Deep Learning-based Automatic Bird Species Identification from Isolated Recordings

Noumida A., Rajeev Rajan
Dept. of Electronics and Communication Engg.,
College of Engineering, Trivandrum.
Kerala, India
noumidaa@gmail.com, rajeev@cet.ac.in

2021 8th International Conference on Smart Computing and Communications (ICSCC) | 978-1-7281-9687-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICSCC51209.2021.9528234

***Abstract*—Birds play an extremely important role in an ecosystem, identifying bird species in audio recordings is challenging and has high research value. This paper aims to develop an effective bird call classification for isolated recordings (single-label) approach using various deep learning architectures, namely convolutional neural networks (CNN), deep neural networks (DNN), and transfer learning schemes. Transfer learning models have been widely used in a variety of deep learning applications. The performance of transfer learning models such as ResNet50, VGG-16, and InceptionResNetV2 has been compared to the acoustic MFCC-DNN methodology. On the Xeno-canto (XC) online bird audio dataset, the presented methods are tested. The dataset comprises ten species with 1078 audio tracks. The classification accuracies of 96.3%, 93.7%, and 91.9% are reported for ResNet50, CNN, and VGG-16, respectively, and outperform with the acoustic signal-based MFCC-DNN methodology.**

***Keywords*—single-label, transfer learning, convolutional neural network.**

I. Introduction

Bird species identification using audio recordings of their calls, songs, and noises is crucial and necessary for avian biodiversity conservation, and it aids ornithologists in detecting the presence of rare bird species in a specific region. Both songs and calls are part of bird vocalisation. Bird songs are viewed as more complicated vocalisations than calls, which are treated as more simple vocalisations. [1]. The challenge with automatic audio recordings is detecting the calls of interest species in lengthy recordings. As a result, developing automatic techniques for bird identification in audio recordings is required. Ornithologists are interested to know if a certain, possibly rare species has appeared in a particular area, as well as how many distinct bird species exist in that area.

Conventional field techniques to identify and track distinct bird species have required much human effort. Global biodiversity information facility (GBIF) [2], which generates multimedia biological datum, also focuses on automatic classification and recognition of bird species from field audio recordings. Deep learning methods can achieve higher detection accuracy on remote monitoring auditory data despite weather noise and a wide diversity of bird call kinds [1]–[4]. In the presented work, we examine the problem of predicting bird species based on their call from a given isolated audio recording. The vocalization and mel-spectrogram of Lapwing are shown in Fig. 1 and Fig. 2, respectively.

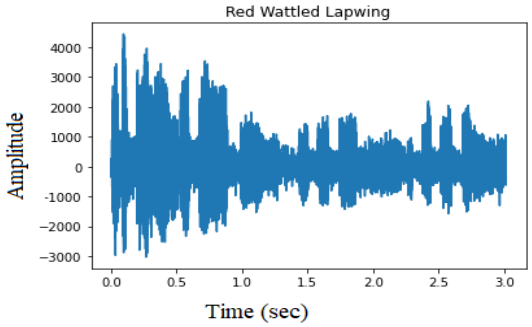

Fig. 1: Bird's vocalization of Lapwing.

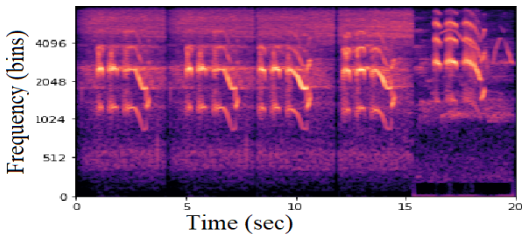

Fig. 2: Melspectrogram of bird's vocalization of Lapwing.

