Bioacoustics
The International Journal of Animal Sound and its Recording

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tbio20

The importance of acoustic background modelling in CNN-based detection of the neotropical White-lored Spinetail (Aves, Passeriformes, Furnaridae)

Thiago M. Ventura, Todor D. Ganchev, Cristian Pérez-Granados, Allan G. de Oliveira, Gabriel de S. G. Pedroso, Marinez I. Marques & Karl-L. Schuchmann

To cite this article: Thiago M. Ventura, Todor D. Ganchev, Cristian Pérez-Granados, Allan G. de Oliveira, Gabriel de S. G. Pedroso, Marinez I. Marques & Karl-L. Schuchmann (2024) The importance of acoustic background modelling in CNN-based detection of the neotropical White-lored Spinetail (Aves, Passeriformes, Furnaridae), Bioacoustics, 33:2, 103-121, DOI: 10.1080/09524622.2024.2309362

To link to this article: https://doi.org/10.1080/09524622.2024.2309362

Published online: 13 Feb 2024.

BIOACOUSTICS
2024, VOL. 33, NO. 2, 103–121
https://doi.org/10.1080/09524622.2024.2309362

Taylor & Francis
Taylor & Francis Group

The importance of acoustic background modelling in CNN-based detection of the neotropical White-lored Spinetail (Aves, Passeriformes, Furnaridae)

Thiago M. Ventura[a,b], Todor D. Ganchev[a,c], Cristian Pérez-Granados[a,d], Allan G. de Oliveira[a,b], Gabriel de S. G. Pedroso[a,b], Marinez I. Marques[a,b,e] and Karl-L. Schuchmann[a,b,e,f,g]

[a]National Institute for Science and Technology in Wetlands (INAU), Federal University of Mato Grosso (UFMT), Cuiabá, Brazil; [b]Institute of Computing, Federal University of Mato Grosso, Cuiabá, Brazil; [c]Department of Computer Science and Engineering, Technical University of Varna, Varna, Bulgaria; [d]Ecology Department, Alicante University, Alicante, Spain; [e]Post-Graduate Program in Zoology (PPGZOO/UFMT), Federal University of Mato Grosso, Cuiabá, Brazil; [f]Vertebrate Department (Ornithology), Zoological Research Museum A. Koenig (ZFMK), Bonn, Germany; [g]Faculty of Mathematics and Natural Sciences, University of Bonn, Bonn, Germany

ABSTRACT

Machine learning tools are widely used in support of bioacoustics studies, and there are numerous publications on the applicability of convolutional neural networks (CNNs) to the automated presence-absence detection of species. However, the relation between the merit of acoustic background modelling and the recognition performance needs to be better understood. In this study, we investigated the influence of acoustic background substance on the performance of the acoustic detector of the White-lored Spinetail (*Synallaxis albilora*). Two detector designs were evaluated: the 152-layer ResNet with transfer learning and a purposely created CNN. We experimented with acoustic background representations trained with season-specific (dry, wet, and all-season) data and without explicit modelling to evaluate its influence on the detection performance. The detector permits monitoring of the diel behaviour and breeding time of White-lored Spinetail solely based on the changes in the vocal activity patterns. We report an advantageous performance when background modelling is used, precisely when trained with all-season data. The highest classification accuracy (84.5%) was observed for the purposely created CNN model. Our findings contribute to an improved understanding of the importance of acoustic background modelling, which is essential for increasing the performance of CNN-based species detectors.

ARTICLE HISTORY
Received 13 January 2023
Accepted 5 January 2024

KEYWORDS
Acoustic activity detection; bird sound recognition; computational bioacoustics; convolutional neural networks; Pantanal; transfer learning

Introduction

Automated acoustic wildlife recognition is challenging due to numerous essential factors that cannot be controlled but are known to influence classification accuracy significantly. Among these are the varying distance between the vocalising individual and the recording device, the singing direction of the target species to

CONTACT Karl-L. Schuchmann klschuchmann@googlemail.com

the microphones, the simultaneous sound production of two or many species, the presence of concurrent sounds from other sources covering the same frequency band, the fast varying changes in the ambient noise level, the composition of environmental noise, and other variables (Pérez-Granados et al. 2019; Zor et al. 2019).

Birds are the group of terrestrial taxa most commonly surveyed using autonomous recording units (reviewed by Sugai et al. 2019), and several studies have used automated bird recognition through various machine learning (ML) methods, such as hidden Markov models (HMMs) (Oliveira et al. 2015; Lasseck 2018), K-nearest neighbours (Tuncer et al. 2021), recurrent neural networks (Noumida and Rajan 2022), and deep learning (DL) (Hidayat et al. 2021). In the past five years, DL methods have been established as a helpful tool to support bird recognition efforts in bioacoustic and ecoacoustic studies (Stowell 2022). The advances in DL have contributed to the improved performance of automated species recognition, and presently, deep neural networks (DNNs) are perceived as the next best candidate for coping with the scalability problem in computational bioacoustic technology.

The broader use of convolutional neural networks (CNNs) in recent years is primarily due to the availability of various open-source implementations and cheap computing resources. Recent advances in CNN-based modelling, including new architectures (Kim 2017) with their enhanced ability to extract features automatically (Murphy 2016; Stowell 2022); automatic feature detection (Tuncer et al. 2021); data augmentation (Knight et al. 2020); combining features to allow the capability of distinguishing between valuable and not useful features (Stowell et al. 2019); and feature extraction to remove or minimise ambient noise impact (Florentin et al. 2020), contributed to addressing real-life problems that were considered intractable earlier.

Previous studies have investigated the impact of different training datasets on bird detection performance. For example, Adavanne et al. (2017) demonstrated the influence of considering different species, variable weather conditions, and habitat characteristics on algorithm performance. Concerning acoustic background variability, Knight et al. (2020) focused on background characteristics and the objects being detected to evaluate the application of distance sampling in the Australian wet tropics. Knight et al. (2020) focused on minimising the problem with multiple features for different training sets through generalisation. Nevertheless, other studies (Anderson et al. 2015) selected specific days of vocalisations instead of multiple data without prior knowledge to distinguish among different background sounds.

In the present work, we evaluate the impact of different acoustic background modelling implementations on the performance of a species-specific CNN-based detector of White-lored Spinetail (*Synallais albilora*, Spinetail hereinafter) vocalisations. Specifically, we assessed the effect of using or not using representative seasonal recordings of the acoustic environment in two main setups: (i) reusing pretrained CNN models that were fine-tuned with transfer learning and (ii) creating a purposely trained detector with and without using season-specific soundscape recordings. We conducted experiments with datasets collected in the two primary season-specific environments (dry and wet) for that purpose. These were compared with implementations where blended dry and wet data were used, and explicit acoustic environment modelling was unavailable. In all

experiments, we evaluated the detection accuracy for Spinetail vocalisations in field recordings collected in the Brazilian Pantanal.

The remaining exposition is structured as follows: In the next section, we describe the biology and vocal behaviour of the Spinetail, the proposed tool, and its training process. Next, we have the experimental results of the influence of the acoustic background modelling and the Diel and seasonal pattern of acoustic activity. In the Conclusions section, we present our final considerations.

