Mapping Deep Learning Approaches for Acoustic Bird Species Classification in the Brazilian Pantanal

Julia Milioranza Gomes[1], **Isis Milena Daron**[1], **Thiago Meirelles Ventura**[1]

[1]Universidade Federal de Mato Grosso

{julia.gomes1, isis.daron}@sou.ufmt.br, {thiago}@ic.ufmt.br

Abstract. *Automated acoustic monitoring of bird species has strong potential for biodiversity conservation, yet key questions remain regarding effective visual representations, deep learning architectures, and preprocessing strategies, particularly in biodiversity-rich regions such as the Brazilian Pantanal. This study analyzes approaches for acoustic bird species classification. Results indicate that Mel-spectrograms and CNNs are the predominant standards, while temporal segmentation is widely adopted as a preprocessing step. Additionally, less than 6% of the reviewed studies focus on Brazil, with none addressing the Pantanal, highlighting the need to combine globally pre-trained architectures with regional datasets for automated biodiversity monitoring.*

1. Introduction

The Pantanal, recognized as the world's largest continuous tropical wetland and a UNESCO World Natural Heritage Site, extends across Mato Grosso, Mato Grosso do Sul, Paraguay, and Bolivia [UNESCO World Heritage Centre 2000]. It hosts around 617 documented bird species and functions as an essential feeding and breeding area for migratory birds [Nunes et al. 2021]. Despite this ecological importance, the biome faces increasing threats from wildfires, deforestation, and climate change, which directly affect bird populations that act as bioindicators of ecosystem health [Justino 2025].

Traditional biodiversity monitoring, based on manual observation and expert auditory identification, is limited by time, spatial coverage, and human analytical capacity. While Artificial Intelligence, particularly Convolutional Neural Networks (CNNs), offers a scalable solution, its application in the Pantanal remains a challenge due to several unresolved issues.

Firstly, most existing tools are trained on North American and European datasets, creating a gap in understanding which visual representations of vocalizations (spectrograms vs. MFCCs) are truly effective for South American species. Furthermore, it is still unclear which deep learning architectures demonstrate superior performance in this specific context and what preprocessing techniques are vital to handle real-world audio noise. Finally, there is a lack of synthesized information regarding automated identification solutions already developed within the Brazilian landscape.

To address these gaps, the objective of this study is to determine the state-of-the-art in deep learning architectures for acoustic bird classification and evaluate their feasibility for the Pantanal ecosystem. This research aims to provide the technical foundation necessary to bridge global AI technologies and regional biodiversity monitoring needs.

2. Methodology

To achieve the proposed objective of identifying the most effective architectures and their applicability to the Pantanal, it was necessary to conduct a rigorous and criteria-based investigation. Therefore, a Systematic Literature Review (SLR) was chosen as the methodological approach. This method follows the guidelines proposed by [Kitchenham 2004], which provide a structured protocol for identifying, evaluating, and synthesizing relevant scientific evidence.

The process was supported by the online tool Parsifal[1], ensuring a systematic management of all stages, from protocol design to data extraction. This included bibliographic record management, duplicate removal, screening, quality assessment, and structured synthesis, ensuring the replicability and transparency of the findings.

2.1. PICOC Framework

The protocol was structured using the PICOC framework [Wohlin et al. 2012], which defines the scope of the review across five dimensions to ensure all aspects of the research problem are covered:

- Population (P): Most abundant bird species of the Pantanal biome;
- Intervention (I): Computational vision models, visual representations of vocalizations (spectrograms, MFCCs), and deep learning techniques;
- Comparison (C): Traditional bird identification methods and other audio classification approaches;
- Outcomes (O): Classification accuracy and precision, generalization capability, and potential for ecological monitoring;
- Context (C): Pantanal.

Based on this framework, the following research questions (RQs) were formulated to guide the review: (RQ1) What are the most effective visual representations of bird vocalizations (spectrograms, MFCCs) used in conjunction with computer vision techniques for avian species classification? (RQ2) Which computer vision models have demonstrated superior performance in classifying bird species based on their acoustic (visualized) features? (RQ3) What preprocessing techniques are commonly applied to raw audio data or its visual representations to improve the accuracy and robustness of computer vision models in avian bioacoustics? (RQ4) What solutions have already been created regarding bird species identification in Brazil?

