World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(RESEARCH ARTICLE)

Check for updates

Multi-label bird species classification using Haar wavelet- based residual convolutional neural network

Noumida A. [1, 3, *] and Rajeev Rajan [2, 3]

[1] *College of Engineering Trivandrum, India.*
[2] *Government Engineering College Idukki, India.*
[3] *APJ Abdul Kalam Technological University, India.*

World Journal of Advanced Engineering Technology and Sciences, 2025, 14(02), 018–025

Publication history: Received on 22 December 2024; revised on 04 February 2025; accepted on 07 February 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.14.2.0043

Abstract

Automatic bird vocalization analysis is advancing research in ecology and conservation. In recent years, numerous studies have employed deep learning models to categorize bird calls. This study examined the efficacy of Haar Wavelet Residual Convolutional Neural Network (WRCNN) for multi-label bird species classification. Initially, Haar wavelet transforms were applied to the mel spectrograms of bird call recordings. These transformed spectrograms were subsequently input into the WRCNN for multi-scale spectral analysis. The model obtained a macro-average F1-score of 0.60, showcasing its potential in multi-label tasks and exhibiting notable improvements over baseline methods. Experiments were conducted utilizing the Xeno-Canto bird sound database.

Keywords: Multi-Label; Sequential; Haar Wavelet; Convolutional Neural Network; Residual Network

1. Introduction

Bioacoustics, the study of animal sounds, explores how vocalizations influence ecology and evolution, particularly in communication, reproduction, and territorial behavior. Birds play a key role in ecosystem health [1], but face threats from human activity, making biodiversity monitoring essential. Research [2] shows that birds can signal environmental shifts due to their wide presence. Identifying bird calls is crucial for minimizing human impact on avian populations [3], as many species serve as pollinators [4], seed dispersers [5], and predators [6]; fluctuations in their numbers can have significant environmental consequences. Bird vocalizations are categorized into two types: calls and songs. Songs are the extended, loud vocal displays produced by male birds, comprising phrases, syllables, and trills. In contrast, calls are brief, unmelodious vocalizations used by both sexes for various purposes, including distress, alarm, flight, warning, feeding, nesting, and flock communication.

The classification of bird calls is challenging due to the wide range of acoustic properties and the difficulty humans face in distinguishing them. Accurate classification requires an analysis of various acoustic features, such as duration and frequency [7]. Recent advancements in deep learning have enabled precise bird call classification models [8]. These models, trained on diverse audio datasets, can identify even the rarest behavioral patterns in avian species [9]. Convolutional Neural Networks (CNNs) have succeeded in multi-label classification but may overlook key spectral details essential for identifying coexisting bird species. To address this, we propose a Haar Wavelet Residual CNN (WRCNN) that applies Haar wavelet transforms to mel spectrograms, enhancing multi-scale recognition of time-frequency patterns. Residual connections retain vital information, strengthening the model's ability to classify complex vocalizations across species.

* Corresponding author: Noumida A.

Traditional acoustic feature extraction methods incorporate various techniques that capture both time and frequency domain characteristics, including average energy, spectral properties, average zero-crossing rate, bandwidth, mel-frequency cepstral coefficients, chromagrams, wavelet coefficients, and Linear predictive coding-derived coefficients. Literature provides some of the speech and audio processing techniques for bird call recognition [10-12]. A study [13] investigated two CNN-based approaches for bird detection on audio signals. The Bulbul model, using diverse field recordings, and the Sparrow model, using smartphone recordings from UK locations, each achieved an AUC of 89% on hidden test sets. Transfer learning models show promise for efficient bird-call classification with limited data [14, 15]. Large-scale bird sound classification has been addressed using various CNNs to extract features from audio recording visualizations [16]. However, time-frequency overlap in these recordings presents challenges for multi-label bird species classification. Researchers have framed this as a multi-instance multi-label problem [17] and employed multi-label classifiers to identify concurrent audio patterns in extended recordings [18]. These approaches have shown that auditory indices can reliably indicate fundamental ecological processes [18].

