICAI 2024), Ratmalana, Sri Lanka, 18–19 December 2024. [CrossRef]
10. Kahl, S.; Wilhelm-Stein, T.; Hussein, H.; Klinck, H.; Kowerko, D.; Ritter, M.; Eibl, M. Large-Scale Bird Sound Classification Using Convolutional Neural Networks. In Proceedings of the CEUR Workshop Proceedings, Dublin, Ireland, 11–14 September 2017; Volume 1866.
11. Bardeli, R.; Wolff, D.; Kurth, F.; Koch, M.; Tauchert, K.H.; Frommolt, K.H. Detecting Bird Sounds in a Complex Acoustic Environment and Application to Bioacoustic Monitoring. *Pattern Recognit. Lett.* **2010**, *31*, 1524–1534. [CrossRef]
12. Digby, A.; Towsey, M.; Bell, B.D.; Teal, P.D. A Practical Comparison of Manual and Autonomous Methods for Acoustic Monitoring. *Methods Ecol. Evol.* **2013**, *4*, 675–683. [CrossRef]
13. Stowell, D.; Wood, M.D.; Pamuła, H.; Stylianou, Y.; Glotin, H. Automatic Acoustic Detection of Birds through Deep Learning: The First Bird Audio Detection Challenge. *Methods Ecol. Evol.* **2019**, *10*, 368–380. [CrossRef]
14. Kahl, S.; Wood, C.M.; Eibl, M.; Klinck, H. BirdNET: A Deep Learning Solution for Avian Diversity Monitoring. *Ecol. Inform.* **2021**, *61*, 101236. [CrossRef]
15. Kogan, J.A.; Margoliash, D. Automated Recognition of Bird Song Elements from Continuous Recordings Using Dynamic Time Warping and Hidden Markov Models: A Comparative Study. *J. Acoust. Soc. Am.* **1998**, *103*, 2185–2196. [CrossRef]
16. Anderson, S.E.; Dave, A.S.; Margoliash, D. Template-based Automatic Recognition of Birdsong Syllables from Continuous Recordings. *J. Acoust. Soc. Am.* **1996**, *100*, 1209–1219. [CrossRef]
17. Trifa, V.M.; Kirschel, A.N.G.; Taylor, C.E.; Vallejo, E.E. Automated Species Recognition of Antbirds in a Mexican Rainforest Using Hidden Markov Models. *J. Acoust. Soc. Am.* **2008**, *123*, 2424–2431. [CrossRef]
18. Stowell, D.; Benetos, E.; Gill, L.F. On-Bird Sound Recordings: Automatic Acoustic Recognition of Activities and Contexts. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1193–1206. [CrossRef]
19. Fagerlund, S.; Laine, U.K. New Parametric Representations of Bird Sounds for Automatic Classification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 8247–8251.
20. Chen, Z.; Maher, R.C. Semi-Automatic Classification of Bird Vocalizations Using Spectral Peak Tracks. *J. Acoust. Soc. Am.* **2006**, *120*, 2974–2984. [CrossRef]

21. Fox, E.J.S.; Roberts, J.D.; Bennamoun, M. Call Independent Individual Identification in Birds. *Bioacoust. Int. J. Anim. Sound Its Rec.* **2008**, *18*, 51–67. [CrossRef]
22. Ranjard, L.; Ross, H.A. Unsupervised Bird Song Syllable Classification Using Evolving Neural Networks. *J. Acoust. Soc. Am.* **2008**, *123*, 4358–4368. [CrossRef]
23. Briggs, F.; Lakshminarayanan, B.; Neal, L.; Fern, X.Z.; Raich, R.; Hadley, S.J.K.; Hadley, A.S.; Betts, M.G. Acoustic Classification of Multiple Simultaneous Bird Species: A Multi-Instance Multi-Label Approach. *J. Acoust. Soc. Am.* **2012**, *131*, 4640–4650. [CrossRef]
24. Priyadarshani, N.; Marsland, S.; Castro, I.; Punchihewa, A. Birdsong Denoising Using Wavelets. *PLoS ONE* **2016**, *11*, e0146790. [CrossRef]
25. Ren, Y.; Johnson, M.T.; Tao, J. Perceptually Motivated Wavelet Packet Transform for Bioacoustic Signal Enhancement. *J. Acoust. Soc. Am.* **2008**, *124*, 316–327. [CrossRef]
26. Chu, W.; Alwan, A. A Correlation-Maximization Denoising Filter Used as an Enhancement Frontend for Noise Robust Bird Call Classification. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Brighton, UK, 6–10 September 2009; pp. 2831–2834.
