IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 31, No. 3, March, 2026, pp. 695-705
Journal homepage: http://iieta.org/journals/isi

EfficientNet-B0 for Automated Bird Species Identification from Spectrogram-Encoded Vocalizations

Nilesh B. Korade[1*], Mahendra B. Salunke[2], Amol A. Bhosle[3], Gayatri G. Asalkar[4], Vivek V. Jog[2], Pradyna A. Vikhar[5], Dhanashri M. Joshi[6], Sunil M. Sangve[7], Manasi Gursale[8]

[1] Department of Computer Science and Engineering (Artificial Intelligence), Vishwakarma Institute of Technology, Pune 411037, India
[2] School of Computer Science and Information Technology, Symbiosis Skills & Professional University, Pune 412101, India
[3] Department of Computer Science and Engineering, MIT Art, Design and Technology University, Pune 412201, India
[4] Department of Computer Science and Engineering (Data Science), Vishwakarma Institute of Technology, Pune 411048, India
[5] Department of Computer Engineering, JSPM's Rajarshi Shahu College of Engineering, Pune 411033, India
[6] School of Engineering, Ajeenkya DY Patil University, Pune 412105, India
[7] Department of Computer Science and Engineering (Software Engineering), Vishwakarma Institute of Technology, Pune 411048, India
[8] Department of Computing, Dublin City University, Dublin City D09 V209, Ireland

Corresponding Author Email: nilesh.korade.ml@gmail.com

https://doi.org/10.18280/isi.310304

Received: 19 July 2025
Revised: 1 November 2025
Accepted: 18 March 2026
Available online: 31 March 2026

Keywords:
bird species identification, spectrogram-based deep learning, EfficientNet-B0, transfer learning, bioacoustics, wildlife conservation

ABSTRACT

Automated bird species identification is essential for biodiversity monitoring and conservation, particularly for rare or elusive species where visual observation is impractical. This study presents an acoustic-based classification framework that converts bird vocalizations into spectrogram representations and applies deep transfer learning for species recognition. A dataset of 436 audio recordings from ten bird species was preprocessed through clipping, noise augmentation, pitch shifting, and time shifting to enhance model robustness. Five deep learning architectures—Convolutional Neural Network (CNN), AlexNet, VGG-19, ResNet-50, and EfficientNet-B0—were trained and evaluated on spectrogram inputs. EfficientNet-B0 consistently outperformed all other models, achieving an accuracy of 93%, precision of 94%, recall of 93%, and F1-score of 93%. The lightweight architecture and compound scaling strategy of EfficientNet-B0 enabled effective feature extraction from time-frequency representations despite the limited dataset size. To enhance practical utility, the classification output is integrated with GPT-4 Omni, which generates contextual information about the identified species, including ecological characteristics and conservation status. This integration transforms the system from a standalone classifier into a field-support tool for researchers, conservationists, and birdwatchers. The proposed framework enables non-invasive, real-time species identification using only acoustic signals, eliminating the need for visual confirmation or close proximity. This approach supports early detection, population monitoring, and biodiversity conservation while minimizing disturbance to natural bird behavior.

1. INTRODUCTION

Everyone on Earth, including animals, plants, and birds, is equally significant in preserving ecological balance and deserves the opportunity for survival. As humans, we commit to ensuring that all species, particularly rare and endangered ones, survive and flourish. Birds serve a significant part in our ecology by pollinating, dispersing seeds, controlling pests, and ensuring the balance of nature. However, many rare bird species are experiencing rapid declines as a result of environmental damage, climate change, and human interference [1]. Tracking and recognizing these rare species are essential for their continued existence. Conventional recognition techniques, such as visual inspection or photography, present substantial obstacles. Many extremely rare birds are on their way to destruction [2]. Early recognition might help in the implementation of timely conservation efforts. Rare bird sightings contribute to biodiversity databases, allowing researchers to study species distribution, population trends, and migration patterns. The rare spotted bird can be reported to renowned conservation organizations such as the Bombay Natural History Society (BNHS) and eBird India so that they can confirm the sighting, update biodiversity records, and take necessary conservation actions such as habitat protection, population monitoring, or species recovery programs [3]. Conventional spotting methods, such as observing birds visually or snapping pictures, face significant difficulties.

