Review

Noise-Resilient Bioacoustics Feature Extraction Methods and Their Implications on Audio Classification Performance: Systematic Review

Geofrey Owino*, MSc; Bernard Shibwabo*, PhD

School of Computing and Engineering Sciences, Strathmore University, Nairobi, Kenya
*all authors contributed equally

Corresponding Author:

Geofrey Owino, MSc
School of Computing and Engineering Sciences
Strathmore University
P.O. Box 75584
Nairobi 00200
Kenya
Phone: 254 721913968
Email: geoffrey.owino@strathmore.edu

Abstract

Background: Bioacoustics classification plays a crucial role in ecological surveillance and neonatal health monitoring. Infant cry analysis can aid early health diagnostics, while ecological acoustics informs conservation. However, the presence of environmental noise, signal variability, and limited annotated datasets often hinders model reliability and deployment. Robust feature extraction and denoising techniques have become critical for improving model robustness, enabling more accurate interpretation of acoustic events across diverse bioacoustic domains under real-world conditions.

Objective: This review systematically evaluates advancements in noise-resilient feature extraction and denoising techniques for bioacoustics classification. Specifically, it explores methodological trends, model types, cross-domain transferability between clinical and ecological applications, and evidence for real-world deployment.

Methods: A systematic review was conducted by searching 8 electronic databases, including IEEE Xplore, ScienceDirect, Web of Science, ACM Digital Library, and Scopus, through December 2024. Eligible studies entailed audio-based classification models and applied empirical or computational evaluations of bioacoustics classification using machine learning or deep learning methods. In addition, studies also included explicit or implicit consideration of noise. Two reviewers independently screened studies, extracted data, and assessed quality. Risk of bias was assessed using a customized tool, and reporting quality was evaluated using the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) checklist.

Results: Of the 5462 records, 132 studies met the eligibility criteria. The majority (112/132, 84.8%) of studies focused on model innovation, with deep learning and hybrid approaches being the most dominant. Feature extraction played a critical role, with 96.2% (127/132) of studies clearly demonstrating feature extraction. Mel frequency cepstral coefficients, spectrograms, and filter bank-based representations were the most common feature representations. Nearly half (62/132, 47%) of the studies incorporated noise-resilient methods, such as adaptive deep models, wavelet transforms, and spectral filtering. However, only 14.4% (19/132) demonstrated real-world deployment across neonatal care and ecological field settings.

Conclusions: The integration of noise-resilient techniques has significantly improved classification performance, but real-world deployment and proper use of denoising strategies in various datasets remain limited. Cross-domain synthesis reveals shared challenges, including dataset heterogeneity, inconsistent reporting, and reliance on synthetic noise. Future work should prioritize harmonized benchmarks, cross-domain generalization, and deployment, as well as opportunities for transferability.

JMIR Biomed Eng 2025;10:e80089; doi: 10.2196/80089

Keywords: bioacoustics classification; noise robustness; feature extraction; denoising techniques; audio signal processing; machine learning; deep learning; real-world deployment

Introduction

Background

Bioacoustics, the study of sound produced by biological organisms, has become an essential tool for understanding ecological dynamics, monitoring biodiversity, and health diagnostics and monitoring [1]. Bioacoustics signals, for instance, birdcalls, marine mammal sounds, human sounds, and infant cries, provide information about species behavior, ecosystem health, and human well-being [2]. In neonatal care, infant cry analysis is explored as a noninvasive marker of health and a potential tool for early diagnostics and caregiver decision support. In ecological monitoring, passive acoustic sensors are increasingly deployed for biodiversity surveillance, species identification, and environmental assessment. Passive acoustic monitoring has been significant in tracking population dynamics and detecting anomalies in biological sound patterns [3].

Bioacoustics signals are also used in health care as noninvasive markers for diagnosing respiratory conditions, neurological disorders, and infections such as sepsis [4]. These signals are increasingly becoming central to digital health. Infant cry analysis is one of the emerging core areas in digital health. It is a practical avenue for early risk triage, remote monitoring, and real-time decision support in neonatal care [5]. Other pathological audio domains, such as lung sound classification for respiratory disease diagnosis, have also been systematically reviewed [6-8]. These reviews reinforce the importance of robust audio pipelines in clinical monitoring. Similarly, acoustic monitoring is crucial for species identification and biodiversity assessments, particularly in remote or inaccessible regions, and is focal to wildlife conservation [9]. Despite rapid progress, both clinical and ecological bioacoustics applications are constrained by one fundamental limitation, noise interference, which undermines the reliability and interpretability of classification models in real-world deployments.

The most persistent challenge in bioacoustics analysis is environmental noise contamination, which degrades signal quality and reduces classification accuracy. Noise arising from human activity, equipment artifacts, and overlapping acoustic sources complicates the extraction of meaningful features. Clinical environments are also acoustically challenged by alarms, caregiver speech, ventilation, and room reverberation. These factors reduce signal quality, thereby limiting the effectiveness of machine learning-based audio classification models [10]. Feature extraction forms the critical interface between raw bioacoustics waveforms and downstream classifiers. While traditional feature extraction techniques remain fundamental in audio classification, they exhibit high noise sensitivity, leading to feature distortion and reduced classification accuracy [11]. Numerous noise-resilient techniques such as wavelet filtering, adaptive spectral subtraction, and hybrid deep neural embeddings have been proposed to tackle these challenges. However, their evaluation remains fragmented and inconsistent across domains. No consensus exists regarding the most effective denoising or feature extraction strategies for bioacoustic classification, nor how these choices influence model deployment or interpretability under realistic noise conditions [10,12].

Persistent research gaps remain in evaluating the effectiveness and generalizability of noise-resilient feature extraction methods across domains. Many studies rely on controlled or synthetic noise settings, limiting ecological and clinical applicability. Benchmark initiatives such as Stowell's roadmap and the BirdSet dataset have advanced standardization in ecoacoustics but do not yet address cross-domain noise resilience or deployment metrics. Reporting of noise protocols and preprocessing remains inconsistent, and evidence of real-world deployment—especially in neonatal and field settings—is scarce. To bridge these gaps, this systematic review aims to (1) map methodological trends in noise-resilient feature extraction and denoising; (2) quantitatively evaluate their impact on classification performance under varying noise conditions; (3) examine evidence for real-world deployment and cross-domain generalization; and (4) identify limitations and future research priorities to advance robust, interpretable, and deployable bioacoustic systems.

Persistent research gaps remain in evaluating the effectiveness of noise-resilient feature extraction methods across different bioacoustics applications [13,14]. Many studies assess models in controlled or synthetic noise conditions, limiting ecological and clinical applicability as models fail to reflect the complexity of real-world acoustic environments [15]. Benchmark initiatives such as Stowell's roadmap explicitly call for community standards and comparable benchmarks in bioacoustics deep learning [13]. "The Benchmark of Animal Sounds," proposed by Hagiwara and colleagues to standardize evaluation across multiple animal-sound datasets [14], and a large-scale dataset for audio classification in avian bioacoustics, "BirdSet," were also created to address dataset fragmentation in avian tasks [16]. However, these efforts remain largely species-specific with no noise protocols, denoising baselines, or clinical (neonatal intensive care unit [NICU]) deployment metrics, underscoring the need for a minimal evaluation to enhance transition to deployment. Existing reviews largely focus on ecoacoustic pipelines and tasks rather than cross-domain noise robustness or deployment in clinical settings [17,18].

Related Work

In addition, reporting of noise protocols and preprocessing is inconsistent, limiting comparability in domains. Evidence on deployment is scarce, with only a minority of studies tested in neonatal or ecological field settings. Finally, little cross-domain synthesis exists to establish whether techniques effective in infant cry analysis generalize to ecological monitoring, and vice versa. To address these gaps, this

systematic review focused on four objectives: (1) mapping methodological trends in feature extraction, denoising, and model development; (2) evaluating classification performance under noisy conditions; (3) assessing evidence for deployment and cross-domain transferability; and (4) synthesizing limitations and future priorities to guide the development of robust, scalable bioacoustics systems.

Bioacoustic recordings across domains are degraded by environmental and clinical noise, limiting the reliability of feature extraction and classification techniques [19]. Noise interference remains a major obstacle in bioacoustics research, stemming from natural background sounds, overlapping vocalizations, human-induced disturbances, and equipment-related artifacts [20]. Low signal-to-noise ratios (SNRs) degrade the clarity of acoustic signals, making it difficult to extract meaningful features [21]. In urban environments, background noise from traffic, industrial activity, and human movement significantly reduces the accuracy of automated species identification. Similarly, in neonatal health care settings, excessive ambient noise negatively affects infant cry-based medical diagnostics, leading to misclassification and reduced sensitivity [22]. This section summarizes literature on feature extraction and denoising techniques to benchmark the gaps for data synthesis.

Feature extraction is a vital phase in bioacoustics classification; it transforms signals into meaningful representations for machine learning and deep learning models. Traditional methods such as Mel frequency cepstral coefficients (MFCCs), spectrograms, and linear predictive cepstral coefficients (LPCCs) have been widely used due to their effectiveness in capturing essential acoustic properties. MFCCs, in particular, have been extensively applied in speech and sound classification tasks due to their ability to model human auditory perception [23]. Spectrogram-based methods provide time-frequency representations, enabling the visualization and analysis of complex vocalization patterns [24]. LPCCs have been used in general acoustics research for feature extraction due to their capacity to model the vocal tract system in speech signals [25]. However, these feature techniques perform well in controlled environments but struggle with real-world noise.

High classification error rates arise when extracted features are distorted by background interference, reverberation, and overlapping signals [10]. These limitations necessitate continuous development of advanced noise-resilient feature extraction techniques as bioacoustics moves toward more complex field applications. Research is moving toward noise-resilient feature extraction methods that integrate signal processing, machine learning, and deep learning–based methods. These methods are objective in feature robustness enhancement, mitigation of noise artifacts, and improving classification accuracy in dynamic environments.

Denoising techniques have been used before feature extraction to enhance signal quality and after feature extraction to enhance model performance. Several techniques have been used extensively; among them, spectral subtraction, Wiener filtering, and wavelet-based denoising are used extensively. Spectral subtraction reduces stationary background noise by estimating the noise spectrum during nonvocalization periods and subtracting it from the noisy signal [26]. However, spectral subtraction can introduce artifacts such as musical noise, which may distort classification results, making it less effective for nonstationary noise [27]. Wiener filtering reduces the mean square error between the estimated clean signal and the observed noisy input, adapting to local SNRs [28]. It has been used successfully in bioacoustics monitoring and medical diagnostics, where background noise levels vary dynamically [4].

Wavelet-based denoising uses wavelet transforms to decompose data into distinct frequency bands. This technique reduces high-frequency noise while maintaining salient biological acoustic properties by selectively attenuating noise components at particular scales [21]. Marine bioacoustics has effectively used wavelet denoising to enhance the detection of low-frequency vocalizations, such as whale sounds, in noisy underwater environments [29]. Adaptive filtering dynamically adjusts its parameters in response to changing noise conditions, making it particularly effective for field-based bioacoustics monitoring [30]. Adaptive filtering has been used in avian bioacoustics, where real-time adjustments help maintain signal clarity despite weather fluctuations and overlapping birdcalls [3].

Advanced neural network architectures have shown significant improvements over conventional techniques for managing noisy bioacoustics data. Recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have been essential in this development. RNNs are well suited to modelling time-based relationships in sequential data, and CNNs excel at extracting spatial characteristics from spectrogram representations of audio signals. These networks improve classification accuracy in diverse acoustic situations by learning noise-invariant representations [31,32]. Hybrid models and their variants have improved classification accuracy in diverse noisy environments. Convolutional recurrent neural networks (CRNNs) combine the advantages of RNNs and CNNs by integrating temporal sequence modeling and spatial feature extraction, enabling CRNNs to efficiently identify intricate patterns in bioacoustics data, even in noisy environments [33].

Generative adversarial networks (GANs) have been used to improve model robustness by generating synthetic training data that simulate real-world noise conditions [34]. GANs allow models to learn from an additional, diverse set of scenarios, refining their generalization capabilities. Additionally, training datasets have been expanded through data augmentation techniques and contextual noise to improve classification performance [35]. Finally, incorporating noise-adaptive attention mechanisms into audio classification models allows selective focus on signal components that are less affected by noise, thereby enhancing classification performance. While these approaches often improve accuracy under noise, latency demands can hinder on-device or field deployment without model compression or edge-aware design.

Evaluation protocols vary widely, with some studies using synthetic overlays with fixed SNR grids while others use in situ recordings with uncontrolled noise. The evaluation metrics and reporting details differ substantially. Underreporting of noise types and inconsistent disclosure of denoising complicate cross-study comparisons and can inflate perceived robustness. Community efforts such as multidataset animal-sound benchmarks and large avian corpora have improved scale and comparability but rarely prescribe explicit noise protocols or denoising baselines [14,16]. Furthermore, there are no NICU-specific deployment metrics. These gaps motivate a tiered evidence strategy, core noise-resilient versus comparator pipelines, and a structured synthesis.

This systematic review aims to summarize existing literature, pinpoint performance patterns, and draw attention to research gaps in the development of classification models for noise-resilient bioacoustics. In order to guide future research toward more scalable, generalizable, and noise-resilient bioacoustics systems, this study attempts to address these issues and offer an organized overview of the topic. Furthermore, we synthesize the effect direction, transferability, and deployment evidence across infant-cry and ecological settings.

Objectives

The core objective of this study is to systematically review and synthesize advancements in noise-resilient bioacoustics feature extraction methods, evaluating their implications on audio classification performance in real-world noise. Specifically, we (1) map methodological trends (features, denoisers, models, and study designs); (2) quantify performance under noisy conditions relative to clean baselines; (3) assess cross-domain transferability and evidence of deployment (clinical, field, or edge); and (4) identify limitations and priorities to guide future research and implementation of robust, scalable bioacoustics classification systems.

To operationalize this objective aim, the review pursued four specific objectives: (1) identify methodological trends in feature extraction, denoising, and machine learning models applied to bioacoustics classification under noise; (2) evaluate performance outcomes reported across studies, including accuracy, precision, recall, and F_1-score, with attention to differences between clinical and ecological domains; (3) assess deployment evidence by analyzing whether and how methods have been tested or implemented in real-world conditions, and to what extent they demonstrate cross-domain robustness; and (4) synthesize limitations and future priorities, highlighting dataset challenges, methodological gaps, and opportunities for advancing noise-resilient bioacoustics analysis.