2.2. Search Strategy and Selection Criteria

A search string was elaborated to encompass the core concepts of the review: artificial intelligence, avian species, classification approaches, and acoustic representations. The following string was applied:

"Artificial Intelligence" AND ("bird" OR "avian") AND ("classification" OR "identification" OR "recognition") AND ("spectrogram" OR "MFCC" OR "sonogram")

To ensure relevance and quality of the selected studies, a set of selection criteria was established and applied during the initial screening phase (title and abstract analysis). Table 1 presents the criteria adopted.

[1]Parsifal `https://parsif.al/`

Inclusion Criteria (IC)	Exclusion Criteria (EC)
Papers published from the year 2021 onwards	Studies not in English
Papers with accessible full text	Duplicate studies
Papers that address the use of computer vision (including Deep Learning/CNNs)	Out of scope studies
	Review papers or theoretical articles
	No methodological transparency

Table 1. Selection criteria adopted in the systematic literature review.

2.3. Quality Assessment

To assess the methodological quality and relevance of the pre-selected studies, six quality assessment questions (QAs) were formulated based on the inclusion criteria and the objectives of the review: (QA1) Is the dataset used clearly described (e.g., number of species, number of samples, audio duration)? (QA2) Are performance metrics (Accuracy, Precision, Recall, F1-Score) clearly defined and reported per species or in an aggregated manner? (QA3) Was the source code publicly available? (QA4) Was the trained model publicly available? (QA5) Are the studied bird species also found in the Pantanal biome? (QA6) Was the study tested with real field data?

Each question was scored on a three-point scale: 0 when the criterion was not met, 0.5 when it was only partially satisfied, and 1 when it was fully addressed. The maximum achievable score per study was therefore 6 points. Only studies scoring above 3 points were retained for in-depth analysis. This scoring mechanism ensured that the final selection comprised studies with sufficient empirical grounding, documented performance metrics, ecological relevance, and methodological transparency.

2.4. Data Synthesis and Analysis

Once the final corpus of studies was established, a data synthesis process was conducted to extract relevant information through a qualitative and comparative analysis of the architectures, visual representations, and preprocessing methods used in the selected papers. The synthesis was structured to provide evidence-based answers to the problems raised in the introduction, allowing the identification of patterns, current limitations, and the feasibility of applying these global deep learning solutions to the Pantanal biome.

This process enabled a structured transition from data collection to analysis. The following section presents the outcomes of the systematic review protocol, including the study selection process, quality assessment results, and synthesized findings.

3. Results

This section presents the mapping obtained through the systematic literature review, as well as the main findings reported in the selected studies regarding visual representations, model architectures and performance, data preprocessing and refinement strategies, and the Brazilian context.

3.1. Study Selection and Quality Flow

The application of the search string across the three databases resulted in 226 candidate articles in total: 14 from IEEE Xplore, 198 from Science Direct, and 14 from Scopus.

As illustrated in the flowchart in Figure 1, after removing duplicates, 219 unique studies remained. The application of the predefined inclusion and exclusion criteria during title and abstract screening reduced this set to 37 articles.

Following the quality assessment protocol described in Section 2.3, only studies that achieved a score higher than 3 out of 6 points were included in the analysis, resulting in a final set of 18 articles. The scores obtained by each paper are presented in https://github.com/MilioranzaJ/Mapping-Deep-Learning-Approaches-for-Acoustic-Bird-Species-Classification-in-the-Brazilian-Pantanal.

