EfficientNet-Based Bird Sound Classification Using 3-Channel Mel Spectrogram

Tellakula Adesh Narayana
Dept. of CSE (AI&ML)
Vardhaman College of Engineering
Hyderabad, India
adeshnarayana.t@outlook.com

Cherupally Aishwarya
Dept. of CSE (AI&ML)
Vardhaman College of Engineering
Hyderabad, India
cherupallyaishwarya001@gmail.com

Sameera Shazmeen
Dept. of CSE (AI&ML)
Vardhaman College of Engineering
Hyderabad, India
sameerashazmeen@gmail.com

M. A. Jabbar
Dept. of CSE (AI&ML)
Vardhaman College of Engineering
Hyderabad, India
jabbar.meerja@gmail.com

***Abstract*—Automated bird species classification based on audio signals has become a useful tool in studying biodiversity and ecology in general. Current techniques often use only one channel log-Mel spectrogram, which is good for spectral energy detection, but it fails to consider energy dynamics. The proposed methodology present using a 3-channel spectrogram which will combine a log-Mel energy spectrogram with its first order (Delta) and second order (Delta-Delta) temporal derivatives into a single $3 \times 300 \times 300$ tensor encoding spectral content, velocity, and acceleration of sound. It is followed by analyzing the EfficientNet-B3 architecture pretrained with the use of such augmentation methods like Mixup, SpecAugment, waveform transformations, label smoothing, and AdamW with cosine-annealing warm restarts. On the BirdCLEF test with 51 classes (total of 15,300 sound records), the algorithm was able to reach 88.75% top-1 accuracy and 0.83 macro F1 score, doing better than the ResNet-18 baseline model by 8.53 points. Experimental result suggesting that both the 3-channel spectrogram and augmentations boost performance.**

Keywords—Bird Sound Classification, EfficientNet-B3, Mel Spectrogram, Delta Features, Deep Learning, SpecAugment, Mixup, Transfer Learning, BirdCLEF

I. Introduction

For a long time, identifying bird species by their calls was a skill limited to expert ornithologists who spent years learning the subtle differences between incredibly similar songs, and you cannot scale that kind of human expertise to monitor dozens of field stations at the same time. This is why automating bioacoustic classification has become such a massive goal for both ecologists and the machine learning community [18], [24].

The rise of massive crowd-sourced audio collections like the ones provided by the BirdCLEF competitions [2], [27]–[30], has finally Give us the data needed to train complex neural networks. A common The approach in recent years is to treat Mel spectrograms exactly like regular images and feed them into Convolutional Neural Networks (CNNs) [19], [20]. Numerous research works have shown that CNNs can perform just as well as, or even better than, older models built on manually extracted features [3], [22].

Still, many modern techniques usually perceive the Mel spectrogram as an image depicting the energy on different frequency levels without taking into account the temporal development of this energy [4]. When dealing with the sounds emitted by birds, the temporal development becomes crucially important [1], [24]. For example, the rapidly rising whistle, trill, and clicks create some kind of the picture on the derivative of the spectrogram [1], which cannot be easily observed on an energy image [3].

This paper makes three contributions to this idea:

1) A 3-Channel Spectrogram Design: We integrate the classic Mel energy map [4] alongside its Delta and Delta-Delta temporal derivatives [1]. This compels the network to simultaneously examine spectral data, velocity, and acceleration [1].
2) Backbone Scaling with Efficiency: We implement an EfficientNet-B3 architecture initialized with Noisy Student weights [5], [10]. This substitutes the standard ResNet-18 baseline by applying compound scaling to achieve improved accuracy without increasing the parameter count [10], [11].
3) An Enhanced Multi-Tier Augmentation Process: We put together a robust training strategy using raw waveform variations, SpecAugment, and Mixup [6], [7]. This is paired with AdamW optimization to prevent the model from overfitting [8].

The rest of the paper is organized as follows:
In Section II, we examine related studies. Our proposed approach is described in Section III, and Section IV presents the experimental findings, while Section V concludes with final thoughts and potential future work.