- Many birds, particularly rare ones, are cautious and easily startled, which renders getting close enough for photography or direct observation challenging or even impossible [4].
- Birds might exist in dense forests, lofty trees, or isolated wetlands, making them challenging to observe visually [5].
- Certain species are only active at night or in the early morning hours, making recognizing them visually impossible without specialized equipment.
- Birds may wander rapidly across branches or fly away unexpectedly, minimizing the time accessible for observation [6].
- Telephoto lenses or binoculars are often needed for effective visual identification; however, these may not always be easily accessible [7].
- Without the assistance of technology, human error or a lack of expertise can result in inaccurate identification.
- Observing birds closely can harm their eggs or habitats, raising ethical questions about traditional bird watching [8].

Even though the collection only includes 436 audio samples from ten different species, this restriction represents a real-world conservation issue where recordings of uncommon or little-studied bird species are naturally hard for researchers to come by. Therefore, investigating a deep-learning framework that can operate well in low-resource environments is the aim of this study.

A collected dataset of bird sound recordings for ten distinct bird species was employed in this study to create and assess the proposed identification framework. The audio recordings were transformed into spectrograms, enabling these image-based models to learn the unique frequency patterns associated with each species of bird. For categorizing these species based on their vocalizations, several deep learning architectures were implemented and evaluated. These involve Convolutional Neural Networks (CNN) for baseline performance, as well as complex pre-trained models like VGG-19, AlexNet, ResNet, and EfficientNet-B0, which are well-known for their robustness in image and spectrogram categorization. EfficientNet-B0 beat the other architectures assessed in terms of classification accuracy and robustness, proving its ability to capture the intricate patterns of bird vocalizations from spectrogram images. The suggested framework enables bird investigators to effectively detect and track bird species entirely on vocalizations, eliminating the need for close proximity or visual verification, leading to non-intrusive assessment that doesn't disturb birds in the environment they inhabit. Although there is restricted availability of labelled bird sound data, the proposed approach reveals promising results by effectively applying deep learning models trained on spectrogram representations, making it feasible even in circumstances with limited data availability.

2. LITERATURE SURVEY

The deep learning framework for bird sound identification presented by Indumathi et al. is for identifying and sorting bird species from their audio fingerprints. The research highlights the crucialness of converting unprocessed bird audio into spectrograms so that the model can discover significant time-frequency patterns linked to various species. To guarantee constant input quality, a curated dataset of labeled bird calls is processed using normalization, Mel-Frequency Cepstral Coefficients (MFCC) extraction, and spectrogram creation. The effects of various CNN configurations, such as differences in convolutional depth, pooling techniques, and kernel counts on accuracy, training cost, and overfitting susceptibility, are methodically assessed. To improve generalization, methods including cross-validation, max-pooling, and regularization are used. In order to capture temporal dynamics in bird vocalizations, the study also investigates the usage of LSTM networks in conjunction with CNNs. The suggested approach yields about 75% prediction accuracy across different amplitudes, epoch settings, and data divisions, according to experimental results. PCA-based dimensionality reduction shortened training times, but it had no discernible effect on identification accuracy [9].

Yang et al. [10] offer an automatic bird category detection model based on picture data enhanced with feature augmentation and contrastive learning approaches. The CUB 200 2011 dataset, which contains 11,788 bird images for 200 species, was utilized to access the framework. The methodology implements a ShuffleNetV2 for efficiency, and the augmentation applies to capture fine-grained and global image details using a multi-scale feature fusion module. To suppress noise and occlusions, the study incorporates an attention-based feature enhancement module. To better distinguish between similar species, the Siamese network for contrastive learning was implemented. The model outperformed existing fine-grained classification algorithms, with a high accuracy of 91.3% and an F1-score of 90.6%. It also maintained superior performance even with limited training data, attaining 65.2% accuracy with only 5% of the training set. The performance was only accessed on the CUB-200 dataset, and the influence of high picture occlusion or environmental noise on classification resilience was not clearly documented.

To address the complexity of ecological monitoring, Wang et al. [11] present a lightweight fine-grained bird classification framework. Using the CUB-200-2011 dataset of 11,788 pictures from 200 species. To compress the model, the method combines attention-guided data augmentation, which crops object and key part regions based on activation maps for improved discriminative feature learning, with decoupled knowledge distillation (DKD). The teacher model DenseNet121 directs the student model ShuffleNetV2, keeping pertinent information with 67% fewer parameters and only 1.2 GFLOPs, achieving 87.6% classification accuracy, closely mirroring the teacher's 89.5% accuracy. Ablation examinations reveal significant improvement in performance offered by attention-guided augmentation, which is a 42% gain over the baseline, and localization-recognition modules, a 5.8% increase.

