Bird Species Classification Using CNN-Based Deep Learning and Advanced Data Augmentation Techniques

Arshad Khan
Department of Computer Science and Engineering
Aliah University
Newtown, Kolkata
mysterywriter175@gmail.com

Khondekar Lutful Hassan
Department of Computer Science and Engineering
Aliah University
Newtown, Kolkata
klh.cse@gmail.com

Moinuddin
Department of Computer Science and Engineering
Aliah University
Newtown, Kolkata
cse.moinuddin@gmail.com

***Abstract*— The study of bird populations is crucial for biodiversity research and conservation. Deep artificial neural networks have revolutionized bird acoustic recognition, with state-of-the-art data augmentation methods employed to enhance and improve classification accuracy and model robustness. This paper proposes a CNN based model for bird species classification. Dataset is taken from BirdClef2024 which focuses on recognizing bird species in the Western Ghats of the Indian Peninsula to assist in conservation efforts. Changes in environmental conditions are closely linked to shifts in species diversity. To classify 182 bird species, the proposed model incorporates well-known data augmentation techniques commonly used in audio classification.**

Keywords—Bird Species, Birds Sound, Audio signal processing, Deep Learning, Bird species classification

I. Introduction

Rise of rapid industrialization has led to a decline in the global biodiversity of both flora and fauna. As the threat of numerous species facing extinction draws near, researchers are scrambling to defuse this threat.

Birds are an excellent indicator of a region's biodiversity change. Monitoring their population is essential to gauze the health of an ecosystem. Traditionally, bird monitoring involved direct observation by experts. But this can be costly, resource intensive and at times may not be viable due to geographical constraints. So, a better way of monitoring bird population is needed.

The solution is using passive acoustic monitoring (PAM). Acoustic monitoring of bird species allows greater efficiency with reduced manpower in monitoring bird population in a specific region. This method of passive monitoring allows humans to monitor biodiversity over a much larger scale, with machine learning algorithms.

This paper aims to delve into the use of various preprocessing and augmentation methods alongside a classifier to make a robust bird species classification model. The paper compares and contrasts the efficacy of different combinations of augmentations. A Convolutional Neural Network (CNN) as the main architecture of the proposed model.

II. Related Work

Deep Learning models have become extremely popular after the use of CNN models by Font et al (2021), which showed the potential of CNNs in bird song recognition. A lot of CNN architectures have evolved over the years, and their strengths have been excellently shown in the annual BirdCLEF competition, where most top scoring teams used these models as a main backbone in their classifiers.

In addition to that, it has been shown that using data augmentation techniques can enhance the accuracy of these models further.

ISBN : 978-81-978522-2-0 https://doi.org/10.37285/bsp.oaisem2025.24

Arunodhayan et al. [3] utilized some common data augmentation methods such as pitch shift and time and frequency masking along with applying background noise, leading to a total of 19 different augmentations. Their paper proposed a model based on DenseNet-161 and ResNet-10. Both DenseNet and ResNet have become a go to architecture for many classification tasks. ViTs were chosen for sound event detection. The augmentations improved model performance to 64.52 % (DenseNet-161), 63.00 % (ResNet-50).

Qingyu et al [7] used random spectrogram mixup, random tanh distortion, and random clipping and splicing as augmentations in their PPNN (Phylogenetic Perspective Neural Network) and attentional CNN model.

In general, classifying birdcalls is a challenging task. These challenges are:

1) The presence of background noises, variation between individuals of the same species and regional variations. In addition to that, isolated birdcalls are very rare, and the soundscape often contain different birdcalls mixed together.
2) Given an audio clip, there is a chance that the clip does not contain any birdcall, which adds noise to the weights during training.
3) Datasets often contain imbalances in the species representation. Rare species have fewer recordings while common species have abundance of recordings.

