Stacked Res2Net-CBAM with Grouped Channel Attention for Multi-label Bird Species Classification

Noumida A.[1], R. Mukund[1], N. Madhavan Nair[1], Rajeev Rajan[2]
[1] College of Engineering Trivandrum
[2] Government Engineering College, Barton Hill
APJ Abdul Kalam Technological University, Thiruvananthapuram, India.

***Abstract*—Identification of bird species through automatic analysis of their vocalizations holds great promise for various fields, such as ecology, conservation monitoring, and vocal behavioural studies. In recent years this has become a research-active area, and many studies have used deep-learning models to classify bird calls. However, small and imbalanced datasets often limit the performance of these models. In this paper, We explore the effectiveness of Res2Net and Convolutional block attention module (CBAM) with Spatial attention (SA) and Grouped Channel attention (GCA) using a sequential aggregation strategy (Se) for multi-label bird classification. The proposed framework which uses fewer parameters than other residual attention frameworks and can be used directly for audio and image classification tasks. Our findings show that our proposed framework is superior to state-of-the-art models with an F1 score of 72.20% using Mel-spectrogram features.**

Keywords: multi-label, sequential, Res2Net, grouped channel attention, spatial attention, Convolutional block attention module

I. Introduction

Bioacoustics is a fascinating field that studies the world of sound in wildlife. It covers vocalizations and hearing mechanisms in various species, including birds, mammals, amphibians, and insects. Additionally, it explores the ecological and evolutionary functions of animal vocalizations, including communication, territory defense, and mate attraction. Unfortunately, birds, which are essential for ecosystem functions [1], are threatened by human activities. Research conducted by [2] has demonstrated that birds are highly sensitive to changes in the environment, such as climate change, and can serve as an indicator of their state. Ornithological studies that identify bird calls, such as alarm calls, flight calls, and mating calls, play a significant role in conservation efforts [3]. Birds also have important ecological roles, such as pollination [4], seed dispersal [5], and predation [6]. Consequently, monitoring bird populations to prevent negative impacts from human structures and devices is crucial.

Birds use the syrinx, a unique organ, to produce a broad spectrum of vocalizations that can be classified into two categories: songs and calls. Bird song is the loud, long vocal display of male birds, composed of syllables, phrases, and trills. Calls are short and simple vocalizations used by both sex and includes distress, alarm, flight, warning, feeding, nest, and flock calls. However, only perching birds (Passeriformes) can sing which means that nearly half of all birds cannot produce songs [7]. Thus, species-level bird identification should be based on calls rather than songs.

The patterns in bird calls are represented by parameters known as acoustic features [8]. The success of recognition depends on the accuracy of these features in representing the calls [9]. Conventional methods for extracting features include frequency and time domain features [10]. Speech and audio processing techniques, such as those mentioned in [11]–[13], have been used for the recognition of bird calls, along with Artificial Neural Networks [14]. There have been many efforts in the literature to classify birds from pre-segmented single-label recordings [15], [16], [27]. A multi-label classification model for finding simultaneous auditory patterns in longer recordings is proposed in [17]. The efficacy of various CNN-derived features [18], [19], [43] and transfer learning models [20], [26] in addressing the challenges associated with obtaining annotated files are explored in previous studies. However, detecting multiple bird species from overlapping recordings remains a major challenge.

Previous research has explored the use of residual networks for audio and image-related tasks [21], [22]. Res2Net [23] increases the receptive field in the residual block and uses a multi-scale feature aggregation approach to improve the performance. Attention mechanisms [24], [25] can boost the representational power of CNNs by honing in on key features and disregarding irrelevant ones [28]–[30]. End-to-end audio classification systems using residual networks along with squeeze-excitation [31], [32] and CBAM [33] give improved accuracy. To reduce the computational complexity of the CBAM module, channel attention is applied to a group of channels (i.e, GCA) instead of all the channels at once. We proposed a stacked Res2Net-CBAM with GCA framework to identify the multiple simultaneous or isolated bird vocalizations present in an audio recording. An augmentation scheme based on SpecAugment [34] has been adopted, and a sliding scheme method [26] has been effectively implemented on Mel-spectrogram features. Our model requires fewer parameters compared to the existing state-of-the-art models. The proposed framework can be used as a public backbone network for complex audio and image classification tasks.