A study [19] proposed wavelet CNNs and evaluated their practical performance in texture classification and image annotation. The experiments revealed that wavelet CNNs can achieve superior accuracy in both tasks compared to existing models while having significantly fewer parameters than conventional CNNs. Another study [20] investigated the Discrete Wavelet Transform (DWT) in the frequency domain and designed a novel Wavelet-Attention (WA) block to implement attention in the high-frequency domain. The work [21] presents a bird call detection method for field recordings that adapts easily to new species and maintains effectiveness despite noise or distance. Using wavelet node reconstruction as a preprocessing filter, it prioritizes high recall to reduce missed calls, which are harder to recover later. In this work, we propose an architecture which combines Haar wavelet transform with residual CNN (WRCNN) for multi-scale feature extraction. The system architecture is described in Section 3, followed by the experimental framework in Section 4. The analysis of the results is presented in Section 5. The paper concludes in Section 6.

2. Materials and methods

We propose Haar Wavelet Residual Convolutional Networks (WRCNN) for detecting multiple overlapping species in field recordings. We used Haar wavelet transforms in combination with residual CNN, using mel spectrograms as input as shown in Figure 1.

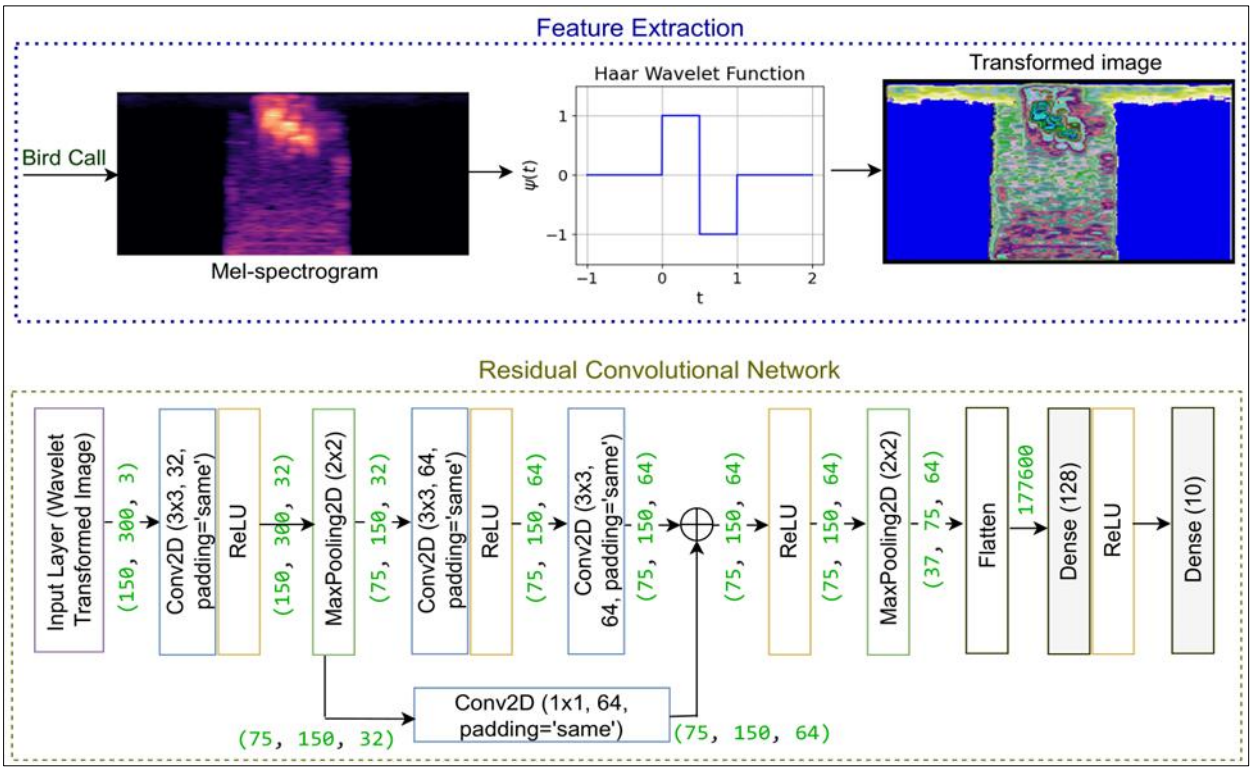

Figure 1 Block diagram of the proposed Haar Wavelet Residual Convolutional Neural Network (WRCNN)