27. Lasseck, M. Bird Song Classification in Field Recordings: Winning Solution for NIPS4B 2013 Competition. In Proceedings of the International Symposium on Neural Information Scaled for Bioacoustics, Lake Tahoe, NV, USA, 10 December 2013; pp. 1–6.
28. Gavali, P.; Banu, J.S. A Novel Approach to Indian Bird Species Identification: Employing Visual-Acoustic Fusion Techniques for Improved Classification Accuracy. *Front. Artif. Intell.* **2025**, *8*, 1527299. [CrossRef]
29. Ramashini, M.; Abas, P.E.; Grafe, U.; De Silva, L.C. Bird Sounds Classification Using Linear Discriminant Analysis. In Proceedings of the 2019 4th International Conference and Workshops on Recent Advances and Innovations in Engineering: Thriving Technologies (ICRAIE), Kedah, Malaysia, 1–6 November 2019; pp. 1–6. [CrossRef]
30. Giannakopoulos, T.; Pikrakis, A. *Introduction to Audio Analysis: A MATLAB Approach*; Elsevier Academic Press: San Diego, CA, USA, 2014; ISBN 9780080993881.
31. Ranjard, L.; Withers, S.J.; Brunton, D.H.; Ross, H.A.; Parsons, S. Integration over Song Classification Replicates: Song Variant Analysis in the Hihi. *J. Acoust. Soc. Am.* **2015**, *137*, 2542–2551. [CrossRef]
32. Lee, C.H.; Han, C.C.; Chuang, C.C. Automatic Classification of Bird Species from Their Sounds Using Two-Dimensional Cepstral Coefficients. *IEEE Trans. Audio Speech Lang. Process.* **2008**, *16*, 1541–1550. [CrossRef]
33. Somervuo, P.; Harma, A.; Härmä, A. Bird Song Recognition Based on Syllable Pair Histograms. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04), Montreal, QC, Canada, 17–21 May 2004; 2004; Volume 5, pp. V-825–V-828. [CrossRef]
34. Chou, C.H.; Liu, P.H.; Cai, B. On the Studies of Syllable Segmentation and Improving MFCCs for Automatic Birdsong Recognition. In Proceedings of the 3rd IEEE Asia-Pacific Services Computing Conference (APSCC), Yilan, Taiwan, 9–12 December 2008; pp. 745–750. [CrossRef]
35. Jinnai, M.; Boucher, N.; Fukumi, M.; Taylor, H. A New Optimization Method of the Geometric Distance in an Automatic Recognition System for Bird Vocalisations. In Proceedings of the Acoustics 2012, Hong Kong, China, 13–18 May 2012; pp. 2439–2445.
36. Neal, L.; Briggs, F.; Raich, R.; Fern, X.Z. Time-Frequency Segmentation of Bird Song in Noisy Acoustic Environments. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; pp. 2012–2015.
37. Frommolt, K.; Bardeli, R.; Gmbh, V.; Clausen, M. Automated Monitoring of Avian Flight Calls during Nocturnal Migration. In Proceedings of the International Expert Meeting on IT-Based Detection of Bioacoustical Patterns, Isle of Vilm, Germany, 7–10 December 2007.
38. Härmä, A.; Somervuo, P. Classification of the Harmonic Structure in Bird Vocalization. In Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, QC, Canada, 17–21 May 2004; Volume 5, pp. 1–4. [CrossRef]
39. De Oliveira, A.G.; Ventura, T.M.; Ganchev, T.D.; De Figueiredo, J.M.; Jahn, O.; Marques, M.I.; Schuchmann, K.L. Bird Acoustic Activity Detection Based on Morphological Filtering of the Spectrogram. *Appl. Acoust.* **2015**, *98*, 34–42. [CrossRef]
40. Jančovič, P.; Kküer, M. Automatic Detection and Recognition of Tonal Bird Sounds in Noisy Environments. *EURASIP J. Adv. Signal Process.* **2011**, *2011*, 982936. [CrossRef]
41. Thakur, A.; Rajan, P. Model-Based Unsupervised Segmentation of Birdcalls from Field Recordings. In Proceedings of the 10th International Conference on Signal Processing and Communication Systems (ICSPCS 2016), Gold Coast, Australia, 19–21 December 2016.
