Regional Adaptation of Deep Learning Models for Bird Species Classification in the Pantanal

Julia Milioranza Gomes[1], **Isis Milena Daron**[1], **Thiago Meirelles Ventura**[1,4], **Allan Gonçalves de Oliveira**[1], **Marinêz Isaac Marques**[2,3,4,5,6], **Karl-Ludwig Schuchmann**[2,5,6,7]

[1] Universidade Federal de Mato Grosso, Instituto de Computação, Cuiabá, MT, Brasil

[2]Universidade Federal de Mato Grosso, Instituto de Biociências, Programa de Pós Graduação em Zoologia, Cuiabá, MT, Brasil.

[3]Universidade Federal de Mato Grosso, Instituto de Biociências, Laboratório de Ecologia e Taxonomia de Artrópodes, Cuiabá, MT, Brasil.

[4]Instituto Nacional de Pesquisa do Pantanal, Cuiabá, MT, Brasil.

[5]Instituto Nacional de Ciência e Tecnologia em Áreas Úmidas, Cuiabá, MT, Brasil.

[6]Computational Bioacoustics Research Unit (CO.BRA), Cuiabá, MT, Brasil

[7]Zoological Research Museum A. Koenig, Bonn, Germany

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

marinez513@gmail.com, klschuchmann@googlemail.com

Abstract. *Despite being one of the world's most biodiverse wetlands, the Pantanal remains underrepresented in global bioacoustic datasets. This study develops a Pantanal-specific bird classification model using transfer learning from BirdNET and Xeno-Canto recordings. An evaluation of acoustic representations (MFCCs, Mel-spectrograms, and linear spectrograms) revealed the limitations of training CNNs from scratch on restricted regional data, establishing the necessity for transfer learning. Consequently, the transfer learning model circumvented these representation bottlenecks and outperformed the original BirdNET, reaching a weighted F1-score of 0.95 on the Pantanal validation set and 0.92 on an external dataset, indicating limited geographic overfitting. Results highlight the effectiveness of regional fine-tuning for automated bird monitoring in tropical environments.*

1. Introduction

Global biodiversity decline urgently threatens ecosystem stability, resilience, and human services [Hooper 2005]. Monitoring strategies capable of capturing population dynamics at scale are essential to support conservation actions. Birds are highly informative biological indicators due to their wide geographic distribution, well-established taxonomy, and sensitivity to environmental disturbances [Gregory and van Strien 2010].

Conventional bird monitoring is limited by fieldwork logistics, expertise requirements, and difficulty surveying remote areas. Automated acoustic monitoring offers a scalable, non-invasive alternative. Enabled by low-cost passive recorders, this approach

facilitates high-resolution and broadscale monitoring of threatened species and ecosystems [Stowell 2022, Duarte et al. 2024]. The main challenge lies in the automatic analysis of these recordings, where deep learning has shown significant progress.

Convolutional Neural Networks (CNNs) applied to audio representations, such as spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs), dominate avian bioacoustics [Purwins et al. 2019]. BirdNET is the most widely adopted open-source model, trained on over 6,500 species mostly from Europe and North America [Kahl 2021, Kahl et al. 2025]. However, its geographic training distribution raises the question of whether it adequately represents complex tropical soundscapes like the Pantanal.

The Pantanal, a UNESCO World Heritage Site and the world's largest tropical wetland, hosts approximately 617 documented bird species [UNESCO World Heritage Centre 2000]. Its soundscapes feature high species richness, frequent vocal overlap, and seasonal acoustic patterns. Pantanal species remain underrepresented globally, and BirdNET's performance on this fauna can vary considerably [Pérez-Granados 2023].

This work evaluates a transfer learning strategy to adapt BirdNET to ten highly recorded Pantanal bird species. We compare acoustic representations and data augmentation strategies, investigating whether regional fine-tuning enhances both local performance and generalizability. Our contributions include: (i) a curated dataset of Pantanal vocalizations from Xeno-Canto; (ii) a controlled comparison of MFCCs, Mel-spectrograms, and linear spectrograms under a standardized CNN, which evidences the challenges of feature extraction in small datasets and justifies the adoption of pre-trained embeddings; (iii) a BirdNET-based transfer learning approach; and (iv) evidence that regional feature adaptation improves species discrimination without local overfitting.