2.1. Dataset and preprocessing

The Xeno-canto bird sound database is utilized for performance evaluation [22]. The dataset used for bird call classification has a sampling rate of 16,000 Hz and an audio resolution of 16-bit in mono format. The dataset encompasses ten bird species: Asian Koel (AK, 26, 121), Blue Jay (BJ, 27, 109), House Crow (HC,27, 111), Mallard Duck (MD, 25, 106), Grey Go-away (GG, 19, 109), Red Lapwing (RL, 24, 104), Eurasian Owl (EO, 25, 107), Indian Peafowl (IP, 29, 103), House Sparrow (HS, 24, 100), and Western Wood Pewee (WW, 24, 108). The number of Xeno-canto files and pre-processed audio files are provided in parentheses.

The raw dataset includes 250 files sourced from Xeno-Canto, with a training set expanded to 1,078 files and an augmented version totaling 2,923 files to enhance model training. For testing, the dataset includes two separate test sets: one with 334 files for two bird species and another with 100 files for three bird species. The training set has 1,078 individual bird calls, while the test sets contain 668 calls for two bird species and 300 calls for three species. Each training file has a duration of 1.5 seconds, while each test file is 9 seconds long, providing a comprehensive dataset structure for model evaluation. Additional training files are generated using specAugment [23].

2.2. Feature extraction

Our methodology applies wavelet transforms to the mel spectrogram, yielding coefficients at various decomposition stages, encompassing approximation and detail coefficients in multiple orientations. The mel spectrogram represents the temporal progression of frequency components, while the wavelet transform provides a multi-scale analysis, elucidating both high and low-frequency information.

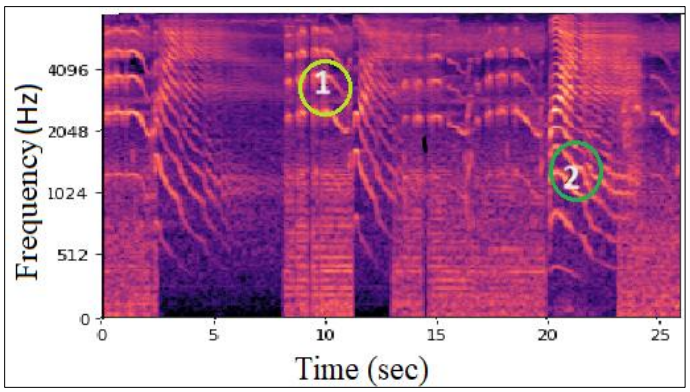

Figure 2 Mel spectrogram of test file with 2 species (1-Red Lapwing, 2- Grey Go-away)

2.2.1. Mel spectrogram

A mel spectrogram provides a visual depiction of how a signal's frequencies evolve over time. It employs a specialized mel-scale filter bank to emphasize the frequency ranges that are most significant for human auditory perception. The mel scale is created to adjust frequency data in a way that more accurately reflects how humans perceive sound. Figure 2 shows the mel spectrogram of audio recording containing multiple species. Data augmentation is implemented utilizing SpecAugment [23], a technique that involves obscuring temporal segments and frequency channels in the mel spectrogram representation. We generated 3344 augmented mel spectrograms. Our calculations employed 224 mel filter banks, a 2048-point FFT, a 2048-sample Hanning window (approximately 128 ms), and a 512-sample hop length (approximately 32 ms).

2.2.2. Haar Wavelet Transform

The Discrete Wavelet Transform (DWT) is applied to each color channel (Red, Green, and Blue) of an RGB image individually, using the Haar wavelet. This method preserves only the low-frequency components (LL sub-band) for each channel, which are then stacked to create a transformed RGB image.

- For each color channel, a 2D DWT is computed using the Haar wavelet, decomposing the image into four sub-bands
    - LL (Approximation): Low-pass filter in both dimensions.
    - LH (Horizontal Detail): Low-pass along rows, high-pass along columns.

- HL (Vertical Detail): High-pass along rows, low-pass along columns.
- HH (Diagonal Detail): High-pass along both dimensions.