The detailed system description is explained in Section II, followed by the experimental setup in Section III. The result analysis is given in Section IV. Finally, the paper is concluded in Section V.

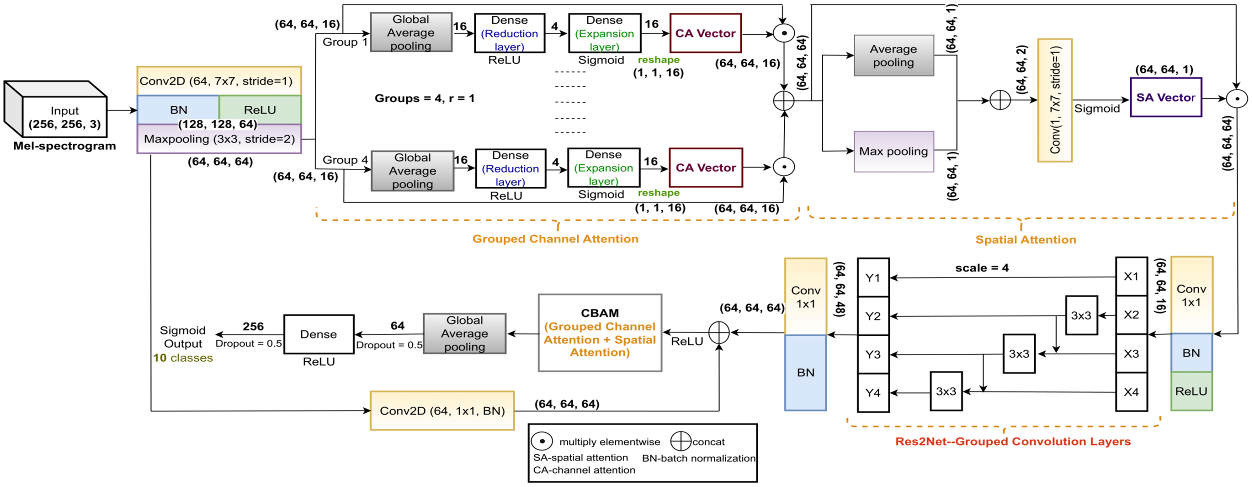

Fig. 1: Block diagram: Proposed Stacked Res2Net-CBAM with GCA with stack, N=1 with 64 filters, reduction ratio=1,groups=4

II. System Description

A. Dataset and Pre-processing

The audio recordings were collected from the bird sound database [35] of the Xeno-Canto Foundation[1]. To ensure consistency, all the files were normalized to a sample rate of 16 kHz, which was the minimum rate among the original files with sample rates ranging from 16 kHz to 44.1 kHz. Further, the audio data was converted to mono with 32-bit resolution. The train and test data statistics is given in Table I,II.

TABLE I: Data statistics for training

Bird Species	Scientific name	Bird Id.	XC files	Calls
House Crow	Corvus splendens	HC	27	111
Mallard Duck	Anas platyrhynchos	MD	25	106
Asian Koel	Eudynamys scolopaceus chinensis	AK	26	121
Eurasian Owl	Bubo bubo	EO	25	107
House Sparrow	Passer domesticus	HS	24	100
Blue Jay	Cyanocitta cristata	BJ	27	109
Red. Lapwing	Vanellus vanellus	RL	24	104
Grey Go-away	Corythaixoides concolor	GG	19	109
Indian Peafowl	Pavo cristatus	IP	29	103
W. Wood Pewee	Contopus sordidulus	WW	24	108
Total			250	1078

Each train file contains a single vocalization of 1.5s duration, while the test data contains time-overlapping vocalizations and multiple bird calls. The selection of these species was based on two rules: (1) the selected species cover a broad range of well-defined bird call structures [36], [37], including chirps, whistles, blocks, warbles, and clicks, to satisfy the generic requirement; and (2) the species should have an adequate number of samples for training and testing the proposed system. In our work, SpecAugment [34] was adopted to overcome data scarcity [38]. This approach masks frequency channels and time frames in the Mel-spectrogram representation. We generated 3344 Mel-spectrograms for the proposed method.