2. Related Work

2.1. Convolutional Neural Network Architectures for Bioacoustics

CNNs analyzing time-frequency representations dominate avian bioacoustics. Multi-representation approaches fusing spectrograms from different transforms (WT, HHT, STFT) consistently outperform single-representation models [Xie et al. 2022]. Subsequent work explored 2D and 3D-CNN architectures on scale frequency maps (SFMs), using sliding-windows to emphasize high-energy regions [Xie and Zhu 2022]. For noisy conditions, lightweight CNNs like CMS-H can outperform transformer-based models (e.g., AST), suggesting Transformers may not fully preserve spectrogram spatial structures [Wang et al. 2023]. Other studies combine self-supervised and supervised learning, such as ISCL [Zhang et al. 2023], or focus on interpretability, like AudioProtoPNet [Heinrich et al. 2025].

Although these approaches advance multiview fusion, self-supervised pretraining, and interpretability, they are primarily evaluated on large-scale datasets of temperate-zone species. The behavior of standardized CNNs on small, imbalanced datasets typical of underrepresented tropical regions remains an open research question, which is addressed in Section 4.1.

2.2. BirdNET as a Feature Extractor

BirdNET is widely adopted for large-scale avian identification [Kahl 2021]. Beyond its end-to-end classification role, recent studies repurposed its internal embeddings as transferable representations. Michaud et al. employed its 512-dimensional embeddings for unsupervised dataset cleaning [Michaud et al. 2023]. Clarfeld et al. used prediction confidence scores to reduce false-positive detections [Clarfeld et al. 2025]. Dematties et al. adopted BirdNET to interpret feature spaces learned by self-supervised architectures like DINO and VICReg [Dematties et al. 2024].

These studies establish BirdNET embeddings as versatile tools. However, none investigates retraining a lightweight fully supervised classification head on a regionally curated dataset while directly benchmarking against the original end-to-end model in both in-domain and out-of-domain settings, as evaluated in Sections 4.2 to 4.4.

2.3. Domain Shift in Bioacoustic Classification

Domain shift due to geographically imbalanced training data remains a critical challenge [Kazeneza et al. 2025]. Although BirdNET is comprehensive, its training data predominantly derive from North America and Europe. This generalization gap is exacerbated by soundscape differences; Heinrich et al. reported substantial declines in the Peruvian Amazon Basin due to extreme acoustic density [Heinrich et al. 2025]. Dematties et al. advocate for self-supervised learning on local recordings to mitigate this [Dematties et al. 2024].

While these studies highlight domain shift, they lack a controlled paired evaluation of regional fine-tuning using the same target species across both in-domain and external regions, which is directly addressed in this work.

3. Materials and Methods

3.1. Dataset

Recordings were collected via the Xeno-Canto API (`xeno-canto.org`), constrained to the Pantanal biome bounding box (14°S–22°S, 55°W–62°W), as illustrated in Figure 1. To avoid performance degradation due to data scarcity, only species with over 30 recordings were retained.



Figure 1. Delimitation of the Pantanal bounding box used for data collection.

Recordings were segmented into 3-second clips (Librosa v0.11.0) to match BirdNET's input length. Manual curation excluded prolonged silence, excessive noise, and multispecies overlap. The dataset was split into training (80%) and validation (20%) subsets, stratified by species, while ensuring that segments from the same original recording file were kept exclusively in either the training or the validation set. The ten selected species are listed in Table 1. The curation pipeline yielded 6,415 valid segments from 17,431 generated across 421 original recordings.

Table 1. Species composition and recorded counts at each processing stage.

Species	Common Name	Original	Postseg.	Postcuration
Cercomacra melanaria	Mato Grosso Antbird	65	3031	1098
Ortalis canicollis	Chaco Chachalaca	54	2995	1210
Synallaxis albilora	White-lored Spinetail	49	1964	714
Crypturellus undulatus	Undulated Tinamou	36	2598	346
Anodorhynchus hyacinthinus	Hyacinth Macaw	55	1632	561
Theristicus caerulescens	Plumbeous Ibis	34	1832	684
Cantorchilus guarayanus	Fawn-breasted Wren	32	1143	761
Campylorhynchus turdinus	Thrush-like Wren	34	727	266
Furnarius rufus	Rufous Hornero	31	560	277
Saltator coerulescens	Blue-gray Saltator	31	949	498

3.2. Audio Representations

Segments were resampled to 22,050 Hz and amplitude-normalized. Three acoustic representations were compared (`hop_length=512`):

- Linear Spectrogram: STFT (`n_fft=1024`), yielding 513×130.
- Mel-Spectrogram: `librosa` function (`n_fft=2048`, `n_mels=128`), yielding 128×130.
- MFCC: Derived using `n_mfcc=40` and `n_fft=2048`, yielding 40×130.