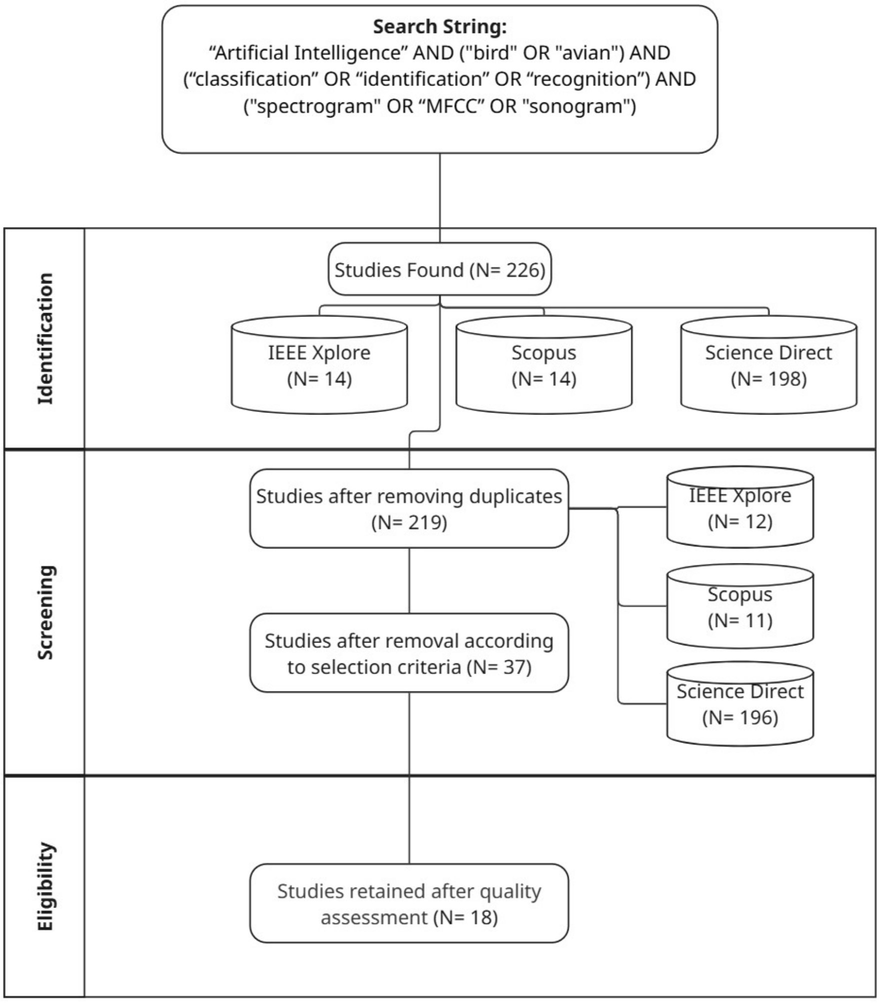

Figure 1. Study selection flowchart, from which the 18 main articles for analysis in this work were selected.

3.2. Visual Representations

Regarding the technical challenges of acoustic feature extraction, the literature reveals that there is no universally superior representation; instead, the choice depends on model architecture and available computational resources.

The most adopted representation among the analyzed studies for deep learning models such as CNNs and Transformers was the Mel-spectrogram. By mimicking auditory perception and reducing redundancy in high frequencies, this representation has become one of the primary inputs for high-performance architectures. Studies such as [Heinrich et al. 2025] and [Zhang et al. 2023] utilized Mel-spectrograms to feed networks. Additionally, [Silva et al. 2025] highlighted the efficiency of generating these representations dynamically via GPU using the Kapre library for end-to-end models.

In contrast, Mel-Frequency Cepstral Coefficients (MFCCs) emerged as the most effective representation in environments with severe background noise and for Edge Computing contexts. A comparative analysis by [Wang et al. 2023] demonstrated that MFCCs possess a superior ability to handle environmental noise compared to other visual representations like Log-Mel or Chroma. Due to their compact nature, MFCCs remain the preferred choice for training traditional, lightweight classifiers such as LightGBM and SVM, enabling high accuracy with minimal computational cost for IoT and agricultural applications [Kazeneza et al. 2025, Delgado-Rajó and Travieso-Gonzalez 2025].

To overcome the limitations of single representations, recent literature proposes hybrid approaches. [Xie et al. 2022] developed a multi-visual fusion model, demonstrating that extracting simultaneous features from Wavelet Transform (WT), Hilbert-Huang Transform (HHT), and STFT maximizes classification precision. Furthermore, for smaller datasets where deep CNNs are prone to overfitting, [Xie and Zhu 2022] proved that Scale-Frequency Maps (SFM) outperform traditional Mel-spectrograms, being highly effective for 2D and 3D-CNN architectures.