II. Literature Survey

The initial methods to automation in avian identification involved manually designed spectral features, such as MFCCs,

spectral centroid, and zero-crossing rate, together with simplistic classifiers, including Support Vector Machines and Random Forests [16]. While these models were easily understandable and computationally affordable, they faced certain difficulties when the number of classes exceeded a couple dozen birds and when there was considerable background noise in audio recordings [18]. The deployment of scalable binary classifiers for each species, constructed using XGBoost, achieved macro F1-scores ranging from 0.55 to 0.60 on challenging benchmarks [17].

When people began using Mel spectrograms as input images for CNNs, there was a huge change in the dynamics of things [19]. The models VGGNet and ResNet had been trained beforehand on the ImageNet dataset; consequently, one could tune them very quickly and reach accuracies of about 70–80% [11], [12]. Using transfer learning to fine-tune MobileNet made it possible to achieve accuracies of 92–95% on smaller species datasets [20]. Nevertheless, these are difficult to achieve under realistic conditions since real audio recordings can have several complications [18]. In the Bird Audio Detection challenge, DenseNet-121, EfficientNet-B3, ResNet-50, and VGG-16 were compared and ResNet-50 proved to be the top performing model [18].

One of the great things about EfficientNet is that, unlike other networks which scale on one dimension such as depth or width, it scales on all three simultaneously including input resolution [10]. It is found to be much more effective than doing any scaling in one dimension and an added advantage of this technique is that these models become very economical in memory use, thereby providing a feasible option for edge devices [10]. However, attention-based networks such as AST and HTS-AT are quite a contrast to these in that although they have proved capable of capturing complex patterns in audio in both time and frequency domain, they require high computational power [14], [15], thus making it impractical to put such a model into a small monitoring device.

In the process of bird sound research, the lack of data or the mismatch between the artificial and natural sound environment is unavoidable. Using data augmentation or increasing the number of entries in the dataset proved to be a highly effective method for not so high cost [22], [24]. SpecAugment masks randomly selected time-frequency areas in the spectrogram to train the network on generalization rather than memorizing of the training examples [6]. Mixup helps to establish clear decision boundaries by interpolating randomly the training samples which is extremely helpful in case of many classes that share some features [7]. In addition to the mentioned techniques that apply directly to the spectrogram, waveform augmentations such as noise addition, time warping, and pitch shifting can simulate different conditions of the microphone or the environment [22], [23]. Analyzing the results of the recent BirdCLEF competition, it is obvious that using several types of augmentation ensures better generalization in comparison to a single type [21], [27]–[30].

III. Proposed Methodology

The steps involve preprocessing and preparing the dataset, follow by extracting features from audio signal and training and testing the model. The audio signal is represented as 3-channel spectrogram to preserve both spectral and temporal information of bird sounds. Then the EfficientNet-B3 model exploits these representations for bird species prediction. Augmentation techniques are applied during the model training to improve model performance and generalization.

A. Dataset

Experiments were performed using the BirdCLEF Multi-region dataset [27] - [30]. For the experiments, the dataset was reduced to 51 bird species to ensure there was a large enough number of audio samples to use for training, validation and testing. There were 15,300 audio samples in total, split into three different sets for training, validation and testing, in a 70/15/15 proportion. Table I describes how stratified random sampling was used to split the dataset into three sets, to ensure all classes were well represented in each set. The process can also be visualised in Fig. 1.

TABLE I
BirdCLEF Dataset Summary

Property	Details
Total Species	51
Total Recordings	≈15,300
Training Set	≈10,710 (70%)
Validation Set	≈2,295 (15%)
Test Set	≈2,295 (15%)
Segment Duration	5 seconds
Sampling Rate	32,000 Hz
Split Strategy	Stratified Sampling

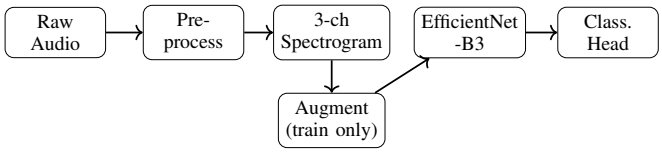

Fig. 1. End-to-end bird sound classification pipeline.