All representations were converted to a decibel scale where applicable.

3.3. Architecture and Transfer Learning

Baseline CNN experiments used a custom architecture comprising 3 convolutional blocks (32, 64, 128 filters; 3×3 kernels, ReLU), followed by max pooling (2×2) and batch normalization. Feature maps were passed to a 128-unit dense layer (50% dropout) and a 10-class softmax output. The network was optimized with Adam (initial learning rate = 0.001) and categorical cross-entropy loss over 50 epochs (batch size=32) using EarlyStopping (patience=8) and ReduceLROnPlateau (factor=0.5, patience=3). For transfer learning, BirdNET v2.4 served as the base model. Segments at 44,100 Hz were zero-padded to 48 kHz (144,000 samples) to meet BirdNET's requirements.

All BirdNET backbone layers were frozen. Three trainable layers were added: `Dense(512, ReLU)`, `Dropout(0.4)`, and `Dense(10, Softmax)`. Training ran identically to the baselines. BirdNET in its original configuration was also evaluated directly.

3.4. Evaluation Protocol

Models were evaluated via accuracy, macro/weighted precision, recall, and F1-score. SpecAugment [Park et al. 2019] was applied during training using two frequency masks (max 15 bins) and two time masks (max 30 frames) per sample. This acts as a powerful regularizer against reliance on narrow time-frequency regions, vital given the limited dataset size.

For external validation, 200 segments per species were collected outside the Pantanal via Xeno-Canto (2,000 segments total). The identical 10-class head was applied to these segments without retraining, testing whether the learned representations generalize to the same species in different geographic contexts.

4. Experimental Results

4.1. Baseline CNN Comparison

Table 2 outlines performance across configurations. Without augmentation, MFCCs achieved 77% accuracy (weighted F1 = 0.79), substantially outperforming linear spectrograms (44% accuracy, F1 = 0.41) and Mel-spectrograms (35% accuracy, F1 = 0.30).

Table 2. Overall performance across all experimental configurations.

Configuration	Accuracy	Macro P	Macro R	Macro F1	Weighted F1
Spectrogram	0.44	0.72	0.39	0.42	0.41
Mel-Spectrogram	0.35	0.68	0.25	0.26	0.30
MFCC	0.77	0.79	0.78	0.76	0.79
Spectrogram + SpecAug	0.78	0.78	0.75	0.75	0.79
Mel-Spec + SpecAug	0.74	0.80	0.71	0.69	0.73
MFCC + SpecAug	0.78	0.78	0.72	0.71	0.76

This result is notable because Mel-spectrograms are the most commonly reported representation in the bioacoustics literature, yet they produced the weakest baseline performance in this dataset. As illustrated in Table 2, we hypothesize this outcome stems from the compact, discriminative nature of MFCCs under limited data and class imbalance. However, alternative factors likely contribute to this disparity, including MFCCs' lower input dimensionality (which may mitigate rapid overfitting), potential mismatches in architecture capacity for larger spectrogram feature spaces, variations in amplitude normalization, and the inherent regularizing effect of the discrete cosine transform.

The application of SpecAugment considerably narrowed the performance gap, bringing both spectrogram variants close to the MFCC baseline, as demonstrated in Table 2. This convergence suggests that the initial underperformance of spectrogram-based representations was partly driven by overfitting to training set patterns rather than by inherent limitations of the representation itself. After augmentation, all three representations achieved comparable weighted F1 scores in the range of 0.73 to 0.79, indicating similar classification quality when controlling for class imbalance.

4.2. Transfer Learning with BirdNET

The original BirdNET achieved 90.22% accuracy (weighted F1 = 0.94), outperforming all CNN baselines. Following transfer learning via feature-extraction, the adapted model reached 95% accuracy (weighted F1 = 0.95), confirming the value of combining large-scale pretraining with regional fine-tuning, as reported in Table 3.