- The coefficients for each color channel can be expressed as:

$coeffs2 = DWT_{haar}(image[:, :, i])$

where i represents the color channel (R, G, or B).

- The Haar wavelet, known for its simplicity, separates signals into local averages and differences. It is defined as:

$$\psi_{haar}(t) = \begin{cases} 1, & \text{for } 0 \le t < 1/2; \\ -1, & \text{for } 1/2 \le t < 1; \\ 0, & \text{otherwise} \end{cases}$$

- This function provides two main filter components:

  - Low-pass filter (approximation): Captures average information.
  - High-pass filter (details): Captures difference information.
- Each channel i of the image is processed as follows

$LL_i, (LH_i, HL_i, HH_i) = DWT_{haar}(image[:, :, i])$

- We retain only the LL component for each color channel, representing the large-scale structure (approximation) of the image. For an RGB image, the calculation proceeds as:

LL_R, LL_G, LL_B

- The LL components from the three color channels are then stacked along the last axis to reconstruct an RGB-like structure suitable for further analysis. This process can be represented by

$transformed\_image = np.stack([LL_R, LL_G, LL_B], axis=-1)$

- The final output is a transformed RGB image containing only the low-frequency components of each color channel:

$transformed\_image(x, y) = [LL_R(x, y), LL_G(x, y), LL_B(x, y)]$

2.3. Architecture

The model processes transformed image with dimensions 150 × 300 × 3, where each RGB channel has been pre-processed with Haar wavelet functions. The proposed Residual Convolutional Network architecture is shown in Figure 1.

2.3.1. Initial Convolutional Layer

The input images first pass through a Conv2D layer with 32 filters, a 3 × 3 kernel size, ReLU activation, and 'same' padding to maintain spatial dimensions. This is followed by a MaxPooling layer with a 2 × 2 window, reducing spatial dimensions by half. The convolution operation is formally defined as:

$$F_{out} = ReLU(W * F_{in} + b)$$

2.3.2. Residual Block for Enhanced Feature Propagation

After the initial convolutional layers, a Residual CNN architecture enables deeper feature extraction by introducing skip connections that preserve gradient flow. To match dimensions, the output of the initial convolution is passed through a 1 × 1 Conv2D layer with 64 filters:

$$\text{shortcut} = W_{1\times 1} * F_{in} + b$$

The main path has two 3 × 3 Conv2D layers with 64 filters each. The first layer uses ReLU activation and 'same' padding. The main path convolutions are represented as:

$$y_{i,j,k}^{(1)} = \text{ReLU}\left(\sum_{p=-1}^{1} \sum_{q=-1}^{1} x_{i+p,\, j+q} \cdot w^{(1)}_{p,q} + b^{(1)}\right)$$

$$y_{i,j,k}^{(2)} = \left(\sum_{p=-1}^{1} \sum_{q=-1}^{1} y_{i+p,\, j+q}^{(1)} . w^{(2)}_{p,q} + b^{(2)}\right.$$

The shortcut and main path outputs are combined through element-wise addition, followed by ReLU activation:

$$\text{residual output} = \text{ReLU}(y_{i,j,k}^{(2)} + \text{shortcut})$$

2.3.3. Final Classification Layers

After applying the chosen attention mechanism, a MaxPooling layer further downsamples spatial dimensions, followed by a Flatten layer to prepare features for dense layers. A dense layer with 128 units and ReLU activation extracts high-level features, and a final dense layer with sigmoid activation computes class probabilities.

2.4. Sequential Aggregation Strategy

Initially, the test audio recordings are sliced into fixed-length segments of 1.5 seconds. The model is then fed with mel spectrograms extracted from these audio segments. The trained network generates a probability score, indicating the likelihood of a bird's presence in each segment. A final score for an audio file is calculated by summing up all the segment-wise probabilities and normalizing the result.

3. Results and discussions

The proposed WRCNN and other existing models were implemented on the Keras-TensorFlow platform. During training, isolated bird vocalizations of 1.5 seconds were used, with all recordings standardized to this length for consistency. For testing, longer bird calls were divided into consecutive 1.5-second segments. We evaluated all models using a sequential aggregation strategy. These models were rigorously trained in a Google Colab notebook for up to 100 epochs, with a batch size of 32. The adam optimizer, categorical cross-entropy loss, and a sigmoid activation function were used. All comparative experiments were conducted under identical operating conditions.