By integrating findings from multiple studies, the review seeks to provide practical recommendations for both academic research and real-world implementations, ensuring the development of more robust, scalable, and adaptive bioacoustics classification systems. Based on these objectives, the following review questions (RQs) were formulated to align closely with the study's scope and focus:

- RQ 1.1: What feature extraction, denoising or enhancement, and machine learning model approaches are used to achieve noise-resilient bioacoustics classification? This question synthesizes traditional signal-processing methods (eg, MFCC, LPCC, and per-channel energy normalization [PCEN]), denoisers (eg, spectral subtraction, Wiener, wavelet, and deep denoisers), and model classes (eg, CNN, RNN, CRNN, and transformers) and documents prevailing study designs.
- RQ 1.2: How do these pipelines perform under noisy conditions compared with clean baselines, and what metrics and noise protocols are reported? This question extracts accuracy, precision, recall, and F_1-score (and area under the receiver operating characteristic curve [AUC] where available); summarizes effect direction (Δ vs clean); and notes noise protocol transparency (type, SNR grids, and synthetic vs in situ).
- RQ 1.3: To what extent have these methods been deployed or prospectively evaluated in real-world settings, and how transferable are they across clinical (infant-cry) and ecological (wildlife) domains? This question examines model evaluation in ward, field, and edge environments; considers scalability and latency constraints; and assesses cross-domain robustness and generalizability.
- RQ 1.4: What limitations and risks of bias recur across studies, and what priorities should guide future work? This question identifies dataset imbalance, synthetic-only noise, reporting gaps (noise type, SNR, and denoising details), and distills priorities such as standardized noise protocols, benchmark design, and real-time or self- or federated-learning approaches.

Methods

Methodological Approach

This study follows a systematic review approach to analyze advancements in noise-resilient bioacoustics feature extraction methods and their implications on audio classification performance. To ensure transparency, reproducibility, and methodological rigor, this review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines for systematic reporting [36,37]. The Methodological Expectations of Cochrane Intervention Reviews (MECIR) standards were also used for study selection and evaluation [38]. The search and analysis were tailored using the PICO (population, intervention, comparison, outcome) framework to focus on studies relevant to the review objectives.

Search Strategy

Information Sources

A comprehensive search was executed across 8 electronic databases—IEEE Xplore, ScienceDirect, Google Scholar, Web of Science Core Collection, ACM Digital Library, Scientific Electronic Library Online, China National Knowledge Infrastructure, and Scopus—yielding 5462 records. The search targeted peer-reviewed journal

and conference papers published through 2024, in English and selected non-English (Spanish, Portuguese, Chinese, and French) languages. The search terms were developed based on the PICO framework in Multimedia Appendix 1, ensuring precision and relevance to the study's scope.

Population (P)

Terms targeting bioacoustics audio data, such as "bioacoustics," "animal vocalizations," "bird calls," "marine mammal sounds," "infant cries," and "biological acoustic signals."

Intervention (I)

Keywords related to noise-resilient feature extraction methods, including "noise-resilient feature extraction," "denoising techniques," "MFCC," "spectrogram," "convolutional neural networks (CNNs)," "recurrent neural networks (RNNs)," "hybrid models," and "attention mechanisms."

Comparison (C)

Keywords related to evaluating the performance of different noise-handling techniques, such as "spectral subtraction," "adaptive filtering," "augmentation," and "attention mechanisms," against baseline approaches.

Outcome (O)

Keywords emphasizing classification performance and robustness, such as "classification accuracy," "precision and recall," "robustness to noise," "scalability," and "real-world applications."

To cater to the non-English studies, the search terms were inadvertently translated into each target language, combined with controlled-vocabulary headings where available. This multilingual strategy ensured maximal coverage of relevant noise-resilient bioacoustics classification studies. In addition to translation, filters were set to yield non-English relevant languages in the specific languages. Boolean operators (AND, OR) were used to combine and refine terms, and search strings were adapted for each database. For instance, the search query for Google Scholar was "bioacoustics" OR "infant cry classification" AND "animal vocalization recognition" AND "feature extraction" AND ("MFCC" OR "spectrogram" OR "wavelet") AND "classification model" AND ("denoising" OR "noise robust" OR "signal enhancement"). The full search query syntax used for each database is provided in Multimedia Appendix 2.

Eligibility Criteria

Inclusion and exclusion criteria were clearly defined to ensure methodological consistency and relevance to the objectives of this review. Studies were considered eligible for inclusion in the review if they involved the classification of bioacoustics signals, such as those from animals, birds, marine mammals, or human infants, using feature extraction methods or denoising techniques in real-world or noisy environments.

To preserve both comprehensiveness and focus, we defined 2 tiers of evidence. Tier A entails all noise-resilient evidence from studies explicitly implementing or evaluating noise-resilient or denoising approaches (eg, spectral subtraction, Wiener filtering, wavelet filtering denoising, and deep learning–based enhancement), and it forms the primary evidence base for assessing robustness. Tier B entails comparator evidence from studies using standard or non–noise-resilient feature extraction techniques (eg, MFCCs and spectrograms) without explicit denoising. These were included to provide baseline comparisons and to highlight the gap since many bioacoustics studies still rely on such methods despite operating under noisy conditions.

Eligible studies had to present empirical or computational results using machine learning or deep learning–based classification models and report at least one standard performance metric such as accuracy, precision, robustness, or generalizability. In addition, the review included both primary and secondary data-based studies, as long as they provided sufficient methodological details regarding feature extraction and classification pipelines.

Studies were excluded if they did not involve biological acoustic signals or if they focused solely on speech or music processing unrelated to ecological or health contexts. Review articles, theoretical discussions without implementation or evaluation, and non–peer-reviewed sources such as preprints, editorials, or technical reports were also excluded. Finally, studies that failed to describe their dataset, feature extraction process, or performance evaluation methods in sufficient detail to permit meaningful analysis were omitted.

Protocol and Registration

The review was registered as required by PRISMA 2020 guidelines in the Open Science Framework (OSF) to enhance transparency. The review protocol was registered on August 16, 2025 (registration ID JKD5Y). The OSF record includes the prespecified objectives, eligibility criteria, data items, and the quantitative synthesis plan. Following peer-review feedback, certain objectives and research questions were refined to reduce overlap and improve clarity. These refinements did not alter the eligibility criteria, search strategy, or dataset. A deviation log has been added to the OSF record to transparently document these revisions without altering the original aims.

Study Selection

The study selection process followed the PRISMA guidelines to ensure transparency, reproducibility, and rigor. All retrieved records from the systematic search were imported into a reference management system, where duplicates were identified and removed. The non-English studies were machine-translated using Google Translate to support screening. Both reviewers independently cross-verified the translations against the original texts to minimize misinterpretation. There was also keen attention to the selected studies to ensure that the original papers were not later published in English to avoid omissions and double entries. The studies underwent a multistage screening process. In the initial stage, two independent reviewers performed a title and abstract screening to assess initial relevance. Any differences were resolved amicably through discussion, resulting in

a consensus mutually agreed upon by both reviewers, with escalation to a third reviewer if required. Studies that clearly failed to meet the inclusion criteria were excluded at this stage, and reasons were recorded.

In the subsequent stage, potentially eligible studies underwent a full-text review. Each study was assessed for methodological clarity, relevance to bioacoustics classification, use of feature extraction techniques, and evaluation in noisy or real-world conditions. Interrater reliability was assessed using Cohen κ at both screening stages. During the title and abstract screening, the reviewers achieved an observed agreement of 90.9%, corresponding to κ=0.79. At the full-text screening stage, the observed agreement was 94.7%, indicating almost perfect agreement with κ=0.89. Discrepancies at both stages were resolved through consensus. Of the 5462 records retrieved, 132 studies met the eligibility criteria and were selected for full review. The study selection process is summarized in a PRISMA flow diagram in the Results section. There are clear details on the screening process from the retrieved studies to the final selection of the sample of 132 studies for inclusion.

Data Extraction

We used a structured data extraction process to ensure reliability and comprehensiveness in capturing relevant study characteristics. A standardized Microsoft Excel spreadsheet was developed to systematically extract key information from each included study. The extraction form was designed to align with the objectives and review questions, capturing both methodological details and performance-related data. Data extracted for each study include:

1. Bibliographic details: authors, title, and year containing basic bibliographic information to uniquely identify and reference the studies.
2. Study design and setting: Including whether the study was experimental, comparative, or simulation-based, together with the domain (clinical infant cry versus ecological) and the context (NICU, field, or lab).
3. Dataset information: whether the dataset used was primary or secondary and its size, class distribution, and description.
4. Feature extraction techniques: Specific methods such as MFCCs, spectrograms, wavelets, and LPCCs, and advanced hybrid approaches, together with their key parameters.
5. Denoising techniques: information on whether denoising was applied and which methods were used, such as spectral subtraction, Wiener filtering, wavelet denoising, and other advanced methods.
6. Models and training: the classifier models used were also identified, such as machine learning, statistical, neural networks, and deep learning models.
7. Performance metrics and contrasts: key performance metrics such as classification accuracy, precision, and F_1-score, together with CIs or statistical tests if reported.
8. Application domain: the area of implementation, such as wildlife monitoring, health care, infant cry analysis, marine mammal detection, or smart sensing.
9. Deployment context: document real-world use, simulation, or proof-of-concept, and also reported challenges such as noise variability, data imbalance, or model generalizability where documented.
10. Where available, each study's future direction or proposed improvements were also extracted to identify research gaps and emerging priorities in noise-resilient bioacoustics analysis.

The extraction was conducted independently by 2 reviewers, with cross-validation to ensure reliability. Missing or unclear information was noted, and where necessary, corresponding authors were contacted for clarification. This structured approach enabled comprehensive synthesis and comparison across studies with diverse methodologies and application contexts.

Data Synthesis and Analysis

The extracted data were analyzed both qualitatively and quantitatively. Qualitative synthesis entailed identification of trends in noise-resilient methods, recurring challenges, emerging technologies, and synthesis of study findings to highlight advancements. On the other hand, quantitative summaries reported frequencies and distributions for classifier model classes, feature families, denoising techniques, deployment contexts, and performance metrics. Given heterogeneity in datasets, noise protocols, and outcomes, formal meta-analysis was not appropriate. We used a structured narrative approach: (1) group studies by feature, denoising family, model class, and domain (clinical vs ecological); (2) contrast performance under noise against baselines when available; (3) examine transferability and deployment evidence; and (4) integrate risk-of-bias and reporting-quality signals into interpretation.

To ensure methodological rigor and transparency, a dual quality assessment approach was adopted, combining both reporting quality and methodological bias evaluation. While the quality rating did not dictate the inclusion of studies, it aimed to present an outline of the reliability and transparency of the selected research. The TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) checklist was used to assess the intelligibility, completeness, and reproducibility of reporting in each study. Five key TRIPOD components were evaluated: title and abstract, introduction, methods, results, and discussion. Each component was scored as either compliant (1) or noncompliant (0), yielding a maximum possible score of 5.

The risk of bias across the studies was assessed to identify potential sources of bias in the reviewed studies. Given the machine learning focus of this review, domain-specific risk of bias criteria was applied to five core areas: (1) bias in data sources and sampling to check whether the data were representative, balanced, and appropriately selected; (2) bias in labeling and ground truth to check whether labels were accurate, consistent, and validated; (3) bias in feature extraction and preprocessing to check whether preprocessing and feature engineering introduced potential artifacts or limitations; (4) bias in model training and evaluation to check whether the training-validation-test split, metrics, and

evaluation protocols were appropriately implemented; and (5) bias in reporting and interpretation of results to check whether performance was selectively reported or overly generalized. Each domain was rated as "low," "moderate," or "high" risk of bias. The overall risk of bias was then derived from these domain-level assessments, with a deliberate distribution.

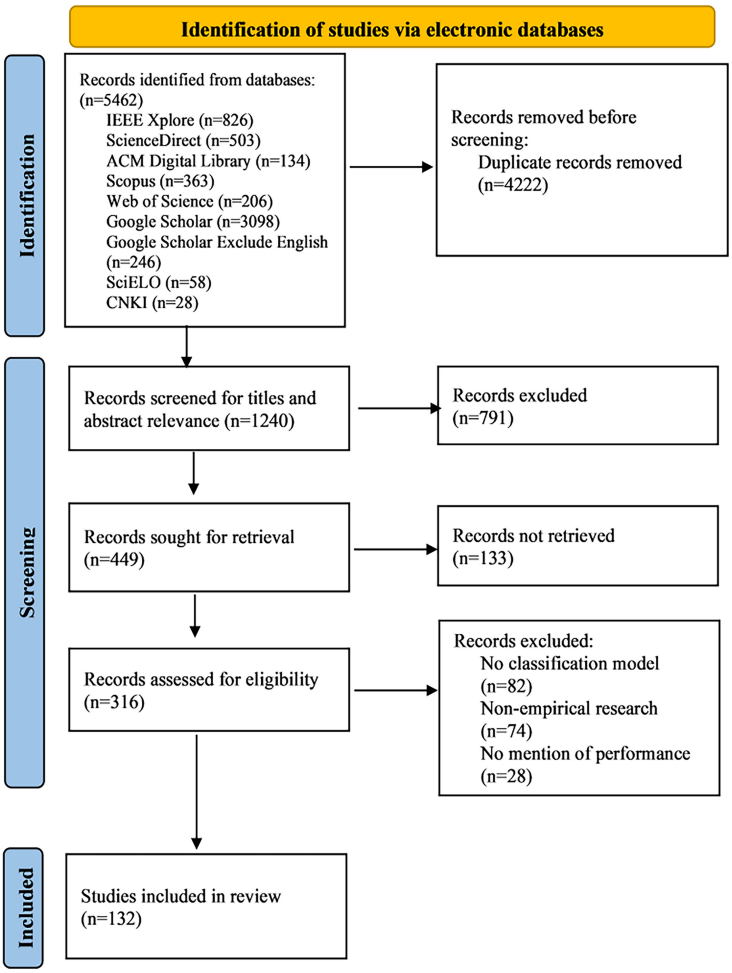
All assessments were conducted independently by 2 reviewers with consensus resolution. Importantly, neither TRIPOD nor risk of bias ratings determined study inclusion; rather, they informed interpretation by highlighting areas of greater or lesser methodological confidence. The numerical results of TRIPOD and risk of bias assessments are reported in the Results section. By combining the TRIPOD framework appraisal with domain-specific risk of bias, the quality assessment provided a robust evaluation of the selected studies' validity and reliability. This comprehensive approach ensured that the findings of this systematic review were built on a foundation of transparent, high-quality research.

Results

Study Selection

This review synthesized 132 studies (Multimedia Appendix 3) published between 2003 and 2024, spanning two primary application domains: ecological monitoring studies (n=80, 60.6%) and clinical infant cry studies (n=52, 39.4%). The study selection process is summarized in a PRISMA flowchart in Figure 1. Studies were further stratified into tier A, comprising noise-resilient pipelines with explicit denoising or robustness testing strategies, 47% (n=62) of studies, and tier B, comprising comparator pipelines without explicit denoising, 53% (n=70) of studies. This distribution highlights both the predominance of ecological applications and the substantial proportion of studies still relying on non–noise-resilient baselines. To establish the reliability of the evidence base, we first summarize the outcomes of the reporting quality (TRIPOD) and risk of bias assessments. Findings are then presented in 5 sections: research focus, methodological trends, performance outcomes, deployment and cross-domain transferability, and limitations with future priorities.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.