III. Database

The dataset provided by this year's BirdCLEF competition (BirdCLEF 2024) has been used in this paper. This year's competition aimed to identify under-studied Indian bird species in the Western Ghats. The dataset's files were sourced from xenocanto.org. All the files were downsampled to 32 kHz. The dataset has over 24000 labelled files, from 182 species. Unlabelled soundscapes from the target domain have also been provided in this dataset.

IV. Proposed work

A. Preprocessing

Audio is loaded using librosa. The first 5 seconds from each audio is taken as input. Audios which are less than 5 seconds are padded with zeros with NumPy. Librosa's melspectrogram function is used to convert the selected audio clip into a melspectrogram. The melspectrogram is normalized and resized.

Additionally, if the number of files of a species is very small, the audio files is cut into 5 sec segments and treated them as separate files.

B. Melspectrogram

Waveforms are very useful for conveying basic information about an audio. Unfortunately, they cannot tell us much about the pitch, frequency, or harmonic content of a recording.

Spectrogram fills in that role. Spectrograms keep time on the X axis but place frequency on the Y axis. They are great at representing the features of an audio, but a melspectrogram can give a lot more than a normal spectrogram.

The configurations of melspectrogram are given in table 1.

ISBN : 978-81-978522-2-0

TABLE I. MELSPECTROGRAM CONFIGURATION

Parameter	Value
Clip Length	5 sec
Sample Rate	32000
Min Frequency	40 Hz
Max Frequency	15000 Hz
Power	2
Mel Bands	128
FFT component	1024
Hop Length	500

C. *Augmentations*

Augmentations are necessary to improve the accuracy of the proposed model, and at the same time prevent overfitting on the data. The following augmentations have been used:

Mixup: Two randomly chosen audio are mixed together by a certain ratio.
Cutmix: Part of the audio is cut and replaced with a different audio sample.
Pitch Shift: The pitch of the audio is shifted by a predefined amount.

Audiomentations library is used to implement pitch shift of the augmentations, while torch and torchaudio is used for the rest of the augmentations.

D. *Post Processing*

Inference is done using the following steps:

The audio from unlabelled soundscapes is divided into 5 second clips and prediction is done on each 5 second chunk.

To match the training domain, the audio is filtered to reduce noise, as the unlabelled soundscaped contain a lot more noise than the training data.

To optimise the predictions, chunk averaging is applied between the scores of the chunks.

ISBN : 978-81-978522-2-0

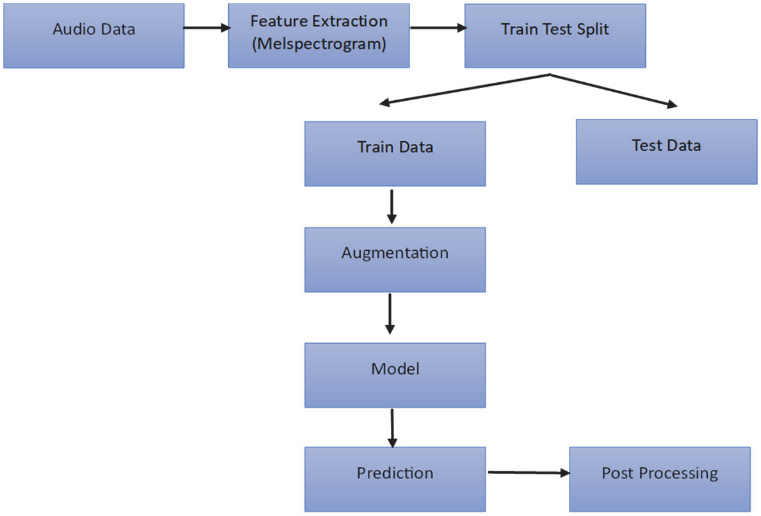

Figure 1: Diagram of Proposed Work

V. Proposed Model

CNN models efficientnet_b2 and resnet_34 has been proposed.

A. Efficientnet_b2

This model is built upon the efficientnet model architecture. It uses 2d convolution layers to extract features from the melspectrogram images. The original input layer is changed to accept an image of one channel only, rather than the default 3 channel (RGB). The output layer, i.e. the fully connected layer is also changed to have an output shape equal to 182, the total number of classes.