Materials and methods

Study species

White-lored Spinetail was selected as the target species because it is a common bird in our study area (Brazilian Pantanal, see next section), and there is very limited knowledge about its ecology (Rubio and Pinho 2008). The Spinetail shows a distribution range restricted to the Pantanal and the borders of Humid Chaco of Bolivia and Paraguay (Smith 2020). We therefore consider it interesting to improve our knowledge of the vocal behaviour of such a restricted range species. Spinetail usually inhabit gallery forests and scrublands near watercourses (Lowen and Bernardon 2010). The only detailed study of the species was also performed in the Brazilian Pantanal and focused on its breeding biology. The authors found that the peak of breeding activity occurred during October, which coincides with the onset of the rainy season in the Brazilian Pantanal, although active nests were found from late July to December (Rubio and Pinho 2008). The species utters two main vocalisation types: songs and calls (Figure 1). The primary vocalisation is the song, a sharp sound composed of two syllables, 'pit-tuiii', with the second note higher (Gwynne et al. 2010). The call of the species is a short sound composed of a unique syllable, 'kiiiii'. The peak frequency of both vocalisations is approximately 2–3 kHz.

Study area and recording protocol

This study was performed in the northeastern part of the Brazilian Pantanal, near the SESC Pantanal (Serviço Social do Comércio, SESC, Poconé, Mato Grosso, Brazil; 16°30′S, 56°25′W). The area is located within the floodplain of the Cuiabá River (Figure 2), which

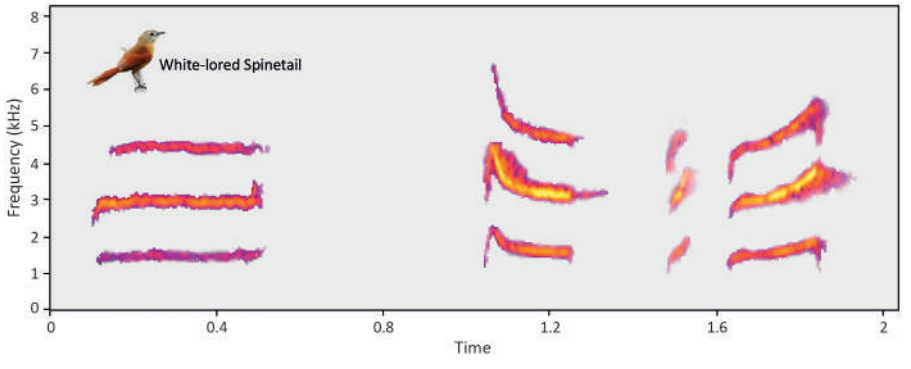

Figure 1. Spectrograms of White-lored Spinetail call (left) and song (right).

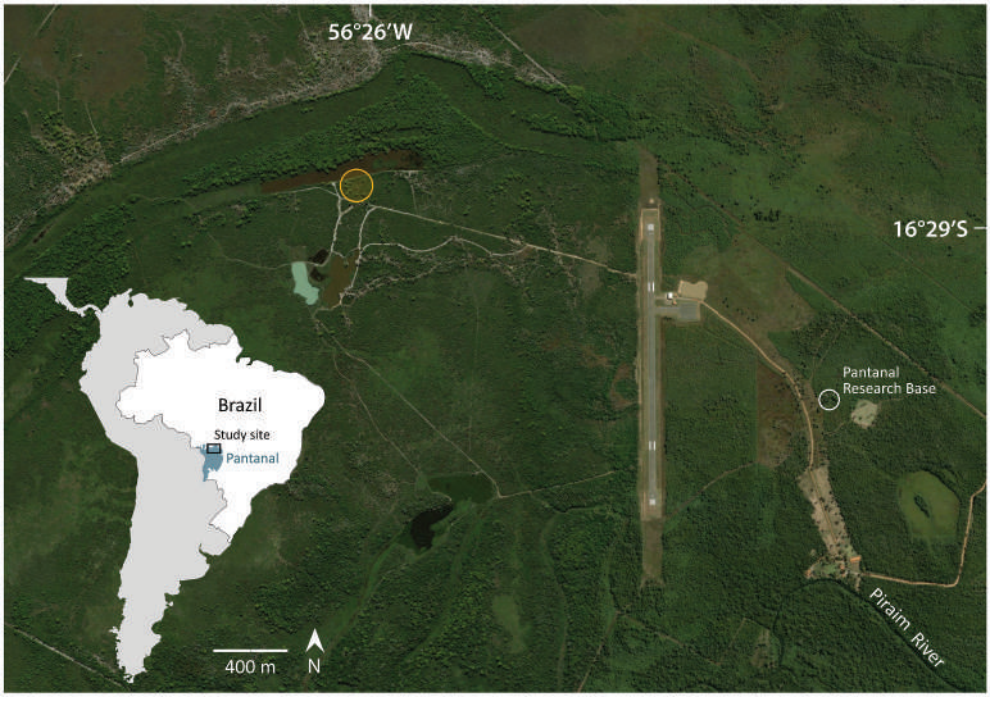

Figure 2. Location of the study area in Brazil and the detailed location of the recording site (yellow circle).

is seasonally inundated during the wet season from October to April due to flooding of the Paraguay River. The dominant vegetation is a mosaic of different forest formations and savannas. The regional climate is tropical and humid, with a mean annual temperature of approximately 24°C and an average annual rainfall of 1,000–1,500 mm (Junk et al. 2006).

One Song Meter SM2 recorder (Wildlife Acoustics, www.wildlifeacoustics.com), which was equipped with two SMX-II omnidirectional microphones (sensitivity −36 ± 4 dB, signal-to-noise ratio >62 dB, frequency response: flat 20 Hz – 20000 Hz) displaced at 180 degrees in the same plane, was used to capture the overall soundscapes. We recorded in the two-channel.wav format for the first 15 minutes of each hour (24 hours per day) with a sampling rate of 48 kHz and 16 bits per sample, storing data on the internal SD memory cards. The recorder was checked weekly to replace batteries and download data. In the current study, we used only the left channel of the SM2 recordings to reduce the amount of data and speed up processing. The SM2 was active from June 2015 to May 2016.

Datasets

We used four categories of audio recordings: (i) annotated excerpts of field recordings containing vocalisations of the target species, which we refer to as the *Spinetail dataset* hereafter; (ii) soundscape recordings, which do not contain the target species and which

Table 1. Summary of datasets in this study.

Dataset	Number of files	Cumulative duration (s)	Average file duration (s)	Recording months	Dataset usage
Spinetail-CALL	205	89.6	0.44	Dec 2015–May 2016	Train and Validation
Spinetail-SYL1	759	255.4	0.34	Dec 2015–May 2016	Train and Validation
Spinetail-SYL2	719	264.7	0.37	Dec 2015–May 2016	Train and Validation
BG-Wet dataset	8	2,271.5	283.9	Oct 2015–Feb 2016	Train and Validation
BG-Dry dataset	8	3,503.4	437.9	Jun-Aug 2015	Train and Validation
Other Birds dataset	1,532	46,109.3	30.1	Xeno-canto*	Train and Validation
Evaluation dataset	22	19,761.1	898.2	Oct-Nov 2015	Detector Evaluation
Monitoring dataset	2,888	2,595,366.0	898.7	Jul-Oct 2015	Activity monitoring

Notations: CALL = call, SYL1 = first syllable of the song, SYL2 = second syllable of the song. *Recordings were extracted from Xeno-Canto.

we refer to as to *acoustic background datasets*; (iii) a collection of sounds of other bird species that are vocally active in the Pantanal area, which we refer to as the *OtherBirds dataset*; and (iv) two sets of soundscape recordings to evaluate the performance of CNN-based detectors or to study the acoustic activity of the Spinetail, which we refer to as the *evaluation dataset* and *monitoring dataset*. We summarise all datasets used in this study in Table 1.