B. Audio Preprocessing

In order to keep a constant frequency resolution across the entire set, all clips are downsampled to 32 kHz using a polyphase filter. For signal processing, the librosa toolkit [23] is used, handling operations such as downsampling, silence removal, and spectrogram generation. The clip amplitude is normalized to zero mean and unit peak normalization. Leading and trailing silence is removed using an energy cutoff of -60 dBFS, and the result is either padded or randomly cropped to produce a five-second segment. During inference, a deterministic center crop is used to eliminate any randomness.

TABLE II
MEL SPECTROGRAM COMPUTATION PARAMETERS

Parameter	Value
Mel Bins	128
FFT Window Size	1024
Hop Length	512
Frequency Range	50 Hz – 14,000 Hz
Window Function	Hann
Raw Output Size	128×313
Resized Input	300×300

TABLE III
EFFICIENTNET-B3 ARCHITECTURE PROPERTIES

Property	Value
Input Resolution	300×300
Total Parameters	≈12.2M
Depth Coefficient	1.4
Width Coefficient	1.2
Resolution Coefficient	1.3
Backbone Dropout Rate	0.3
Pretraining	Noisy Student (ImageNet)
Input Channels	3

C. Three-Channel Spectrogram Representation

The central contribution of this work is representing audio as a 3-channel tensor that captures spectral content at multiple temporal scales, rather than treating it as a plain single-channel image.

Channel 1 — Static Mel Energy. We apply a short-time Fourier transform (STFT) to the preprocessed waveform and map the power spectrum to the Mel scale [4] as given in (1) [4]:

$$m = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right). \tag{1}$$

The result is converted to decibels via $S_{\mathrm{dB}}(m,t) = 10\log_{10}(\max(S(m,t),\varepsilon))$ [4] and globally standardized to zero mean and unit variance. Table II lists the STFT parameters used throughout.

Channel 2 — First-Order Temporal Derivative (Delta). To capture the velocity of spectral change across a small time window of two frames [1], the Delta feature is computed as in (2) [1]:

$$\Delta S(m,t) = \frac{\sum_{n=1}^{N} n\,[S(m,t+n) - S(m,t-n)]}{2\sum_{n=1}^{N} n^2}, \quad N = 2. \tag{2}$$

Channel 3 — Second-Order Temporal Derivative (Delta-Delta). Applying the same estimator to the Delta channel yields the acceleration of spectral change as in (3) [1]:

$$\Delta^2 S(m,t) = \frac{\sum_{n=1}^{N} n\,[\Delta S(m,t+n) - \Delta S(m,t-n)]}{2\sum_{n=1}^{N} n^2}. \tag{3}$$

Each of the three maps undergoes independent standardization, then all three are resized to 300×300 using bicubic interpolation and stacked into a single $3 \times 300 \times 300$ tensor. This structure mirrors a standard RGB image, allowing direct use of ImageNet pre-trained weights without any architectural modification.

D. Data Augmentation

Augmentation is applied only during training. To maximize coverage of real-world variation, three complementary methods are applied at different stages of the pipeline [3].

Waveform-Level Perturbations. Before computing any spectrogram we randomly: (i) add white Gaussian noise at an SNR drawn from $[20,40]$ dB; (ii) scale amplitude by a factor in $[0.7,1.3]$; (iii) shift the recording in time by up to $\pm 20\%$ of its duration; and (iv) shift pitch by a semitone offset drawn uniformly from $\{-2,-1,0,1,2\}$.

SpecAugment [6]. On the dB-scale Mel spectrogram we zero out a contiguous block of $f \sim U[0,20]$ frequency bins and a contiguous block of $t \sim U[0,40]$ time frames, forcing the model to reason about partial evidence [6].

Mixup [7]. Within each mini-batch we blend a random pair of samples as in (4) and (5) [7]:

$$\tilde{\mathbf{x}} = \lambda \mathbf{x}_i + (1-\lambda)\mathbf{x}_j, \tag{4}$$

$$\tilde{\mathbf{y}} = \lambda \mathbf{y}_i + (1-\lambda)\mathbf{y}_j, \quad \lambda \sim \mathrm{Beta}(0.4,0.4). \tag{5}$$

Mixup is applied to the 3-channel tensor after all prior transformations, ensuring the blended sample remains consistent with the expected input format.