Xie and Zhu [12] proposed a deep learning-based framework for bird species categorization using acoustic data, with an emphasis on early fusion of deep features. The CLO-43DS dataset consists of flight call recordings from forty-three North American wood-warbler species utilized for research. To address the heterogeneity in recording environments and enhance model resilience, audio signals are transformed into Mel-spectrograms, and data augmentation techniques are used, including mixup and pitch shift. Deep features were extracted using five pretrained CNN architectures: VGG16, ResNet50, EfficientNetB0, MobileNetV2, and Xception, which were then concatenated using an early fusion technique and classified with a linear SVM. Among the models

examined, the fusion of VGG16 and MobileNetV2 obtained the best-balanced accuracy of 94.89%, outperforming individual models. As the system performed well, limitations such as class imbalance and poor generalization in noisy or unfamiliar contexts were acknowledged. The proposed method has potential for non-intrusive bird monitoring and automated classification in ecological studies.

Wang et al. [13] presented a deep learning framework that recognizes an extensive variety of bird species using audio file inputs. The study used a dataset of over 70,000 bird-call audio samples for 264 species from Xeno-Canto; the researchers used a combination of Mel-spectrogram and MFCC as input features. These features were processed using an LSTM network enhanced with coordinate attention mechanisms, achieving a mAP value of 77.43%, indicating its potential to handle large-scale, multi-class bird species identification work. The study also emphasized the significance of feature selection and normalization, stating that integrating normalized Mel-spectrograms and MFCCs increased accuracy by almost 4%, reaching 74.94%. Comparative assessments with classic machine learning models like SVM and RF revealed that the suggested deep learning approach outperformed them. and demonstrates the potential of complex neural network designs for large-scale avian biodiversity monitoring using passive acoustic approaches.

Revadekar et al. [14] conducted an extensive evaluation of deep learning models for bird acoustic classification, using the BirdCLEF 2022 dataset from Xeno-Canto. The study evaluated the performance of five pre-trained CNN architectures, such as Xception, InceptionV3, ResNet50, EfficientNet, and VGG16, with two custom-designed CNN models. The input features were derived from MFCCs and spectrograms, with bird sounds treated as pictures for visual pattern recognition using CNNs. The first custom CNN model comprised three convolutional layers with ReLU activation, followed by max-pooling and dropout layers to prevent overfitting, and finally fully connected layers for categorization. The second custom model includes batch normalizing layers for faster convergence and stability. These architectures were specifically created for the acoustic characteristics of bird calls and tuned for a small number of parameters to allow for deployment on edge devices. Custom models beat usual pre-trained models in some circumstances, with accuracies of 80.11% and 76.94%, respectively. The study reveals that with the correct layout, lightweight custom CNNs may accurately distinguish bird species from audio and act as efficient instruments for bioacoustics monitoring in real-time field situations.

Noumida et al. [15] implemented a hierarchical attention-based bidirectional GRU model for classifying multi-label bird species from field acoustic recordings. The study makes use of the BirdCLEF 2021 dataset, comprising 80,000 audio recordings encompassing 397 bird species. The model takes MFCC as input features to capture the exact spectral aspects of bird vocalizations. The BiGRU architecture accurately represents temporal connections in bird songs, and the attention mechanism points out essential audio parts for each species. The final multi-label predictions were produced using a short-time aggregation strategy. The approach proposed outperformed baseline CNN and RNN-based models, with a mean LRAP of 0.654 and an mAP of 0.537. The framework illustrates potential for non-intrusive and expandable bird species identification in natural habitats, but the system's performance declines in recordings with multiple calling species and high environmental noise. The model requires significant computational resources during inference.

AMResNet, an autonomous bird-sound recognition model constructed to tackle the challenges of species identification in complex acoustic environments, is presented by Xiao et al. [16]. The BirdCLEF dataset, which comprises tens of thousands of annotated bird audio recordings spanning hundreds of species, is used in the study to provide a broad but extremely unbalanced collection of samples. The model can learn discriminative time-frequency representations of bird calls by converting each audio clip into a log-mel spectrogram. In order to effectively capture small auditory differences between species, AMResNet integrates residual learning with attention techniques. The model outperforms conventional CNN architectures on the same dataset, with an overall accuracy of about 85%, according to experimental evaluation. The model's efficacy declines for species with few records, and background noise in real-world acoustic data continues to be an issue, the researchers point out. Despite these limitations, AMResNet shows great promise for automated, large-scale biodiversity monitoring and provides a good foundation for further advancements in bird-sound recognition.