B. Feature Extraction

TABLE II: Dataset specification

	Class	Count (Bird Files)	# Calls
1	Audio Files (Train)	1078	1078
2	Audio Files (Test)		
	Calls with two species	334	668
	Calls with three species	100	300
	Total	1512	**2046**

Experienced bird watchers can recognize almost all species by their sounds. However, auditory abilities are often inferior to the visual capabilities of a person. A Mel-spectrogram is a visual representation that displays how frequencies in a signal change over time. We selected 224 Mel filter banks, 2048-point FFT, Hanning window of 2048 samples (approx 128 ms) and a hop-length of 512 samples (approx 32 ms) in our computation. Figure 2 shows the Mel-spectrogram of bird vocalizations with multiple bird sounds.

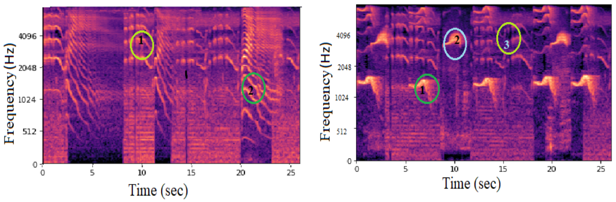

Fig. 2: Mel-spectrograms with multiple birds in a single audio file. Repetitive patterns shown in circles: 2 species (left), 3 species (right)

C. Proposed Framework

In this section, we describe the proposed Stacked Res2Net-CBAM [33] framework with SA and GCA for multi-label bird species classification using Mel-spectrogram features. The overall system is illustrated in Figure 1.

The Mel-spectrogram input (256×256×3) is first passed through a convolutional block. The proposed framework with stack number, N=1 includes:

Step 1: Grouped Channel Attention (GCA) block: The purpose of this block is to emphasize informative channels and suppress less informative ones. The input tensor is split into a specified number of groups $(G_1, G_2, ..., G_N)$, and for each group, a similar set of operations is applied. First, global average pooling generates a vector representation of the group's feature map (Z_i). Z_i represents mean activation,

[1] www.xeno-canto.org (Xeno-canto)

H and W represent height and width respectively, $G_{i,h,w}$ represent the activation of the h-th row and w-th column of feature map i. W_1, W_2 represents weight matrix of the first and second layer of the neural network respectively.

$$Z_i = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} G_{i,h,w} \tag{1}$$

This vector is then passed through a reduction layer with ReLU activation, followed by an expansion layer with sigmoid activation (σ). These layers reduce the dimensionality of the feature map and emphasize the most informative channels.

$$S_i = \sigma(W_2(ReLU(W_1(Z_i)))) \tag{2}$$

The resulting CA vector is reshaped to match the input shape and multiplied element-wise with the original group feature map to generate an enhanced output.

$$O_i = S_i \cdot G_i \tag{3}$$

This process is repeated for each group, and the outputs are concatenated and passed through a SA block.

$$O = concat(O_1, O_2, ..., O_N) \tag{4}$$

Step 2: Spatial Attention block (SA): The SA block includes two pooling layers, average pooling (A) and max pooling (M), and the resulting tensors are concatenated along the channel axis.

$$A = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} X_{h,w,:} \tag{5}$$

$$M = \max_{h,w}(X_{h,w,:}) \tag{6}$$

$$C = concat(A, M) \tag{7}$$

Next, a convolutional layer with sigmoid activation is applied to the concatenated tensor to generate a SA vector.