42. Wa Maina, C. Audio Diarization for Biodiversity Monitoring. In Proceedings of the IEEE AFRICON Conference 2015, Addis Ababa, Ethiopia, 14–17 September 2015. [CrossRef]
43. Frommolt, K.H.; Tauchert, K.H. Applying Bioacoustic Methods for Long-Term Monitoring of a Nocturnal Wetland Bird. *Ecol. Inform.* **2014**, *21*, 4–12. [CrossRef]

44. Stowell, D.; Plumbley, M.D. Automatic Large-Scale Classification of Bird Sounds Is Strongly Improved by Unsupervised Feature Learning. *PeerJ* **2014**, *2*, e488. [CrossRef]
45. Leng, Y.R.; Dat Tran, H. Multi-Label Bird Classification Using an Ensemble Classifier with Simple Features. In Proceedings of the 2014 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA 2014), Siem Reap, Cambodia, 9–12 December 2014.
46. Sharma, G.; Umapathy, K.; Krishnan, S. Trends in Audio Signal Feature Extraction Methods. *Appl. Acoust.* **2020**, *158*, 107020. [CrossRef]
47. Fagerlund, S. Bird Species Recognition Using Support Vector Machines. *EURASIP J. Adv. Signal Process.* **2007**, *2007*, 038637. [CrossRef]
48. Somervuo, P.; Härmä, A.; Fagerlund, S. Parametric Representations of Bird Sounds for Automatic Species Recognition. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 2252–2263. [CrossRef]
49. Selin, A.; Turunen, J.; Tanttu, J.T. Wavelets in Recognition of Bird Sounds. *EURASIP J. Adv. Signal Process.* **2007**, *2007*, 051806. [CrossRef]
50. Acevedo, M.A.; Corrada-Bravo, C.J.; Corrada-Bravo, H.; Villanueva-Rivera, L.J.; Aide, T.M. Automated Classification of Bird and Amphibian Calls Using Machine Learning: A Comparison of Methods. *Ecol. Inform.* **2009**, *4*, 206–214. [CrossRef]
51. Franzen, A.; Gu, I.Y.H.H. Classification of Bird Species by Using Key Song Searching: A Comparative Study. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Washington, DC, USA, 8 October 2003; Volume 1, pp. 880–887. [CrossRef]
52. Li, W.; Lv, D.; Yu, Y.; Zhang, Y.; Gu, L.; Wang, Z.; Zhu, Z. Multi-Scale Deep Feature Fusion with Machine Learning Classifier for Birdsong Classification. *Appl. Sci.* **2025**, *15*, 1885. [CrossRef]
53. Tachibana, R.O.; Oosugi, N.; Okanoya, K. Semi-Automatic Classification of Birdsong Elements Using a Linear Support Vector Machine. *PLoS ONE* **2014**, *9*, 92584. [CrossRef]
54. Andreassen, T.; Surlykke, A.; Hallam, J. Semi-Automatic Long-Term Acoustic Surveying: A Case Study with Bats. *Ecol. Inform.* **2014**, *21*, 13–24. [CrossRef]
55. Lopes, M.T.; Gioppo, L.L.; Higushi, T.T.; Kaestner, C.A.A.; Silla, C.N.; Koerich, A.L. Automatic Bird Species Identification for Large Number of Species. In Proceedings of the 2011 IEEE InternationalSymposium on Multimedia (ISM 2011), Dana Point, CA, USA, 5–7 December 2011; pp. 117–122. [CrossRef]
56. Lasseck, M. *Improved Automatic Bird Identification through Decision Tree Based Feature Selection and Bagging*; CEUR-WS.org: Bonn, Germany, 2015; Volume 1391.
57. Papadopoulos, T.; Roberts, S.; Willis, K. Detecting Bird Sound in Unknown Acoustic Background Using Crowdsourced Training Data. *arXiv* **2015**, arXiv:1505.06443.
58. Grill, T.; Schluter, J. Two Convolutional Neural Networks for Bird Detection in Audio Signals. In Proceedings of the 25th European Signal Processing Conference (EUSIPCO 2017), Kos, Greece, 28 August–2 September 2017; pp. 1764–1768.
59. Ramashini, M.; Abas, P.E.; Mohanchandra, K.; De Silva, L.C. Robust Cepstral Feature for Bird Sound Classification. *Int. J. Electr. Comput. Eng. (IJECE)* **2022**, *12*, 1477–1487. [CrossRef]
60. Badi, A.; Ko, K.; Ko, H. Bird Sounds Classification by Combining PNCC and Robust Mel-Log Filter Bank Features. *J. Acoust. Soc. Korea* **2019**, *38*, 39–46. [CrossRef]
61. Sprengel, E.; Jaggi, M.; Kilcher, Y.; Hofmann, T. Audio Based Bird Species Identification Using Deep Learning Techniques. In Proceedings of the Working Notes of CLEF 2016 (Cross Language Evaluation Forum), Evora, Protugal, 5–8 September 2016; pp. 547–559.