Reporting Quality and Risk of Bias

The TRIPOD checklist revealed that all reviewed studies (n=132) demonstrated excellent reporting standards, achieving a perfect compliance score (5/5, 100%). This clearly indicates that titles and abstracts, introductions, methods, results, and discussions were consistently reported in line with transparency standards. This reflects a strong cultural shift in the bioacoustics and audio classification community toward structured and reproducible reporting practices. While TRIPOD compliance was universal, a high score largely captures surface-level reporting standards (eg, presence of sections and completeness of description) rather than deeper methodological rigor. In practice, studies varied in how clearly they justified feature extraction choices, described preprocessing, or documented evaluation protocols. This suggests that, although reporting has become standardized, interpretive caution is still required when assessing methodological robustness.

The risk of bias evaluation across all included studies revealed strong methodological rigor overall, with most domains rated as low risk. However, a small proportion of studies exhibited moderate or high risks in specific areas. The risk of bias assessment results across the studies in each domain are presented in Table 1.

Table 1. Risk of bias values across various domains.

Risk of bias	Low	Moderate	High
Bias in data sources and sampling	127	3	2
Bias in labeling and ground truth	127	3	2
Bias in feature extraction and preprocessing	126	2	4
Bias in model training and evaluation	125	5	2
Bias in reporting and interpretation of results	130	0	2
Overall risk of bias	116	10	6

Bias in data sources and sampling was rated low in 96.2% (127/132) of the studies, indicating that the studies used clearly documented datasets with appropriate sampling strategies. However, 2.3% (3/132) [39-41] and 1.5% (2/132) [42,43] were rated as moderate and high variability due to a lack of clear discussion on sample size and sample selection strategies.

Bias in labeling and ground truth was rated low in 96.2% (127/132) of the studies, reflecting strong adherence to consistent annotation practices. Bias in feature extraction and preprocessing was rated low in 95.5% (126/132) of the studies, suggesting a high degree of transparency in preprocessing protocols. However, 1.5% (2/132) [44,45] and 3.4% (4/132) [12,46-48] were rated as moderate and high, largely due to a lack of justification for chosen features and unclear preprocessing steps.

Bias in model training and evaluation was rated low in 94.7% (125/132) of the studies, demonstrating widespread adoption of sound training practices. A small proportion (5/132, 3.8%) [49-53] were rated as moderate, while 1.5% (2/132) [35,54] were rated as high, due to improper validation schemes and related design weaknesses.

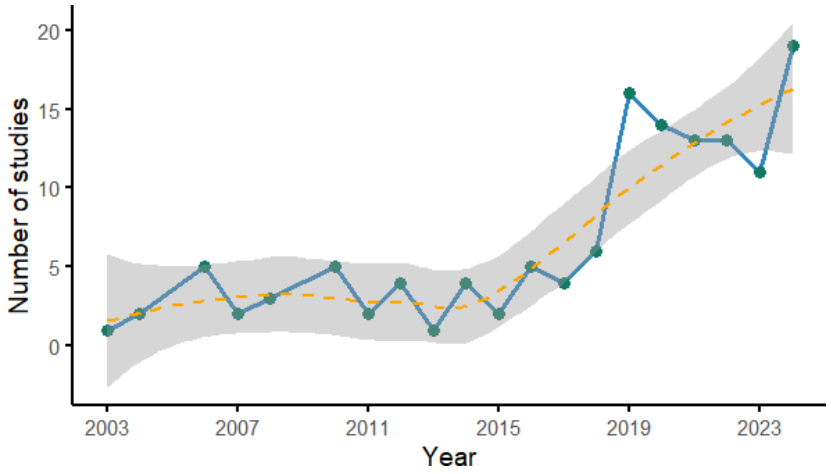
Bias in reporting and interpretation was rated low in 98.5% (130/132) of the studies, indicating that most studies provided transparent and well-supported results. However, 1.5% (2/132) [55,56] were rated high, mainly due to lack of clarity in reporting key results.

Overall risk of bias was rated low in 87.9% (116/132) of the studies, highlighting the generally high methodological rigor across the reviewed literature. A notable proportion of 7.6% (10/132) [40-44,50-52,57,58] were rated as moderate, while 4.5% (6/132) [12,35,48,54-56] were rated as high, often due to cumulative concerns across multiple bias domains. Detailed per-study risk of bias ratings can be found in Multimedia Appendix 4. Although the corpus is predominantly low risk of bias, the small cluster of moderate or high ratings concentrates in preprocessing justification and evaluation rigor. In subsequent results, we interpret performance and robustness claims with greater weight placed on low-risk studies and flag results from studies with methodological gaps where relevant.

Trends and Research Focus

Research in noise-resilient bioacoustics has expanded rapidly since 2019, with most contributions centered on model innovation, while noise robustness and deployment remain underrepresented. The reviewed studies reflect a growing momentum in the field of noise-resilient bioacoustics, demonstrated by a pronounced upward trend in publications over the past decade. The annual trend distribution in Figure 2 illustrates steady growth, with a notable increase in the number of publications from 2019 onward, peaking in 2024 with 14.4% (19/132). This surge coincides with the uptake of deep learning and larger annotated datasets.

Figure 2. Trend of the number of publications per year.



Overall, 65.2% (86/132) were published between 2019 and 2024, highlighting a recent surge in research activity motivated by advancements in machine learning, particularly deep learning architectures, and an increased availability of publicly accessible, annotated acoustic datasets. This rapid expansion underscores the field's responsiveness to technological innovation and its potential for addressing practical challenges. The field is shifting adeptly from merely experimental exploration to a mainstream research agenda in ecological monitoring, wildlife conservation, and infant cry monitoring.

In terms of contribution types, the vast majority of studies (112/132, 84.8%) focused on model innovation, primarily through the design of novel architectures and algorithms for bioacoustics classification. These included deep learning approaches such as CNNs, RNNs, CRNNs, and transformer-based models. Hybrid frameworks combined traditional signal processing techniques, for example, MFCCs and spectral features, with neural networks. Within this category, some studies emphasized architectural novelty, for example, attention mechanisms and temporal-context modeling, while others explored optimization strategies such as regularization, hyperparameter tuning, or multimodal feature fusion. Feature selection and engineering were addressed in 43.9% (58/132) of studies, emphasizing the role of extracting relevant and informative features to improve classification accuracy.

Noise robustness and generalization were explicitly explored in 28.8% (38/132) of studies, which incorporated denoising techniques, noise-aware training, and evaluation across diverse acoustic environments to improve real-world performance. Finally, only 14.4% (19/132) of studies reported deployment-focused applications, demonstrating implementations in wildlife conservation zones, smart farming, NICUs, and edge-based monitoring systems.

The field remains heavily weighted toward architectural innovation, with robustness testing and deployment underrepresented. This imbalance highlights a translational gap—methodological advances are plentiful—but their practical application in real-world bioacoustics is still emerging.

Methodological Landscape

The methodological landscape across the reviewed studies showcases a strong emphasis on empirical evaluation, consistent with the practical and performance-driven nature of noise-resilient bioacoustics research. Every study was categorized as experimental, involving the development, training, and testing of machine learning and signal processing models on bioacoustics datasets. The models were carefully developed, and their performance was evaluated to assess model generalization. In addition to an experimental foundation, 36.4% (48/132) were comparative studies, systematically benchmarking multiple models or feature extraction pipelines under controlled noise conditions. These studies were instrumental in benchmarking traditional versus advanced techniques and identifying optimal configurations for noisy environments.

A small subset (20/132, 15.2%) of studies were also descriptive, providing detailed explanations of the models they implemented alongside empirical evaluations. This is vital for the growing research and learning era. New researchers are able to learn from what has already been done to implement improvements. Across all methodological types, studies demonstrated a commitment to reproducibility, with datasets and detailed parameter settings provided. However, the lack of standardized evaluation frameworks and consistent reporting practices remains a limitation, hindering comparability across studies. It is therefore evident that empirical research has matured broadly, but there is a continuing need for standardized methodologies to enhance comparability and real-world applicability.

Feature Extraction Techniques

Feature extraction was nearly universal across 96.2% (127/132) of studies, with cepstral features forming the foundation of most bioacoustic classification pipelines. Spectral, temporal, and wavelet-based features served complementary roles. The distribution of the feature extraction methods across the studies per domain is presented in Table 2. The percentage distribution of each feature category in relation to the domain, as well as the tier category,

is also presented to show relative variation between the domains.

Table 2. Distribution of feature extraction methods across the studies per domain (N=132).

	Number of studies, n (%)				
Feature type	Tier A	Tier B	Infant cry	Ecology	Total
Cepstral features	36 (27.3)	37 (28)	39 (29.5)	34 (25.8)	73 (55.3)
Filter bank and spectral representations	36 (27.3)	29 (22)	8 (6.1)	57 (43.2)	34 (25.8)
Spectral features	26 (19.7)	15 (11.4)	11 (8.3)	30 (22.7)	41 (31.1)
Temporal or time domain features	17 (12.9)	21 (15.9)	21 (15.9)	17 (12.9)	38 (28.8)
Prosodic features	8 (6.1)	8 (6.1)	5 (3.8)	11 (8.3)	16 (12.1)
Wavelet features	9 (6.8)	1 (0.8)	4 (3)	6 (4.5)	10 (7.6)

Cepstral features were the predominant category used (73/132, 55.3%), with MFCCs alone appearing in 50.8% (67/132) of studies. These features were widely favored for their ability to capture perceptually relevant sound components, closely aligned with human auditory perception. Variants such as LPCCs, constant-Q cepstral coefficients, and gammatone cepstral coefficients, often enhanced with derivatives (Δ, ΔΔ) and feature fusion strategies, were also used. Of the 132 studies, infant cry consisted of 39 (29.5%) studies, while ecology consisted of 34 (25.8%) studies. The distribution was nearly equal, indicating that the use of cepstral features in both domains was broadly comparable across applications and tiers.

Filter bank and spectral representations were also common, being used in 34 (25.8%) of the 132 studies. However, the use of these features was skewed toward the ecological domain, showing that ecological studies used these presentations in their modeling. Spectral features (38/132, 28.8%), including spectral centroid, roll-off, and entropy, quantified frequency energy distributions and were valuable for detecting anomalies in vocalizations. Similarly, the use of spectral features was skewed toward ecological application. Temporal features (41/132, 31.2%), such as zero-crossing rate, root mean square energy, and voicedness, captured time-domain behaviors and proved particularly useful in infant cry analysis for identifying cry phases and sharp transitions. Prosodic features (16/132, 12.1%) focused on pitch and intonation contours, offering insights into emotional or health-related states. Wavelet-based features (10/132, 7.6%), derived from transformations such as the discrete wavelet transform or the wavelet packet transform, were used to capture transient and nonstationary characteristics, particularly enhancing noise robustness in ecological monitoring tasks.

A small subset (n=5) [27,56,59-61] of the studies did not explicitly report a predefined feature extraction step but instead relied on the model architecture itself to learn and extract relevant features directly from the raw waveform. These approaches typically use end-to-end deep learning models, such as raw waveform CNNs, which are designed to learn spectral and temporal representations directly from the audio signal during training. These techniques are essential when developing a fully automated pipeline. However, end-to-end waveform learning without explicit feature extraction raises concerns regarding interpretability, computational cost, and data requirements.

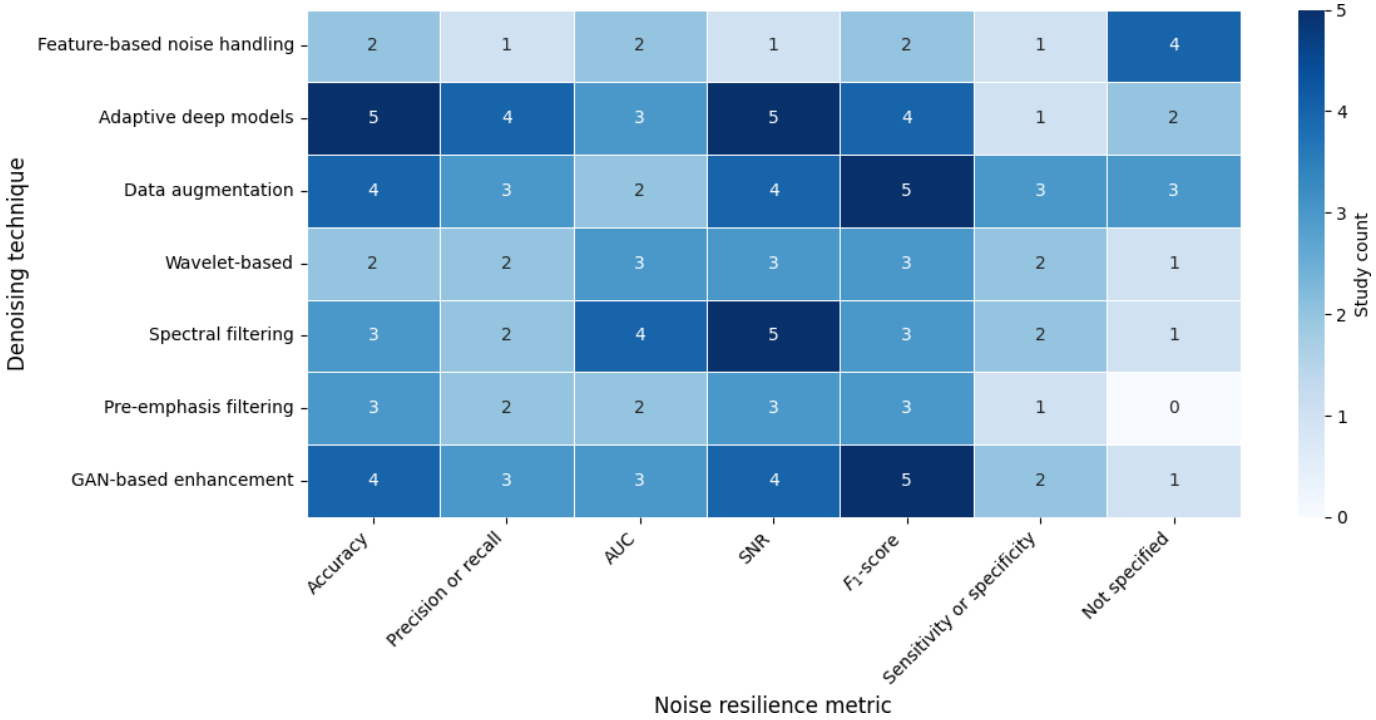
When comparing domains, infant cry studies leaned heavily on cepstral and prosodic features, reflecting the speech-like and emotionally driven nature of cries. MFCCs and intonation contours were most frequently used to capture subtle variations in vocal tone linked to health or emotional states. In contrast, ecological monitoring studies applied a broader mix of spectral, temporal, and wavelet features to represent the diversity of animal calls and environmental soundscapes. These choices highlight the domain-driven adaptation of feature extraction. It was evident that studies in tier A concentrate on filter bank, log mel, and spectral descriptors. These feature families align perfectly with denoising applications. Tier B studies, however, inflate cepstral feature use since there was no explicit denoising.