$$S = \sigma(Conv(C)) \tag{8}$$

The σ ensures that the attention vector values are between 0 and 1 where 1 indicates that the corresponding spatial feature is most relevant. The input tensor at the SA block is then element-wise multiplied with the SA vector (S) to emphasize the relevant spatial features and suppress the less relevant ones.

$$O = S \cdot O \tag{9}$$

Step 3: Res2Net block: The resulting tensor is then passed through a Res2Net block with a series of convolutional layers to perform feature extraction. Specifically, a convolutional layer with a 1x1 kernel and stride is applied to reduce the number of filters.

$$S = Conv(X); \qquad C_1 = Conv(O) \tag{10}$$

Then, the result is passed through a series of grouped convolutional layers(with scale= 4), each of which has a 3x3 kernel and stride of 1.

$$C_2 = grouped\_conv(C_1) \tag{11}$$

Finally, another 1x1 convolutional layer is applied to restore the original number of filters. A shortcut connection (S) is applied through a 1x1 convolutional layer to match the number of filters in the input.

$$C_3 = Conv(C_2); \quad O = C_3 + S; \quad O = ReLU(O) \tag{12}$$

Finally, a CBAM block [33] is stacked to further capture both channel and spatial dependencies of the resulting feature map.

Step 4: Applying CBAM:

$$GCA = grouped\_channel\_attention(O) \tag{13}$$

$$SA = spatial\_attention(GCA) \tag{14}$$

We repeat the steps from 1 to 4 for N stacks (Here, N=8 with filter size of 64, 64, 128, 128, 256, 256, 512, 512). After N stacks, a global average pooling layer is added to aggregate the spatial information and output a fixed-size feature vector. This vector is then fed to dense layers with ReLU (dropout=0.5) followed by sigmoid classifier for final prediction.

$$P_1 = GlobalAvgPool(SA)) \tag{15}$$

D. Sequential Aggregation Strategy (Se)

We propose a sequential aggregation strategy for detecting multiple species in overlapping recordings. The audio recordings are sliced into fixed-length segments, and the Mel-spectrogram of each segment is extracted and inputted to the models. The model, trained on ten classes, predicts the probability of each of the 10 species. The network then predicts the probability of each species present in a segment. Since there may be multiple species present in a single audio clip, multiple sigmoid outputs from each segment are aggregated and normalized, and the nodes with the highest probability values are considered the target species.

III. Experimental setup

We experimented with Res2Net+SA, Res2Net+CA, Res2Net+CBAM, and the proposed model based on Keras-TensorFlow. Additionally, we retrained deep CNN models such as Res2Net, InceptionV3, IncepResNetV2, EfficientNetB3, and some existing models using sequential aggregation strategy. These models were trained on a GPU P100 Kaggle notebook for a maximum of 300 epochs with a batch size of 32. The model's target was optimized using Adam with categorical cross-entropy loss, and sigmoid activation was applied at the output layer.

All comparative experiments shared the same operating environment, and hyperparameters, and utilized the same train and test set.

IV. Result Analysis

In our experiments shown in Table III, the proposed model outperforms all other sequential attention models with an F1 score of 72.20%. When comparing the species-specific results, all the bird species are showing class-wise results greater than 50% . The class-wise performance of the Eurasian Owl, Mallard Duck, Red-wattled Lapwing and Blue Jay is also