3.3. Performance of Architectures and Models

To address the uncertainty regarding model performance in bioacoustics, the selected studies were analyzed to identify which computer vision architectures have demonstrated the best results for classifying bird species from visual acoustic representations. This analysis is central to understanding the state-of-the-art models that could be effectively applied to the Pantanal's biodiversity. Of the 18 reviewed studies, eight proposed and trained their own customized CNN architectures.

Among these, notable examples include the AudioProtoPNet [Heinrich et al. 2025], an interpretable model with a ConvNeXt backbone, the CMS-H [Wang et al. 2023], which combines static and dynamic hierarchical modeling of bird song features, and the ISCL [Zhang et al. 2023], which integrates self-supervised learning with a modified ResNet. In addition, the PNW-Cnet [Ruff et al. 2021], a six-layer CNN for threatened species monitoring, and a lightweight CNN for low-cost IoT devices [Delgado-Rajó and Travieso-Gonzalez 2025] highlight the diversity of architectural proposals. Four studies evaluated standard pretrained architectures. EfficientNetB0, EfficientNetB4, and MobileNetV2 [Kazeneza et al. 2025] were explored in knowledge distillation, with EfficientNetB4 achieving accuracy close to 99%, although these metrics depend on dataset conditions and validation protocols. Similarly, ECAPA-TDNN combined with ResNet34 [Hu et al. 2025] employed temporal delay networks as an alternative to traditional convolutions.

BirdNET appeared as a central reference in three studies, reflecting its consolidation as a widely adopted baseline. [Ware et al. 2023] demonstrated its capacity to expand species detection in large-scale monitoring compared to manual listening alone. [Clarfeld et al. 2025] evaluated a two-stage detection framework applied to distinct classifiers, reporting true positive rates of 71.5% and 17.0% in the first stage. Logistic regression-based post-processing improved separation in both cases, reaching 84.5% and 89.8%, indicating that false positives are a broader challenge in automated wildlife detection pipelines. [Michaud et al. 2023] used BirdNET to compare against unsupervised clustering (DBSCAN), demonstrating its versatility as a benchmarking baseline. Additionally, [Dematties et al. 2024] employed BirdNET as an auxiliary validation mechanism for self-supervised models (DINO and VICReg).

Beyond deep learning, some studies investigated traditional machine learning models in scenarios prioritizing computational efficiency. [Sun et al. 2024] reported 94.09% accuracy using Random Forest with sparse signal representation via OMPA, while SVM and LightGBM also achieved competitive results in agricultural pest detection with lower computational cost. Overall, although CNNs remain the dominant approach, hybrid and lightweight models continue to be relevant for deployment in resource-constrained environments.

3.4. Data Preprocessing and Refinement Strategies

The analysis of the selected studies further identified the preprocessing techniques commonly applied to raw audio data or its visual representations to improve accuracy and robustness in avian bioacoustics. Table 2 summarizes the frequency of occurrence of each technique across the 18 selected studies.

Author and Year	**Specific Parameters**	**Seg.**	**Res.**	**Noi.**	**Aug.**	**Norm.**	**Pad.**
[Silva et al. 2025]	22 kHz, Band-pass, Zero-pad	X	X	X			X
[Heinrich et al. 2025]	32 kHz, SpecAugment, Z-score	X	X		X	X	X
[Hu et al. 2025]	32 kHz, Background noise mix	X	X		X	X	
[Xie and Zhu 2022]	16 kHz, Wavelet denoising	X	X	X			
[Sun et al. 2024]	OMPA/Sparse, Min-Max scaling	X		X		X	
[Zhang et al. 2023]	Mixup, time-shift	X			X		
[Michaud et al. 2023]	44.1 kHz sampling	X	X				
[Delgado-Rajó and Travieso-Gonzalez 2025]	RMS thresholding	X		X			
[Xie et al. 2022]	5s windows	X					
[Doren et al. 2025]	Fixed-length segments	X					
[Wang et al. 2023]	3s windows	X					
[Dematties et al. 2024]	5s windows	X					
[Wu et al. 2023]	1s windows	X					
[Kazeneza et al. 2025]	2s windows	X					
[Duarte et al. 2024]	Fixed-length segments	X					
[Clarfeld et al. 2025]	BirdNET default (3s)	X					
[Ruff et al. 2021]	12s windows	X					
[Ware et al. 2023]	BirdNET default (3s)	X					
Total	—	**18**	**5**	**4**	**3**	**3**	**2**

Table 2. Taxonomy of preprocessing techniques applied per study and their technical parameters. Seg. = Temporal Segmentation; Res. = Resampling; Noi. = Noise Filtering; Aug. = Data Augmentation; Norm. = Normalization; Pad. = Padding.