Spinetail dataset

The Spinetail dataset is representative of the acoustic emissions of the target species and was used to train and validate the CNN models. It consists of 1683 short snippets with tagged vocalisations of the target species. These snippets are excerpts of field recordings manually cut from soundscape recordings collected with the SM2 device between December 2015 and May 2016. These vocalisations are distributed among three categories: call (205 instances), the first syllable in a song (759 instances), and the second syllable in a song (719 instances). In total, there are approximately 10 minutes and 10 seconds. The average duration of each snippet is 0.36 seconds, where the first syllable of a song has an approximate duration of 0.34 seconds, and the second has 0.37 seconds. The Spinetail calls (Figure 3) are slightly longer, with an approximate duration of 0.44 seconds.

OtherBirds dataset

The *OtherBirds dataset* is a collection of nontarget sounds, which we used as negative examples during the training and validation of the CNN models to strengthen the selective capability of the Spinetail detectors. It consists of sound emissions of 40 bird species vocally active in the Pantanal area. Following Ventura et al. (2015), we retrieved recordings from the Xeno-Canto Archive (XENO-CANTO 2022). These were mainly processed excerpts from soundscapes or field recordings made with a narrow-angle directed microphone tagged for specific bird species. The *OtherBirds dataset* consists of 1532 recordings that contain the vocalisation of at least one avian species different from

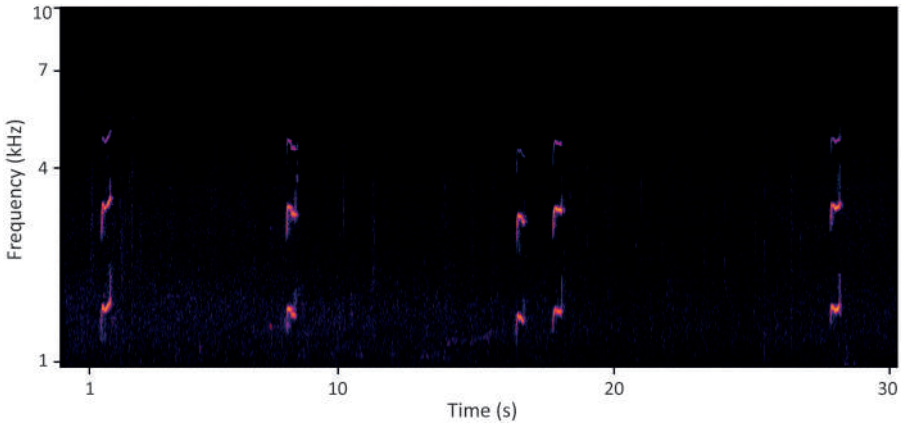

Figure 3. Example of spectrogram obtained from audio recording of the Spinetail.

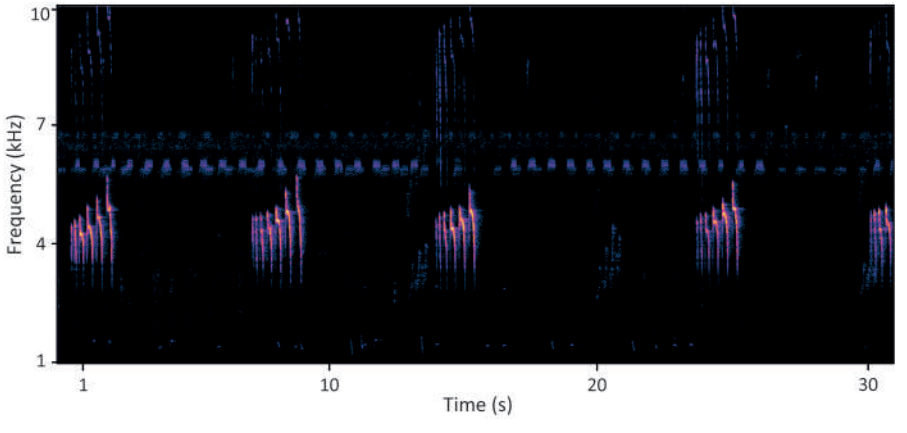

Figure 4. Example of a spectrogram obtained from audio recording from the *OtherBirds dataset*.

our target species (Figure 4). In some cases, the dominant vocalisations in these files were accompanied by sounds from other sources and contained episodes with a mixture of distant sounds from the environment. Here, we used these recordings 'as is', i.e. without further processing, because the ambient interference is negligible in their majority. Therefore, we consider the *OtherBirds dataset* representative of the bird species vocally active in the Pantanal area but not of the acoustic environment in which they were recorded. The cumulative duration of all recordings is approximately 12.8 hours, and the average duration of each audio file is approximately 30.1 seconds.

Datasets used for the acoustic background representation

The datasets used for the acoustic background representation are subsets of recordings that do not contain sounds of the target species (Figures 5 and 6), which were used as examples of the nontarget class to train the CNN-based detectors. These recordings are excerpts of soundscapes collected with the SM2 device and were selected as representative of the acoustic environment of the

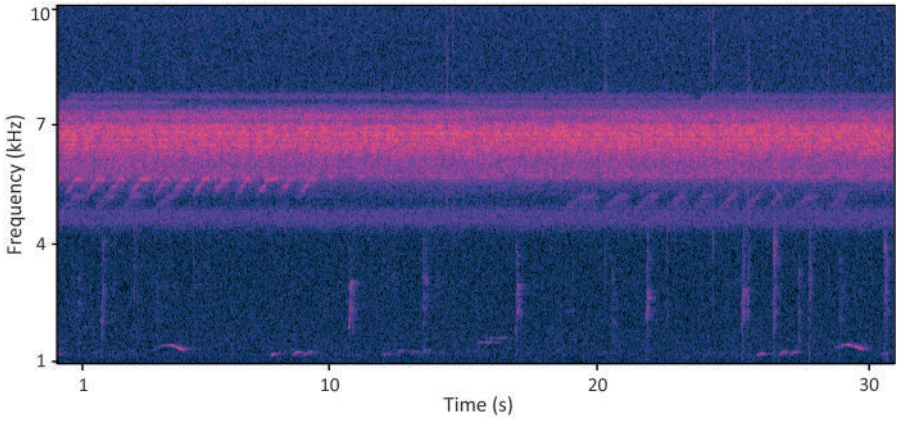

Figure 5. Example of a spectrogram obtained from an audio recording of a forest biome during the wet season.