Feature extraction emerged as a cornerstone of noise-resilient bioacoustics classification. Cepstral features dominate current practice, while spectral, temporal, and prosodic features provide complementary insights. Wavelets offer noise-robust representations, and end-to-end models mark an emerging direction toward automation. Together, these approaches illustrate a balance between established feature engineering and exploratory deep learning–based representation learning.

Denoising Techniques

Nearly half (62/132, 47%) of the reviewed studies presented denoising application in the modeling pipeline tier A, and more than half (76/132, 57.6%) of the studies presented use of a noise-resilient metric to assess model robustness. While traditional signal processing methods remain common, advanced deep learning–based denoising is gaining traction, though still underrepresented. A visual representation of the distribution of these methods is presented in Figure 3.

Figure 3. Distribution of noise-resilient metrics and denoising application across the studies. AUC: area under the receiver operating characteristic curve; GAN: generative adversarial network; SNR: signal-to-noise ratio.



Darker shading indicates higher study counts. Adaptive deep models and GAN-based enhancement were most frequently paired with robust evaluation metrics such as F_1-score and SNR degradation, while classical approaches (eg, pre-emphasis and spectral filtering) relied more on accuracy and AUC alone. Studies omitting denoising often reported only accuracy, highlighting a reporting gap between baseline pipelines and noise-resilient methods.

Traditional denoising approaches rooted in classical signal processing were used in 25% (33/132) of studies. These techniques included pre-emphasis filters to suppress low-frequency noise, spectral subtraction, Butterworth high-pass filters, and windowing techniques. Adobe Audition and WavePad Sound Editor were also used for manual noise reduction and audio cleanup. Transformations such as the fast Fourier transform, discrete wavelet transform, and wavelet packet transform were leveraged to enhance feature robustness against noise, together with energy-based descriptors like root mean square energy, zero-crossing rate, and segmentation techniques that also supported noise minimization.

Advanced deep learning–based denoising techniques were used in 16.7% (22/132) of the studies, marking a shift toward more adaptive and context-aware noise handling. These approaches included the use of stage-wise GANs for structured denoising [62], PCEN for real-time noise suppression [3], and deep CNNs trained with a pretext to enhance resilience [63]. A portion of the studies used contextual metadata-aware CNNs [56], dimensionality reduction via YAMME [50], or custom neural denoisers like DS-Denoiser [63] and Burn Layer noise injection strategies [48].

In 12.9% (17/132) of studies, noise resilience was achieved indirectly through strategic feature design and training methodologies rather than explicit denoising. These included data augmentation with controlled noise injection [64], spectrogram normalization [65], entropy-based descriptors, and frame-based segmentation to reduce the impact of transient background noise [66]. Several studies introduced false-positive distractors during training to improve model discrimination [35,55], while others used SNR-aware evaluation metrics [35,67] and principal component analysis to filter out irrelevant variation [68].

Noise-resilient metrics included standard evaluation tools such as AUC, F_1-score, precision, recall, and accuracy, often reported across multiple SNR levels (eg, 100 dB, 3 dB, 0 dB, and –3 dB) to capture degradation effects [35,69,70]. Some studies used equal error rate or Earth Mover's Distance to assess alignment between predictions and ground truth under distortion [65]. Studies also introduced custom fitness metrics that weighted false positives caused by noise more heavily or used domain-specific indicators like perceptual evaluation of speech quality and false alarm rates [56]. These metrics were crucial for evaluating not just raw classification accuracy, but also how robustly the models maintained performance in realistic and adverse audio conditions.

Denoising strategies also diverged across domains. Infant cry studies often applied classical noise-reduction methods such as spectral subtraction and Wiener filtering to handle consistent background noise in hospitals or home environments. More recent works explored denoising autoencoders to improve robustness in clinical deployment. Ecological monitoring, by contrast, dealt with far more heterogeneous noise sources, including overlapping species, wind, and rain. As a result, adaptive filtering and multiband denoising approaches were common, enabling resilience to highly variable outdoor acoustic conditions. The ecological field

has gone into extreme detail to ensure features for model development are not affected by environmental noise.

It is evident that the clinical field is heavily dependent on classical denoising and occasionally AUC and F_1-score metrics, while robustness testing was less frequent. However, in ecology, there is greater use of deep or indirect denoising and systematic evaluation across SNR levels, reflecting highly variable outdoor noise. Noise-resilient evaluation was largely confined to tier A pipelines. Studies that skipped denoising (tier B) also rarely reported robustness metrics, inflating apparent performance. Infant cry pipelines showed limited robustness testing, while ecology studies drove innovation in both denoising and noise-resilient evaluation frameworks.

Classifier Architectures and Performance

A diverse range of classification models was used across the reviewed studies, reflecting both the evolution of machine learning techniques and the complexity of bioacoustics data. The distribution of classifier architectures is presented in Table 3.

Table 3. Distribution of classifier architectures per domain (N=132).

	Number of studies, n (%)				
Model family	Tier A	Tier B	Infant cry	Ecology	Total
Traditional machine learning	27 (20.5)	20 (15.1)	23 (17.4)	24 (18.2)	47 (35.6)
CNN[a]	18 (13.6)	21 (15.9)	7 (5.3)	32 (24.2)	39 (29.5)
CRNN[b] or hybrid	2 (1.5)	2 (1.5)	2 (1.5)	2 (1.5)	4 (3)
Deep neural network	13 (9.8)	13 (9.8)	7 (5.3)	19 (14.4)	26 (19.7)
Classical neural network	2 (1.5)	10 (7.6)	11 (8.3)	1 (0.8)	12 (9.1)
Transformer	0 (0)	1 (0.8)	0 (0)	1 (0.8)	1 (0.8)

[a]CNN: convolutional neural network.
[b]CRNN: convolutional recurrent neural network.

Traditional machine learning architectures dominated the reviewed literature in both ecological monitoring and infant cry analysis. More than half of the studies (70/132, 53%) reported accuracies ≥90%, with CNN-based approaches most frequently associated with high performance. Traditional models such as support vector machines (SVMs), *k*-nearest neighbors, decision trees, Gaussian mixture models, and Naive Bayes were used in 48.5% (64/132) of the studies, with SVMs being used in 24.2% (32/132) of the reviewed studies. These models typically relied on handcrafted features like MFCCs and LPCCs and showed decent performance under low-noise or controlled conditions but often struggled in the presence of complex noise or overlapping signals.

In contrast, deep learning models appeared in 53% (70/132) of the studies and formed the dominant category. CNNs, RNNs, long short-term memory, and their hybrids (eg, CRNNs) were frequently used since they can automatically learn features from raw data. Advanced models—ResNet, EfficientNet, and DenseNet—offered high performance with transfer learning advantages. The CNN model was used in 42.4% (56/132) of the studies.

Classical neural networks, including multilayer perceptrons, time-delay neural networks, and probabilistic neural networks, were seen in 22.7% (30/132) of studies, while 24.2% (32/132) used hybrid or ensemble models, such as CNN + RNN architectures or transformer-based pipelines. These advanced approaches were particularly suited for handling real-world noise, variability in signal patterns, and generalizing across datasets, making them ideal for deployment in bioacoustics monitoring systems.

The performance of these models was centered on classification accuracy, with several studies also reporting precision, recall, and F_1-score. Most studies (96/132, 72.7%) reported classification accuracies exceeding 85%, with 53% (70/132) achieving 90% or higher. High-accuracy models were typically based on deep learning architectures, particularly CNNs, CRNNs, and transformer variants. Accuracy was generally enhanced when models incorporated noise-aware training, denoising preprocessing, or attention mechanisms.

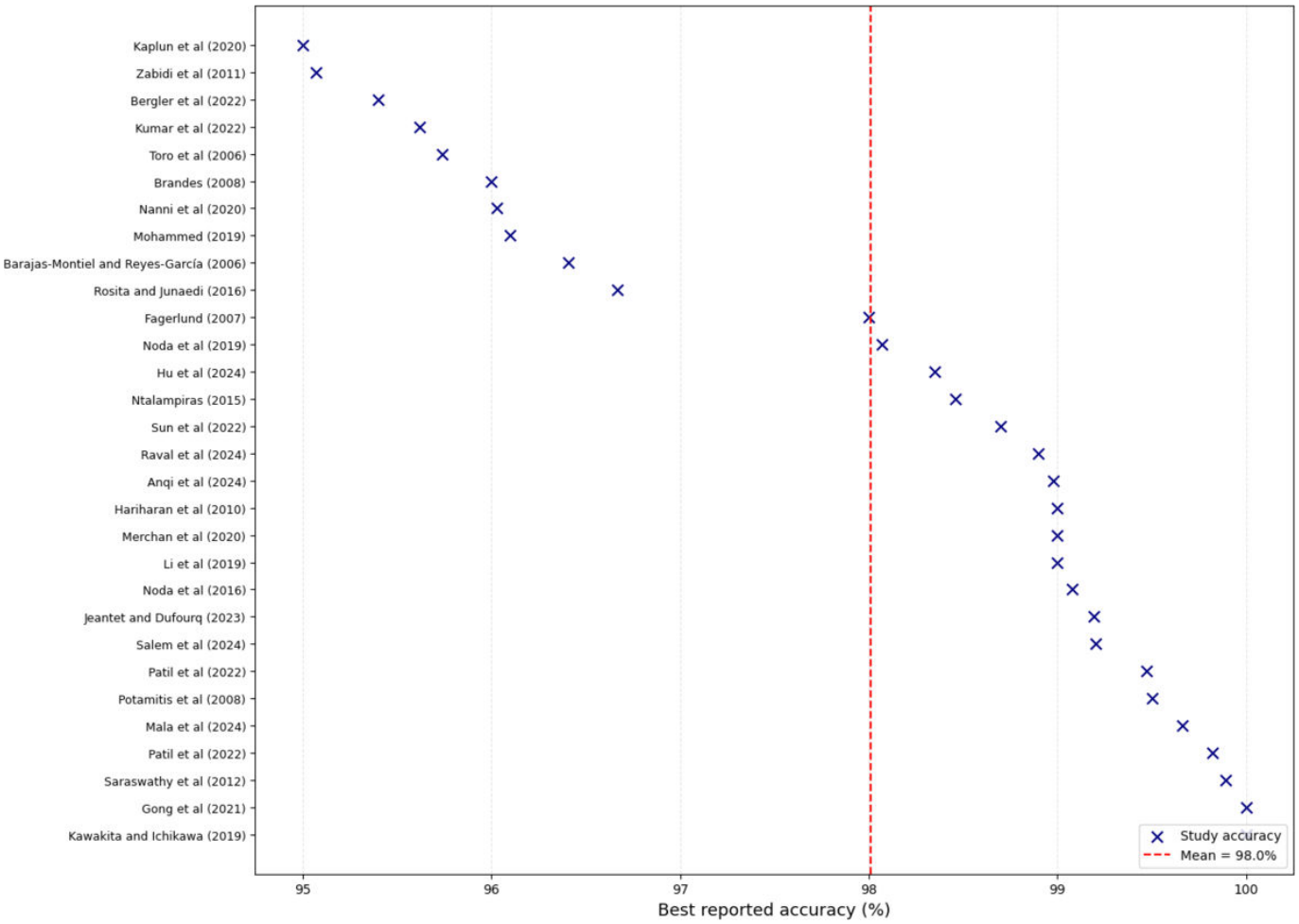
Models using traditional machine learning techniques (eg, SVMs and decision trees) tended to report lower accuracies, often in the 70%-85% range, especially when tested under real-world acoustic conditions. However, in low-noise or synthetic scenarios, these models performed comparably well. In studies that evaluated precision and recall, scores were typically balanced, often above 0.8, especially in binary classification tasks. However, multiclass classification scenarios showed slightly reduced precision in species-rich datasets, often due to class imbalance or overlapping vocalizations.

Studies leveraging ensemble methods or hybrid networks showed some of the best overall performance, with AUC values as high as 0.96 and accuracy consistently above 92% when evaluated on diverse and noisy bioacoustics datasets. Notably, some studies used post hoc statistical analysis such as the Nemenyi test, ANOVA, or CIs to validate model significance across different noise conditions or experimental configurations.

A forest plot of the best-reported accuracies in Figure 4 illustrates the performance clustering of the majority between 95% and 100%. This reflects the strong classification potential of modern bioacoustics models across domains. The clustering near 99% indicates a ceiling effect in reported results. Most of these values originate from tier

B pipelines evaluated under clean or synthetic conditions, while tier A pipelines tested under noisy ecological conditions reported more variable results (approximately 75%-95%). This discrepancy highlights that reported best-case accuracies often reflect optimized conditions rather than real-world robustness.