Temporal segmentation was the most adopted technique, present in all 18 studies. Segment durations ranged from 1 second to 30 seconds depending on study objectives, though windows of 1, 2, 3, and 5 seconds were most common. This standardization ensures consistent input dimensions for visual representations and facilitates batch-based deep learning training. For audio segments shorter than the defined window, zero-padding or circular padding were employed to maintain fixed dimensionality, though these appeared in only two studies.

Resampling to a standardized sampling rate was applied in five studies, with rates varying widely from 11,025 Hz to 48,000 Hz. The choice of rate was often determined by model specifications (BirdNET requires 48 kHz) or by the target species frequency range.

Noise filtering and denoising appeared in four studies, implemented through various strategies: band-pass filtering to isolate frequency ranges of interest, stationary noise subtraction, wavelet-based thresholding, and sparse representation-based signal reconstruction using MPA and OMPA algorithms. These approaches reflect a concern with isolating the biological signal from environmental interference, which is particularly relevant in field recording scenarios.

Data augmentation was reported in three studies, primarily through SpecAugment, a technique that applies random zero-masks to frequency bands and time frames on spectrograms. Additional time-domain strategies included background noise injection, time-shifting, gain adjustment, and mixup. Finally, amplitude normalization, including maximum absolute value scaling or Z-score standardization, was applied in three studies to standardize signal scale across heterogeneous recordings.

3.5. Mapping Solutions in the Brazilian Context

The fourth research question investigated the extent to which deep learning solutions for bird acoustic classification have been developed specifically for the Brazilian ecological context. The results expose a critical gap: of the 18 analyzed studies, only one (5.6%), [Silva et al. 2025], was designed entirely around Brazilian data and ecological conditions. The remaining 17 articles (94.4%), by contrast, were conducted with data from other geographic regions, predominantly Europe, North America, and Asia.

Despite this geographic disparity, a more nuanced reading of the taxonomic composition of datasets reveals that 12 of the 18 studies (66.7%) included species that also occur in Brazil, suggesting technical applicability to the national fauna even when the research was not designed with that intention. This taxonomic overlap manifests in three main contexts: shared Amazonian fauna [Heinrich et al. 2025] used recordings from the Peruvian Amazon, a region sharing numerous species with the Brazilian Amazon); migratory birds [Ware et al. 2023], [Doren et al. 2025], and [Duarte et al. 2024] detected Nearctic migrants that winter in Brazil, such as the Lesser Yellowlegs, the Solitary Sandpiper, and the Osprey); and cosmopolitan or introduced species widely distributed in Brazil (House Sparrow, Rock Dove, Barn Swallow), present in eight studies.

The sole exception oriented to the national ecological context, [Silva et al. 2025], proposes an end-to-end model to analyze the impact of anthropogenic noise on the Southern House Wren (Troglodytes aedon musculus) in urban Bauru, São Paulo. Although methodologically rigorous and ecologically meaningful, this study does not address the specific challenges of the Pantanal or other species-rich Brazilian biomes, such as extraordinary tropical biodiversity, acoustically complex multi-species soundscapes, seasonal variation in vocal activity, or the absence of labeled training data for many endemic species. Taken together, these observations confirm that the automated acoustic monitoring of Pantanal avifauna remains a largely unaddressed research challenge, constituting a significant opportunity for future work.

4. Discussion

The results of this research converge toward a consistent understanding of the current state of the art, while also exposing structural gaps with direct implications for biodiversity monitoring in the Pantanal. Across the research questions, a clear tension emerges between the technical maturity of global AI solutions and their effective applicability to the ecological particularities of tropical South American biomes.