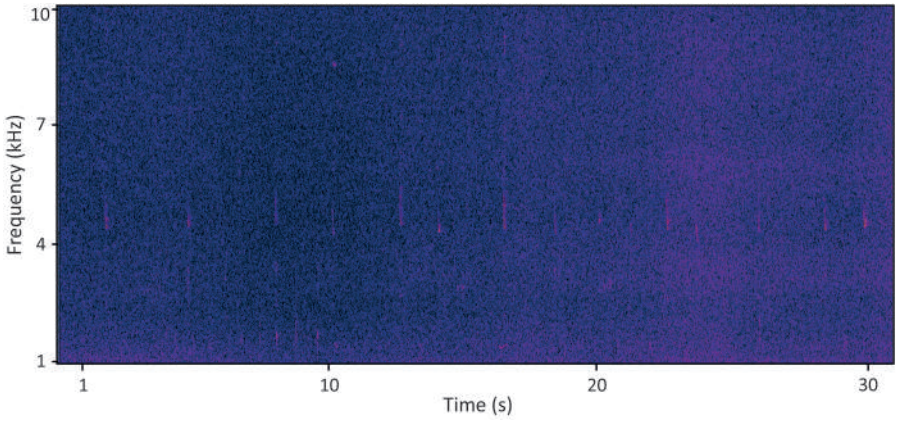

Figure 6. Example of a spectrogram obtained from an audio recording of a savanna biome during the dry season.

study area – savanna and forest biomes during the dry and wet seasons. Each acoustic condition is represented by four soundscape recordings grouped by Pantanal seasonality. In the experiments, we consider two season-specific datasets: (i) the wet season background dataset, referred to as BG-Wet, which consists of 8 soundscape recordings collected between October 2015 and February 2016, with a total duration of 2271.5 seconds and an average duration of approximately 283.9 seconds; and (ii) the dry season background dataset, referred to as BG-Dry, which consists of 8 soundscape recordings collected between June and August 2015, with a total duration of 3503.4 seconds and an average duration of approximately 437.9 seconds. In addition, we considered the combined dataset containing all 16 recordings, with a total duration of 1 hour and 36 minutes. We refer to it as the Wet+Dry background dataset BG-Dry+Wet hereafter.

Evaluation dataset

The *Evaluation dataset*, which consists of soundscape recordings annotated for Spinetail vocalisations, was used to assess the efficiency of background representations depending on the season and the detector architecture and training. It consists of 22 SM2 soundscape recordings collected in the study area between October and November 2015. Each recording has a duration of 14 minutes and 55 seconds of audio, and the cumulative duration of the dataset is 5.49 hours. In 16 of these recordings, we observed the presence of Spinetail, and the remaining six contained sounds of various other species and sounds due to natural phenomena. We timestamped only the Spinetail vocalisations. These were used to evaluate the recognition performance of the CNN-based detectors with different background representations. Experimental results are discussed in the Influence of the acoustic background modelling section (see Discussion).

Monitoring dataset

The *Monitoring dataset* consists of soundscape recordings, which are not tagged. These were used to study the hourly and daily patterns of vocal behaviour of the Spinetail in the Pantanal area during different stages of its lifecycle. It consists of 2888 SM2 soundscape recordings collected for the first 15 minutes of each hour between July and October 2015, with a cumulative duration of over 720 hours of audio. The observed behaviour of the target species is discussed in the Discussion section Diel and seasonal pattern of acoustic activity.

CNN-based detector

All datasets were processed, following the common protocol outlined in the Audio processing section, to segment the audio into frames of equal sizes, which were subsequently transformed into spectrograms. These spectrograms were treated as colour images and fed to the CNNs as described in sections *Transfer learning-based CNN model* and *Purposely developed CNN architecture*.

Spinetail audios with an average duration of 0.36 seconds were prescreened to estimate the minimum duration m recording in the dataset, which was 0.15 seconds. In the subsequent processing, we used only the first m seconds of each snippet in the Spinetail dataset.

The recordings in the *OtherBirds* dataset are of different lengths with an average duration of approximately 30.1 seconds, and they contain well-discerned sounds of various birds or background noise only. To reduce the chance of selecting noise-only segments from each of the *OtherBirds* recordings, we kept two snippets of m seconds, representing the initial 40%-60% of the audio.

The soundscape recordings from the wet and dry season background datasets were segmented into 500 snippets with a duration of m seconds each.

The soundscape recordings from the evaluation and monitoring datasets were segmented and evaluated every m seconds.

Audio processing

Once the audio frames with equal duration were obtained, the spectrogram was computed through a 512-point Fast Fourier Transform. Considering the sampling rate of 48

kHz, we used a Hamming window of 480 samples sliding with overlaps of 360 samples. From the spectrogram, we selected the frequency range [3 kHz, 4 kHz] as it was experimentally found to be advantageous after an extensive investigation of various subbands. This specific frequency range conveys the critical portion of the Spinetail calls spectrum and most of the energy of its song sounds.

In general, the Spinetail vocalisations cover the frequency range of 1 kHz – 10 kHz, which overlaps with the sounds of other bird species (Zhao et al. 2019) and many other animals. However, using the frequency range of [3 kHz, 4 kHz], we retained most of the information for Spinetail sounds while inhibiting most of the background noise. The magnitude spectrogram of the selected frequency subband was used as a colour image fed into the CNN input.

Transfer learning-based CNN model

Training machine learning models is a complex and time-consuming process. In different scenarios, it is necessary to use a larger dataset and adjust the model's parameters to be able to be used outside the training environment. A common method to minimise these problems is transfer learning, which consists of using a pretrained set of weights to adapt to the target task. Commonly, the weights of the models are trained with a large dataset, reducing the time necessary to look for an architecture and set of parameters for the specified dataset (Diment and Virtanen 2017).

The success of ResNet-based transfer learning in acoustic presence-absence detection for various animal species (e.g. Shiu et al. 2020; Ruan et al. 2022) encouraged us to evaluate this approach in recognising Spinetail vocalisations. In the present study, we used transfer learning in the model 152-layer ResNet (He et al. 2016), which was known to be trained with a considerable volume of data from the Imagenet Database and is reportedly capable of classifying various types of data, including bird sounds (Kahl et al. 2017).

The ResNet-152 model was initially trained with colour images (He et al. 2016). In the current research, we used a learning rate of 0.001 with the Adam optimiser for 30 epochs. We kept the colour image setup since it was reported to be more advantageous than other setups that rely on species recognition based on greyscale spectrograms (Dufourq et al. 2022).

In the following, we consider the soundscape spectrograms as a colour image, with a resolution of 224×224 pixels, represented by RGB-colour channel images. Once the spectrograms were computed, we fed them into the ResNet-152 model and froze the weights of all layers except the output layer adapted to our classes and the proposed model.

Purposely developed CNN architecture

We trained a CNN-based detector for Spinetail, using all datasets and audio preprocessing described above. The proposed CNN has six trainable layers, including four convolutional layers, one fully connected layer with Swish, and an output softmax layer (Figure 7). This CNN architecture was obtained after extensive experimentation with combinations of convolutional, pooling, and dropout layers to avoid overfitting. The experimentation was focused on either modifying the layers or incrementing the layers, concerning the number of layers and processing units at each layer (e.g. 2 convolutional

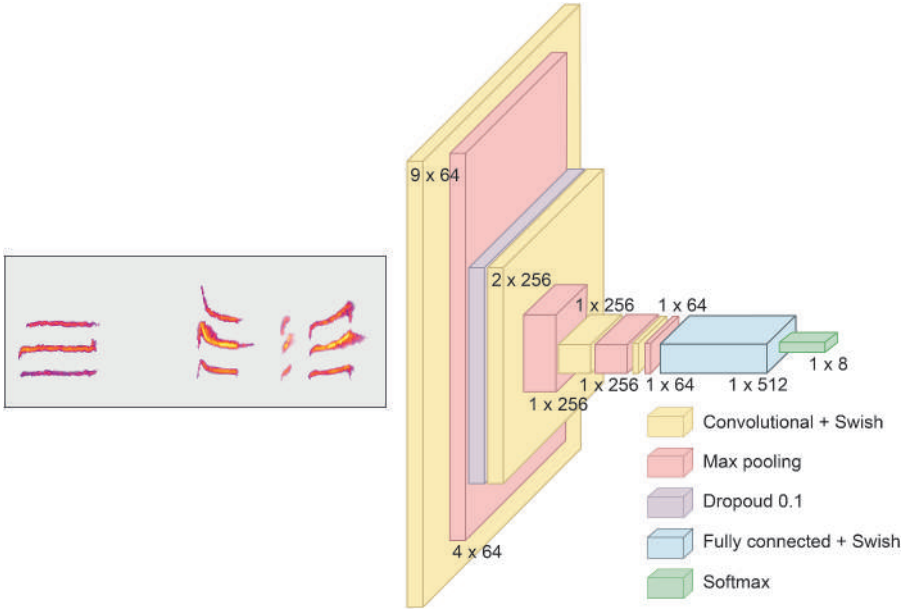

Figure 7. The proposed CNN architecture is tuned to detect the acoustic activity of White-lored Spinetail.

with 64 feature maps or 1 convolutional with 128 feature maps), and activation functions (e.g. Hyperbolic Tangent, ReLU, and Swish) to extract the spectrogram's main characteristics. The dense final layer was trained to make the classification.