Based on an evaluation of recent work in the field of bird species classification using deep learning, the main observation is that picture-based categorization of bird species, which is extensively employed, poses significant limitations, especially in real-world ecological contexts. Capturing clear pictures of birds is typically difficult because of their constant movement, distance, and desire to escape when approached. Rare or endangered birds are often elusive and live in dense or inaccessible surroundings, making photographing them very impossible without causing disturbance. Furthermore, similar-looking species can lead to incorrect categorization due to slight visual variations that are difficult to detect in varying lighting or background conditions. These restrictions reduce the efficiency of image-based systems in practical scenarios, emphasizing the necessity for noninvasive alternatives to species identification, such as sound-based classification, that do not need eye contact or disturb the species' natural activity.

3. METHODOLOGY

The process for identifying bird species is illustrated in Figure 1.

In the present study, we offer an automated approach for classifying bird species using acoustic data. The iBC53 dataset from Kaggle is used for model training and testing. We preprocess each audio clip by clipping it to isolate the relevant portion and then convert it into spectrograms to extract time-frequency information. During the preprocessing stage, data augmentation and normalization approaches are employed to enhance model resilience. In order to simulate environmental disturbances, additive Gaussian noise was introduced with a variance range of 0.001–0.01. Pitch alteration was conducted within ±2 semitones, while time-shifting was done at random within ±0.3 seconds to increase variability while maintaining natural sound qualities. In order to ensure that the target bird call remained detected while adding real variability, Gaussian white noise and environmental noise (wind, gentle chirps, and forest ambiance) were introduced at signal-to-noise ratios (SNR) between 15 dB and 30 dB. Several deep learning models, including CNN, AlexNet, VGG-19, ResNet-50, and

EfficientNet-B0, are trained on the processed data. Each model's performance has been assessed using precision, recall, accuracy, and the F1-score, and the model with the best performance across all of these metrics is selected. The framework is constructed as an assistant to travellers visiting jungle or woodland areas, which allows them to do real-time bird species identification from recorded bird calls. The recorded bird call is pre-processed and transformed into a spectrogram for classification. Once a bird call is captured and classified, the identified species name is forwarded to GPT-4 Omni to deliver additional details such as information regarding the identified species, the rarity of the species, the population count, and other relevant data to enhance the user's knowledge and experience.

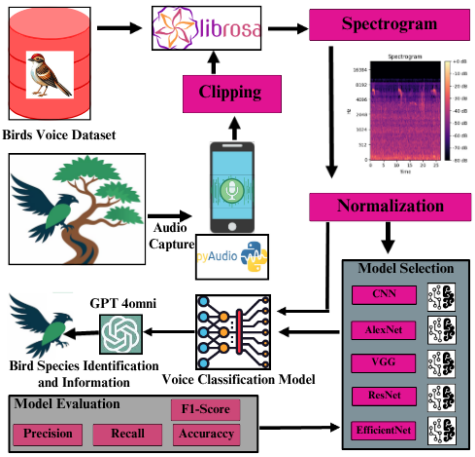
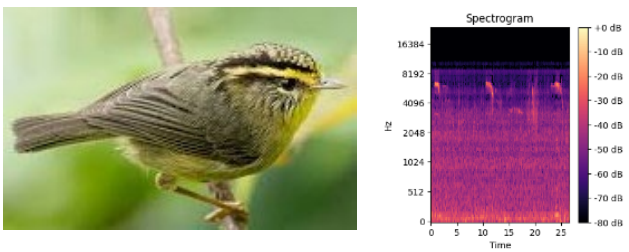