Figure 3 illustrates the precision, recall, and F1-score achieved in our experiments. Figure 4 displays the confusion matrices for the proposed model, applied to the target dataset with two and three species. The best-performing WRCNN attained a macro-average precision, recall, and F1-score of 0.65, 0.62, and 0.60, respectively. In comparison, the baseline CNN yielded values of 0.50 for precision and recall, with an F1-score of 0.45.

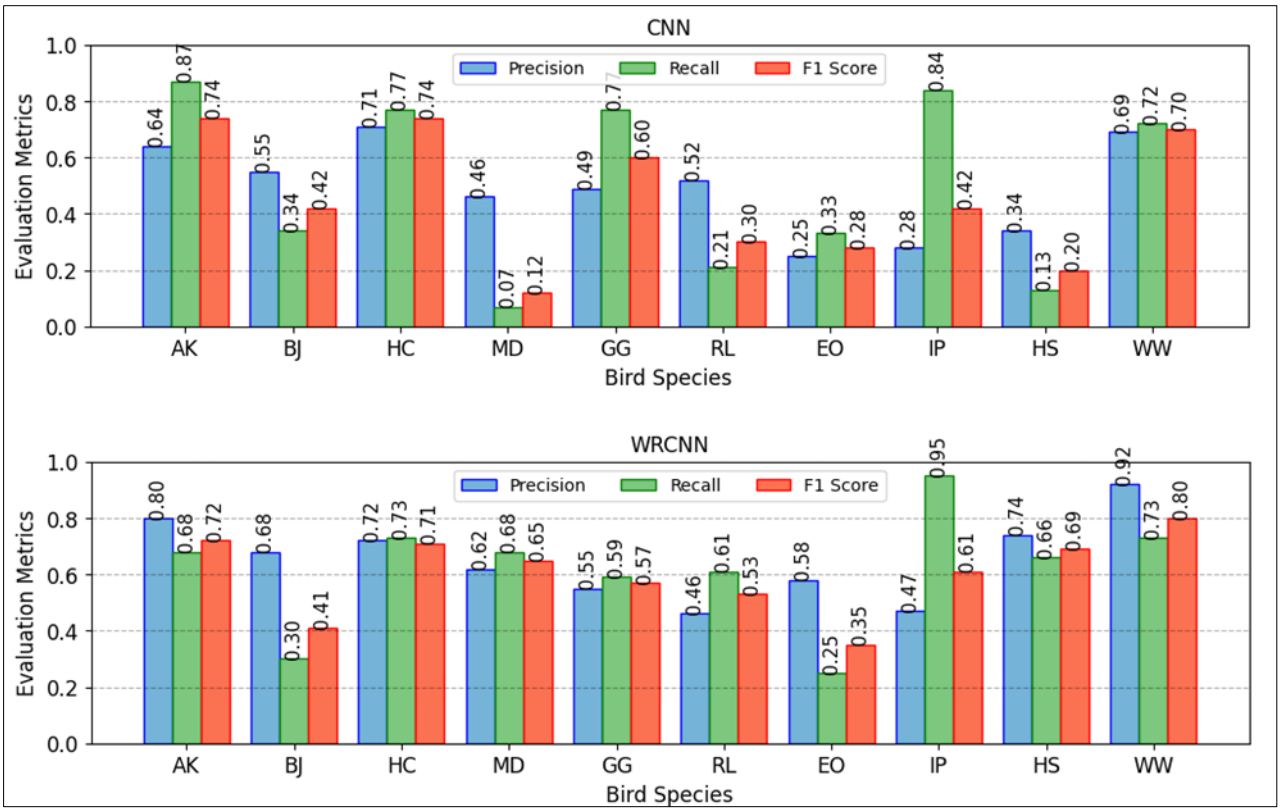

Figure 3 Precision, Recall and F1-Score for the experiments

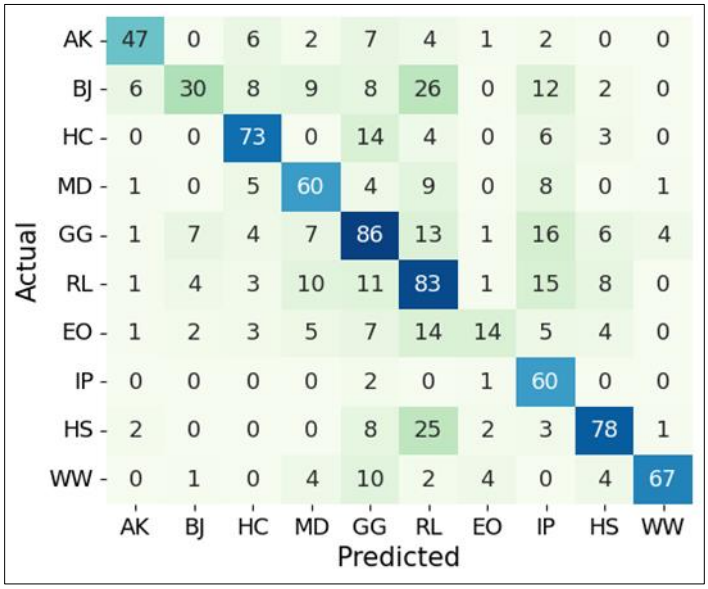

Figure 4 Confusion matrix of the proposed WRCNN

The proposed model demonstrated significant improvements in species-specific performance, particularly for the Mallard Duck, Red-wattled Lapwing, Eurasian Owl, and House Sparrow. While the baseline CNN achieved F1-scores below 50% for nearly six species, the proposed WRCNN reduced this number, showing F1-scores below 50% only for Blue Jay and Eurasian Owl. Overall, the WRCNN effectively minimized misclassification errors, highlighting its advantage over the baseline model in accurately classifying diverse bird species.

Table 1 Performance Comparison with Existing Methods

No.	Method	Precision	Recall	F1-Score
1	Grill et al. [Model 1] [13]	0.50	0.50	0.45
2	Grill et al. [Model 2] [13]	0.51	0.48	0.48
3	Efremova et al. [15]	0.61	0.54	0.53
4	Yang et al. [24]	0.65	0.58	0.58
5	Proposed WRCNN	0.65	0.62	0.60

The performance comparison in Table 1 highlights the effectiveness of our proposed Wavelet Residual Convolutional Network architecture using Haar wavelet transforms, in achieving superior results over several existing models in bird species classification. Compared to earlier models such as