The model is imported with the default pretrained weights and is trained on the five second audio clips. For inference, the same clip length is used. This model is able to extract deep features due to being a deep neural network.

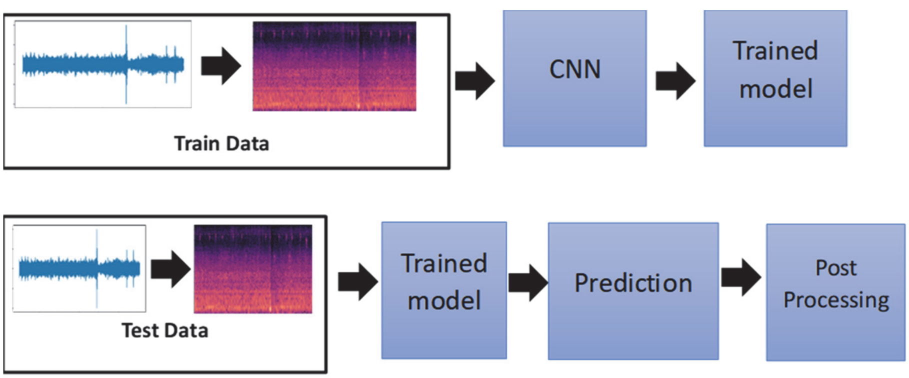

Figure 2: Classifier Pipeline

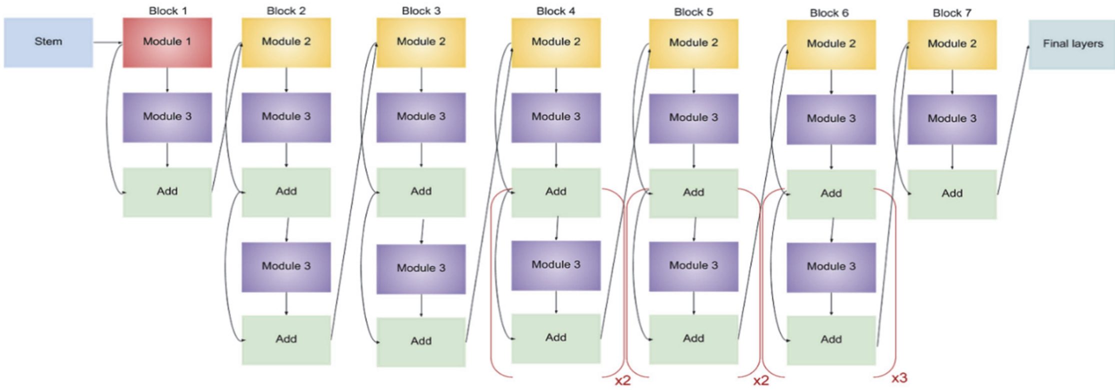

Figure 3: EfficientNet B2 Architecture

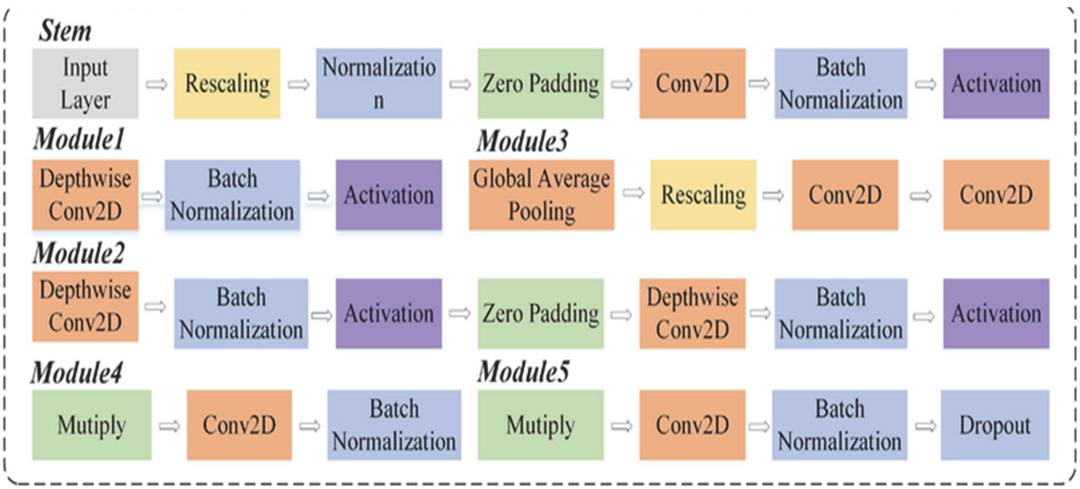

Figure 4: EfficientNet B2 Architecture

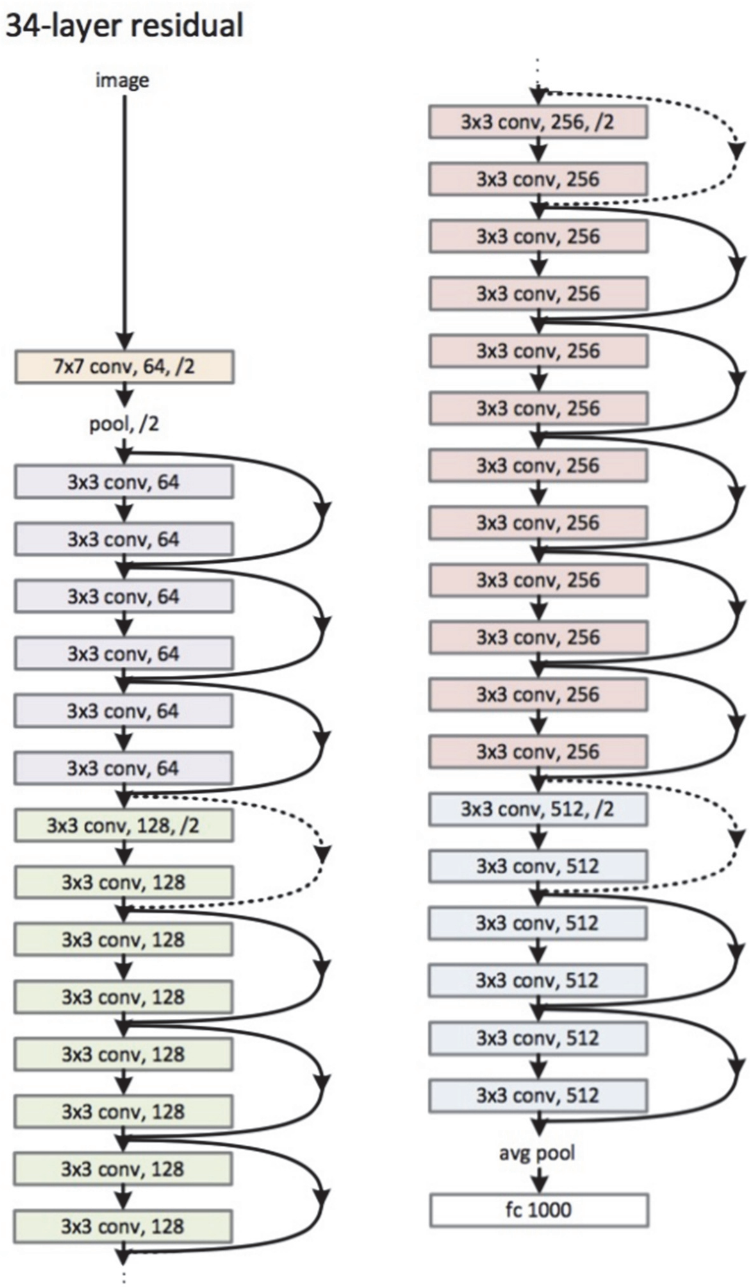

Figure 5: ResNet34 Architecture

ISBN : 978-81-978522-2-0

B. ResNet34

This model is implemented in the same way as the efficientnet model. The input and output layers are modified to take in an image of one channel and give an output of 182 classes. This model is faster than efficientnet_b2 in training since it is smaller. It has residual blocks in its architecture which gives it an advantage over efficientnet model. The skip connections ensure the gradient is not lost during backpropagation. Feature extraction is more robust from the melspectrograms.