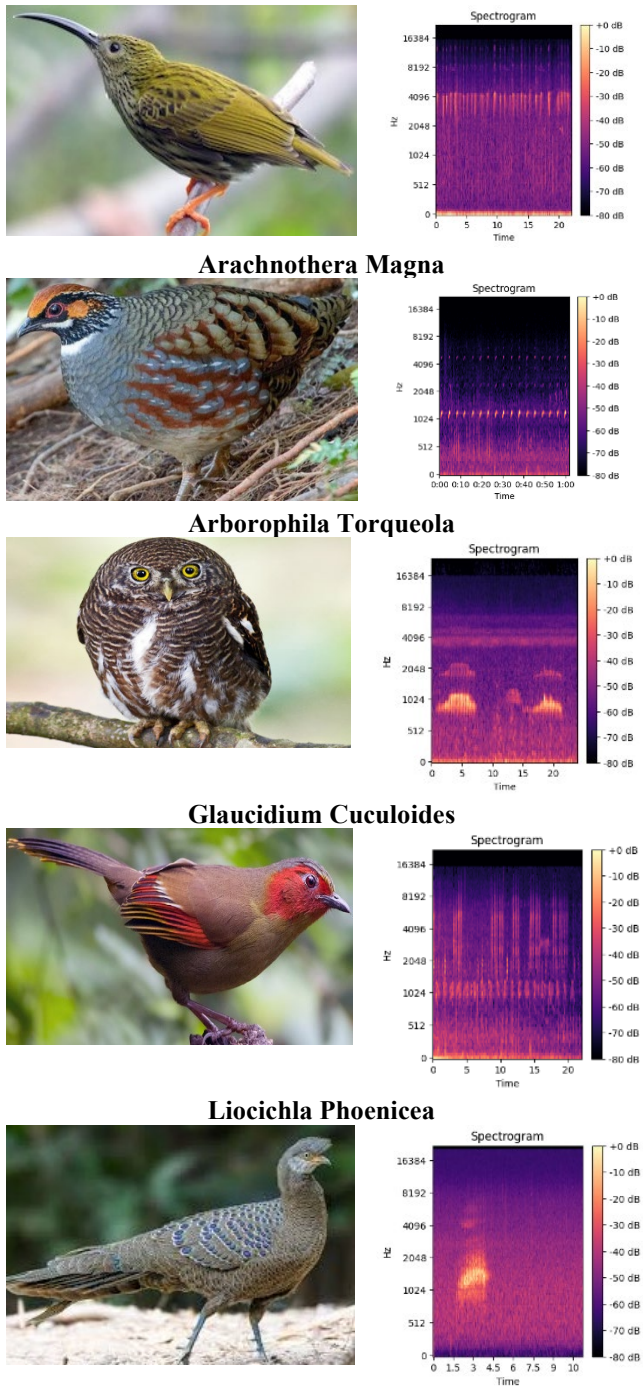
Figure 1. Architecture of the proposed bird species identification system

3.1 Dataset

The framework makes use of audio recordings of bird calls from ten distinct bird species collected from the iBC53 dataset on Kaggle [17], converting YouTube bird videos to audio clips and the Xeno-Canto Library [18]. In order to prepare the data for model training, each audio clip was pre-processed by clipping to remove extraneous sections. The raw audio was then transformed into a time-frequency domain using spectrogram representations, allowing relevant features to be extracted for categorization. Data augmentation techniques such as noise addition, time-shifting, and pitch modifications were used to artificially expand the dataset and improve model generalization. Finally, normalization was used to scale the spectrogram values and ensure uniform input across all models. This comprehensive dataset allowed the training and assessment of models, ensuring a thorough comparative analysis of model performance. The representation of bird vocalizations through spectrograms, along with the corresponding bird images and names, is presented in Figure 2.



Alcippe Cinerea



Arachnothera Magna

Arborophila Torqueola

Glaucidium Cuculoides

Liocichla Phoenicea

Polyplectron Bicalcara

Figure 2. Visual and acoustic representation of rare bird species

3.2 Feature extraction using spectrogram

The raw audio signal collected from bird vocalizations must be transformed into an appropriate format in order to classify bird species by deep learning models. Spectrograms serve as an effective 2D time-frequency representation of a bird's call 1D audio waveform. It allows the model to incorporate both frequency and temporal variations found in different bird species. A spectrogram is a two-dimensional graph with a third dimension represented by colour. The horizontal X-axis represents time, which moves from left to right, while the vertical Y-axis represents frequency, which moves from low at the bottom to high at the top [19]. In the spectrogram representation, the darker or deeper colours specify high

amplitude or strong energy at specific frequencies, corresponding to loud, prominent bird calls with stronger frequency components. The fainter or lighter colours represent low amplitude or weak energy, corresponding to softer, quieter sounds or background noise [20]. The audio stream x(t), where t represents time, is first separated into simple overlapping frames through windowing functions, usually Hamming or Hann windows. The Short-Time Fourier Transform (STFT) serves as a foundation for building spectrograms and is used on each frame to determine the frequency components within that window. The spectrogram S(t, f) can be mathematically defined as below in Eq. (1) [21, 22].