Table 3. BirdNET (original) versus the regionally adapted model on the Pantanal validation set.

Model	Accuracy	Weighted F1
MFCC + SpecAug (best baseline)	0.78	0.76
BirdNET (original, no fine-tuning)	0.90	0.94
BirdNET + Transfer Learning (adapted)	0.95	0.95

This substantial improvement of 17 percentage points over the best baseline (MFCC + SpecAug, at 78% accuracy) highlights the limitations of training CNN architectures from scratch on relatively small regional datasets. Even with effective regularization strategies, models trained from scratch often struggle to learn robust feature representations from limited examples. In contrast, leveraging BirdNET's pretrained embeddings provided a rich, high-level acoustic feature space right from the first epoch. Furthermore, the feature-extraction strategy proved to be highly computationally efficient. By keeping the deep backbone completely frozen, the trainable parameters were restricted to the shallow classification head, strictly mitigating the risk of overfitting on the Pantanal dataset. The adapted model converged rapidly, achieving its optimal validation performance at epoch 12 before early stopping was triggered. This demonstrates that regional fine-tuning does not require massive computational resources or extensive training times to yield highly significant performance gains.

4.3. Per-Species Performance and Confusion Analysis

Figure 2 compares the per-species F1-scores for the CNN baselines without data augmentation, revealing MFCC's consistent superiority across most taxonomic groups. This advantage is particularly evident for species with complex or high-frequency vocalizations. For instance, the performance gap is striking for *Synallaxis albilora* and *Cantorchilus guarayanus*, where Mel-spectrograms yielded F1-scores of only 0.11 and 0.22, respectively, compared to 0.79 and 0.82 achieved by the MFCC representations. As previously noted, the Mel scale's compression of higher frequency bands disproportionately degrades the representation of these specific vocalizations. In contrast, the cepstral coefficients provide a much more robust baseline representation across the entire evaluated taxonomic spectrum, maintaining higher discriminative power even for challenging classes before any data augmentation is applied.

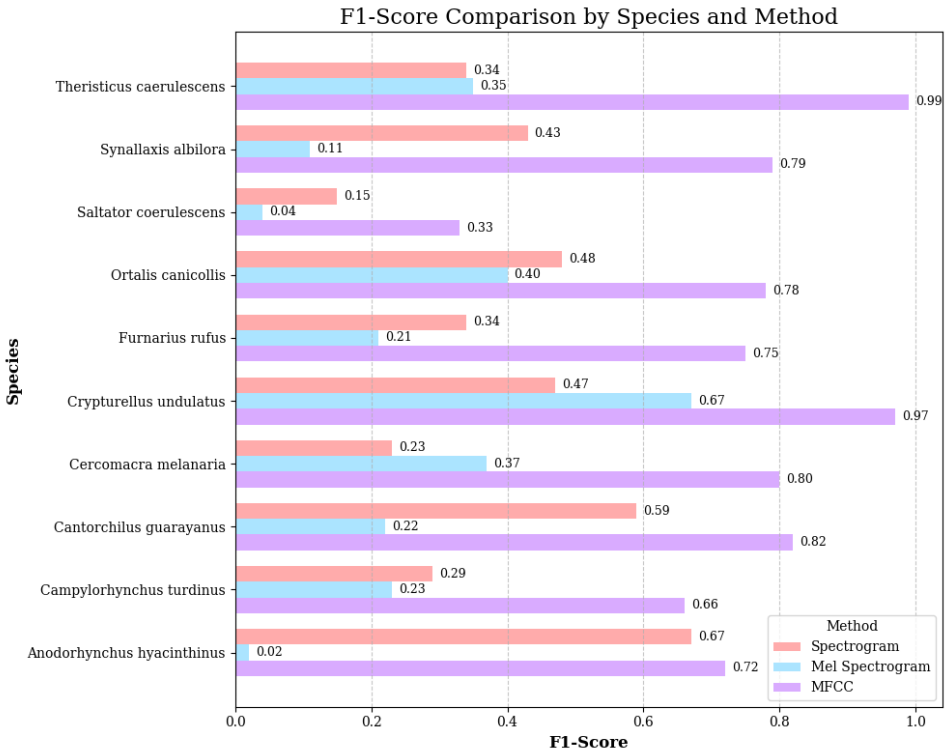


Figure 2. Per-species F1-score comparison across the three audio representations, evaluated without data augmentation.