C. Loss functions

Following loss functions have been used:

CE Loss: Cross-Entropy Loss, also known as log loss, is a popular loss function used for classification tasks.

Focal Loss: Focal Loss was introduced by Lin et al. (2017) to address the limitations of CE Loss in handling class imbalance. Focal Loss adds a modulating factor to the Cross-Entropy Loss, focusing more on hard-to-classify examples.

In the end, CE Loss was chosen as it was easier to understand and implement.

D. Optimizer

Adam optimizer is used in model training. Adam, a stochastic optimization algorithm, is effective due to its adaptive learning rates and it is computationally efficient.

E. Scheduler

OneCycleLR is used as a scheduler for model training. This scheduler uses a technique where the learning rate starts at an initial value, increases to a maximum value, and then decreases to a minimum value lower than the initial value. This helps in faster convergence and avoids getting stuck in local minima.

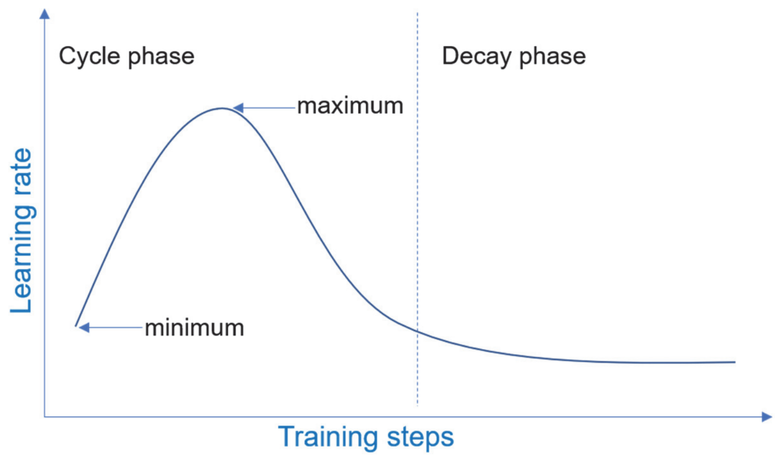

Figure 6: Graph of Learning Rate in OneCycleLR scheduler

VI. RESULTS

The performance of the various models is compared using the validation scores. Interestingly, EfficientNet had higher score than the ResNet model when trained without any augmentations. Its assumed this is because the EfficientNet model has more layers in it than ResNet. The validation loss also was lower than in ResNet.

ISBN : 978-81-978522-2-0

TABLE II. LOSS COMPARISON (WITHOUT AUGMENTATION)

Model	Loss	Accuracy
ResNet34	1.66	0.64
EfficientNet_B2	1.44	0.68

TABLE III. LOSS COMPARISON (WITH AUGMENTATION)

Model	Loss	Accuracy
ResNet34	1.62	0.66
EfficientNet_B2	1.74	0.61

Using augmentations had inconsistent results. The cause is not known, and it is attributed to the lack of experience in fine-tuning the augmentations.

Pitch Shift and Time Shift do not improve the accuracy, but sometimes showed a worse performance when used. Its suspected that changing the pitch and time changes the information of the bird call in the audio, which reduces model accuracy.

Other than that, CutMix and Mixup show a good improvement when used together, but it caused overfitting. Only when the probability, cutmix ratio and mixup factor was reduced that there was some improvement. More experimentation is needed to figure out why it isn't working as expected.