$$S(t,f) = \left| \sum_{n=-\infty}^{\infty} x(n)\omega(n-t)e^{-j2\pi fn} \right|^2 \quad (1)$$

where, $x(n)$ represents the input audio signal, $w(n)$ represents the window function applied, f represents frequency, $e^{-j2\pi fn}$ represents the Fourier kernel, and $|.|^2$ represents the power spectrum [23]. The STFT converts each time segment into its corresponding frequency components represented in Eq. (2).

$$STFT\{x(t)\}(t,f) = \int_{-\infty}^{\infty} x(\tau)\omega(\tau-t)e^{-j2\pi f\tau}\, d\tau \quad (2)$$

For better feature representation, often the logarithmic scale is applied, as represented in Eq. (3).

$$S_{log}(\mathrm{t},\mathrm{f}) = log\big(1 + \mathrm{S}(\mathrm{t},\mathrm{f})\big) \quad (3)$$

3.3 EfficientNet-B0

The study implements EfficientNet-B0 architecture for successfully categorizing bird species using spectrogram representations of bird calls. EfficientNet-B0, proposed by Google AI in 2019, employs a revolutionary compound scaling approach to precisely scale network depth, width, and resolution with a correctly weighed set of coefficients. In contrast to typical deep networks, which simply add more layers or filters, EfficientNet optimizes the use of computational resources, resulting in a lightweight but highly accurate model. The EfficientNet-B0 version, the simplest and best-performing type, has around 5.3 million parameters, making it ideal for tasks where data availability is low. Training deep neural networks on short datasets can be problematic due to the possibility of overfitting and poor generalization. EfficientNet-B0 proves extremely beneficial in such cases due to its lightweight architecture, transfer learning capability, and efficient feature extraction [24]. This framework offers good classification performance even with low training sets through combining spectrogram-based feature extraction with the efficiency of EfficientNet-B0, demonstrating its usefulness in identifying bird species from audio data. Table 1 presents EfficientNet-B0 architecture details such as Block, Input-Output Size, Filter, Kernel, Stride, etc.

Table 1. EfficientNet-B0 architecture details

Stage	Layer Name/ Block	Input Size	Output Size	Filter/ Kernel	Stride
1	Conv3 × 3	224 × 224 × 3	112 × 112 × 32	3 × 3 Conv, 32 filters	2
2	MBConv1	112 × 112 × 32	112 × 112 × 16	3 × 3 DWConv, 16 filters	1
3	MBConv6	112 × 112 × 16	56 × 56 × 24	3 × 3 DWConv, 24 filters	2
4	MBConv6	56 × 56 × 24	28 × 28 × 40	5 × 5 DWConv, 40 filters	2
5	MBConv6	28 × 28 × 40	14 × 14 × 80	3 × 3 DWConv, 80 filters	2
6	MBConv6	14 × 14 × 80	14 × 14 × 112	5 × 5 DWConv, 112 filters	1
7	MBConv6	14 × 14 × 112	7 × 7 × 192	5 × 5 DWConv, 192 filters	2
8	MBConv6	7 × 7 × 192	7 × 7 × 320	3 × 3 DWConv, 320 filters	1
9	Conv1x1	7 × 7 × 320	7 × 7 × 1280	1 × 1 Conv, 1280 filters	1
10	Pool & FC	7 × 7 × 1280	1 × 1 × 1280 → FC → Output Classes	Global AvgPool + Dense Layer	-

Compound Scaling: Traditional scaling approaches affect only one dimension of a network (depth, width, or resolution). EfficientNet employs a compound coefficient (ϕ) to scale all dimensions concurrently, as described in Eq. (4). The constants α, β, and γ are chosen using a small grid search, whereas ϕ determines the additional resources available for scaling [25].

$$\begin{aligned}\text{Depth (d)} = \alpha^{\phi}, Width\ (w) = \beta^{\phi}, \text{Resolution} \\ = (\mathrm{r}) = \gamma^{\phi}\end{aligned} \quad (4)$$