The first convolutional layer contained 64 filters with a kernel size of 3×3, followed by a pooling layer of size 2×2, which slides every two frames, acquiring the maximum number for each window. We applied a 1-D convolution instead of 2-D convolutional layers when classifying or identifying birds, which is a common approach similar to that in Lasseck (2018) and Kiranyaz et al. (2019). In the drop-out layer that follows, 10% of the weights are dropped by forcing their value to 0. The subsequent convolution and pooling layers use 256, 256, and 64 filters. Despite the identical number of filters and size of windows, the next convolutional and pooling layers do not pad the data; thus, the data dimensions remain unchanged. Finally, the output of the fourth pooling layer is flattened to a fully connected layer of 512 units and a dense final layer for classification in eight classes. This is implemented as a softmax layer so that it outputs eight values that represent the probability for each class defined above – four for the acoustic background (savanna wet, savanna dry, forest dry, forest wet), three for the Spinetail vocalisations (call, first syllable of the song, and second syllable of the song), and category OtherBirds, which represents the sounds of other typical bird species for the same region that were expected to inhabit or migrate through the studied area. This approach is helpful because it provides additional information that allows us to distinguish different vocalisations of the Spinetail and different acoustic contexts. The latter may be helpful when processing unlabelled recordings, for which it is not apparent during which seasons were recorded, in which biome, and whether the acoustic presence-absence of Spinetail occurs together with the presence-absence of any other bird species.

Experimental protocol

Since the present study focuses on evaluating the merit of the acoustic background representation, in the following, we consider the CNN-based detector a binary classifier with two possible outcomes of the detection process: either Spinetail acoustic activity was detected or not. Thus, we introduce the postprocessing of the CNN output to obtain in one category any of the Spinetail vocalisations (call, first syllable of the song, and second syllable of the song) and every other sound (*OtherBirds* and acoustic background) to verify whether a Spinetail vocalisation was detected.

The CNN-based detectors were trained in four setups using different subsets of audio to model the acoustic background: (i) using only season-specific background data, which resulted in independent wet and dry season models; (ii) combining the available data from both seasons; and (iii) no explicit acoustic background modelling.

Data augmentation was used before the training by decreasing the linear magnitude of each audio recording at a random rate of 40%-60%, duplicating the data in the training set.

The performance of the CNN detectors was evaluated in terms of accuracy (Equation. 1), precision Equation. 2), and recall (Equation. 3), where *TP* is the number of accurate detections, *FP* is the number of false positives, *FN* is the number of false-negatives, and *N* is the total number of instances.

$$Accuracy = \frac{TP + TN}{N} 100[\%]; \tag{1}$$

$$Precision = \frac{TP}{TP + FP} 100[\%]; \tag{2}$$

$$Recall = \frac{TP}{TP + FN} 100[\%]; \tag{3}$$

Considering the binary presence-absence detection setup, we also generated detection error tradeoff (DET) curves due to the capability to evaluate the tradeoff between false alarms and miss probability errors (Martin et al. 1997). The DET curves were also used to compute the optimal decision cost point (DCTopt) for each CNN detector, which is the equilibrium point of the curve metric between high miss probability and high false alarm probability. For this metric, lower values indicate better performance.

Table 2. Performance of the proposed CNN- and ResNet-152-based detectors for different training configurations.

	Accuracy (%)				Precision (%)				Recall (%)			
Model	no-BG	BG-Wet	BG-Dry	BG-Wet +Dry	no-BG	BG-Wet	BG-Dry	BG-Wet+ Dry	no-BG	BG-Wet	BG-Dry	BG-Wet +Dry
Proposed CNN	81.2	76.9	80.2	84.5	40.8	34.7	39.3	48.4	55.4	61.3	58.6	50.4
ResNet-152	83.7	83.8	49.9	80.0	35.3	25.4	17.2	20.6	9.9	4.1	61.1	11.7

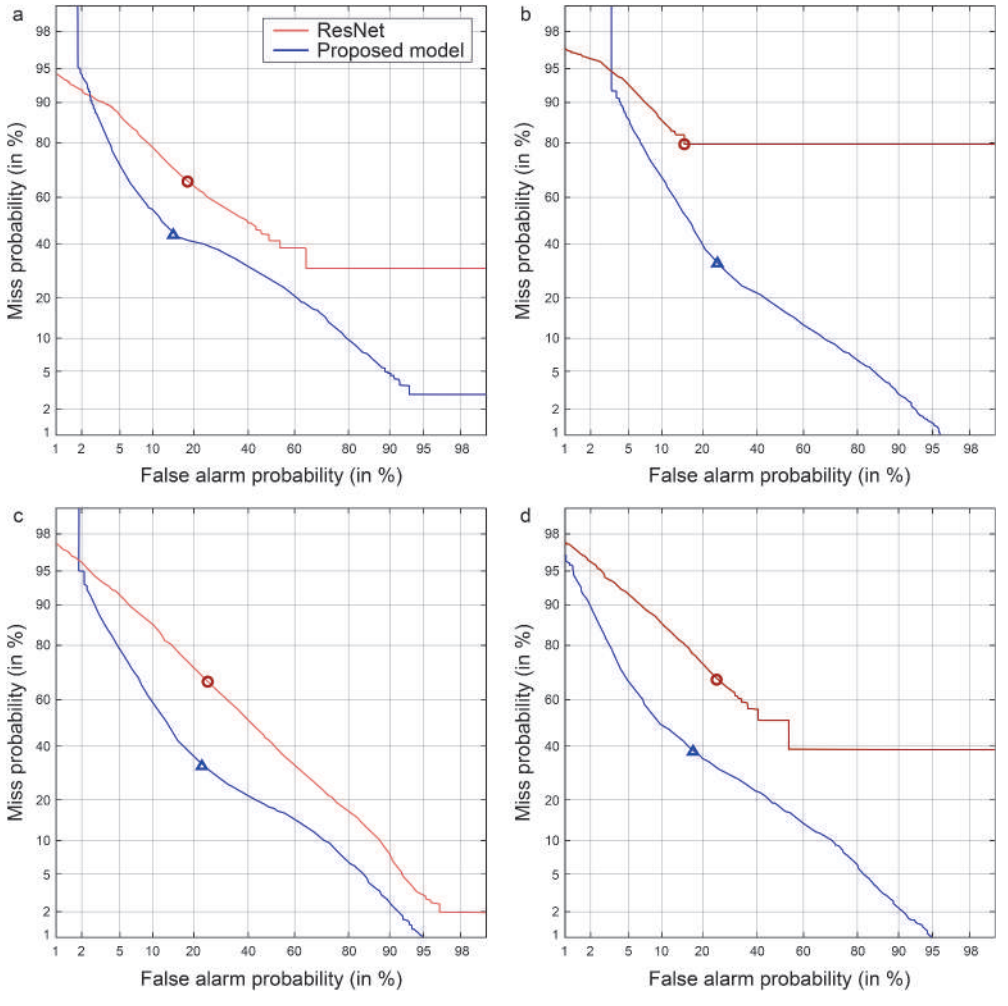

Figure 8. DET plots for the White-lored Spinetail detector when implemented with ResNet-152 (red colour) and the proposed CNN architecture (blue colour). The plots show results with (a) no explicit background modelling, no-BG; (b) wet season data alone, BG-Wet; (c) dry-season data alone, BG-Dry; and (d) combined wet+dry season data, BG-Wet+Dry.