Regarding visual representations, spectrogram-based inputs are firmly consolidated as the dominant paradigm in avian bioacoustics. Mel-spectrograms stand out as the most frequently adopted representation, followed by linear spectrograms and MFCCs, indicating a preference for time–frequency structures that balance acoustic detail and computational efficiency, particularly in combination with CNN architectures.

In terms of model architectures, the field is strongly dominated by deep learning approaches. Custom CNNs and high-performance pretrained models, such as BirdNET and EfficientNetB4, achieve very high classification accuracy, reinforcing the consolidation of deep learning as the central methodological paradigm. Nevertheless, traditional models such as Random Forest and SVM remain competitive in contexts where computational resources are limited, highlighting the continued relevance of lightweight and hybrid solutions for real-world deployment.

With respect to preprocessing, the findings reveal a strong emphasis on standardization and input conditioning. Temporal segmentation appears consistently as a structural step to ensure fixed-size inputs, while resampling, normalization, and denoising are widely employed to reduce variability caused by heterogeneous recording conditions. In general, preprocessing is treated less as a space for methodological innovation and more as a necessary stage to stabilize training and improve performance consistency.

Most critically, the lack of Brazilian-contextualized research, with only 5.6% of the selected studies focusing on national ecological conditions, represents a significant geographic and scientific gap. Although a substantial portion of the reviewed datasets includes species that also occur in Brazil, taxonomic overlap alone does not substitute for models trained on regionally representative data. This limitation becomes particularly relevant in the Pantanal, where ecological complexity and increasing environmental pressures demand locally adapted and ecologically grounded monitoring solutions.

5. Conclusion

This work mapped the state of the art in deep learning and computer vision approaches applied to acoustic bird species classification, with an lens toward its applicability to the Brazilian Pantanal context. The analysis of 18 selected studies reveals that Mel-spectrograms are the most widely used acoustic representation, CNN-based architectures dominate the field, and temporal segmentation is a universal preprocessing step. Beyond these technical patterns, however, the review also uncovers a critical research gap: only one study was specifically designed for the Brazilian ecological context, and none addressed the Pantanal's acoustic environment, biodiversity, and endemic species directly.

Taken together, these findings support the conclusion that while technically mature AI tools exist for bird acoustic classification, their application to the Pantanal is constrained by geographic bias in training data, absence of regionally adapted models, and limited standardization of preprocessing practices for tropical soundscapes. In this regard, combining transfer learning from globally pre-trained models with region-specific datasets emerges as the most promising pathway toward accurate automated monitoring. To advance in this direction, future work should prioritize dataset expansion for Pantanal species, exploration of partial fine-tuning strategies, development of standardized evaluation benchmarks for Neotropical bioacoustics, and the integration of ecological validation to ensure that technical performance translates into meaningful conservation outcomes.

6. Acknowledgments

We would like to express our gratitude to the Computational Bioacoustics Research Unit (CO.BRA) at the Federal University of Mato Grosso for their support and contributions to this research.

References

Clarfeld, L. A., Gieder, K. D., Abrams, R., Bernier, C., Cahill, J., Staats, S., Wixsom, S., and Donovan, T. M. (2025). Two-stage models improve machine learning classifiers in wildlife research: A case study in identifying false positive detections of ruffed grouse. *Ecological Informatics*, 89.

Delgado-Rajó, F. A. and Travieso-Gonzalez, C. M. (2025). Flexible hybrid edge computing iot architecture for low-cost bird songs detection system. *Ecological Informatics*, 90.

Dematties, D., Rajani, S., Sankaran, R., Shahkarami, S., Raut, B., Collis, S., Beckman, P., and Ferrier, N. (2024). Acoustic fingerprints in nature: A self-supervised learning approach for ecosystem activity monitoring. *Ecological Informatics*, 83.

Doren, B. M. V., DeSimone, J. G., Firth, J. A., Hillemann, F., Gayk, Z., Cohen, E., and Farnsworth, A. (2025). Social associations across species during nocturnal bird migration. *Current Biology*, 35:898–904.e4.

Duarte, A., Weldy, M. J., Lesmeister, D. B., Ruff, Z. J., Jenkins, J. M., Valente, J. J., and Betts, M. G. (2024). Passive acoustic monitoring and convolutional neural networks facilitate high-resolution and broadscale monitoring of a threatened species. *Ecological Indicators*, 162.