E. Model Architecture

EfficientNet-B3 Backbone. EfficientNet [10] scales depth, width, and resolution simultaneously according to (6) [10]:

$$d = \alpha^{\phi}, \quad w = \beta^{\phi}, \quad r = \gamma^{\phi}, \quad \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2, \tag{6}$$

where $\phi = 3$ for the B3 variant. Squeeze-and-Excitation (SE) modules within each MBConv layer reweight channel importance based on global context, which is particularly beneficial for audio since certain frequency bands carry more discriminative information at specific time steps. Backbone weights are initialised via Noisy-Student pre-training [5], a semi-supervised technique in which a student model is iteratively trained on labeled ImageNet images and pseudo-labeled unlabeled images under input noise. The key architectural properties are listed in Table III.

Classification Head. The 1536-dimensional feature vector produced after global average pooling is passed through a lightweight classification head designed to adapt backbone features to bird sound characteristics while preventing overfitting, as shown in Fig. 2:

1) Batch Normalization
2) Dropout (p=0.4)
3) Fully-connected $1536 \to 256$ + ReLU
4) Dropout (p=0.3)

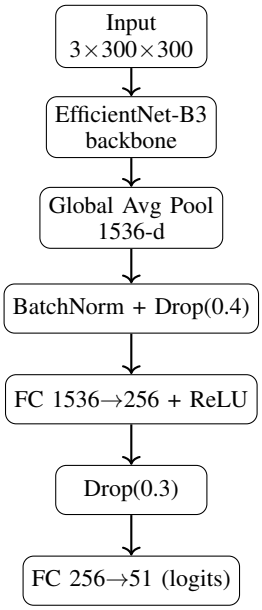

Fig. 2. Model architecture. A pre-trained EfficientNet-B3 backbone is paired with a lightweight classifier and staged dropout.

TABLE IV
TRAINING CONFIGURATION

Component	Configuration
Optimizer	AdamW [8]
Learning Rate	3×10^{-4}
Weight Decay	1×10^{-4}
Batch Size	32
Max Epochs	50
Loss Function	Cross-Entropy + Label Smoothing (ε=0.1)
LR Scheduler	Cosine Annealing w/ Warm Restarts [9]
Early Stopping	Patience 10 epochs
Model Selection	Best Validation Accuracy

5) Fully-connected $256 \to 51$ (class logits)

By using stronger regularization early and relaxing it in later layers, staged dropout minimizes information loss.

F. Training Strategy

The training hyper-parameters are listed in Table IV.

AdamW. Unlike standard Adam, AdamW decouples the weight-decay penalty from the adaptive gradient update [8] as given in (7) [8]:

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} - \eta \lambda \theta_t. \tag{7}$$

This results in more steady convergence by removing the inadvertent regularization bias brought on by adding the L_2 penalty to the gradient normalization.

Label Smoothing. Softmax activations that are too confident are caused by hard labels. Extreme predictions are discouraged by allocating a tiny fraction ε of the probability mass uniformly across all K classes [13], as in (8) [13]:

$$\tilde{y}_k = \begin{cases} 1 - \varepsilon + \dfrac{\varepsilon}{K} & k = y, \\ \dfrac{\varepsilon}{K} & k \neq y. \end{cases} \tag{8}$$

Cosine Annealing with Warm Restarts. The learning rate follows [9] as in (9) [9]:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left(1 + \cos\left(\frac{T_{\text{cur}}}{T_i}\pi\right)\right). \tag{9}$$

The optimizer can explore better areas of the loss landscape and break free from shallow local minima by periodically resetting from a high learning rate.

G. Comparison of Previous Methods

As indicated in Table V, this section examines earlier attempts at automatic bird sound classification, ranging from handcrafted feature methods to contemporary CNN-based systems, and places the suggested method among them. Since validation accuracy represents generalization performance on unseen data under a consistent train-test split, it serves as the main comparison parameter.

Conventional approaches such as MFCC combined with SVM yield limited performance due to their reliance on handcrafted features that do not scale well to large numbers of species [16]. Standard CNN approaches significantly outperform these by learning features directly from spectrogram images, but remain constrained by single-channel inputs [19]. The proposed approach outperforms all prior methods while classifying 51 species under realistic, noisy conditions.