MBConv Block: The MobileNetV2 architecture serves as the foundation for the Mobile Inverted Bottleneck Convolution (MBConv), which significantly enhances accuracy and computational performance. The MBConv block consists of an expansion phase that expands feature dimensions, depthwise convolution that collects spatial features effectively, a squeeze-and-excitation (SE) module that recalibrates channel-wise importance, a projection phase that decreases dimensions back to the original size, and an optional skip connection. The complete operation of the MBConv block is stated in Eq. (5) [26].

$$\begin{aligned}Y_{out} = X + BN + \Bigg(\Big(\Big(SE\Big(BN\Big(Swish\Big(BN(X \\ * W_{exp})\Big) * W_{dw}\Big)\Big)\Big) * W_{proj}\Bigg)\end{aligned} \quad (5)$$

where, W_{exp}, W_{dw}, and W_{proj} indicate the expansion, depthwise, and projection weights, respectively. Swish is the nonlinear activation function, *BN* represents batch normalization, and *SE* represents the squeeze-and-excite module.

Squeeze-and-Excitation (SE) Block: The SE block is an attention mechanism incorporated into the MBConv structure that extracts the beneficial features while suppressing unimportant ones, which leads to increasing the network's representational capacity and requiring low computational

expense. During the squeeze phase, global spatial information from each channel is pooled using global average pooling, transforming the feature map $U \in R^{HxWxC}$ into a channel descriptor $s \in R^{C}$, as illustrated below in Eq. (6) [27].

$$Z_c = \frac{1}{H\ x\ W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{i,j,c} \quad for\ c=1,2,...C \tag{6}$$

During the excitation phase, this channel descriptor moves through a gating mechanism that includes two fully connected layers with a non-linear activation (ReLU) and a sigmoid activation that generates channel-wise modulation weights represented in Eq. (7) [28].

$$s = \sigma\big(W2 \cdot \delta(W1 \cdot z)\big) \tag{7}$$

where, $W1 \in R^{c/r \times C}$, $W2 \in R^{C \times C/r}$, δ is ReLU, σ is the sigmoid activation function, and r is the reduction ratio.

Batch Normalization (BN): BN is a popular regularization and normalizing approach that improves the stability and speed of deep neural network training. It addresses the issue of internal covariate shift by normalizing each layer's input characteristics, ensuring that the distribution of activations is stable during training. Batch Normalization calculates the mini-batch mean (μ) and variance (σ^2) from an input mini-batch $B = \{x_1, x_2, \ldots, x_m\}$ for a specific feature channel, as shown below in Eq. (8) and (9) [29].

$$\mu = \frac{1}{m} \sum_{i=1}^{m} x_i \tag{8}$$

$$\sigma^2 = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu)^2 \tag{9}$$

Eq. (10) is then used to normalize each input x so it has a zero mean and unit variance. To prevent division by zero, a minor constant (ϵ) is added.

$$\widehat{X} = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \tag{10}$$

Swish activation: To boost the representational capacity of deep neural networks, a smooth, non-monotonic Swish activation function is designed. Swish allows slight negative activations for propagation, which strengthens gradient flow and model expressiveness, particularly in deep architectures like EfficientNet represented in Eq. (11).

$$Swish(x) = x \cdot \sigma(x) = \frac{x}{1 + e^{-x}} \tag{11}$$

3.4 GPT-4 Omni

GPT-4 Omni expands the system's usefulness through substantially producing contextual data about the predicted species, such as ecological characteristics, rarity indicators, and conservation relevance. Through this integration, the classifier is transformed from a stand-alone prediction model into an all-encompassing field-support tool that helps researchers, tourists, and conservation workers by offering insightful, timely interpretations. Without changing the fundamental classification results, the addition improves the system's usability and practicality.

4. RESULT AND DISCUSSION

The system for classifying bird sounds with spectrograms was set up and run on Google Colab, which uses GPU acceleration (NVIDIA T4) to speed up the training and testing of deep learning models. The required audio files and datasets have been collected and saved on Google Drive in order to offer easy integration with the Colab environment. The dataset used in this research study comprises ten bird categories with 436 sound samples partitioned into 80:10:10 ratio for training, validation, and testing, respectively. The dataset was prepared by collecting publicly available resources, including the iBC53 dataset from Kaggle, converting YouTube bird videos to audio, and using the Xeno-canto bird sound library. Audacity, a popular and efficient open-source audio editor, was used to standardize and separate the voice recordings to ensure proper clipping of significant bird sounds. To capture real-time audio input and make live predictions during deployment, the PyAudio module was used. Deep learning models to classify spectrograms of bird calls were constructed using Python's Scikit-learn and TensorFlow/Keras library. To produce extensive species descriptions based on the model's output class, OpenAI's GPT-4 Omni model was used. The input spectrogram images were scaled to 224 × 224 pixels to satisfy the model's input specifications.