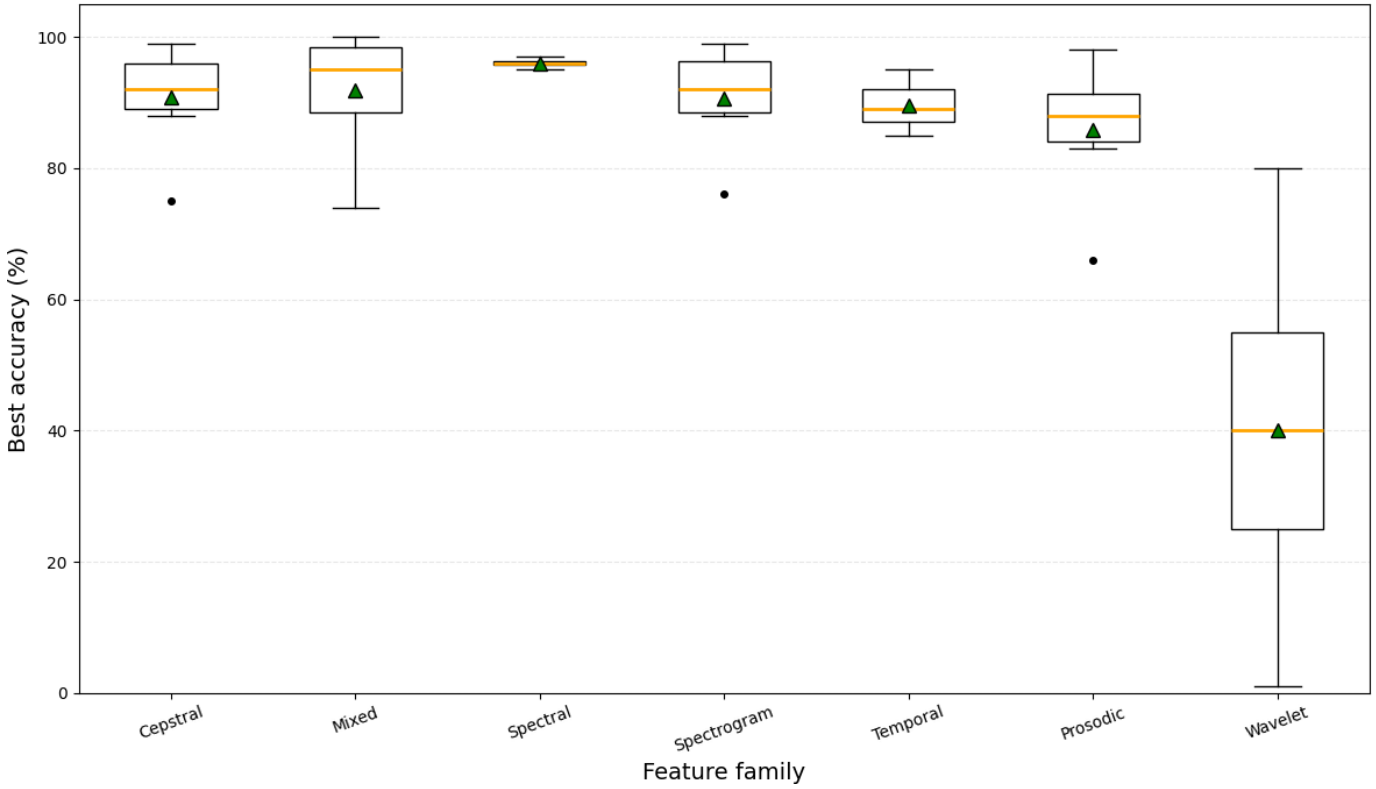
Figure 4. Forest plot of best-reported accuracies reporting the top 30 studies [27,51,58,64,66-91].



Infant cry studies often reported >95% accuracies under controlled conditions, while wildlife monitoring required more extensive preprocessing or noise-handling strategies to achieve comparable results. These findings highlight both the promise of bioacoustics classification and the need for standardized reporting of performance variability across noise levels and datasets.

Figure 5 compares the best-reported accuracies across feature families. Cepstral, spectrogram-based, and mixed feature sets clustered above 90%, confirming their central role in bioacoustics classification. However, tier A pipelines achieved these results under noisy conditions when using spectrogram or log-mel representations, while tier B pipelines often reported inflated accuracies from cepstral-only inputs under clean settings. Temporal features produced moderately strong outcomes but showed greater variance, particularly in infant cry studies. Wavelet-based features exhibited the greatest spread (0%-80%), reflecting their experimental use in ecological tier A pipelines for transient, nonstationary noise. These results suggest that while cepstral and spectrogram-based features remain the most reliable overall, robustness under realistic noise depends on whether the pipeline incorporates explicit tier A resilience strategies.

Figure 5. Distribution of accuracy against feature family.



In tier A pipelines, deep learning models dominated, particularly CNNs and CRNNs, which together accounted for nearly two-thirds of ecological studies. These were typically paired with noise-resilient features such as log-mel spectrograms or PCEN, enhancing robustness across variable acoustic conditions. By contrast, tier B pipelines were skewed toward traditional machine learning, and classical neural networks were most often applied with MFCCs. These models frequently reported strong results in clean or synthetic conditions, but robustness to real-world noise was rarely evaluated.

In domain comparison, infant cry studies leaned heavily on interpretable and computationally efficient approaches, with traditional machine learning used in 44.2% (23/52) of pipelines and classical neural networks in 21.2% (11/52), while CNNs were fewer, at 13.5% (7/52). Most of these pipelines were tier B baselines, reflecting a focus on clinical interpretability and resource efficiency over robustness. Ecological studies, in contrast, showed stronger adoption of CNNs, used in 40% (32/80) of pipelines, and deep neural networks, used in 23.8% (19/80), particularly within tier A pipelines. Transformers were rare and appeared only in ecology and tier B pipelines in 1.25% (1/80), reflecting early experimentation with sequence models. It is therefore evident that robust tier A ecological pipelines favored deep CNN and CRNN models with noise-resilient features, while infant cry pipelines remained anchored in tier B baselines combining MFCCs with traditional machine learning or classical neural networks. This contrast highlights a trade-off between robustness and interpretability across domains.

Performance reporting was dominated by classification accuracy, though many studies supplemented it with F_1-score, precision, recall, or AUC. Most studies (96/132, 72.7%) reported accuracies ≥85%, with more than half (70/132, 53%) ≥90%. High-performing models were typically deep learning architectures (CNNs, CRNNs, and transformers). Tier A pipelines consistently tested performance under noisy conditions and reported smaller accuracy drops across SNR levels (typically 5%-10%). Tier B pipelines rarely incorporated noise protocols and often reported inflated best-case accuracies (>95%), reflecting performance under clean or synthetic conditions rather than realistic robustness.

Infant cry studies frequently reported very high accuracies (>95%), but these were predominantly from tier B baselines using MFCC + traditional machine learning or CNN in controlled NICU or home environments. Few infant cry studies tested performance in truly noisy or cross-population conditions, limiting confidence in their generalizability. Ecological studies, by contrast, showed a wider performance spread (approximately 75%-95%), reflecting more diverse datasets, taxa, and recording environments. Tier A ecological pipelines that incorporated denoising and spectrogram or PCEN features frequently exceeded 90% accuracy, but results were more variable due to dataset complexity and nonstationary noise. It is evident that reported accuracies cluster near ceiling values, but these reflect tier B clean-condition pipelines more than tier A robustness evidence. Infant cry studies appear stronger on paper but are less often validated under noise, whereas ecological tier A pipelines, though more variable, provide the most convincing demonstrations of resilience under realistic acoustic conditions.

Quantitative Analysis of Performance

Of the 132 included studies, 82.6% (n=109) reported classification accuracy, while 21.2% (n=28) reported F_1-scores. Accuracy was the dominant performance indicator, particularly in traditional and early deep learning approaches, whereas F_1-score appeared more often in recent studies emphasizing class balance in imbalanced datasets. Across all studies, the mean accuracy was 89.47% (SD 30.82%) with a median of 93.64%, ranging from 2% to 343%, while the mean F_1-score was 93.04% (SD 7.86%) with a median of 95.88%, spanning 71%-100%. These results indicate generally high predictive capability across bioacoustic classification models, though the wide variation in accuracy reflects methodological diversity in dataset size, preprocessing techniques, and evaluation strategies. The overall distribution of quantitative findings is presented in Table 4.

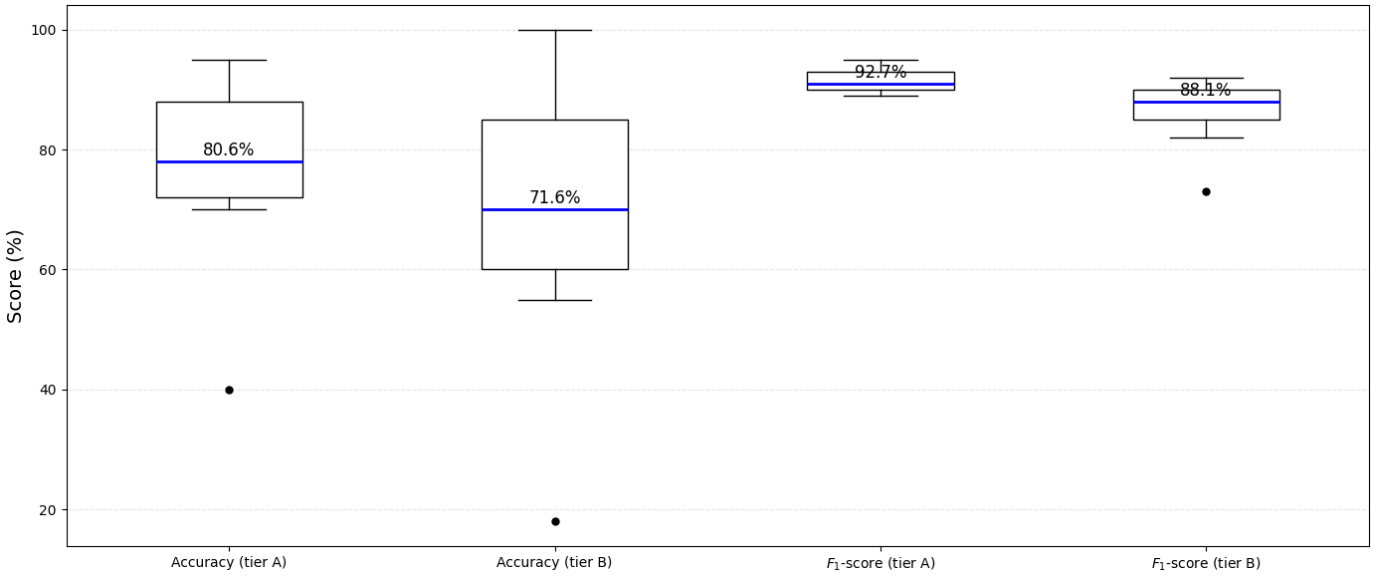
Table 4. Summary of performance metrics across tier A and tier B studies.

Metric	Tier	Studies, n	Mean, %	Median, %	SD	Min	Max	95% CI (lower-upper)	*P* value (versus tier A)
Accuracy	A	52	83.64	90.5	22.51	2	100	77.6-89.7	Reference
Accuracy	B	57	94.79	94	36.19	20	343	85.3-104.3	.04
F_1-score	A	9	93.49	92.6	5.35	87	100	89.1-97.8	Reference
F_1-score	B	19	92.83	98	8.93	71	100	88.6-97.1	.67

Statistical comparison revealed a significant difference in accuracy between the 2 tiers (P=.04), confirming that tier B models achieve superior accuracy overall. However, no significant difference was observed in F_1-scores (P=.67), suggesting that while denoising enhances general classification accuracy, it does not consistently alter the precision-recall trade-off.

Tier A models achieved a higher mean accuracy (94.79%) compared to tier B (83.64%), suggesting the benefit of integrating denoising and noise-resilient feature extraction methods. However, tier B exhibited greater variability (SD 36.19) than tier A (SD 22.51), indicating that while noise-resilient models often achieve superior results, their performance may depend heavily on implementation quality and dataset characteristics. The F_1-scores of both tiers were relatively consistent, averaging 93.49% for tier A and 92.83% for tier B, implying that noise handling primarily improves robustness rather than precision-recall balance. The comparative distribution of performance metrics across tiers A and B is illustrated in Figure 6. The box plot summarizes the spread and central tendency of both accuracy and F_1-score values, highlighting that denoising generally elevates overall performance yet increases score variability. The clear clustering of F_1-scores around the upper quartile further confirms the stability of the precision-recall balance across studies.

Figure 6. Distribution of model accuracy and F_1-score across tier A and tier B studies.



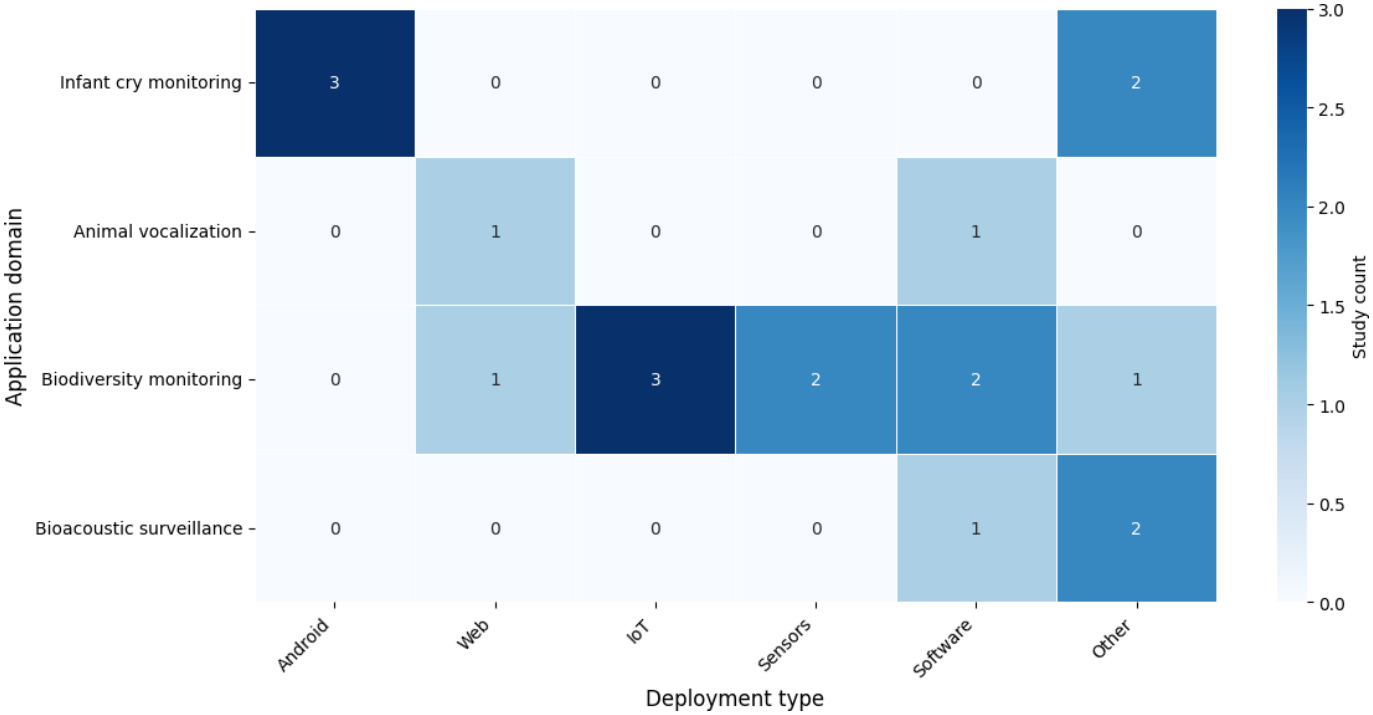
In domain-specific comparison, infant cry research reported a higher overall mean accuracy of 92.8% and a mean F_1-score of 94.6%, with relatively low variability, reflecting the controlled recording settings, smaller class counts, and limited background interference characteristic of clinical datasets. In contrast, ecological studies exhibited broader score dispersion, with accuracy values ranging from 70% to 95% and F_1-scores between 80% and 96%, indicating greater heterogeneity due to environmental noise, overlapping species vocalizations, and larger taxonomic class sets. Tier B (noise-resilient) ecological studies achieved modest gains in mean accuracy (+4.2%) compared to tier A, though with higher standard deviation, underscoring the impact of denoising complexity in natural soundscapes. Conversely,

infant cry models benefited less from denoising interventions, maintaining stable performance even under tier A configurations.