62. Xie, J.; Hu, K.; Zhu, M.; Yu, J.; Zhu, Q. Investigation of Different CNN-Based Models for Improved Bird Sound Classification. *IEEE Access* **2019**, *7*, 175353–175361. [CrossRef]
63. Baowaly, M.K.; Sarkar, B.C.; Walid, M.A.A.; Ahamad, M.M.; Singh, B.C.; Alvarado, E.S.; Ashraf, I.; Samad, M.A. Deep Transfer Learning-Based Bird Species Classification Using Mel Spectrogram Images. *PLoS ONE* **2024**, *19*, e0305708. [CrossRef]
64. Heinrich, R.; Rauch, L.; Sick, B.; Scholz, C. AudioProtoPNet: An Interpretable Deep Learning Model for Bird Sound Classification. *Ecol. Inform.* **2025**, *87*, 103081. [CrossRef]
65. Mundry, R.; Sommer, C. Tonal Vocalizations in a Noisy Environment: An Approach to Their Semi-Automatic Analysis and Examples of Its Application. *An. Acad. Bras. Cienc.* **2004**, *76*, 284–288. [CrossRef]
66. Ganchev, T.; Mporas, I.; Jahn, O.; Riede, K.; Schuchmann, K.L.; Fakotakis, N. Acoustic Bird Activity Detection on Real-Field Data. In *Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7297, pp. 190–197.
67. Stowell, D.; Plumbley, M.D. Audio-Only Bird Classification Using Unsupervised Feature Learning. In Proceedings of the CEUR Workshop Proceedings, Rome, Italy, 24–25 November 2014; Volume 1180, pp. 673–684.
68. Evangelista, T.L.F.; Priolli, T.M.; Silla, C.N., Jr.; Angelico, B.A.; Kaestner, C.A.A. Automatic Segmentation of Audio Signals for Bird Species Identification. In Proceedings of the 2014 IEEE International Symposium on Multimedia (ISM 2014), Taichung, Taiwan, 10–12 December 2014; pp. 223–228. [CrossRef]

69. Rauch, L.; Heinrich, R.; Moummad, I.; Joly, A.; Sick, B.; Scholz, C. Can Masked Autoencoders Also Listen to Birds? *arXiv* **2025**, arXiv:2504.12880.
70. Xie, J.; Zhu, M. Acoustic Classification of Bird Species Using an Early Fusion of Deep Features. *Birds* **2023**, *4*, 138–147. [CrossRef]
71. Murugaiya, R.; Abas, P.E.; De Silva, L.C. Probability Enhanced Entropy (PEE) Novel Feature for Improved Bird Sound Classification. *Mach. Intell. Res.* **2022**, *19*, 52–62. [CrossRef]
72. Selouani, S.-A.S.-A.; Kardouchi, M.; Hervet, É.; Roy, D.; Hervet, E.; Roy, D. Automatic Birdsong Recognition Based on Autoregressive Time-Delay Neural Networks. In Proceedings of the 2005 ICSC Congress on Computational Intelligence Methods and Applications, Istanbul, Turkey, 15–17 December 2005; pp. 1–6. [CrossRef]
73. Ramashini, M.; Abas, P.E.; De Silva, L.C. A Novel Approach of Audio Based Feature Optimisation for Bird Classification. *J. Sci. Technol.* **2021**, *29*, 2383–2407. [CrossRef]
74. Hasan, N.I. Bird Species Classification And Acoustic Features Selection Based on Distributed Neural Network with Two Stage Windowing of Short-Term Features. *arXiv* **2022**, arXiv:2201.00124.
75. Wei, Y.C.; Chen, W.L.; Tuanmu, M.N.; Lu, S.S.; Shiao, M.T. Advanced Montane Bird Monitoring Using Self-Supervised Learning and Transformer on Passive Acoustic Data. *Ecol. Inform.* **2024**, *84*, 102927. [CrossRef]
76. Xeno-Canto: Sharing Bird Sounds from Around the World. 2019. Available online: https://xeno-canto.org/ (accessed on 15 November 2026).
77. Macaulay Library—A Scientific Archive for Research, Education, and Conservation, Powered by You. Available online: https://www.macaulaylibrary.org/?utm_source=chatgpt.com (accessed on 18 October 2026).