Results

The evaluation results for different detectors were performed with continuous audio recordings containing 15-minute soundscapes. Table 2 shows the performance of the models for various background modelling setups for the acoustic detectors based on the proposed CNN architecture and ResNet-152 model with transfer learning.

We generated the DET curves (Figure 8) to better understand better the potential usefulness of the different CNN-based models independent of the selected FA-miss tradeoffs (false alarm in the X-axis and miss probability in the Y-axis). Each plot's red circle and blue triangle indicate each detector's optimal decision cost point, DCTopt. The optimal decision cost points in all setups, which were computed following Martin et al. (1997), are shown in Table 3.

Table 3. Actual optimal decision cost value, DCTopt, for ResNet-152 and the proposed CNN in the different setups. The lowest optimal decision cost value for each architecture is shown in bold font.

	DCTopt			
Model	no-BG	BG-Wet	BG-Dry	BG-Wet+Dry
ResNet-152	**0.652**	0.725	0.709	0.721
Proposed CNN	0.454	0.452	0.435	**0.432**

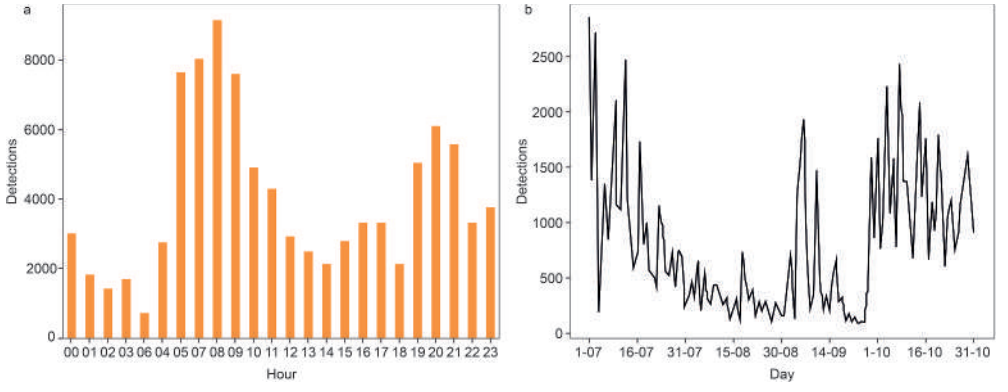

Figure 9. White-lored Spinetail (*S. albilora*) detections: (a) hourly and (b) daily activity patterns.

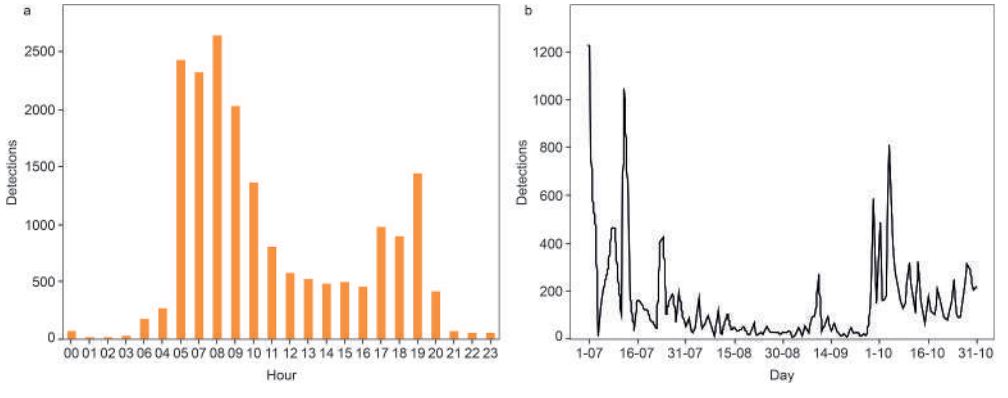

Figure 10. Positive dB diurnal White-lored Spinetail (*S. albilora*) detections: (a) by the hour and (b) by day.

Finally, to describe the diel and seasonal acoustic activity patterns of Spinetail, we used the detector results obtained with the proposed CNN architecture that was trained with the combined BG-Wet+Dry background dataset because of its advantageous precision. We analysed the diel and seasonal patterns of the Spinetail for several months, covering the wet season of 2015, by (i) using all detections made by our model (Figure 9) and (ii) converting linear magnitude spectrograms to db-scaled spectrograms using only those detections with the strongest signals, i.e. amplitudes with positive dB in this case (Figure 10). Vocalisations with lower amplitudes were not consistently detected due to the floating ambient noise level.

Based on the results shown in Figures 9 and 10, in both cases, the hourly detections of Spinetail showed a double peak of vocal activity, one around sunrise and a second lower peak around sunset. The main difference between both approaches (considering all detections and only those with positive dB) was found during the nocturnal period. There were a large number of nocturnal detections when considering all detections, while almost no nocturnal detections occurred when considering only detections with positive dB. The seasonal pattern of detections was also similar following both approaches. Based on all detections, we observed the maximum vocal activity during the first fortnight of July, a small peak around mid-September, and high and constant vocal activity during October (Figure 9). The peak of detections around mid-September was minimised when we only analysed the detections with loud signals (Figure 10).

Discussions

Influence of the acoustic background modeling

We observed higher detection accuracy and precision for the proposed CNN architecture when using the combined BG-Wet+Dry datasets to model the acoustic background (Table 2). This combination provides more data for training, contributing to the model's performance. This combination also increases the imbalance between classes, although there is also an imbalance in other combinations, which may explain the lower performance.

Both metrics decreased when tested with no explicit background modelling, no-BG, or only soundscapes from one of the seasons. Meanwhile, the recall probability of detecting the Spinetail vocalisations increased in these cases, i.e. the proposed CNN architecture improved the recall values when trained with season-specific background data: BG-Wet and BG-Dry.

Despite the observed inferior performance, when trained with different subsets of background data, the ResNet-152 model improved the overall performance when not using the acoustic background datasets. When the background datasets BG-Wet and BG-Dry were used alone, we observed dissimilar performance: BG-Dry had lower error rates in precision.

Based on the DCTopt scores (Figure 8 and Table 3), we observed that in all setups, the proposed CNN architecture provides lower values for the optimal decision cost function and, therefore, outperformed the detector based on ResNet-152 with transfer learning.