Various endeavors may be visible in the literature for the classification of bird species, from pre-segmented single-label acoustic sound recordings [5]–[7]. Bird species recognition through unsupervised modeling of individual bird syllables and duration modeling is explored in [8]. An iterative maximum likelihood procedure was introduced in the above work, to train the individual HMMs for each species' syllables. As a baseline system, an HMM-based detector with a general model learned from all syllables was created [9]. Conventional hidden Markov models with a probability density function at each state are employed for bird species recognition in

[1]https://en.wikipedia.org/wiki/Bird-vocalization
[2]www.gbif.org

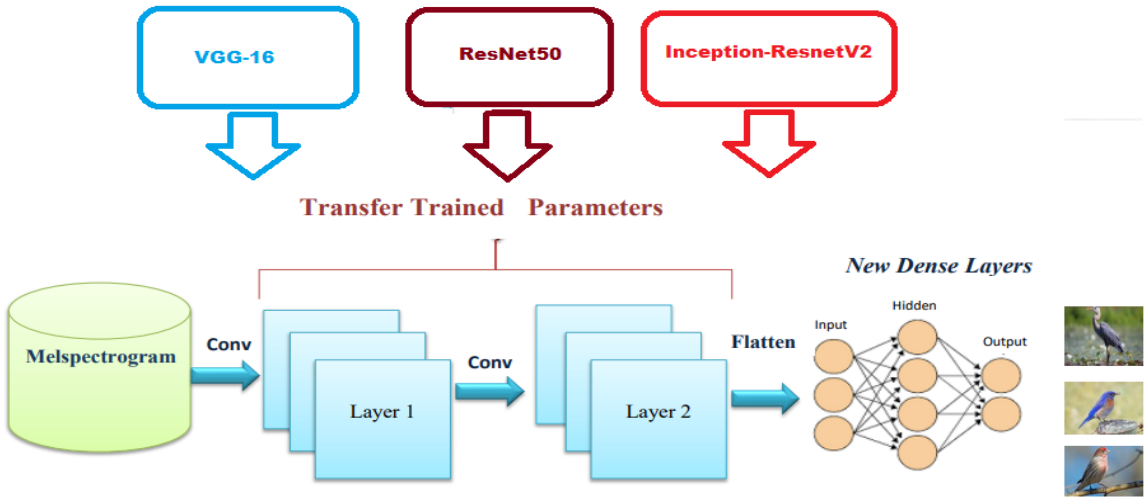

Fig. 3: The proposed scheme for single-label bird species identification system

[10]. The application of dynamic temporal warping to the automated analysis of continuous animal vocalisation recordings is discussed in [11]. Splitting the audio recording into time chunks, passing it to the CNN and receiving output for each segment is experimented in [12].Due to the difficulty of acquiring annotated training calls, the use of transfer learning in CNN could be advantageous in the field of bird call classification. A data-efficient bird call classification is performed using the Convolutional Neural Network(CNN) Transfer learning approach in [13].

ResNet50, a deep convolutional neural network structure for bird species recognition is implemented in [14] reached 60%-72% accuracy. The technique in [15] uses a deep learning based convolutional network approach to predict the most dominant foreground bird species within the acoustic scene with a precision score of 0.686. In the proposed work, we identify bird species from isolated recordings using various transfer learning schemes and study the performance.

In Section II, a brief description of the proposed framework and classification based on several transfer learning schemes is illustrated; in Section III, performance assessment is done; in Section IV, the findings are analysed; and finally, in Section V, the conclusion is given.

II. System Description for Single-Label Bird Identification

Bird species identification has been done from mel-spectrograms of audio files using CNN-based architectures. We mainly focus on single-label isolated audio recordings. In addition to the proposed CNN architecture, transfer learning schemes adopted from ResNet50, InceptionResNetV2, and VGG-16 have also been experimented with in the proposed study. The experimental framework is shown in Fig.3.

A. Feature Extraction

Melspectrograms are computed with frame size 30 ms and hop size 10 ms with 256 bins for the CNN architecture. The MFCC-DNN framework uses the librosa python package to compute 40-dimensional MFCCs with a frame size of 30ms and a hop size of 10ms. Automatic speech/speaker detection and acoustic-based applications commonly use MFCCs.In MFCC, the frequency bands are logarithmically positioned. Therefore it approximates the response of the human auditory system more closely than any other system.

TABLE I: CNN architecture.