TABLE III: Precision (P), recall (R), and F1 score (in %) of the experiments

Species name	**Se-Res2Net**			**Se-Res2Net+SA**			**Se-Res2Net+CBAM**			**Proposed model**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
House Crow	82.60	76.00	79.17	86.75	72.00	78.69	76.85	83.00	79.81	77.12	73.04	78.32
Mallard Duck	51.80	48.86	50.29	50.00	48.86	49.43	58.24	60.23	59.22	73.13	76.09	75.35
Asian Koel	48.74	84.06	61.70	59.14	79.71	67.90	60.66	78.26	67.92	71.87	75.60	76.12
Eurasian Owl	47.37	32.73	38.71	51.72	27.27	35.71	59.37	34.55	43.68	66.07	41.05	55.98
House Sparrow	72.34	57.14	63.85	77.12	76.47	76.79	83.33	79.43	76.92	77.65	62.67	78.30
Blue Jay	51.72	44.55	47.87	65.48	54.46	59.46	65.08	40.59	50.00	62.18	51.02	59.55
Red-wattled Lapwing	59.09	57.35	58.21	64.29	66.18	65.22	73.98	66.91	70.27	88.54	61.00	74.98
Grey go-away	62.23	61.38	61.81	56.55	65.52	60.70	52.66	68.27	59.46	60.12	80.30	71.36
Indian Peafowl	54.64	84.13	66.25	60.00	76.19	67.13	55.17	76.19	64.00	47.83	94.06	69.85
Western Wood-Pewee	75.90	68.48	72.00	77.01	72.83	74.86	79.76	72.83	76.13	85.00	74.67	81.94
Macro Average	61.00	61.47	**60.00**	64.81	63.95	**63.60**	66.51	66.03	**64.74**	71.00	68.95	**72.20**

significantly improved. The proposed approach is better at handling variations in the audio recordings of the bird calls, such as variations in background noise or the presence of other calls. The results clearly show that the proposed Stacked Res2Net-CBAM with GCA using sliding window analysis improved the detection performance of multiple overlapping bird species, which is a significant challenge.

We explore the impact of the number of groups (G) and reduction ratio (r) in the GCA block as shown in Table IV. Intuitively, a larger group size gives a wider model that can give more attentive features. We kept the number of layers, epochs, optimizer and learning rate unchanged while varying the number of groups to explore its influence on classification results. The best F1 score of 72.20 is observed for G=4 and r=8 combination.

TABLE IV: F1 score (%) with varying reduction ratio and number of groups in the GCA block (Best values in bold)

Groups (G)	Reduction ratio (r)			
	2	4	8	16
2	60.37	61.68	62.78	63.86
4	58.14	68.30	**72.20**	70.18
8	61.32	62.00	64.12	64.70
16	58.18	61.82	64.83	64.16

From Table V, our proposed model has fewer trainable parameters (1.36 M) and the time consumption (285.32 s) is significantly reduced. In Figure 3, we can observe that increasing the number of groups from 1 to 16 in GCA block leads to fewer network parameters. However, higher accuracy is expected with more groups. The results in IV indicate that F1 score is improved with just four groups, and no further improvement with increase in the number of groups.

TABLE V: Comparison of proposed models in terms of Complexity (Best values in bold)

Models	**Parameters (M-Million)**	**Time per epoch (s)**
Se-Res2Net	1.95 M	641.554
Se-Res2Net+SA	1.58 M	770.770
Se-Res2Net+CA	2.19 M	1313.480
Se-Res2Net+CBAM	1.75 M	647.379
Proposed model	**1.36 M**	**285.320**

We conducted ablation experiments on Res2Net block [23] with different scales, including 1, 2, 4, 8, and 16. For scales= 1, 2, there were 4 and 8 groups, respectively, which reduced model capacity but also decreased complexity. For scales= 8, 16, there were 32 and 64 groups, respectively, which improved the performance but increased the complexity and overfitting risk. However, we found that using a scale of 4 provided the best balance between performance and computational efficiency, enabling optimal multi-scale feature learning. Thus, we used a scale of 4 in our experiments.