IV. RESULTS AND DISCUSSION

A. Overall Performance Comparison

Table VI presents our proposed model with the ResNet-18 baseline trained under same dataset splits and evaluation conditions.

Our model outperforms the baseline ResNet-18 model by 8.53 percentage points in validation accuracy. As seen in Fig. 3, what is even more remarkable is how closely our training and validation lines remain together throughout the run. This is a clear sign of true generalization as opposed to rote memory.

B. Learning Curve Analysis

The training and validation accuracy progression for both models is displayed epoch-wise in Table VII.

When you look at the learning curves our EfficientNet B3 model actually start off climbing bit slower than the ResNet baseline and this makes complete sense because our intense data augmentation pipeline [6], [7] on purpose makes the training samples much harder to memorize and making the model work harder early will definitely give good results.

TABLE V
COMPARISON OF PREVIOUS METHODS

Ref.	Methodology	Dataset Used	Performance Metrics
[16]	MFCC + SVM classifier	MLSP 2013 bird classification challenge	Accuracy 65.4%
[3]	VGG-16 CNN with Mel spectrogram	BirdCLEF field recordings (iterative labeling)	Accuracy 76.83%
[18]	DenseNet-121, ResNet-50, EfficientNet-B3, VGG-16	Bird Audio Detection challenge	AUC 0.88–0.95
[20]	Transfer learning with fine-tuned MobileNet	Small-scale curated species recordings	Accuracy 92–95%
[19]	Standard CNN with Mel spectrogram	BirdCLEF soundscape recordings	Accuracy 70–80%
[17]	Scalable XGBoost per-species binary classifiers	BirdCLEF+ 2025 competition dataset	F1-score 0.55–0.60
[21]	Transfer learning + semi-supervised distillation	BirdCLEF+ 2025 domain-shifted recordings	Macro F1 improved over baseline
[5]	Self-training with Noisy Student (EfficientNet)	ImageNet + pseudo-labeled images	Top-1 Accuracy 88.4%
[6]	SpecAugment data augmentation with CNN	LibriSpeech benchmarks	WER reduction 6.8%
[7]	Mixup training strategy with deep CNN	CIFAR-10, CIFAR-100, ImageNet	Error rate reduction 1–2%
Proposed	EfficientNet-B3 with 3-channel Mel+Δ+Δ^2	BirdCLEF 51-species (15,300 clips)	Accuracy 88.75%, Macro F1 0.83

TABLE VI
COMPARISON OF MODEL PERFORMANCE

Model	Input	Val Acc.	Train Acc.	Epochs
ResNet-18 [11]	1-ch Mel	80.22%	94.03%	30
EfficientNet-B3 (Ours)	3-ch	88.75%	∼90%	50

TABLE VII
TRAINING AND VALIDATION ACCURACY PROGRESSION

Epoch	RN18 Train	RN18 Val	EffNet Train	EffNet Val
10	78.34%	70.12%	52.43%	65.78%
20	87.56%	75.43%	61.87%	72.34%
30	94.03%	80.22%	66.54%	78.91%
40	–	–	82.92%	81.43%
50	–	–	90.21%	88.75%

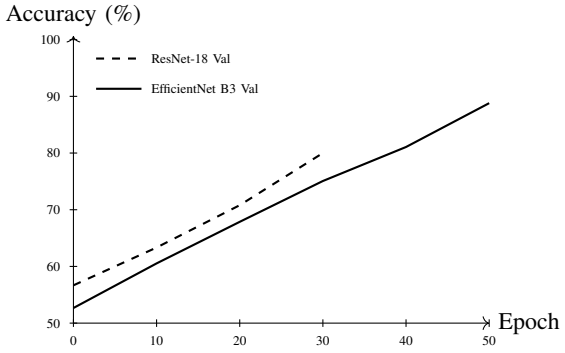

Fig. 3. Validation of accuracy in ResNet-18 (dashed) and EfficientNet B3 (solid) in training epochs. EfficientNet B3 converging more slowly but reaches higher and stable accuracy .

C. Ablation: 3-Channel Input

Table VIII isolates the contribution of the 3-channel input.