Input size	Description
3x256x256	Melspectrogram
32x256x256	7x7 Convolution, 32 filters
32x128x128	3x3 Max-pooling
64x128x128	3x3 Convolution, 64 filters
64x64x64	3x3 Max-pooling
128x64x64	3x3 Convolution, 128 filters
128x32x32	3x3 Max-pooling
256x32x32	3x3 Convolution, 256 filters
256x16x16	3x3 Max-pooling
512x16x16	3x3 Convolution, 512 filters
512x8x8	3x3 Max-pooling
32768	Flatten
128	Dense
64	Dense
10	Softmax

B. Classification Phase

Deep learning methods, namely CNN, DNN, and transfer learning models such as ResNet50, Inception-ResNetV2 and VGG-16 are implemented. The deep learning convolutional neural network-based models are based on processing mel spectrogram. The proposed CNN architecture is shown in Table I; Convolutional

Neural Networks has been efficaciously investigated as one of the representation learning strategies that permit a machine to automatically find the patterns required for classification tasks [16]–[18]. Instead of stacked layers, CNN uses alternating convolution layers and pooling layers to deliver the advantage of spatial invariance and local information extraction. For a uniform input to CNN, all the bird calls are processed to have the same length.

Transfer learning is the process of learning the model on a specified dataset and using the feature parameters in the target dataset. The three models addressed in this paper are pre-trained on an ImageNet database consisting of more than 1000 different categories and then the model parameters are transferred to the birds call database for training. In this work, weight parameters of the network-specific layers in each model are transferred and fine-tuned the model by learning the actual data. Fig. 3. shows the method of transfer learning based on CNN models.

III. Performance Assessment

A. Data set

Test records are gathered from broadly used Xeno-canto bird audio database [19] of xeno-canto foundation [3]. Table II portrays the dataset (sampling rate, 16000Hz) utilized for the proposed methodology. The train statistics comprised of 644 isolated sound recordings of ten bird species. A total of 434 records are utilized for testing. The files are refined in such a way that only one vocalisation of 1.5 second duration is there in each sound record.

TABLE II: Statistics of Train data

SL.No	Class	Calls
1	House Crow	66
2	Mallard Duck	63
3	Asian Koel	72
4	Eurasian Owl	64
5	House Sparrow	60
6	Blue Jay	66
7	Red-wattled Lapwing	62
8	Grey go-away	66
9	Indian Peafowl	61
10	Western Wood Pewee	64
	Total	644

B. Experimental Framework

The CNN model was trained by optimizing the categorical cross-entropy between predictions and targets with adaptive moment estimation (Adam). The neural network used a deep architecture where multiple convolution and max-pooling layers are collectively stacked to study an affluent set of attributes. We increase the number of feature maps for the convolution layer by a factor of two after every layer, starting from 32 up to 512, followed by the Rectified linear unit activation.

The training of the model is executed with 50 epochs and 32 batch size using the softmax activation function. 10% data was used for validation during experimentation,.

Three deep convolutional neural network models are retrained in our study: VGG-16, ResNet50, Inception-ResNetV2. The VGG-16 architecture uses 13 convolutional and 3 fully connected dense layers. The convolutional layers in VGG-16 are of kernel size 3×3 and stride size 1, and the pooling layers are of size 2×2 with stride size 2. The size of the activation map is reduced by half after each pooling layer. The final activation map is having 7×7 with 512 channels. ResNet50, a variant of the ResNet model that utilizes residual learning, is a 50 layer deep CNN architecture having 48 Convolutional layers, 1 Max-pooling layer, and 1 Average-pooling layer. InceptionResNetv2, a variant of Inceptionv3 is a very deep convolutional network architecture having 164 layers that build on the architecture of the Inception family but incorporates residual interconnections. The deep CNN

TABLE III: Classification accuracy

Sl.No	Method	Accuracy (%)
1	Melspectrogram-CNN	93.7
2	MFCC-DNN	87.7
3	VGG-16	91.9
4	ResNet50	**96.3**
6	InceptionResNetV2	87.0