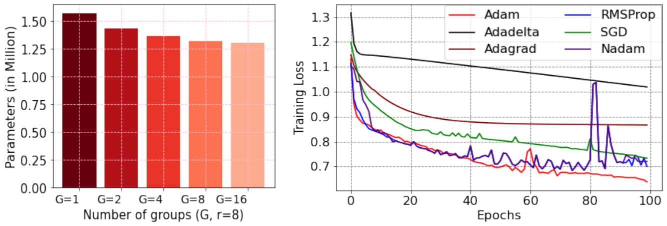

Fig. 3: Comparison of model parameters in terms of varying number of groups in GCA block(left), optimization algorithms(right)

As shown in Figure 3, by employing adam optimizer, model's training loss drops more quickly, exhibiting its ability to rapidly pick up on the essential characteristics from the bird vocalization. Adam performs admirably for our model and as the gradients get sparser toward the end of optimization, marginally outperforms RMSprop.

TABLE VI: Performance comparison with existing methods

Method	**P(%)**	**R(%)**	**F1(%)**
T. Grill et al. [Model1] [39]	50.15	50.81	45.12
T. Grill et al. [Model2] [39]	51.20	48.68	48.00
D. B. Efremova et al. [20]	60.97	54.00	53.10
G. Gupta et al. [43] [CNN+GRU]	67.97	65.00	66.98
F. Yang et al. [40][SENet]	65.00	58.12	58.00
Se-EfficientNetB3	56.94	57.00	55.32
Se-InceptionV3	55.70	54.11	51.22
Se-IncepResNetV2	63.98	60.99	59.79
Se-Res2Net [23]	61.00	61.47	**60.00**
Se-Res2Net+SA	64.81	63.95	**63.60**
Se-Res2Net+CA	62.00	60.10	**59.31**
Se-Res2Net+CBAM(SA+CA)	66.51	66.03	**64.74**
Proposed model (G=4, r=8)	71.00	68.95	**72.20**

The comparison of various algorithms using the Xeno-Canto dataset(multi-label) in terms of precision, recall, and F1 score is listed in Table VI. T. Grill et al. [39] compared two approaches (Global and local) to detect the presence of bird calls in audio recordings. D. B. Efremova et al. [20] used ResNet-50 to measure the efficacy of bird call classification and reports an F1 score of 53.10% . We implemented the CNN+GRU part of [43] and obtained 66.98% results. In [40], the SENet is utilized to facilitate the network's ability to dynamically recalibrate features on a per-channel basis. Comparing all pre-trained models with the Res2Net model [23] in Table VI, Se-Res2Net gives the best performance among the 4 pre-trained models used. Hence, we decided to utilize the Res2Net block and added the CBAM block [33] with GCA to it in a specified manner and as expected it further boosted the overall performance. Also, we experimented Res2Net with the spatial, channel, and CBAM independently. The F1 score for our best-performing model using sequential aggregation strategy is 72.20%, which is 19%, 27%, 24%, 14%, and 5% superior to the existing models [20], [39], [40], [43].

V. Conclusion

The paper presents the identification of bird species using Stacked Res2Net-CBAM with GCA framework in a multi-label scenario. The results show that the proposed model is superior to existing start-of-the-art models and achieves an F1 score of 72.20% with fewer network parameters.

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] Hackett, Shannon J., et al. "A phylogenomic study of birds reveals their evolutionary history." *science* 320.5884 (2008): 1763-1768.

[8] Fagerlund, Seppo. "Automatic recognition of bird species by their sounds." *Finlandia: Helsinki University Of Technology* (2004).

[9] Guyon, Isabelle, et al. "An introduction to feature extraction." *Feature Extraction:Foundations and Applications* (2006): 1-25.

[10] Vaca-Castano, Gonzalo, and Domingo Rodriguez. "Using syllabic mel cepstrum features and k-nearest neighbors to identify anurans and birds species." *IEEE Workshop On Signal Processing Systems*, 2010.

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

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

[13] 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.

[14] Schrama, T., M. Poot, M. Robb, and H. Slabbekoorn, Automated recording, detection and identification of nocturnal flight calls: Results of a pilot study during autumn migration in the Netherlands. *Journal of Ornithology*, 147.5(2006): 248-248.