78. Warblr. Available online: https://www.warblr.co.uk/ (accessed on 4 June 2026).
79. Lostanlen, V.; Salamon, J.; Farnsworth, A.; Kelling, S.; Bello, J.P. Birdvox-Full-Night: A Dataset and Benchmark for Avian Flight Call Detection. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 266–270. [CrossRef]
80. BirdSet: A Benchmark Dataset Collection for Bird Sound Classification. Available online: https://huggingface.co/datasets/DBD-research-group/BirdSet (accessed on 15 November 2026).
81. Rauch, L.; Schwinger, R.; Wirth, M.; Heinrich, R.; Huseljic, D.; Herde, M.; Lange, J.; Kahl, S.; Sick, B.; Tomforde, S.; et al. BirdSet: A Large-Scale Dataset for Audio Classification in Avian Bioacoustics. *arXiv* **2025**, arXiv:2403.10380.
82. Borror, D.J. Interspecific Variation in Passerine Bird Songs. *Wilson Bull.* **1961**, *73*, 57–78.
83. Konishi, M. Birdsong: From Behavior to Neuron. *Annu. Rev. Neurosci.* **1985**, *8*, 125–170. [CrossRef]
84. Ryan, M.J.; Brenowitz, E.A. The Role of Body Size, Phylogeny, and Ambient Noise in the Evolution of Bird Song. *Am. Nat.* **1985**, *126*, 87–100. [CrossRef]
85. Settles, B. *Computer Sciences Department Active Learning Literature Survey*; University of Wisconsin-Madison Department of Computer Sciences: Madison, WI, USA, 2009.
86. Snell, J.; Swersky, K.; Zemel, T.R. Prototypical Networks for Few-Shot Learning. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4080–4090.
87. Eldridge, A.; Guyot, P.; Moscoso, P.; Johnston, A.; Eyre-Walker, Y.; Peck, M. Sounding out Ecoacoustic Metrics: Avian Species Richness Is Predicted by Acoustic Indices in Temperate but Not Tropical Habitats. *Ecol. Indic.* **2018**, *95*, 939–952. [CrossRef]
88. Zhang, L.; Towsey, M.; Zhang, J.; Roe, P. Classifying and Ranking Audio Clips to Support Bird Species Richness Surveys. *Ecol. Inform.* **2016**, *34*, 108–116. [CrossRef]
89. Using Paired Acoustic Sampling to Enhance Population Monitoring of New Zealand's Forest Birds | JSTOR. Available online: https://www.jstor.org/stable/26775033 (accessed on 18 October 2026).
90. Abrahams, C.; Geary, M. Combining Bioacoustics and Occupancy Modelling for Improved Monitoring of Rare Breeding Bird Populations. *Ecol. Indic.* **2020**, *112*, 106131. [CrossRef]
91. Freitas, B.; Bas, Y.; Robert, A.; Doutrelant, C.; Melo, M. Passive Acoustic Monitoring in Difficult Terrains: The Case of the Principe Scops-Owl. *Biodivers. Conserv.* **2023**, *32*, 3109–3132. [CrossRef]
92. Tolliver, J.D.M.; Moore, A.A.; Green, M.C.; Weckerly, F.W. Coastal Texas Black Rail Population States and Survey Effort. *J. Wildl. Manag.* **2019**, *83*, 312–324. [CrossRef]
93. Sanders, C.E.; Mennill, D.J. Acoustic Monitoring of Nocturnally Migrating Birds Accurately Assesses the Timing and Magnitude of Migration through the Great Lakes. *Condor Ornithol. Appl.* **2014**, *116*, 371–383. [CrossRef]
94. Fiss, C.J.; Lapp, S.; Cohen, J.B.; Parker, H.A.; Larkin, J.T.; Larkin, J.L.; Kitzes, J. Performance of Unmarked Abundance Models with Data from Machine-Learning Classification of Passive Acoustic Recordings. *Ecosphere* **2024**, *15*, e4954. [CrossRef]
95. Winiarski, J.M.; Whitmore, S.A.; Wood, C.M.; Eiseman, J.P.; Netoskie, E.C.; Bieber, M.E.; Kramer, H.A.; Kelly, K.G.; McGinn, K.; Thompson, C.; et al. Passive Acoustic Monitoring Can Provide Insights into Occupancy Dynamics and Impacts of Disturbance for At-Risk Species. *Ecol. Appl.* **2026**, *36*, e70177. [CrossRef]
96. Salamon, J.; Bello, J.P. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Process. Lett.* **2017**, *24*, 279–283. [CrossRef]

97. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; Lempitsky, V. Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.
98. Bendale, A.; Boult, T. Towards Open Set Deep Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1563–1572.
99. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.