Table 4 details the adapted model's performance. The transfer learning model achieved F1-scores above 0.88 for all ten species, with seven species exceeding 0.90. The best individual results were obtained for *Cercomacra melanaria* (F1 = 0.99), *Crypturellus undulatus* (F1 = 0.99), and *Cantorchilus guarayanus* (F1 = 0.97), which are species characterized by acoustically distinctive vocalizations in the Pantanal soundscape. *Anodorhynchus hyacinthinus* (F1 = 0.95) and *Ortalis canicollis* (F1 = 0.94) also performed strongly.

Table 4. Per-species classification report for the BirdNET + Transfer Learning model.

Species	Precision	Recall	F1-score	Support
Anodorhynchus hyacinthinus	0.99	0.92	0.95	255
Campylorhynchus turdinus	0.93	0.89	0.91	75
Cantorchilus guarayanus	0.95	1.00	0.97	157
Cercomacra melanaria	0.99	0.99	0.99	150
Crypturellus undulatus	0.99	0.99	0.99	97
Furnarius rufus	1.00	0.79	0.88	106
Ortalis canicollis	0.89	1.00	0.94	197
Saltator coerulescens	0.86	0.98	0.91	49
Synallaxis albilora	0.95	0.78	0.86	27
Theristicus caerulescens	0.84	0.97	0.90	32

The two species with comparatively lower F1-scores were *Synallaxis albilora* (0.86) and *Furnarius rufus* (0.88). For *Synallaxis albilora*, the score is linked to limited validation data (n=27) and known challenges in modeling acoustic backgrounds for this neotropical species [Ventura et al. 2024]. For *Furnarius rufus*, 12 validation segments were classified as *Ortalis canicollis*, while nine were distributed among *Saltator coerulescens*, *Synallaxis albilora*, and *Theristicus caerulescens*, as revealed by the confusion matrix in Figure 3. This suggests partial overlap in spectro-temporal characteristics. A similar pattern occurred for *Theristicus caerulescens* (F1 = 0.90).

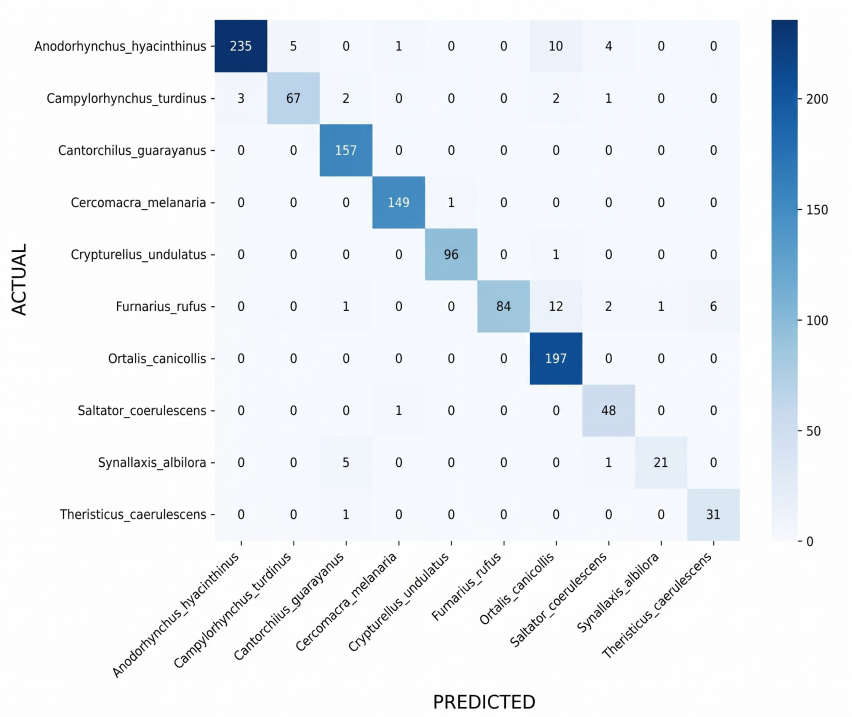


Figure 3. Confusion matrix of the BirdNET + Transfer Learning model evaluated on the Pantanal validation set.