Deployment and Application Domains

A small number of studies (19/132, 14.3%) discussed the potential deployment of models in real-world environments. Figure 7 shows how these studies were deployed in various bioacoustic domains. It shows that infant cry deployments were primarily smartphone-based caregiver tools, whereas ecological applications emphasized the Internet of Things (IoT) and sensor networks for biodiversity monitoring. Bioacoustic surveillance deployments were rare. Importantly, almost all deployment cases arose from tier A studies, underscoring that robustness evidence is a prerequisite for translation.

Figure 7. Distribution of deployment across various bioacoustic domains. IoT: Internet of Things.



A significant portion (7/132, 5.3%) involved tool-based deployments, including Android apps [92,93], web interfaces [64,88], and user-friendly platforms for real-time monitoring and human verification [13,71,94,95]. IoT and embedded systems were featured in 4.5% (6/132) of studies, leveraging low-power devices [27,49,96,97], sensor networks [3,72], and mobile hardware for field applications. Open-source software [98] solutions such as DeepSqueak [63] and ORCA-SPOT [99] were used in 2.3% (3/132) of studies, and 1.5% (2/132) of studies used limited offline toolkits such as MATLAB's Neural Network Toolbox [100,101]. The deployed studies were distributed across the various applications of bioacoustics classification technologies. Infant monitoring systems (5/132, 3.8%) focused on detecting cries associated with health conditions or needs, supporting early diagnosis and caregiver response.

Animal vocalization monitoring (2/132, 1.5%) aimed to detect and classify specific species calls, contributing to behavioral and ecological research. Biodiversity monitoring (9/132, 6.8%) represented the largest category, with deployments targeting broad-scale species tracking, conservation efforts, and habitat assessment in diverse ecosystems. Lastly, bioacoustics surveillance (3/132, 1.7%) focused on monitoring environmental soundscapes for human-induced or unusual acoustic events, supporting real-time situational awareness and management in protected or sensitive areas.

Infant cry deployments (5/132, 3.8%) focused on caregiver and clinical support, such as smartphone apps and hospital monitoring tools. These pipelines emphasized real-time cry detection for diagnosis and caregiver response but were limited by data privacy, ethical constraints, and the need for interpretability. Ecological deployments (14/132, 10.6%) concentrated on scalability, leveraging IoT and embedded systems for biodiversity monitoring, conservation surveillance, and edge-based species detection. Tools such as DeepSqueak and ORCA-SPOT exemplified open-source tier A systems tailored to diverse and noisy outdoor environments.

Deployment patterns further underline the distinct priorities of each domain. Infant cry research emphasized caregiver support through hospital tools and smartphone apps, focusing on real-time cry detection and monitoring for clinical or home use. Ecological monitoring prioritized scalability, leveraging IoT sensor networks, embedded low-power devices, and open-source tools such as ORCA-SPOT for biodiversity tracking. Whereas infant cry deployments aim for individualized, human-centered decision support, ecological deployments are oriented toward large-scale, automated monitoring across ecosystems. It was also

evident that deployed studies are a portion of tier A, underscoring that noise robustness is a precondition for real-world deployment. Deployment in the real world is directly related to denoising, and studies with no implicit denoising did not translate to deployment.

Despite these promising efforts, a primary limitation reported across studies was the lack of large, high-quality, standardized datasets for both clinical and ecological domains. This gap restricted generalization, with most deployments validated in narrow or pilot settings. Failures and constraints were often tied to dataset variability, hardware limitations, and energy efficiency trade-offs, underscoring the need for more robust field trials, benchmark datasets, and harmonized evaluation protocols to achieve sustainable real-world applicability.

Beyond technical feasibility, deployment in sensitive domains requires attention to ethical, interpretability, and infrastructural concerns. In neonatal care, noise-resilient models must safeguard patient privacy and provide transparent outputs that clinicians and caregivers can trust. Similarly, ecological monitoring systems need explainable decisions to ensure transparency in conservation policy and sustainability of automated surveillance. These considerations highlight that deployment success depends not only on accuracy but also on responsible integration into clinical and environmental workflows.

Challenges and Future Direction

Despite significant advancements in noise-resilient bioacoustics classification, several recurring challenges continue to hinder progress. A primary limitation reported across studies was the lack of large, high-quality, and standardized datasets, mentioned in 34.8% (46/132) of studies. Researchers relied on small datasets, which limited the generalizability of findings and the ability to compare models across studies. Diverse audio samples were unavailable for various species in varying recording environments, therefore restricting models from performing reliably in real-world scenarios. In addition to the limited data available, datasets were small and imbalanced, which contributed greatly to biased and overfit models.

Noise interference and acoustic variability were mentioned as a challenge in 20.5% (27/132) of studies. Studies highlighted the difficulty of extracting clean signals in field conditions, especially with background noise from human activity, equipment, or other animals. Despite some attempts using denoising and noise-aware training, many models struggled to maintain robustness under nonstationary and low signal-to-noise conditions. Additionally, inconsistencies in labeling arising from semisupervised annotations introduced noise into ground truth data, reducing model reliability.

Several deep learning approaches, especially CNNs and hybrid models, required high-performance computing resources, posing a challenge for real-time deployment. High computational costs and dependence on platform-specific tools also posed barriers to scalable and accessible deployment. Deployment was observed in very few studies despite the advancements in technology. This is due to hindrances by a lack of platform compatibility, difficulty integrating models into systems, and challenges related to real-time processing, energy efficiency, concerns over hardware requirements, and user-friendliness.

A portion of the studies (11/132, 8.3%) reported cases of overfitting, especially due to limited data for complex models, while 6.1% (8/132) reported inconsistencies in data due to variability in signal quality by recording instruments. Other challenges reported were domain transfer challenges with models trained on one species, lack of open set recognition, and false positives in some models.

Infant cry studies were constrained by small, private datasets due to ethical and privacy concerns, limiting cross-population generalizability. Ecological monitoring faced challenges with data imbalance, as rare species were underrepresented, and annotation required expert input. Both domains therefore underscore the urgent need for larger, standardized, and openly available datasets, but with differing solutions: ethical data-sharing frameworks for infant cries versus coordinated biodiversity databases for ecological monitoring.

Looking forward, many studies have emphasized the need to expand datasets across taxa, habitats, and call types, especially for underrepresented classes such as infant cries from pathological conditions to rare animal vocalizations [13,102,103]. Researchers also recommend developing semisupervised and unsupervised labeling strategies to reduce annotation burden, improving noise robustness through signal enhancement modules [73,104]. In addition, transfer learning, domain adaptation, and transformer-based architectures were proposed for better generalization [74,75]. Several studies proposed real-time deployment strategies, calling for lightweight, energy-efficient models suitable for edge computing environments [55,63].

Finally, researchers highlighted the importance of open-set recognition, anomaly detection in dynamic acoustic environments, and model interpretability, especially in health care or conservation settings. Incorporating animal-independent denoising mechanisms, optimizing data augmentation for species-specific acoustics, and refining clustering techniques for individual or dialect-level recognition were among the key future directions. Together, these efforts aim to make bioacoustics systems more scalable, reliable, and ecologically meaningful, ultimately enabling widespread deployment in biodiversity monitoring, pest detection, and early diagnosis of health conditions.

Discussion

Principal Findings

This systematic review synthesizes evidence from 132 studies on noise-resilient bioacoustics classification and provides an integrated perspective across methodologies, performance outcomes, and deployment contexts. The central finding is that high reported accuracies do not necessarily equate

to robustness. Instead, robustness emerges from tier A pipelines—those combining explicit denoising or resilience testing with modern feature representations and architectures. By contrast, tier B pipelines, though numerous, often reported near-perfect accuracies under clean conditions but rarely progressed toward real-world deployment. This distinction frames our interpretation of the evidence against the review objectives.

Methodological Advances

Recent years have seen a shift from handcrafted features and statistical classifiers toward deep architectures capable of capturing temporal and spectral dependencies. CRNNs, CNNs, and in some cases, transformers consistently outperformed classical machine learning under noise [76,105], echoing trends in both ecoacoustics [17] and audio enhancement research [10]. However, our synthesis shows that the real methodological gap lies not in model availability but in evaluation design. Tier B pipelines often prioritized architectural novelty but omitted robustness testing, inflating performance claims. Tier A studies, while fewer, demonstrated that rigorous evaluation across SNR levels or noise-injected datasets yields more credible, if variable, results [106]. This confirms that methodological progress in bioacoustics must be judged not only by model choice but also by the framework of validation.

Feature Extraction and Denoising

Feature use reflected domain priorities: infant cry pipelines emphasized cepstral and prosodic features for speech-like cues [107], while ecological pipelines favored spectrogram, filter bank, and wavelet features to capture diverse soundscapes [77]. Importantly, our review shows that feature choice alone was insufficient as robustness depended on pairing features with denoising or noise-aware training. Classical filters, for example, Wiener and spectral subtraction, were common in infant cry studies [108], while ecology led the adoption of deep denoisers, augmentation strategies, and PCEN [3,21]. These practices align with broader advances in audio processing [10] but remain inconsistently applied. The insight here is that robustness is not feature-intrinsic but emerges from the integration of features, denoising, and evaluation metrics.

Performance Outcomes

Reported accuracies clustered around 95%-100%, creating the impression of ceiling-level performance. Yet, these results were largely driven by tier B pipelines tested in clean conditions, especially in infant cry datasets [47,109]. Tier A studies, particularly in ecological monitoring, reported more variable accuracies (approximately 75%-95%) because they were evaluated under realistic noise conditions [64,104]. This variance is not a weakness but evidence of genuine robustness testing. It highlights the risk of publication bias: inflated best-case results dominate the literature, while average-case resilience is underreported. Interpreting these outcomes, therefore, requires caution. The broader implication is that progress in bioacoustics cannot be judged by peak accuracy alone, but by the consistency of performance under noise.

Deployment and Translation

Deployment was reported in only 14.3% (19/132) of studies, nearly all from tier A pipelines. Infant cry applications emphasized mobile apps and caregiver support tools [110,111], prioritizing interpretability and immediacy but facing constraints around data privacy and ethics. Ecological deployments leveraged IoT networks, sensors, and open-source platforms to enable scalable biodiversity monitoring [61,98,112]. Bioacoustics surveillance deployments were rare. The absence of tier B baselines in deployment confirms that robustness is a prerequisite for translation. Domain-specific contrasts are clear: neonatal pipelines must prioritize ethical safeguards and clinician trust, while ecological pipelines require scalability, automation, and energy efficiency.

Limitations and Future Direction

Limited standardized, high-quality datasets in both clinical and ecological domains restricted comparability. Although we included non-English studies, reliance on automated translation may have introduced subtle interpretive inaccuracies, though independent reviewer checks mitigated this risk. The inclusion of these non-English studies did not change the direction of findings, as their reported outcomes were consistent with the broader evidence base.

To advance sustainable integration, bioacoustics research should (1) standardize evaluation by adopting shared benchmarks, harmonized SNR protocols, and open datasets across taxa and infant populations [13,14]; (2) strengthen robustness methods, extending lightweight denoisers, augmentation strategies, and federated learning to support real-world generalization [77]; (3) tailor deployment strategies, interpretability, and privacy-preserving approaches for neonatal monitoring [78,113]; and (4) foster cross-domain transfer: ecological augmentation strategies can inform infant cry robustness, while clinical interpretability standards can guide ecological applications [64,98].

The translational relevance of these findings extends beyond research. In clinical contexts, robust infant cry classification could support early diagnostics and caregiver decision-making. In ecology, noise-resilient monitoring systems can enhance biodiversity surveillance and conservation policy [98,112]. Future studies should explicitly bridge domains, evaluating not only technical performance but also usability, interpretability, and sustainability in deployment.

Conclusions

This review demonstrates that progress in bioacoustics classification is shaped less by the abundance of models than by the rigor of robustness evaluation. Tier A pipelines that incorporated explicit denoising and resilience testing provided the most credible evidence of real-world applicability, while tier B baselines, though often reporting high accuracies, rarely translated into deployment. Domain-specific contrasts further underscore that infant cry pipelines must prioritize interpretability and privacy, whereas ecological systems require scalable, energy-efficient designs.

Looking ahead, 3 levels of priority emerge. Immediate priorities include the creation of standardized, noise-augmented benchmark datasets and consistent reporting of preprocessing and denoising protocols. Short-term goals involve systematic evaluation of feature-model pairings across infant cry and ecological applications, coupled with pilot deployment studies in neonatal and field monitoring settings. Longer-term priorities focus on scaling deployment through cross-domain generalization methods (eg, transfer learning and federated learning), the development of lightweight edge-ready models, and the integration of interpretability and privacy safeguards for sustainable adoption.

By integrating insights on feature extraction, denoising, model architectures, and deployment, this review advances a cross-domain understanding of noise-resilient bioacoustics and provides a roadmap for future research. Moving beyond peak accuracies toward consistent robustness across diverse acoustic conditions will be key to translating methodological advances into reliable digital health and biodiversity conservation tools, with noise resilience as the cornerstone of sustainable impact.