The behaviour of the presence-absence detector based on the proposed CNN architecture shows the highest DCTopt, i.e. the lowest performance, for the case with no explicit background modelling. In contrast, when no explicit modelling of the acoustic background was used, the ResNet-152-based detector showed the lowest optimal decision cost, i.e. the highest decision performance. This result can be explained by the differences in the architecture of these models and the datasets used for training. The ResNet-152-based detector model has more layers and must be trained with more data, even when applying transfer learning. Background modelling makes the problem harder and requires additional data. Thus, it was observed that for the ResNet-152-based detector, additional training with background data is not beneficial.

Analysing the detection scores concerning their deviation from the normal distribution, we observe (Figure 8) that the ResNet-152 curves are far too distant from a straight line, i.e. they do not adhere to the normal distribution of probabilities. The latter makes the detector's behaviour less predictable and more challenging to understand and trust. We observed that the output scores roughly followed the normal distribution in all setups for the proposed CNN architecture. The latter is evident in the less-curved DET plots compared to ResNet-152. The season-specific explicit background modelling with the combined wet + dry dataset, BG-Wet+Dry, improved the detector behaviour predictability because the observed output probabilities were closer to the expected normal distribution.

Diel and seasonal pattern of acoustic activity

The double approach employed in our study (using all detections or only those with positive dB) provided similar diel and seasonal patterns of vocal activity of the species. Nonetheless, we found slight differences, especially during the nocturnal period, when few detections occurred when considering only detections with positive dB. These findings suggest the possibility that several false positives may have occurred when considering all detections in our dataset. These results are in agreement with prior research that already proved a negative relationship between the sound level of the desired signal and the accuracy of the CNNs. Overall, the observed vocal activity of the target species was maximum during the first hours after sunrise and close to sunset, consistent with the typical bimodal pattern of vocal activity described for most passerines (Catchpole and Slater 2008; Gil and Llusia 2020). When we only considered those detections with loud signals, the diel pattern of vocal activity was almost completely restricted to the daytime and with much-reduced activity during the central hours of the day. We found that the seasonal pattern of vocal activity of the species showed two peaks separated by approximately one month and a half. This observation is consistent with the multi-brooded character of the Spinetail (Rubio and Pinho 2008), since it may indicate two breeding attempts in early July and early October. The silent interval in between it may be related to the incubating, nestling, and fledging periods when passerines' vocal activity is usually reduced (e.g. Lampe and Espmark 1987; Amrhein et al. 2002). Our findings provide the first description of the vocal behaviour of the Spinetail. We are aware that our results are based on a single locality and use data collected during a limited period. Therefore, further research is needed before drawing any broad conclusions about the vocal behaviour of the Spinetail.

Conclusions

Contrary to the intuitive expectation that the extra effort invested in correctly representing the acoustic background environment will improve performance, our results demonstrated that the CNN-based detector behaviour depends on the model creation approach. The CNNs trained in environmentally richer contexts that combine wet and dry season recordings were observed to outperform the detectors trained with season-specific background data or without background modelling. When the bird detector is based on pretrained CNNs fine-tuned with

transfer learning, the highest classification accuracy is observed without season-specific adaptation. Here, only one species (Spinetail) was evaluated in a single region. As much as data from different climatic periods were used for detector training, only two months in the year were used for model evaluation. Nevertheless, the experimental results indicate the importance of background modelling for obtaining a good bird detector. In this research, we depended on the assumption of using OtherBirds recordings free of ambient noise and thus did not interfere with the wet and dry acoustic environments. This holds for most Xeno-Canto data; however, it must be considered when other sources of negative examples are used.

The best result was observed with the proposed CNN model trained with all-season data. We observed an improvement in recognition accuracy (an increase of 3.3%) and precision (an increase of 7.6%) in exchange for a decrease in recall (of 5%). The CNN-based detector enabled us to study the vocal behaviour of the Spinetail for several months during the wet season and analyse the diel and seasonal patterns of its acoustic activity for the first time.

The acoustic detector used in this study provides the means for further research on the vocal behaviour of the neotropical White-lored Spinetail. Next, we aim to process the bulk of Pantanal recordings with wider spatial and temporal coverage to gain deeper insights and a conclusive understanding of the Spinetail species behaviour.

Finally, further research on acoustic background modelling is needed to advance the passive acoustic monitoring methods and technology, primarily for handling regional and seasonal variabilities in acoustic environments, as these impose the main challenge to the automated acoustic detection of sound emitting species.

Acknowledgements

The authors would like to thank Luiz Vicente da S. Campos Filho, Pouso Alegre Farm, and Director Waldir Wolfgang Valutky of the SESC Pantanal, Mato Grosso, for allowing us to conduct our studies on their lands, and Ana Silvia Tissiani for graphical support. This study is part of the Biodiversity Monitoring Project Sounds of the Pantanal – The Pantanal Automated Acoustic Biodiversity Monitoring of INAU/CO.BRA, Cuiabá, Mato Grosso, Brazil

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) under Grant [CAPES-01]; Instituto Nacional de Ciência e Tecnologia em Áreas Úmidas (INAU/UFMT/CNPq); Centro de Pesquisa do Pantanal (CPP); and Brehm Funds for International Bird Conservation (BF), Germany.

ORCID

Thiago M. Ventura http://orcid.org/0000-0002-3758-5466

Todor D. Ganchev http://orcid.org/0000-0003-0384-4033
Cristian Pérez-Granados http://orcid.org/0000-0003-3247-4182
Allan G. de Oliveira http://orcid.org/0000-0003-4827-1048
Gabriel de S. G. Pedroso http://orcid.org/0000-0003-0700-7398
Marinez I. Marques http://orcid.org/0000-0002-9890-8505
Karl-L. Schuchmann http://orcid.org/0000-0002-3233-8917

Ethical statement

A licence to collect was obtained from Brazilian legal authorities (ICMBIO-SISBIO process: 39095-KLS).

References

Adavanne S, Drossos K, Cakir E, Virtanen T. 2017. Stacked convolutional and recurrent neural networks for bird audio detection. In: 25th European Signal Processing Conference (EUSIPCO); Aug 28–Sep 2; Kos (EL): IEEE Computer Society. p. 1729–1733. doi: 10.23919/EUSIPCO.2017.8081505.

Amrhein V, Korner P, Naguib M. 2002. Nocturnal and diurnal singing activity in the nightingale: correlations with mating status and breeding cycle. Anim Behav. 64:939–944. doi: 10.1006/anbe.2002.1974.

Anderson AS, Marques TA, Shoo LP, Williams SE. 2015. Detectability in audio-visual surveys of tropical rainforest birds: the influence of species, weather and habitat characteristics. PLoS One. 10(6):e0128464. doi: 10.1371/journal.pone.0128464.

Catchpole CK, Slater PJ. 2008. Bird song: biological themes and variations. 2nd ed. Cambridge (UK): Cambridge University Press.

Diment A, Virtanen T. 2017. Transfer learning of weakly labelled audio. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA); Out. p. 6–10. doi: 10.1109/WASPAA.2017.8169984.

Dufourq E, Batist C, Foquet R, Durbach I. 2022. Passive acoustic monitoring of animal populations with transfer learning. Ecol Inform. 70:101688. doi: 10.1016/j.ecoinf.2022.101688.

Florentin J, Dutoit T, Verlinden O. 2020. Detection and identification of European woodpeckers with deep convolutional neural networks. Ecol Inform. 55:101023. doi: 10.1016/j.ecoinf.2019.101023.