Grill et al. [13] [Model 1] and [Model 2], Efremova et al. [15], and SENet [24] which achieve F1-scores of 0.45, 0.48, 0.58, and 0.53 respectively, our WRCNN model provides substantial improvements. The SENet is employed in \cite{yang} to enable the network to perform dynamic channel-wise feature re-calibration. The WRCNN outperforms these benchmarks with an F1-score of 0.60 indicating the utility of wavelet transformations in enhancing feature representation

4. Conclusion

Our study demonstrates the effectiveness of WRCNN in multi-label bird species classification. By applying Haar wavelet transforms to mel spectrograms, we introduce a multi-scale analysis to spectral features. The proposed WRCNN achieves a macro average F1-score of 0.60 showcasing the potential of this approach for multi-label classification tasks.

Compliance with ethical standards

Acknowledgments

The authors thank Kerala State Council for Science, Technology and Environment (KSCSTE), Kerala, India, for the financial assistance during the research.

Disclosure of conflict of interest

I Noumida A and Rajeev Rajan declare no conflicts of interest or competing interests related to the publication of this manuscript.

References

[1] Carignan, Vincent, and Marc-André Villard. "Selecting indicator species to monitor ecological integrity: a review." Environmental Monitoring and Assessment 78 (2002): 45-61.

[2] Virkkala, Raimo, and Aleksi Lehikoinen. "Patterns of climate-induced density shifts of species: Poleward shifts faster in northern boreal birds than in southern birds." Global Change Biology 20.10, 2014.

[3] Clemmons, Janine R., and Richard Buchholz. Behavioral approaches to conservation in the wild. Cambridge University Press, 1997.