Acknowledgments

Generative artificial intelligence tools were used only for language polishing and structural improvements under author supervision. All content was reviewed and verified by the authors prior to submission.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The datasets supporting this systematic review are available in Multimedia Appendix 3, and more details are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

PICO framework of the research questions. PICO: population, intervention, comparison, outcome.
[DOCX File (Microsoft Word File), 15 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Databases and search terms used in the search and selection of reviewed studies.
[DOCX File (Microsoft Word File), 16 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Data synthesis.
[XLSX File (Microsoft Excel File), 103 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Databases and search terms used in the literature search.
[XLSX File (Microsoft Excel File), 33 KB-Multimedia Appendix 4]

Checklist 1

PRISMA checklist.
[DOCX File (Microsoft Word File), 61 KB-Checklist 1]

Checklist 2

TRIPOD checklist.
[DOCX File (Microsoft Word File), 21 KB-Checklist 2]

References

1. Blumstein DT, Mennill DJ, Clemins P, et al. Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. J Appl Ecol. Jun 2011;48(3):758-767. [doi: 10.1111/j.1365-2664.2011.01993.x]
2. Sueur J, Farina A. Ecoacoustics: the ecological investigation and interpretation of environmental sound. Biosemiotics. Dec 2015;8(3):493-502. [doi: 10.1007/s12304-015-9248-x]

3. Lostanlen V, Cramer A, Salamon J, et al. BirdVoxDetect: large-scale detection and classification of flight calls for bird migration monitoring. IEEE/ACM Trans Audio Speech Lang Process. 2024;32:4134-4145. [doi: 10.1109/TASLP.2024.3444486]
4. Chen X, Hu M, Zhai G. Cough detection using selected informative features from audio signals. In: 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). IEEE; 2021. [doi: 10.1109/CISP-BMEI53629.2021.9624379]
5. Shayegh SV, Tadj C. Deep audio features and self-supervised learning for early diagnosis of neonatal diseases: sepsis and respiratory distress syndrome classification from infant cry signals. Electronics (Basel). 2025;14(2):248. [doi: 10.3390/electronics14020248]
6. Alqudah AM, Moussavi Z. A review of deep learning for biomedical signals: current applications, advancements, future prospects, interpretation, and challenges. Comput Mater Contin. 2025;83(3):3753-3841. [doi: 10.32604/cmc.2025.063643]
7. Sfayyih AH, Sabry AH, Jameel SM, et al. Acoustic-based deep learning architectures for lung disease diagnosis: a comprehensive overview. Diagnostics (Basel). May 16, 2023;13(10):1748. [doi: 10.3390/diagnostics13101748] [Medline: 37238233]
8. Sfayyih AH, Sulaiman N, Sabry AH. A review on lung disease recognition by acoustic signal analysis with deep learning networks. J Big Data. 2023;10(1):101. [doi: 10.1186/s40537-023-00762-z] [Medline: 37333945]
9. Aide TM, Corrada-Bravo C, Campos-Cerqueira M, Milan C, Vega G, Alvarez R. Real-time bioacoustics monitoring and automated species identification. PeerJ. 2013;1:e103. [doi: 10.7717/peerj.103] [Medline: 23882441]
10. Xie J, Colonna JG, Zhang J. Bioacoustic signal denoising: a review. Artif Intell Rev. Jun 2021;54(5):3575-3597. [doi: 10.1007/s10462-020-09932-4]
11. Gibbons A, King E, Donohue I, et al. Generative AI-based data augmentation for improved bioacoustic classification in noisy environments. arXiv. Preprint posted online on Dec 2, 2024. [doi: 10.48550/arXiv.2412.01530]
12. Gupta G, Kshirsagar M, Zhong M, Gholami S, Ferres JL. Comparing recurrent convolutional neural networks for large scale bird species classification. Sci Rep. Aug 24, 2021;11(1):17085. [doi: 10.1038/s41598-021-96446-w] [Medline: 34429468]
13. Stowell D. Computational bioacoustics with deep learning: a review and roadmap. PeerJ. 2022;10:e13152. [doi: 10.7717/peerj.13152] [Medline: 35341043]
14. Hagiwara M, Hoffman B, Liu JY, Cusimano M, Effenberger F, Zacarian K. BEANS: the benchmark of animal sounds. In: 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2023. [doi: 10.1109/ICASSP49357.2023.10096686]
15. Priyadarshani N, Marsland S, Castro I. Automated birdsong recognition in complex acoustic environments: a review. J Avian Biol. May 2018;49(5). [doi: 10.1111/jav.01447]
16. Raucha L, Schwinger R, Wirth M, et al. BirdSet: a large-scale dataset for audio classification in avian bioacoustics. arXiv. Preprint posted online on Mar 15, 2024. [doi: 10.48550/arXiv.2403.10380]
17. Nieto-Mora DA, Rodríguez-Buritica S, Rodríguez-Marín P, Martínez-Vargaz JD, Isaza-Narváez C. Systematic review of machine learning methods applied to ecoacoustics and soundscape monitoring. Heliyon. Oct 2023;9(10):e20275. [doi: 10.1016/j.heliyon.2023.e20275]
18. Kohlberg AB, Myers CR, Figueroa LL. From buzzes to bytes: a systematic review of automated bioacoustics models used to detect, classify and monitor insects. J Appl Ecol. Jun 2024;61(6):1199-1211. [doi: 10.1111/1365-2664.14630]
19. Mutanu L, Gohil J, Gupta K, Wagio P, Kotonya G. A review of automated bioacoustics and general acoustics classification research. Sensors (Basel). Oct 31, 2022;22(21):8361. [doi: 10.3390/s22218361] [Medline: 36366061]
20. Apol CA, Valentine EC, Proppe DS. Ambient noise decreases detectability of songbird vocalizations in passive acoustic recordings in a consistent pattern across species, frequency, and analysis method. Bioacoustics. May 3, 2020;29(3):322-336. [doi: 10.1080/09524622.2019.1605310]
21. Kiskin I, Zilli D, Li Y, Sinka M, Willis K, Roberts S. Bioacoustic detection with wavelet-conditioned convolutional neural networks. Neural Comput Appl. Feb 2020;32(4):915-927. [doi: 10.1007/s00521-018-3626-7]
22. Ji C, Mudiyanselage TB, Gao Y, Pan Y. A review of infant cry analysis and classification. EURASIP J Audio Speech Music Process. Dec 2021;2021(1):8. [doi: 10.1186/s13636-021-00197-5]
23. Muda L, Begam M, Elamvazuthi I. Voice recognition algorithms using Mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. arXiv. Preprint posted online on Mar 22, 2010. [doi: 10.48550/arXiv.1003.4083]
24. Knight EC, Poo Hernandez S, Bayne EM, Bulitko V, Tucker BV. Pre-processing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks. Bioacoustics. May 3, 2020;29(3):337-355. [doi: 10.1080/09524622.2019.1606734]

25. Dhonde SB, Jagade SM. Feature extraction techniques in speaker recognition: a review. Int J Recent Technol Mech Electr Eng. 2015;2(5):104-106. URL: https://www.ijrmee.org/download/feature-extraction-techniques-in-speaker-recognition--a-review-1433397000.pdf [Accessed 2025-11-18]
26. Boll S. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans Acoust, Speech, Signal Process. Apr 1979;27(2):113-120. [doi: 10.1109/TASSP.1979.1163209]
27. Kumar R, Gupta M, Ahmed S, Alhumam A, Aggarwal T. Intelligent audio signal processing for detecting rainforest species using deep learning. Intell Autom Soft Comput. 2022;31(2):693-706. [doi: 10.32604/iasc.2022.019811]
28. Scalart P, Filho JV. Speech enhancement based on a priori signal to noise estimation. In: 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. IEEE; 1996:629-632. [doi: 10.1109/ICASSP.1996.543199]
29. Vickers W, Milner B, Risch D, Lee R. Robust North Atlantic right whale detection using deep learning models for denoising. J Acoust Soc Am. Jun 2021;149(6):3797-3812. [doi: 10.1121/10.0005128] [Medline: 34241455]
30. Napier T, Ahn E, Allen-Ankins S, Schwarzkopf L, Lee I. Advancements in preprocessing, detection and classification techniques for ecoacoustic data: a comprehensive review for large-scale passive acoustic monitoring. Expert Syst Appl. Oct 2024;252:124220. [doi: 10.1016/j.eswa.2024.124220]
31. Piczak KJ. Environmental sound classification with convolutional neural networks. In: 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE; 2015. [doi: 10.1109/MLSP.2015.7324337]
32. Tang G, Liang R, Xie Y, Bao Y, Wang S. Improved convolutional neural networks for acoustic event classification. Multimed Tools Appl. Jun 2019;78(12):15801-15816. [doi: 10.1007/s11042-018-6991-4]
33. Sang J, Park S, Lee J. Convolutional recurrent neural networks for urban sound classification using raw waveforms. In: 26th European Signal Processing Conference (EUSIPCO). IEEE; 2018. [doi: 10.23919/EUSIPCO.2018.8553247]
34. Madhu A, K. S. EnvGAN: a GAN-based augmentation to improve environmental sound classification. Artif Intell Rev. Dec 2022;55:6301-6320. [doi: 10.1007/s10462-022-10153-0]
35. Salamon J, Bello JP, Farnsworth A, Kelling S. Fusing shallow and deep learning for bioacoustic bird species classification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2017:141-145. [doi: 10.1109/ICASSP.2017.7952134]
36. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. Mar 29, 2021;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
37. Page MJ, Moher D, Bossuyt PM, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ. Mar 29, 2021;372:n160. [doi: 10.1136/bmj.n160] [Medline: 33781993]
38. Higgins JPT, Lasserson T, Chandler J, Tovey D, Churchill R. Methodological Expectations of Cochrane Intervention Reviews. Cochrane; 2016. URL: https://methods.cochrane.org/sites/methods.cochrane.org/files/uploads/Cochrane%20MECIR_Standards%20FINAL%20booklet_web_version.pdf [Accessed 2025-11-26]
39. Castro J, Vargas-Masís R, Alfaro-Rojas D. Entendiendo el Desempeño Variable en el Marco de Trabajo MIL Profundo para la Detección Acústica de Aves Tropicales. Rev Tecnol En Marcha. 2020;33(5):49-54. [doi: 10.18845/tm.v33i5.5075]
40. Sabitha R, Poonkodi P, Kavitha MS, Karthik S. Premature infant cry classification via deep convolutional recurrent neural network based on multi-class features. Circuits Syst Signal Process. Dec 2023;42(12):7529-7548. [doi: 10.1007/s00034-023-02457-5]
41. Castro Ramírez A. Construcción de una red neuronal artificial para clasificar cantos de aves: una aplicación de la inteligencia artificial a la biología [Master's thesis]. Universidad de Costa Rica; 2006. URL: https://repositorio.sibdi.ucr.ac.cr/handle/123456789/1088 [Accessed 2025-11-18]
42. Benitez Labori GJ, Escobedo Beceiro DI. Clasificación del llanto en neonatos utilizando una red neuronal artificial con parámetros acústicos cuantitativos. Orange J. 2021;3(5):4-9. [doi: 10.46502/issn.2710-995X/2021.5.01]
43. Mala BM, Darandale SS. Effective infant cry signal analysis and reasoning using IARO based leaky Bi-LSTM model. Comput Speech Lang. 2024;86:101621. [doi: 10.1016/j.csl.2024.101621]
44. Reyes-García CA, Torres-García AA, Ruiz-Díaz MA. Extracción de Características Cualitativas del Llanto de Bebé y su Clasificación para la Identificación de Patologías Utilizando Modelos Neuro-Difusos [Article in Spanish]. Proc Nat Cong Biomed Eng. 2018;5(1):106-109. URL: https://memoriascnib.mx/index.php/memorias/article/view/582 [Accessed 2025-11-18]
45. Chen D, Lin J, Yi X, et al. Classification of underwater acoustic signals based on wavelet packet time-frequency map features and convolutional neural network [in Chinese]. Tech Acoust. 2021;40(3):336-340. [doi: 10.16300/j.cnki.1000-3630.2021.03.006]
46. Bashiri A, Hosseinkhani R. Infant crying classification by using genetic algorithm and artificial neural network. Acta Med Iran. 2020;58(10):531-539. [doi: 10.18502/acta.v58i10.4916]

47. Chang CY, Bhattacharya S, Raj Vincent PMD, Lakshmanna K, Srinivasan K. An efficient classification of neonates cry using extreme gradient boosting-assisted grouped-support-vector network. J Healthc Eng. 2021;2021:7517313. [doi: 10.1155/2021/7517313] [Medline: 34804460]
48. Hassan E, Elbedwehy S, Shams MY, Abd El-Hafeez T, El-Rashidy N. Optimizing poultry audio signal classification with deep learning and burn layer fusion. J Big Data. 2024;11(1):135. [doi: 10.1186/s40537-024-00985-8]
49. Mahmoud AM, Swilem SM, Alqarni AS, Haron F. Infant cry classification using semi-supervised k-nearest neighbor approach. In: 13th International Conference on Developments in eSystems Engineering (DeSE). IEEE; 2020:305-310. [doi: 10.1109/DeSE51703.2020.9450239]
50. Andono PN, Shidik GF, Prabowo DP. Bird voice classification based on combination feature extraction and reduction dimension with the k-nearest neighbor. Int J Intell Eng Syst. 2022;15:28-39. [doi: 10.22266/ijies2022.0228.24]
51. Merchan F, Guerra A, Poveda H, Guzmán HM, Sanchez-Galan JE. Bioacoustic classification of Antillean manatee vocalization spectrograms using deep convolutional neural networks. Appl Sci (Basel). 2020;10(9):3286. [doi: 10.3390/app10093286]
52. Ruiz-Munoz JF, Orozco-Alzate M. Dissimilarity-based classification for bioacoustic monitoring of bird species. In: 2011 IEEE IX Latin American Robotics Symposium and IEEE Colombian Conference on Automatic Control (LARC). IEEE; 2011. [doi: 10.1109/LARC.2011.6086822]
53. Han X, Peng J. Bird sound classification based on ECOC-SVM. Appl Acoust. 2023;204:109245. [doi: 10.1016/j.apacoust.2023.109245]
54. Gómez Bellido J, Luque Sendra A, Carrasco Muñoz A. Ingeniería de características para clasificación de señales sonoras. Presented at: Proceedings of the 24th International Congress on Project Management and Engineering; Jul 7-9, 2020; Universitat Politècnica de València, Spain.
55. LeBien J, Zhong M, Campos-Cerqueira M, et al. A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. Ecol Inform. Sep 2020;59:101113. [doi: 10.1016/j.ecoinf.2020.101113]
56. Zhang C, He K, Gao X, Guo Y. Automatic bioacoustics noise reduction method based on a deep feature loss network. Ecol Inform. May 2024;80:102517. [doi: 10.1016/j.ecoinf.2024.102517]
57. Zabidi A, Mansor W, Khuan LY, Yassin IM, Sahak R. The effect of F-ratio in the classification of asphyxiated infant cries using multilayer perceptron neural network. In: 2010 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES). IEEE; 2010:126-129. [doi: 10.1109/IECBES.2010.5742213]
58. Toro N, Giraldo Gómez SF, Salazar Jiménez T. Reconocimiento de especies de anuros por sus cantos, en archivos de audio, mediante técnicas de procesamiento digital de señales [Article in Spanish]. Sci Tech. 2006;3(32). URL: https://revistas.utp.edu.co/index.php/revistaciencia/article/view/6193 [Accessed 2025-12-09]
59. Gorin A, Subakan C, Abdoli S, Wang J, Latremouille S, Onu C. Self-supervised learning for infant cry analysis. In: 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW). IEEE; 2023. [doi: 10.1109/ICASSPW59220.2023.10193421]
60. Schall E, Kaya II, Debusschere E, Devos P, Parcerisas C. Deep learning in marine bioacoustics: a benchmark for baleen whale detection. Remote Sens Ecol Conserv. Oct 2024;10(5):642-654. [doi: 10.1002/rse2.392]
61. Gatto BB, Colonna JG, dos Santos EM, et al. Discriminative singular spectrum classifier with applications on bioacoustic signal recognition. arXiv. Preprint posted online on Mar 18, 2021. [doi: 10.48550/arXiv.2103.10166]
62. Li P, Roch MA, Klinck H, et al. Learning stage-wise GANs for whistle extraction in time-frequency spectrograms. IEEE Trans Multimed. 2023;25:9302-9314. [doi: 10.1109/TMM.2023.3251109]
63. Romero-Mujalli D, Bergmann T, Zimmermann A, Scheumann M. Utilizing DeepSqueak for automatic detection and classification of mammalian vocalizations: a case study on primate vocalizations. Sci Rep. Dec 27, 2021;11(1):24463. [doi: 10.1038/s41598-021-03941-1] [Medline: 34961788]
64. Jeantet L, Dufourq E. Improving deep learning acoustic classifiers with contextual information for wildlife monitoring. Ecol Inform. Nov 2023;77:102256. [doi: 10.1016/j.ecoinf.2023.102256]
65. Laplante JF, Akhloufi MA, Gervaise C. Deep learning for marine bioacoustics and fish classification using underwater sounds. In: 2022 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE). IEEE; 2022:288-293. [doi: 10.1109/CCECE49351.2022.9918242]
66. Fagerlund S. Bird species recognition using support vector machines. EURASIP J Adv Signal Process. Dec 2007;2007(1):038637. [doi: 10.1155/2007/38637]
67. Hu R, Hu K, Wang L, et al. Using deep learning to classify environmental sounds in the habitat of western black-crested gibbons. Diversity (Basel). 2024;16(8):509. [doi: 10.3390/d16080509]