Gil D, Llusia D. 2020. The bird dawn chorus revisited. In: Aubin T. Mathevon N, editors. Coding strategies in vertebrate acoustic communication. Zurich (CH): Springer Nature; p. 45–90.

Gwynne JA, Ridgely RS, Argel M, Tudor G. 2010. Birds of Brazil, the Pantanal & Cerrado of central Brazil. São Paulo (SP): Horizonte.

He K, Zhang X, Ren S, Sun J. 2016. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; Jun 26–Jul 1; Las Vegas (NV): Computer Vision Foundation.

Hidayat AA, Cenggoro TW, Pardamean B. 2021. Convolutional neural networks for scops owl sound classification. Procedia Comput Sci. 179:81–87. doi: 10.1016/j.procs.2020.12.010.

Junk WJ, Cunha CN, Wantzen KM, Petermann P, Strüsmann C, Marques MI, Adis J. 2006. Biodiversity and its conservation in the Pantanal of Mato Grosso, Brazil. Aquat Sci. 68(3):278–309. doi: 10.1007/s00027-006-0851-4.

Kahl S, Wilhelm-Stein T, Hussein H, Klinck H, Kowerko D, Ritter M, Eibl M. 2017. Large-scale bird sound classification using convolutional neural networks. In: Cappellato L, Ferro N, Goeuriot L, Mandl T, editors. CLEF 2017. Conference and Labs of the Evaluation Forum; Sep 11–14; Dublin (IRL): CEUR Workshop Proceedings; p. 143–157.

Kim P. 2017. Convolutional neural network. MATLAB deep learning. Berkeley (CA): Apress; p. 121–147. doi: 10.1007/978-1-4842-2845-6_6.

Kiranyaz S, Ince T, Abdeljaber O, Avci O, Gabbouj M. 2019. 1-D Convolutional neural networks for signal processing applications. ICASSP 2019 – 2019. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 12–17; Brighton (UK): The Institute of Electrical and Electronics Engineers, Signal Processing Society; p. 8360–8364. doi: 10.1109/ICASSP.2019.8682194.

Knight EC, Poo Hernandez S, Bayne EM, Bulitko V, Tucker BV. 2020. Pre-processing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks. Bioacoustics. 29:337–355. doi: 10.1080/09524622.2019.1606734.

Lampe HM, Espmark YO. 1987. Singing activity and song pattern of the redwing *Turdus iliacus* during the breeding season. Ornis Scand. 18(3):179–185. doi: 10.2307/3676764.

Lasseck M. 2018. Audio-based bird species identification with deep convolutional neural networks. In: Cappellato L, Ferro N, Nie J-Y Soulier L, editors. Working Notes of CLEF 2018. Conference and Labs of the Evaluation Forum; Sep 10–14; Avignon (FR): CEUR Workshop Proceedings.

Lowen J, Bernardon G. 2010. Seeing Pantanal specialities along the transpantaneira. Neotrop Birding. 6(1):47–54. doi: 10.1590/S1519-566X2011000100007.

Martin A, Doddington G, Kamm T, Ordowski M, Przybocki M. 1997. The DET curve in assessment of detection task performance. In: Kokkinakis G, Fakotakis N, Dermatas E, editors. EUROSPEECH 1997. European Conference on Speech Communication and Technology; Sep 22–25; Rhodes (EL): ISCA International Speech Communication Association.

Murphy J. 2016. An overview of convolutional neural network architectures for deep learning. Microway Inc. 1–22. https://www.microway.com/download/whitepaper/An_Overview_of_Convolutional_Neural_Network_Architectures_for_Deep_Learning_fall2016.pdf.

Noumida A, Rajan R. 2022. Multi-label bird species classification from audio recordings using attention framework. Appl Acoust. 197:108901. doi: 10.1016/j.apacoust.2022.108901.

Oliveira AG, Ventura TM, Ganchev TD, de Figueiredo JM, Jahn O, Marques MI, Schuchmann K-L. 2015. Bird acoustic activity detection based on morphological filtering of the spectrogram. Appl Acoust. 98:34–42. doi: 10.1016/j.apacoust.2015.04.014.

Pérez-Granados C, Bota G, Giralt D, Albarracín J, Traba J. 2019. Cost-effectiveness assessment of five audio recording systems for wildlife monitoring: differences between recording distances and singing direction. Ardeola. 66(2):311–325. doi: 10.13157/arla.66.2.2019.ra4.

Ruan W, Wu K, Chen Q-C, Zhang C-G. 2022. ResNet-based bio-acoustics presence detection technology of Hainan gibbon calls. Appl Acoust. 198:108939. doi: 10.1016/j.apacoust.2022.108939.

Rubio TC, Pinho JB. 2008. Biologia reprodutiva de *Synallaxis albilora* (Aves: Furnariidae) no Pantanal de Poconé, Mato Grosso. Pap Avulsos Zool. 48(17):181–197. doi: 10.1590/S0031-10492008001700001.

Shiu Y, Palmer KJ, Roch MA, Fleishman E, Liu X, Nosal E-M, Helble T, Cholewiak D, Gillespie D, Klinck H. 2020. Deep neural networks for automated detection of marine mammal species. Sci Rep. 10(1):607. doi: 10.1038/s41598-020-57549-y.

Smith P. 2020. Azara's spinetails II: what is No. 239 "Cola aguda cola sanguina"? Ornithol Res. 28(3):191–194. doi: 10.1007/s43388-020-00021-2.

Stowell D. 2022. Computational bioacoustics with deep learning: a review and roadmap. PeerJ. 10:e13152. doi: 10.7717/peerj.13152.

Stowell D, Petrusková T, Šálek M, Linhart P. 2019. Automatic acoustic identification of individuals in multiple species: improving identification across recording conditions. J R Soc Interface. 16:20180940. doi: 10.1098/rsif.2018.0940.

Sugai LSM, Silva TSF, Ribeiro JW Jr, Llusia D. 2019. Terrestrial passive acoustic monitoring: review and perspectives. BioScience. 69:15–25. doi: 10.1093/biosci/biy147.

Tuncer T, Akbal E, Dogan S. 2021. Multileveled ternary pattern and iterative ReliefF based bird sound classification. Appl Acoust. 176:107866. doi: 10.1016/j.apacoust.2020.107866.

Ventura TM, de Oliveira AG, Ganchev TD, de Figueiredo JM, Jahn O, Marques MI, Schuchmann K-L. 2015. Audio parameterization with robust frame selection for improved bird identification. Expert Syst Appl. 42(22):8463–8471. doi: 10.1016/j.eswa.2015.07.002.

XENO-CANTO. Sharing wildlife sounds from around the world. 2022. [accessed 2022 Nov 18]. https://xeno-canto.org .

Zhao Z, Xu Z, Bellisario K, Zeng R, Li N, Zhou W, Pijanowski BC. 2019. How well do acoustic indices measure biodiversity? Computational experiments to determine effect of sound unit shape, vocalization intensity, and frequency of vocalization occurrence on performance of acoustic indices. Ecol Indic. 107:105588. doi: 10.1016/j.ecolind.2019.105588.

Zor C, Awais M, Kittler J, Bober M, Husain S, Kong Q, Kroos C. 2019. Divergence based weighting for information channels in deep convolutional neural networks for bird audio detection. ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 12–17; Brighton (UK): The Institute of Electrical and Electronics Engineers, Signal Processing Society; p. 3052–3056. doi: 10.1109/ICASSP.2019.8682483.