[15] Thakur, Anshul, et al. "Deep Convex Representations: Feature Representations for Bioacoustics Classification." *Interspeech*(2018):2127-2131.

[16] Noumida, A., and Rajeev Rajan. "Deep learning-based automatic bird species identification from isolated recordings." *International Conference on Smart Computing and Communications (ICSCC)*(2021): 252-256.

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

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

[19] A. Noumida, R. Mukund, N. Madhavan Nair and Rajeev Rajan, "Multi-label Bird Species Classification Using Ensemble of Pre-trained Networks," *International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, 2023, pp. 644-649.

[20] 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.

[21] He, Kaiming, et al. "Deep residual learning for image recognition." *IEEE Conference on Computer Vision and Pattern Recognition*(2016): 770–778.

[22] Kim, Taejun, Jongpil Lee, and Juhan Nam. "Sample-level CNN architectures for music auto-tagging using raw waveforms." *IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP)*.(2018):366–370.

[23] Gao, Shang-Hua, et al. "Res2net: A new multi-scale backbone architecture." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.2 (2019): 652-662.

[24] Noumida, A., and Rajeev Rajan. "Multi-label bird species classification from audio recordings using attention framework." *Applied Acoustics* 197 (2022): 108901.

[25] Noumida, A., and Rajeev Rajan."Multi-label Bird Species Classification Using Hierarchical Attention Framework," *IEEE 19th India Council International Conference (INDICON)*, 2022.

[26] Rajan, Rajeev, and A. Noumida. "Multi-label bird species classification using transfer learning." *IEEE International Conference on Communication, Control and Information Sciences (ICCISc)*. Vol. 1, 2021.

[27] Rajan, Rajeev, Johnson, Jisna Abdul Kareem, Noumida. "Bird Call Classification Using DNN-Based Acoustic Modelling." *Circuits, Systems, and Signal Processing* 41, 2669–2680, 2022.

[28] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." *IEEE Conference on Computer vision and Pattern Recognition*(2018): 7132–7141.

[29] Zhao, Hengshuang, Jiaya Jia, and Vladlen Koltun. "Exploring self-attention for image recognition." *IEEE/CVF Conference on Computer Vision and Pattern Recognition.* (2020):10073–10082.

[30] Ramachandran, Prajit, et al. "Stand-alone self-attention in vision models." *Advances in Neural Information Processing Systems* 32 (2019).

[31] Naranjo-Alcazar, Javier, Sergi Perez-Castanos, Pedro Zuccarello, and Maximo Cobos. "Acoustic scene classification with squeeze-excitation residual networks." *IEEE Access* 8 (2020): 112287-112296.

[32] Yang, Jeong Hyeon, Nam Kyun Kim, and Hong Kook Kim. "Se-resnet with gan-based data augmentation applied to acoustic scene classification." *DCASE Workshop*. 2018.

[33] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." *European conference on Computer Vision (ECCV)*. 2018.

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

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

[36] Brandes, T. Scott. "Automated sound recording and analysis techniques for bird surveys and conservation." *Bird Conservation International* 18.S1 (2008): S163-S173.

[37] Duan, Shufei, et al. "Acoustic component detection for automatic species recognition in environmental monitoring." *International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, 2011.

[38] Kaya, Mahmut, and Hasan Şakir Bilge. "Deep metric learning: A survey." *Symmetry* 11.9 (2019): 1066. DOI:10.3390/sym11091066.

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

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

[41] Maegawa, Yuko, et al. "A new survey method using convolutional neural networks for automatic classification of bird calls." *Ecological Informatics* 61 (2021): 101164.

[42] Ntalampiras, Stavros, and Ilyas Potamitis. "Acoustic detection of unknown bird species and individuals." *CAAI Transactions on Intelligence Technology* 6.3 (2021): 291-300.

[43] Gupta, Gaurav, et al. "Comparing recurrent convolutional neural networks for large scale bird species classification." *Scientific Reports* 11.1 (2021): 17085.