68. Gong CSA, Su CHS, Chao KW, Chao YC, Su CK, Chiu WH. Exploiting deep neural network and long short-term memory methodologies in bioacoustic classification of LPC-based features. PLOS ONE. 2021;16(12):e0259140. [doi: 10.1371/journal.pone.0259140]
69. Rosita YD, Junaedi H. Infant's cry sound classification using mel-frequency cepstrum coefficients feature extraction and backpropagation neural network. In: 2nd International Conference on Science and Technology-Computer (ICST). IEEE; 2016:160-166. [doi: 10.1109/ICSTC.2016.7877367]
70. Brandes TS. Feature vector selection and use with hidden Markov models to identify frequency-modulated bioacoustic signals amidst noise. IEEE Trans Audio Speech Lang Process. 2008;16(6):1173-1180. [doi: 10.1109/TASL.2008.925872]
71. Noda JJ, Travieso CM, Sanchez-Rodriguez D, Dutta MK, Singh A. Using bioacoustic signals and support vector machine for automatic classification of insects. In: 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN). IEEE; 2016:656-659. [doi: 10.1109/SPIN.2016.7566778]
72. Potamitis I, Ganchev T, Fakotakis N. Automatic bioacoustic detection of *Rhynchophorus ferrugineus*. Presented at: 16th European Signal Processing Conference (EUSIPCO 2008); Aug 25-29, 2008; Lausanne, Switzerland. URL: https://www.eurasip.org/Proceedings/Eusipco/Eusipco2008/papers/1569101704.pdf [Accessed 2025-11-18]
73. Salem SI, Shirayama S, Shimazaki S, Oki K. Ensemble deep learning and anomaly detection framework for automatic audio classification: insights into deer vocalizations. Ecol Inform. Dec 2024;84:102883. [doi: 10.1016/j.ecoinf.2024.102883]
74. Kawakita S, Ichikawa K. Automated classification of bees and hornet using acoustic analysis of their flight sounds. Apidologie (Celle). Feb 2019;50(1):71-79. [doi: 10.1007/s13592-018-0619-6]
75. Sun Y, Midori Maeda T, Solís-Lemus C, Pimentel-Alarcón D, Buřivalová Z. Classification of animal sounds in a hyperdiverse rainforest using convolutional neural networks with data augmentation. Ecol Indic. Dec 2022;145:109621. [doi: 10.1016/j.ecolind.2022.109621]
76. Patil HA, Patil AT, Kachhi A. Constant Q cepstral coefficients for classification of normal vs. pathological infant cry. In: 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2022:7392-7396. [doi: 10.1109/ICASSP43922.2022.9746946]
77. Nanni L, Brahnam S, Lumini A, Maguolo G. Animal sound classification using dissimilarity spaces. Appl Sci (Basel). 2020;10(23):8578. [doi: 10.3390/app10238578]
78. Zabidi A, Mansor W, Lee YK, Yassin IM, Sahak R. Binary particle swarm optimization for selection of features in the recognition of infants cries with asphyxia. In: 2011 IEEE 7th International Colloquium on Signal Processing and Its Applications. IEEE; 2011:272-276. [doi: 10.1109/CSPA.2011.5759886]
79. Mala S, Smita S, Darandale S. Effective analysis and inference of infant cry signals using an IARO-based leaky Bi-LSTM model. Comput Speech Lang. 2024;82:101637. [doi: 10.1016/j.csl.2024.101637]
80. Hariharan M, Chee LS, Yaacob S. Infant cry classification using wavelet packet decomposition and support vector machine. J Med Syst. 2010;34(6):965-975. [doi: 10.1007/s10916-010-9591-z]
81. Li R, Garg S, Brown A. Identifying patterns of human and bird activities using bioacoustic data. Forests. 2019;10(10):917. [doi: 10.3390/f10100917]
82. Anqi G, Yukun L, Xinwen Y, et al. Soundscape composition and acoustic activity assessment of *Nomascus hainanus* habitat. Chin J Ecol. 2024;43(3):1-12. [doi: 10.13292/j.1000-4890.202403.023]
83. Raval M, Chauhan P, Rahevar M, Thakkar A. Bioacoustic bird monitoring: a deep learning solution for effective biodiversity conservation. In: 2024 International Conference on Data Science and Network Security (ICDSNS). IEEE; 2024. [doi: 10.1109/ICDSNS62112.2024.10691115]
84. Ntalampiras S. Automatic acoustic classification of infant cries. J Audio Eng Soc. 2014;62(10):658-665. [doi: 10.17743/jaes.2015.0025]
85. Noda JJ, Travieso-González CM, Sánchez-Rodríguez D, Alonso-Hernández JB. Acoustic classification of singing insects based on MFCC/LFCC fusion. Appl Sci (Basel). 2019;9(19):4097. [doi: 10.3390/app9194097]
86. Barajas-Montiel SE, Reyes-Garcia CA. Identifying pain and hunger in infant cry with classifiers ensembles. In: International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06). IEEE; 2006:770-775. [doi: 10.1109/CIMCA.2005.1631561]
87. Mohammed YA. Infant cry recognition system: a comparison of system performance based on CDHMM and ANN. Int J Auton Adapt Commun Syst. 2019;12(1):1-17. [doi: 10.4018/IJAPUC.2019010102]
88. Bergler C, Smeele SQ, Tyndel SA, et al. ANIMAL-SPOT enables animal-independent signal detection and classification using deep learning. Sci Rep. 2022;12:21966. [doi: 10.1038/s41598-022-26429-y]

89. Kaplun D, Voznesensky A, Romanov S, Andreev V, Butusov D. Classification of hydroacoustic signals based on harmonic wavelets and a deep learning artificial intelligence system. Appl Sci (Basel). 2020;10(9):3097. [doi: 10.3390/app10093097]
90. Saraswathy J, Hariharan M, Vijean V, Yaacob S, Khairunizam W. Performance comparison of Daubechies wavelet family in infant cry classification. In: 2012 IEEE 8th International Colloquium on Signal Processing and Its Applications. IEEE; 2012. [doi: 10.1109/CSPA.2012.6194767]
91. Patil AT, Kachhi A, Patil HA. Subband teager energy representations for infant cry analysis and classification. In: 2022 30th European Signal Processing Conference (EUSIPCO). IEEE; 2022:1313-1317. [doi: 10.23919/EUSIPCO55093.2022.9909974]
92. Etz T, Reetz H, Wegener C. A classification model for infant cries with hearing impairment and unilateral cleft lip and palate. Folia Phoniatr Logop. 2012;64(5):254-261. [doi: 10.1159/000343994] [Medline: 23182951]
93. Anjali G, Sanjeev S, Mounika A, Suhas G, Reddy GP, Kshiraja Y. Infant cry classification using transfer learning. In: 2022 IEEE Region 10 Conference (TENCON). IEEE; 2022:1-7. [doi: 10.1109/TENCON55691.2022.9977793]
94. Batist CH, Dufourq E, Jeantet L, Razafindraibe MN, Randriamanantena F, Baden AL. An integrated passive acoustic monitoring and deep learning pipeline for black-and-white ruffed lemurs (*Varecia variegata*) in Ranomafana National Park, Madagascar. Am J Primatol. Apr 2024;86(4):e23599. [doi: 10.1002/ajp.23599] [Medline: 38244194]
95. Kheddache Y, Tadj C. Identification of diseases in newborns using advanced acoustic features of cry signals. Biomed Signal Process Control. 2019;50:35-44. [doi: 10.1016/j.bspc.2019.01.010] [Medline: 33281921]
96. Weerasena H, Jayawardhana M, Egodage D, et al. Continuous automatic bioacoustics monitoring of bird calls with local processing on node level. In: 2018 IEEE Region 10 Conference. IEEE; 2018:235-239. [doi: 10.1109/TENCON.2018.8650196]
97. Dhakne D, Kuduvan VM, Palhade A, Kanjwani T, Kshirsagar R. Bird species identification using audio signal processing and neural networks. Int J Res Appl Sci Eng Technol. 2022;10:4002-4005. [doi: 10.22214/ijraset.2022.43309]
98. Lostanlen V, Salamon J, Farnsworth A, Kelling S, Bello JP. Robust sound event detection in bioacoustic sensor networks. PLOS ONE. 2019;14(10):e0214168. [doi: 10.1371/journal.pone.0214168] [Medline: 31647815]
99. Bergler C, Schröter H, Cheng RX, et al. ORCA-SPOT: an automatic killer whale sound detection toolkit using deep learning. Sci Rep. Jul 29, 2019;9(1):10997. [doi: 10.1038/s41598-019-47335-w] [Medline: 31358873]
100. Reyes Galaviz OF, Reyes Garcia CA, et al. Infant cry classification to identify hypoacoustics and asphyxia with neural networks. In: Monroy R, Arroyo-Figueroa G, Sucar LE, editors. MICAI 2004: Advances in Artificial Intelligence. Springer; 2004:69-78. [doi: 10.1007/978-3-540-24694-7_8]
101. Cano Ortiz SD, Escobedo Beceiro DI, Ekkel T. A radial basis function network oriented for infant cry classification. In: Sanfeliu A, Martínez Trinidad JF, Carrasco Ochoa JA, editors. Progress in Pattern Recognition, Image Analysis and Applications. Springer; 2004:374-380. [doi: 10.1007/978-3-540-30463-0_46]
102. Liu L, Li W, Wu X, Zhou BX. Infant cry language analysis and recognition: an experimental approach. IEEE/CAA J Autom Sinica. 2019;6(3):778-788. [doi: 10.1109/JAS.2019.1911435]
103. Clark ML, Salas L, Baligar S, et al. The effect of soundscape composition on bird vocalization classification in a citizen science biodiversity monitoring project. Ecol Inform. Jul 2023;75:102065. [doi: 10.1016/j.ecoinf.2023.102065]
104. Baptista PB, Antunes C. Bioacoustic classification framework using transfer learning. Presented at: Proceedings of the Modelling Decisions for Artificial Intelligence (MDAI 2021); Sep 27-30, 2021; Umeå, Sweden. URL: https://web.ist.utl.pt/claudia.antunes/artigos/bonito2021mdai.pdf [Accessed 2025-11-26]
105. Manikanta K, Soman KP, Manikandan MS. Deep learning based effective baby crying recognition method under indoor background sound environments. In: 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS). IEEE; 2019. [doi: 10.1109/CSITSS47250.2019.9031058]
106. Osmani A, Hamidi M, Chibani A. Machine learning approach for infant cry interpretation. In: 2017 IEEE 29th International Conference on Tools With Artificial Intelligence (ICTAI). IEEE; 2017:182-186. [doi: 10.1109/ICTAI.2017.00038]
107. Chittora A, Patil HA. Classification of pathological infant cries using modulation spectrogram features. In: 2014 9th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE; 2014:541-545. [doi: 10.1109/ISCSLP.2014.6936626]
108. Farsaie Alaie H, Abou-Abbas L, Tadj C. Cry-based infant pathology classification using GMMs. Speech Commun. Mar 2016;77:28-52. [doi: 10.1016/j.specom.2015.12.001] [Medline: 27524848]
109. Sharma K, Gupta C, Gupta S. Infant weeping calls decoder using statistical feature extraction and Gaussian mixture models. In: 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE; 2019:1-6. [doi: 10.1109/ICCCNT45670.2019.8944527]

110. Abou-Abbas L, Tadj C, Fersaie HA. A fully automated approach for baby cry signal segmentation and boundary detection of expiratory and inspiratory episodes. J Acoust Soc Am. Sep 2017;142(3):1318-1331. [doi: 10.1121/1.5001491] [Medline: 28964073]
111. Ali MZM, Mansor W, Lee YK, Zabidi A. Asphyxiated infant cry classification using Simulink model. In: 2012 IEEE 8th International Colloquium on Signal Processing and Its Applications (CSPA). IEEE; 2012:491-494. [doi: 10.1109/CSPA.2012.6194778]
112. Mac Aodha O, Gibb R, Barlow KE, et al. Bat detective—deep learning tools for bat acoustic signal detection. PLOS Comput Biol. 2018;14(3):e1005995. [doi: 10.1371/journal.pcbi.1005995]
113. Farsaie Alaie H, Tadj C. Cry-based classification of healthy and sick infants using adapted boosting mixture learning method for Gaussian mixture models. Model Simul Eng. 2012;2012:1-10. [doi: 10.1155/2012/983147]

Abbreviations

AUC: area under the receiver operating characteristic curve
CNN: convolutional neural network
CRNN: convolutional recurrent neural network
GAN: generative adversarial network
IoT: Internet of Things
LPCC: linear predictive cepstral coefficient
MECIR: Methodological Expectations of Cochrane Intervention Reviews
MFCC: Mel frequency cepstral coefficient
NICU: neonatal intensive care unit
OSF: Open Science Framework
PCEN: per-channel energy normalization
PICO: population, intervention, comparison, outcome
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RNN: recurrent neural network
RQ: review question
SNR: signal-to-noise ratio
SVM: support vector machine
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

Edited by Javad Sarvestan; peer-reviewed by Ahmad H Sabry, Indu Sekhar Samanta; submitted 04.Jul.2025; final revised version received 09.Nov.2025; accepted 10.Nov.2025; published 16.Dec.2025

Please cite as:
Owino G, Shibwabo B
Noise-Resilient Bioacoustics Feature Extraction Methods and Their Implications on Audio Classification Performance: Systematic Review
JMIR Biomed Eng 2025;10:e80089
URL: https://biomedeng.jmir.org/2025/1/e80089
doi: 10.2196/80089

© Geofrey Owino, Bernard Shibwabo. Originally published in JMIR Biomedical Engineering (http://biomsedeng.jmir.org), 16.Dec.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Biomedical Engineering, is properly cited. The complete bibliographic information, a link to the original publication on https://biomedeng.jmir.org/, as well as this copyright and license information must be included.