TABLE IV: Confusion matrix for InceptionResNetV2

	AK	BJ	HC	MD	GG	RL	EO	IP	HS	WW
AK	47	0	0	0	0	0	0	2	0	0
BJ	0	25	8	0	0	0	0	10	0	0
HC	0	0	44	1	0	0	0	0	0	0
MD	0	0	9	32	0	1	0	0	1	0
GG	1	2	3	0	37	0	0	0	0	0
RL	1	1	0	0	2	37	0	0	0	1
EO	1	0	0	0	0	0	42	0	0	0
IP	0	0	0	0	0	0	0	42	0	0
HS	0	0	1	2	0	2	1	0	32	2
WW	1	0	0	0	0	0	0	0	1	42

TABLE V: Confusion matrix for ResNet50

	AK	BJ	HC	MD	GG	RL	EO	IP	HS	WW
AK	49	0	0	0	0	0	0	0	0	0
BJ	0	39	0	0	0	0	0	0	0	4
HC	0	0	44	1	0	0	0	0	0	0
MD	0	0	3	40	0	0	0	0	0	0
GG	0	1	1	0	40	0	0	0	1	0
RL	0	0	0	0	0	42	0	0	0	0
EO	0	0	0	0	0	0	42	0	0	1
IP	0	0	0	0	0	0	0	42	0	0
HS	0	1	0	2	0	0	1	0	36	0
WW	0	0	0	0	0	0	0	0	0	44

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

TABLE VI: PRECISION , RECALL , AND F1 SCORE

SL.No	Bird Class	MFCC DNN			CNN			VGG16			ResNet50			InceptionResNetV2		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	Asian Koel	0.75	0.82	0.78	1.00	1.00	1.00	0.96	0.98	0.97	1.00	1.00	1.00	0.92	0.96	0.94
2	Blue Jay	0.78	0.93	0.85	0.71	0.93	0.81	0.97	0.97	0.97	0.95	0.91	0.93	0.89	0.58	0.70
3	House Crow	0.89	0.89	0.89	0.95	0.90	0.97	0.70	0.89	0.78	0.92	0.98	0.95	0.67	0.9	0.80
4	Mallard Duck	0.77	0.93	0.84	0.96	1.00	0.98	0.82	0.77	0.79	0.93	0.93	0.93	0.91	0.74	0.82
5	Grey go-away	0.95	0.91	0.93	0.93	0.91	0.92	1.00	0.91	0.95	1.00	0.93	0.96	0.95	0.86	0.90
6	Red-wattled Lapwing	0.98	1.00	0.99	0.91	0.76	0.83	0.95	1.00	0.97	1.00	1.00	1.00	0.92	0.88	0.90
7	Eurasian Owl	1.00	0.56	0.72	1.00	0.95	0.98	0.93	0.93	0.93	0.97	0.98	0.98	0.97	0.97	0.97
8	Indian Peafowl	1.00	1.00	1.00	1.00	0.97	0.98	0.97	1.00	0.99	1. 00	1.00	1.00	0.78	1.0	0.87
9	House Sparrow	0.80	0.87	0.83	0.97	0.82	0.90	0.97	0.75	0.84	0.97	0.90	0.93	0.94	0.80	0.86
10	Western Wood Pewee	1.00	0.89	0.94	0.98	1.00	1.00	1.00	1.00	1.00	0.90	1.00	0.95	0.93	0.95	0.94