Heinrich, R., Rauch, L., Sick, B., and Scholz, C. (2025). Audioprotopnet: An interpretable deep learning model for bird sound classification. *Ecological Informatics*, 87.

Hu, T., Yuan, M., Li, J., Wang, J., Wang, L., and Zhang, H. (2025). Avian vocalizations in huangmaohai sea-crossing channel: Automatic birdsong recognition and ecological impact analysis based on deep learning. *Biological Conservation*, 305.

Justino, S. T. P. e. a. (2025). Monitoring environmental degradation and spatial changes in vegetation and water resources in the brazilian pantanal. *Sustainability*, 17(1):51.

Kazeneza, M., Bosman, A. S., Amenyedzi, D. K., Hanyurwimfura, D., Ndashimye, E., and Vodacek, A. (2025). Balancing complexity and performance of machine learning models for avian pests sound detection in agricultural environments. *IEEE Access*, 13:105813–105827.

Kitchenham, B. (2004). Procedures for performing systematic reviews. Technical Report TR/SE-0401, Department of Computer Science, Keele University, Keele, UK.

Michaud, F., Sueur, J., Cesne, M. L., and Haupert, S. (2023). Unsupervised classification to improve the quality of a bird song recording dataset. *Ecological Informatics*, 74.

Nunes, A. P., Posso, S. R., da Frota, A. V. B., Vitorino, B. D., Laps, R. R., Donatelli, R. J., Straube, F. C., Pivatto, M. A. C., de Oliveira, D. M. M., Carlos, B., de Melo, A. V., Tomas, W. M., de Freitas, G. O., de Souza, R. A. D., Benites, M., Mamede, S., and Moreira, R. S. (2021). Birds of the pantanal floodplains, brazil: Historical data, diversity, and conservation. *Papeis Avulsos de Zoologia*, 61.

Ruff, Z. J., Lesmeister, D. B., Appel, C. L., and Sullivan, C. M. (2021). Workflow and convolutional neural network for automated identification of animal sounds. *Ecological Indicators*, 124.

Silva, L. D. A., Colonna, J. G., Gatto, B. B., and Protazio, J. M. (2025). Impacts of anthropogenic noise on the house wren's song: An xai approach to bioacoustic insights. In *2025 IEEE Symposium on Trustworthy, Explainable and Responsible Computational Intelligence, CITREx 2025*. Institute of Electrical and Electronics Engineers Inc.

Sun, Z., Zhang, M., Liu, J., Wu, Q., Wang, J., and Wang, G. (2024). Research on filtering and classification method for white-feather broiler sound signals based on sparse representation. *Engineering Applications of Artificial Intelligence*, 127.

UNESCO World Heritage Centre (2000). Pantanal Conservation Area.

Wang, Y., Chen, A., Li, H., Zhou, G., Yi, J., and Zhang, Z. (2023). A hierarchical birdsong feature extraction architecture combining static and dynamic modeling. *Ecological Indicators*, 150.

Ware, L., Mahon, C. L., McLeod, L., and Jetté, J. F. (2023). Artificial intelligence (birdnet) supplements manual methods to maximize bird species richness from acoustic data sets generated from regional monitoring. *Canadian Journal of Zoology*, 101:1031–1051.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in Software Engineering*. Springer.

Wu, S. H., Ko, J. C. J., Lin, R. S., Tsai, W. L., and Chang, H. W. (2023). An acoustic detection dataset of birds (aves) in montane forests using a deep learning approach. *Biodiversity Data Journal*, 11.

Xie, J. and Zhu, M. (2022). Sliding-window based scale-frequency map for bird sound classification using 2d- and 3d-cnn. *Expert Systems with Applications*, 207.

Xie, S., Lu, J., Liu, J., Zhang, Y., Lv, D., Chen, X., and Zhao, Y. (2022). Multi-view features fusion for birdsong classification. *Ecological Informatics*, 72.

Zhang, C., Li, Q., Zhan, H., Li, Y. F., and Gao, X. (2023). One-step progressive representation transfer learning for bird sound classification. *Applied Acoustics*, 212.