[4] Stiles, F. Gary. "Ecological and evolutionary implications of bird pollination." American Zoologist 18.4 (1978): 715-727.

[5] Howe, Henry F., and Judith Smallwood. "Ecology of seed dispersal." Annual Review of Ecology and Systematics 13.1 (1982): 201-228.

[6] Marquis, Robert J., and Christopher J. Whelan. "Insectivorous birds increase growth of white oak through consumption of leaf-chewing insects." Ecology 75.7 (1994): 2007-2014.

[7] Virtanen, Tuomas, Mark D. Plumbley, and Dan Ellis, eds. "Computational analysis of sound scenes and events". Cham: Springer International Publishing, 2018.

[8] Félix Michaud, Jérôme Sueur, Maxime Le Cesne, Sylvain Haupert, "Unsupervised classification to improve the quality of a bird song recording dataset," Ecological Informatics, Volume 74, 2023, 101952, ISSN 1574-9541, https://doi.org/10.1016/j.ecoinf.2022.101952.

[9] Erhan Akbal, Sengul Dogan, Turker Tuncer, "An automated multispecies bioacoustics sound classification method based on a nonlinear pattern: Twine-pat," Ecological Informatics, Volume 68, 2022, 101529, ISSN 1574-9541, https://doi.org/10.1016/j.ecoinf.2021.101529.

[10] Stowell, Dan, et al. "Bird detection in audio: a survey and a challenge." International Workshop on Machine Learning for Signal Processing (MLSP): 1-6.

[11] Gelling, Douwe. "Bird song recognition using gmms and hmms." Master Project Dissertation (2010):1-46.

[12] Thakur, Anshul, et al. "Local compressed convex spectral embedding for bird species identification." The Journal of the Acoustical Society of America 143.6 (2018): 3819-3828.

[13] Grill, Thomas, and Jan Schlüter. "Two convolutional neural networks for bird detection in audio signals." 25th European Signal Processing Conference (EUSIPCO) IEEE, 2017.

[14] Y.-P. Huang and H. Basanta, "Recognition of Endemic Bird Species Using Deep Learning Models," IEEE Access, vol. 9, pp. 102975-102984, 2021.

[15] Efremova, Dina B., Mangalam Sankupellay, and Dmitry A. Konovalov. "Data-efficient classification of birdcall through convolutional neural networks transfer learning." Digital Image Computing: Techniques and Applications (DICTA), (2019): 1-8.

[16] Kahl, Stefan, et al. "Large-Scale Bird Sound Classification using Convolutional Neural Networks." Working Notes of CLEF 1866 (2017).

[17] Briggs, Forrest, Balaji Lakshminarayanan, Lawrence Neal, Xiaoli Z. Fern, Raviv Raich, Sarah JK Hadley, Adam S. Hadley, and Matthew G. Betts. "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach." The Journal of the Acoustical Society of America 131.6 (2012): 4640-4650.

[18] Zhang, Liang, et al. "Using multi-label classification for acoustic pattern detection and assisting bird species surveys." Applied Acoustics 110 (2016): 91-98.

[19] Fujieda, Shin, Kohei Takayama, and Toshiya Hachisuka. "Wavelet convolutional neural networks." arXiv preprint arXiv:1805.08620 (2018).

[20] Zhao, Xiangyu, Peng Huang, and Xiangbo Shu. "Wavelet-Attention CNN for image classification." Multimedia Systems 28, no. 3 (2022): 915-924.

[21] Priyadarshani, Nirosha, Stephen Marsland, Julius Juodakis, Isabel Castro, and Virginia Listanti. "Wavelet filters for automated recognition of birdsong in long-time field recordings." Methods in Ecology and Evolution 11, no. 3 (2020): 403-417.

[22] Vellinga, Willem-Pier, Planqu, Robert, "The xeno-canto collection and its relation to sound recognition and classification.", Working Notes of CLEF (2015).

[23] Park, Daniel S., et al. "Specaugment: A simple data augmentation method for automatic speech recognition.", Interspeech (2019).

[24] Fan Yang and Ying Jiang and Yue Xu. "Design of Bird Sound Recognition Model Based on Lightweight" IEEE Access 10 (2022):85189-85198.