4.4. Cross-Region Evaluation

To evaluate geographic overfitting, models were tested on 2,000 recordings collected outside the Pantanal, as mapped in Figure 4.

Table 5. Cross-region evaluation results for BirdNET (original) and the transfer learning model.

	Pantanal			External		
Model	F1 (wtd)	Acc.	Gain	F1 (wtd)	Acc.	Accuracy Gain
BirdNET (original)	0.94	90.22%	—	0.94	88.75%	—
BirdNET + Transfer	0.95	95.00%	+4.78 pp	0.92	92.00%	+3.25 pp

As detailed in Tabl 5, the adapted model achieved 92.00% accuracy externally, representing a 3.25 percentage-point increase over BirdNET (88.75%). However, this

gain in overall accuracy was accompanied by a slight decrease in the weighted F1-score (from 0.94 to 0.92). This indicates a trade-off: while the adapted model correctly identifies a higher total number of external segments, its precision-recall balance across different species exhibits minor variations outside the training domain.

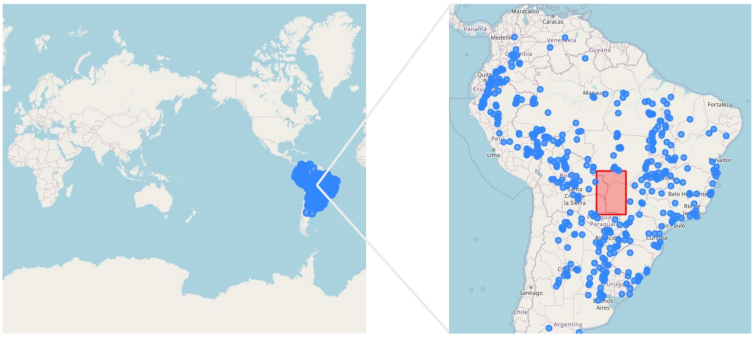


Figure 4. Geographic distribution of recordings used for external validation.

5. Discussion

5.1. Practical Implications of the Accuracy Gain

The 4.78 percentage-point improvement in accuracy resulting from regional adaptation corresponds to a substantial reduction in classification errors within automated monitoring pipelines, which often process thousands of 3-second audio segments in a single monitoring session. For example, a model operating at 95% accuracy generates approximately half as many errors as a model operating at 90% accuracy, thereby providing more reliable inputs for downstream ecological analyses, including species richness estimation, population density assessment, and vocalization activity monitoring. In an ecologically sensitive biome such as the Pantanal, where false negatives may result in the systematic underdetection of rare or threatened species, such improvements have important implications for the reliability and accuracy of biodiversity monitoring outcomes.

Notably, the species-level results presented in Section 4.3 indicate that these performance gains are not uniformly distributed across taxa. Instead, the largest improvements are observed for species whose vocalizations are underrepresented in BirdNET's original training dataset. From an ecological perspective, this outcome is more meaningful than an equivalent increase in overall accuracy, as it helps reduce systematic biases against species that receive limited representation in globally curated datasets.

5.2. The Importance of training with regional data

The results of this study provide empirical support for a broader principle in applied bioacoustics: general-purpose models trained on globally aggregated data are not interchangeable with regionally adapted models in biodiversity-rich tropical contexts. The original BirdNET achieved a weighted F1 of 0.94 and 90.22% accuracy, which is already impressive for a model not specifically designed for Pantanal fauna; nevertheless, this reflects the geographic bias documented by Pérez-Granados [Pérez-Granados 2023] and Kazeneza et

al. [Kazeneza et al. 2025]. The consistent 4 to 5 percentage-point gap between the original and adapted models, replicated across both in-domain and out-of-domain evaluations, suggests that this bias is not simply a matter of missing species labels but reflects a deeper misalignment in the learned acoustic feature space.

The feature extraction strategy adopted in this work, which involves freezing all the BirdNET backbone layers and training only a shallow classification head, is notable for its efficiency. It requires neither large volumes of labeled data nor extensive computational resources. The model achieved its best validation performance at epoch 12 before early stopping was triggered, underscoring that even a brief fine-tuning phase on a curated regional dataset is sufficient to produce meaningful realignment of the classification boundary. These findings have practical implications for conservation programs in other underrepresented biomes: the barrier to regional adaptation is lower than is often assumed, provided that a modest volume of quality-curated vocalizations is available.