ResNet had an improvement in its accuracy, while EfficientNet's accuracy decreased. Even though both were trained up to the same epochs. Its assumed that the EfficientNet's deep layers caused it to pick up the noise more than the features, which caused a decrease in accuracy score.

TABLE IV. COMPARISON TABLE

Paper	Accuracy
Arunodhayan et al. [3]	64.52 %
Proposed Method (ResNet34)	66%

VII. CONCLUSION

This paper demonstrates the use of CNN models and augmentation techniques in birdcall recognition. Augmentation techniques when used after finetuning, had a positive impact on our models. Results suggest EfficientNet was more prone to overfitting and picking up noise than ResNet.

This paper concludes that model architecture containing ResNet is more resilient to the background noise and is better suited for recognizing bird species in complex acoustic environments of the real world. Using its residual networks, it can overcome the vanishing gradient problem.

CNNs have huge potential in solving this bird classification problem. Both model architecture, and augmentation techniques have an impact on the robustness of the model.

A lot of models were left out due to time constraints, which the future experiments can implement in order to find a balance between speed and performance. There are also other augmentations and postprocessing techniques that could be implemented in the future. A lot of experiments are needed to finetune the augmentations.

Training took a lot of time for efficientnet, which is needed to be mitigated in the future. A fast trainable model will be very promising, as it can be trained and deployed earlier. Furthermore, the preprocessing pipeline could be improved to reduce resource utilization..

ACKNOWLEDGMENT

I, Arshad Khan would like to thank Prof. Khondekar Lutful Hassan and Moinuddin, Aliah University, for their insightful discussions and suggestions, which greatly enriched the quality of this work during the preparation of this technical paper.

REFERENCES

[1] Hong, L. (2023). Acoustic Bird Species Recognition at BirdCLEF 2023: Training Strategies for Convolutional Neural Network and Inference Acceleration using OpenVINO. In CLEF (Working Notes) (pp. 2043-2050).

[2] Kahl, S., Denton, T., Klinck, H., Reers, H., Cherutich, F., Glotin, H., ... & Joly, A. (2023, September). Overview of BirdCLEF 2023: Automated Bird Species Identification in Eastern Africa. In CLEF (Working Notes) (pp. 1934-1942).

[3] Kumar, A. S., Schlosser, T., Kahl, S., & Kowerko, D. (2024). Improving learning-based birdsong classification by utilizing combined audio augmentation strategies. Ecological Informatics, 82, 102699.

[4] Miyaguchi, A., Cheung, A., Gustineli, M., & Kim, A. (2024). Transfer Learning with Pseudo Multi-Label Birdcall Classification for DS@ GT BirdCLEF 2024. arXiv preprint arXiv:2407.06291.

[5] Porwal, A. (2024). Bird-Species Audio Identification, Ensembling of EfficientNet-B0 and Pre-trained EfficientNet-B1 model.

[6] Baowaly, M. K., Sarkar, B. C., Walid, M. A. A., Ahamad, M. M., Singh, B. C., Alvarado, E. S., ... & Samad, M. A. (2024). Deep transfer learning-based bird species classification using mel spectrogram images. *PloS one*, *19*(8), e0305708.

[7] Wang, Q., Song, Y., Du, Y., Yang, Z., Cui, P., & Luo, B. (2024). Hierarchical-taxonomy-aware and attentional convolutional neural networks for acoustic identification of bird species: A phylogenetic perspective. Ecological Informatics, 80, 102538.

[8] Xeno-canto : Sharing wildlife sounds from around the world. (n.d.). https://xeno-canto.org

[9] Sura, H. (2021, December 13). Audio Classification using Librosa and Pytorch - Hasith Sura - Medium. Medium. https://medium.com/@hasithsura/audio-classification-d37a82d6715

[10] Nandi, P. (2022, August 2). CNNs for audio Classification - towards data science. Medium. https://towardsdatascience.com/cnns-for-audio-classification-6244954665ab

[11] BirdCLEF 2024 | Kaggle. (n.d.). https://www.kaggle.com/competitions/birdclef-2024