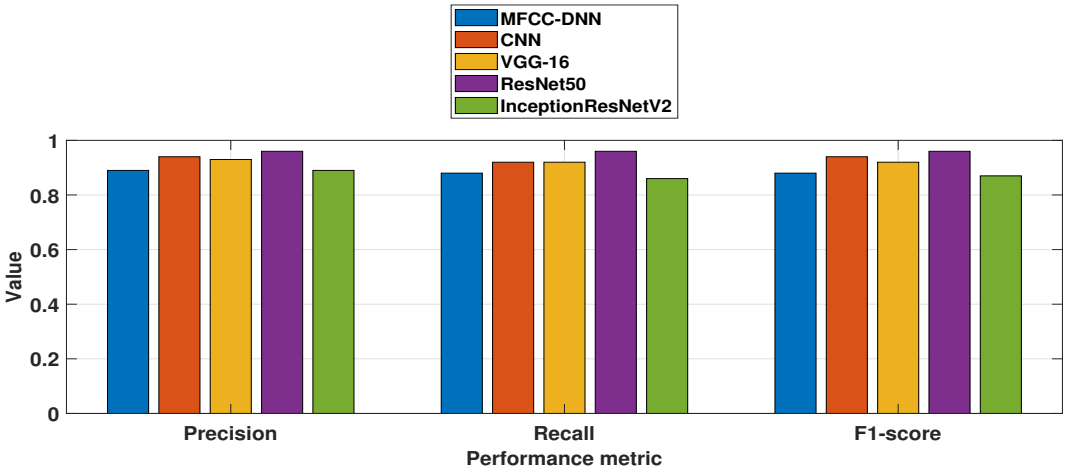

Fig. 4: The performance metrics for five phases

network is trained on 1.28M images in thousand different classes. Implementation of the transfer learning schemes is done by stacking 3 dense layers, with the number of perceptrons expanding in powers of 2 starting from 32 units. All the transfer learning models are trained with a batch size of 32 for 300 epochs using Adam optimizer, after hyperparameter tuning.Precision, recall, and F1-score are used to assess the performance, which is then compared to that of the MFCC-DNN technique.

For MFCC-DNN framework, 40 dim MFCCs are computed utilizing the librosa library in the front-end. The Deep Neural Network model consists of 2 hidden layers, each with 256 neurons, followed by ReLU activation and a 0.5 dropout value. The output layer has ten neurons corresponding to 10 classes followed by the softmax activation . The training of the model is carried out in 400 epochs with a batch size of 32 using Adaptive Moment Estimation optimizer.

IV. RESULTS ANALYSIS

Overall results of the experiment are shown in Table III. Overall classification accuracy of 96.3%, 91.9%, 93.7%, 87.7%, 87% is obtained for ResNet50, VGG-16, CNN, MFCC-DNN, InceptionResNetV2 respectively. It is observed that there is an improvement of 8.6% for the best performing ResNet50 models over the MFCC-DNN model. As per the experiments reported in [14], ResNet50 achieved an accuracy of 60%-72%. In our experiments, ResNet50 yields the highest accuracy of 96.3%. The performance of CNN and VGG-16 is at par with that of ResNet50, with a clear margin over DNN.

Table IV. and V., respectively, show the confusion matrix for the InceptionResNetV2 model and the best performing ResNet50 model for the target dataset. It is clear that the ResNet50 models significantly reduce misclassification errors among classes. The class-wise accuracy of Eurasian Owl is less than 75% for the MFCC-DNN approach. For VGG-16 and CNN, no bird shows accuracy less than 75%. All of the classes in the best-performing ResNet50 report accuracy of greater than 90%. In the transfer learning-based ResNet50, misclassification errors of Asian Koel and Eurasian Owl were significantly reduced. Notable improvement is observed for the class, Asian Koel. It is worth noting that class-wise accuracy of three species is 100%.

Table VI. shows the precision, recall, and F1 score of the experiments. The MFCC-DNN technique's average precision, recall, and F1 measure are 0.89, 0.88, and 0.88, respectively. Precision, recall, and F1 score are reported as 0.96, 0.96, and 0.96, respectively, by ResNet50. Fig.4.shows the macro average precision, recall, and F1 measure for the four approaches. Average F1 measure for DNN, CNN, ResNet50, VGG-16 and InceptionResNetV2 based framework is 0.88, 0.94, 0.96, 0.92 and 0.87, respectively. It is worth mentioning that, the best performance of the deep learning strategies is achieved with reasonably less amount of training data as compared to other deep learning models. Deep-learning convolutional neural networks are adapted and fine-tuned to detect the presence of bird species in recordings in this paper. They were originally designed for image classification.