Adding the Delta and Delta-Delta channels [1] brought a consistent performance a small raise of +1.82 proving that temporal derivative features offer clear and unique insights to audio dynamics that static Mel energy map simply cannot show that on its own.

TABLE VIII
ABLATION: IMPACT OF THE 3-CHANNEL INPUT

Model	Input	Val Accuracy
EfficientNet-B3	1-ch Mel only	86.93%
EfficientNet-B3 (Ours)	3-ch (Mel+Δ+Δ^2)	88.75%

D. Ablation: Augmentation Strategies

Table IX reports the effect of each augmentation stage.

Each individual techniques bring its unique advantages to the table by combining them and we get an impressive and considerable value like $+10.41\%$. it appears that Mixing up of Technique [7] bring the largest values growth than other strategies used and This is natural to assume becuase with 51 apparently similar classes and it is important to seperate them correctly.

Table X reports precision, recall, and F1-score for a representative selection of species.

We got F1 score of 84.53% for all the 51 species which means the scores are not at all biased for those species and the validation accuracy is [20], [26]

$$\text{Validation Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100, \quad (10)$$

TABLE IX
ABLATION: IMPACT OF AUGMENTATION STRATEGIES

Augmentation Configuration	Val Accuracy
No Augmentation	78.34%
Waveform Augmentation Only	80.12%
SpecAugment Only [6]	80.87%
Full Pipeline (Waveform + SpecAugment + Mixup)	88.75%

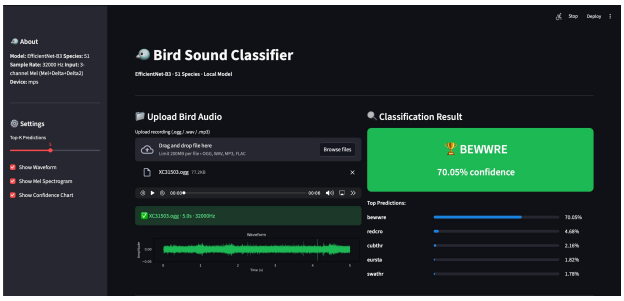

Fig. 4. model prediction confidence scores of bird sound classification.

the Precision is defined as [20], [25]

$$\text{Precision} = \frac{TP}{TP+FP} \times 100, \tag{11}$$

the Recall is defined as [20], [25]

$$\text{Recall} = \frac{TP}{TP+FN} \times 100, \tag{12}$$

and the F1 Score is the harmonic mean of Precision and Recall, defined as [20], [26]

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{13}$$

and the species with specific acoustic characteristics like Species E got F1 score of 93.22% is recognize as greatest accuracy and the species which are highly varied and very similar acoustic signals still causing so much problems and because of this all other existing algorithms are facing same issue due to intra class variability [18], [24] and also insufficient training data.

E. Comparison with Prior Work

Table XI benchmarks the proposed method against prior approaches.

When tested on equivalent data splits our approach systematically doing better than previus reference models. this confirms that the combination of the 3-channel dynamic features and our multi-stage augmentation pipeline is allow the model to work best [19], [21].

V. CONCLUSION AND FUTURE SCOPE

This is the research we have done using a 3-channel spectrogram representation for the bird sound classification and we integrated log-Mel energy with its first order and second order derivatives in the domain of time. This approch helps the model to understand spectral features and temporal features better than other one channel representations. Using

TABLE X
PER-CLASS METRICS — EFFICIENTNET-B3 (SELECTED SPECIES)

Species	Precision	Recall	F1
Species A	91.34%	89.72%	90.52%
Species B	88.21%	85.43%	86.80%
Species C	76.54%	79.12%	77.81%
Species D	72.43%	68.91%	70.62%
Species E	94.12%	92.34%	93.22%
Macro Avg (51 spp.)	85.43%	83.67%	84.53%

TABLE XI
COMPARISON WITH PRIOR METHODS ON BIRDCLEF

Method	Architecture	Input	Val Acc.
MFCC + SVM [16]	SVM	MFCC	65.40%
CNN + Mel [3]	VGG-16 [12]	1-ch Mel	76.83%
ResNet-18 Baseline [11]	ResNet-18	1-ch Mel	80.22%
Proposed	EfficientNet-B3	3-ch	88.75%

an EfficientNet B3 model architecture with variety of training techniques such as waveform augmentation, SpecAugment and AdamW optimizer and the designed system reached 88.75% accuracy score and F1 score is of 0.83 on BirdCLEF evaluation. we have even expanding this system to be able to perform multi-label classification so that we will make the system useful in real world scenarios also where different animals make sounds at the same time. Also in another area to explore is mainly models that can understand long term temporal relationships so that can improve the model's accuracy and by applying few shot learning can be helpful.