5.3. Cross-Regional Generalization Analysis

A central concern in regional adaptation is whether fine-tuning produces a model that has genuinely learned more discriminative acoustic representations, or one that has merely memorized environment-specific patterns and has lost effectiveness elsewhere. The cross-region evaluation results presented in Section 4.4 address this concern directly.

The fact that the adapted model maintains a 3.25 percentage-point advantage in absolute accuracy on external recordings, even with the slight trade-off in the weighted F1-score, indicates that the adaptation largely preserved robust, species-specific representations. This transparent assessment of the model's capabilities reflects a commitment to the responsible and inclusive use of artificial intelligence in conservation. Rather than overfitting and failing silently in new environments, the model demonstrates generalizable improvements that can reliably support automated environmental monitoring across diverse habitats, contributing to biodiversity preservation as a common good.

6. Conclusion

This study investigated whether regional adaptation of BirdNET v2.4 could improve the discrimination of ten Pantanal bird species and whether the resulting improvements would generalize beyond the target region. Three main findings support the effectiveness of this approach. First, among CNN models trained from scratch, MFCC representations achieved superior performance compared with linear and Mel-spectrograms without data augmentation; the application of SpecAugment substantially reduced these differences, resulting in comparable weighted F1-scores across all three representations (0.73–0.79) and highlighting its value as a regularization strategy for small and imbalanced datasets. Second, a lightweight transfer-learning approach-freezing the BirdNET backbone and training only a shallow classification head-increased accuracy from 90.22% to 95% on the Pantanal validation set, with gains concentrated in species underrepresented in BirdNET's original training data. Third, the adapted model maintained a 3.25 percentage-point accuracy advantage over the original BirdNET on an external dataset of the same species collected outside the Pantanal, confirming that adaptation improved representations in a manner that generalizes across regions.

The primary limitation of this study is restricted data availability: the ten species analyzed represent less than 2% of the approximately 617 bird species documented in

the Pantanal. Future work should expand taxonomic coverage by incorporating passive acoustic monitoring data across Pantanal subregions and seasons, explore partial fine-tuning by unfreezing upper BirdNET layers, apply class-balancing techniques for acoustically similar species pairs such as *Furnarius rufus* and *Theristicus caerulescens*, and evaluate well-established architectures (such as EfficientNet or MobileNet) as baselines instead of custom configurations.

In conclusion, lightweight feature-extraction-based adaptation of a large-scale bioacoustic model constitutes a practical and effective strategy for improving automated bird monitoring in complex tropical soundscapes, and supports the adoption of regionally specialized models as a complement to general-purpose tools in conservation programs for underrepresented biomes such as the Pantanal.

Acknowledgments

We would like to extend our thanks to the Computational Bioacoustics Research Unit (CO.BRA) at the Federal University of Mato Grosso for their valuable 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.

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.

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.

Gregory, R. D. and van Strien, A. (2010). Wild Bird Indicators: Using Composite Population Trends of Birds as Measures of Environmental Health. *Ornithological Science*, 9(1):3 – 22.

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

Hooper, D. U. e. a. (2005). Effects of biodiversity on ecosystem functioning: a consensus of current knowledge. *Ecological Monographs*, 75:3–35.

Kahl, S., Wood, C., and Klinck, H. (2025). Birdnet model v2.4 (version 1) [figure].

Kahl, S. e. a. (2021). Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61.

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. 13.

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.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (2019). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Interspeech 2019*, pages 2613–2617.

Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y., and Sainath, T. (2019). Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219.

Pérez-Granados, C. (2023). Birdnet: applications, performance, pitfalls and future opportunities. *Ibis*, 165:1068–1075.

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

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

Ventura, T. M., Ganchev, T. D., Pérez-Granados, C., de Oliveira, A. G., de S. G. Pedroso, G., Marques, M. I., and Schuchmann, K. L. (2024). The importance of acoustic background modelling in cnn-based detection of the neotropical white-lored spinetail (aves, passeriformes, furnaridae). *Bioacoustics*, 33:103–121.

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.

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.