We would like to extend the proposed scheme for multi-label classifications with multiple overlapping bird vocalizations present in the audio recording. In such cases, a sliding window analysis may be required to identify multiple bird species for sequential processing of audio recordings. In addition, data augmentation methods need to be incorporated to overcome the scarcity of data for training the networks. The presented approach illustrates the potential of pre-trained weights-based experimental methods for the specified task.

V. Conclusion

In the current years, many deep learning models have been applied to various image classification tasks, since deep learning models have a good performance in image classification. The paper reports a relative analysis on bird call classification using various deep learning algorithms, specifically CNN, DNN, and transfer learning. The performance is evaluated using a dataset with 10 species. The results show that ResNet50, CNN, and VGG-16 outperforms the DNN and prove to be far better. We were able to achieve 96.3% accuracy of single-label bird call recognition using ResNet50.

References

[1] D. Stowell, M. Wood, H. Pamuła, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning: The first bird audio detection challenge," *Methods in Ecology and Evolution*, vol. 10, 10 2018.

[2] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin, "Bird detection in audio: A survey and a challenge," *in Proceedings of IEEE International Workshop on Machine Learning for Signal Processing Salerno, Italy*, pp. 1–6, 2016.

[3] D. Gelling, "Bird song recognition using GMMs and HMMs," *Masters Project Dissertation, Department of Computer Science,University of Sheffield*, 2001.

[4] A. Thakur, V. Abrol, P. Sharma, and P. Rajan, "Local compressed convex spectral embedding for bird species identification," *The Journal of the Acoustical Society of America*, vol. 143, pp. 3819–3828, 2018.

[5] I. Potamitis, S. Ntalampiras, and K. R. Olaf Jahn, "Automatic bird sound detection in long real-field recordings: Applications and tools," *Applied Acoustics*, pp. 1–9, 2014.

[6] A. Thakur, V. Abrol, P. Sharma, and P. Rajan, "Compressed convex spectral embedding for bird species classification," *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 261–265, 2018.

[7] ——, "Deep convex representations: Feature representations for bioacoustics classification," *in Proceedings of International Conference on Spoken Language Processing*, pp. 2127–2131, 2018.

[8] P. Jancovic, M. Kokue, M. Zakeri, and M. Russell, "Bird species recognition using HMM-based unsupervised modelling of individual syllabls with incoparated duration modelling," *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 559–563, 2016.

[9] A. Harma, "Automatic identification of bird species based on sinusoidal modeling of syllables," *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. V–545, 2003.

[10] P. Jančovič, M. Köküer, and M. Russell, "Bird species recognition from field recordings using HMM-based modelling of frequency tracks," *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8252–8256, 2014.

[11] S. E. Anderson, A. S. Dave, , and D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings," *The Journal of Acoustical Society of America*, pp. 1209–1219, 1996.

[12] E. Knight, K. Hannah, G. Foley, C. Scott, R. Brigham, and E. Bayne, "Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs," *Avian Conservation and Ecology*, vol. 12, no. 2, 2017.

[13] D. B. Efremova, M. Sankupellay, and D. A. Konovalov, "Data-efficient classification of birdcall through convolutional neural networks transfer learning," *Digital Image Computing: Techniques and Applications*, pp. 1–8, 2019.

[14] M. Sankupellay and D. Konovalov, "Bird call recognition using deep convolutional neural network, resnet-50," *in Proceedings of ACOUSTICS 2018, Adelaide, Australia*, pp. 1–8, 2018.

[15] E. Sprengel, M. Jaggi, Y. Kilcher, and T. Hofmann, "Audio based bird species identification using deep learning techniques," *in Proceedings of CLEF*, 2016.

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[17] M. Sukhavasi and S. Adappa, "Music theme recognition using CNN and self-attention," *preprint arXiv:1911.07041*, 2019.

[18] D. Ghosal and M. H. Kolekar, "Music genre recognition using deep neural networks and transfer learning." *in Proceedings of Interspeech*, pp. 2087–2091, 2018.

[19] W.-P. Vellinga and R. Planqué, "The xeno-canto collection and its relation to sound recognition and classification," *Working Notes of CELF*, 01 2015.