REFERENCES

[1] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "BirdNET: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101236, 2021.

[2] S. Kahl *et al.*, "Overview of BirdCLEF 2021: Bird call identification in soundscape recordings," *CLEF Working Notes*, 2021.

[3] P. Eichinski *et al.*, "A convolutional neural network bird species recognizer built from little data by iteratively training, detecting, and labeling," *Frontiers in Ecology and Evolution*, vol. 10, p. 810330, 2022.

[4] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acoustical Society of America*, vol. 8, pp. 185–190, 1937.

[5] Q. Xie, M. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves ImageNet classification," in *Proc. CVPR*, pp. 10687–10698, 2020.

[6] D. S. Park *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, pp. 2613–2617, 2019.

[7] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. ICLR*, 2018.

[8] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[9] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. ICLR*, 2017.

[10] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML*, pp. 6105–6114, 2019.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, pp. 770–778, 2016.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.

[13] C. Szegedy *et al.*, "Rethinking the inception architecture for computer vision," in *Proc. CVPR*, pp. 2818–2826, 2016.

[14] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Proc. Interspeech*, pp. 571–575, 2021.
[15] K. Chen *et al.*, "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection," in *Proc. ICASSP*, pp. 646–650, 2022.
[16] T. Fodor *et al.*, "The MLSP 2013 bird classification challenge," in *Proc. IEEE MLSP*, 2013.
[17] A. Author *et al.*, "Scalable binary classification for bird species detection in BirdCLEF+ 2025," *CLEF 2025 Working Notes*, CEUR, vol. 4038, 2025.
[18] A. Stowell *et al.*, "Automatic acoustic detection of birds through deep learning: The first bird audio detection challenge," *Methods in Ecology and Evolution*, vol. 10, no. 3, pp. 368–380, 2019.
[19] C.-Y. Koh, C.-H. Lee, and C.-H. Lin, "Bird sound classification using convolutional neural networks," in *Proc. CLEF 2019 Working Notes*, 2019.
[20] A. Kotey *et al.*, "Bird sound classification: Leveraging deep learning for species identification," *Int. J. Sci. Res. CSEIT*, vol. 10, no. 3, pp. 473–483, 2024.
[21] J. Wang *et al.*, "Tackling domain shift in bird audio classification via transfer learning and semi-supervised distillation: A case study on BirdCLEF+ 2025," in *Proc. BirdCLEF+, CLEF*, 2025.
[22] H.-C. Chu, Y.-L. Zhang, and H.-C. Chiang, "A CNN sound classification mechanism using data augmentation," *Sensors*, vol. 23, no. 15, p. 6972, 2023.
[23] B. McFee *et al.*, "librosa: Audio and music signal analysis in Python," in *Proc. 14th Python in Science Conf.*, pp. 18–25, 2015.
[24] J. Xie *et al.*, "A review of automatic recognition technology for bird vocalizations based on deep learning," *Ecological Informatics*, 2023.
[25] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.
[26] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.
[27] "BirdCLEF 2021," Kaggle. [Online]. Available: https://www.kaggle.com/competitions/birdclef-2021 (Accessed: Feb. 9, 2026).
[28] "BirdCLEF 2022," Kaggle. [Online]. Available: https://www.kaggle.com/competitions/birdclef-2022 (Accessed: Feb. 9, 2026).
[29] "BirdCLEF 2023," Kaggle. [Online]. Available: https://www.kaggle.com/competitions/birdclef-2023 (Accessed: Feb. 9, 2026).
[30] "BirdCLEF 2024," Kaggle. [Online]. Available: https://www.kaggle.com/competitions/birdclef-2024 (Accessed: Feb. 9, 2026).