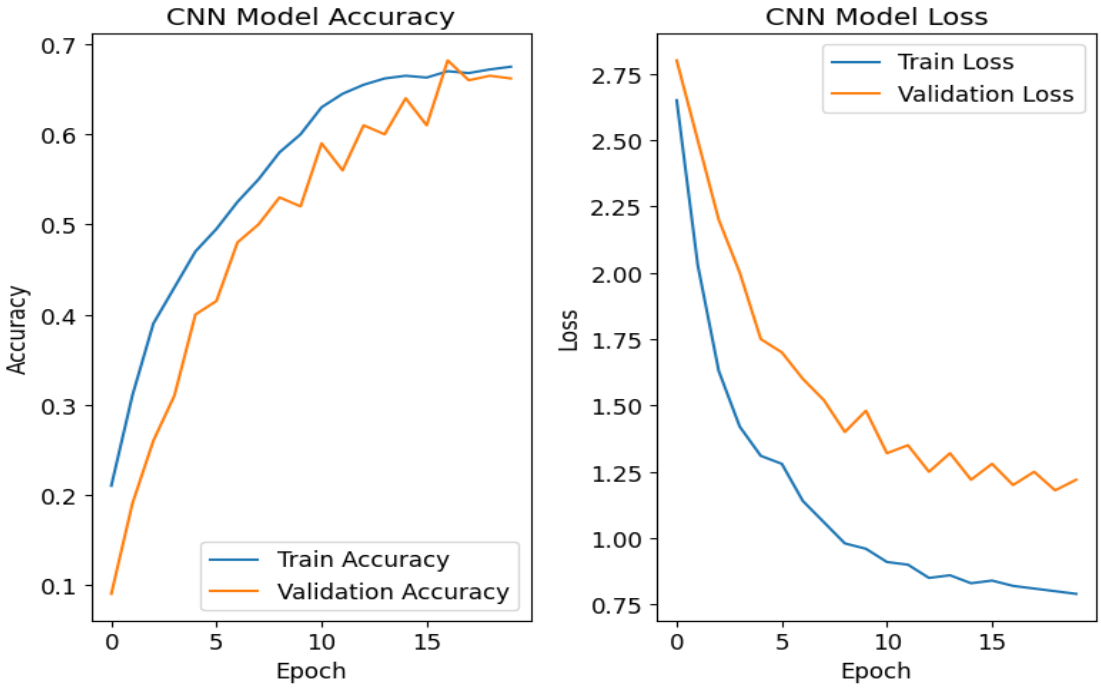
All models are trained for 20 epochs using the Adam optimizer with a learning rate of 0.0001 and a batch size of 32, and the cross-entropy loss function was used as the problem required multi-class classification. While AlexNet, VGG-19, ResNet-50, and EfficientNet-B0 adhere to their standard ImageNet-pretrained configurations with improved final layers customized for 10-class classification, the CNN architecture consists of three convolutional blocks (Conv–ReLU–MaxPooling). Model weights were updated using the Adam optimizer, which was chosen for its adaptable learning capabilities and excellent performance in deep learning applications. The softmax activation function is used as the final classification layer to generate probability distributions for each of the ten classes. Figure 3 presents an analysis comparing accuracy and loss for both training and validation phases to evaluate the performance of several models trained on bird call spectrograms.

The basic CNN model obtained a training accuracy of 0.675 and a validation accuracy of 0.662, demonstrating an insufficient ability to correctly classify bird call spectrograms. AlexNet improved with 0.78 training and 0.70 validation accuracy but did not converge after 20 epochs, demonstrating substantial variation in training and validation performance. VGG-19 performed better, with 0.79 training and 0.75 validation accuracy, indicating more potent generalization. ResNet-50 improved the results with 0.85 training and 0.82 validation accuracy, exhibiting good learning capacity. EfficientNet-B0 beat all other models, with training and validation accuracy of 0.92, showing greater convergence, minimum overfitting, and superior classification abilities on bird call spectrograms.

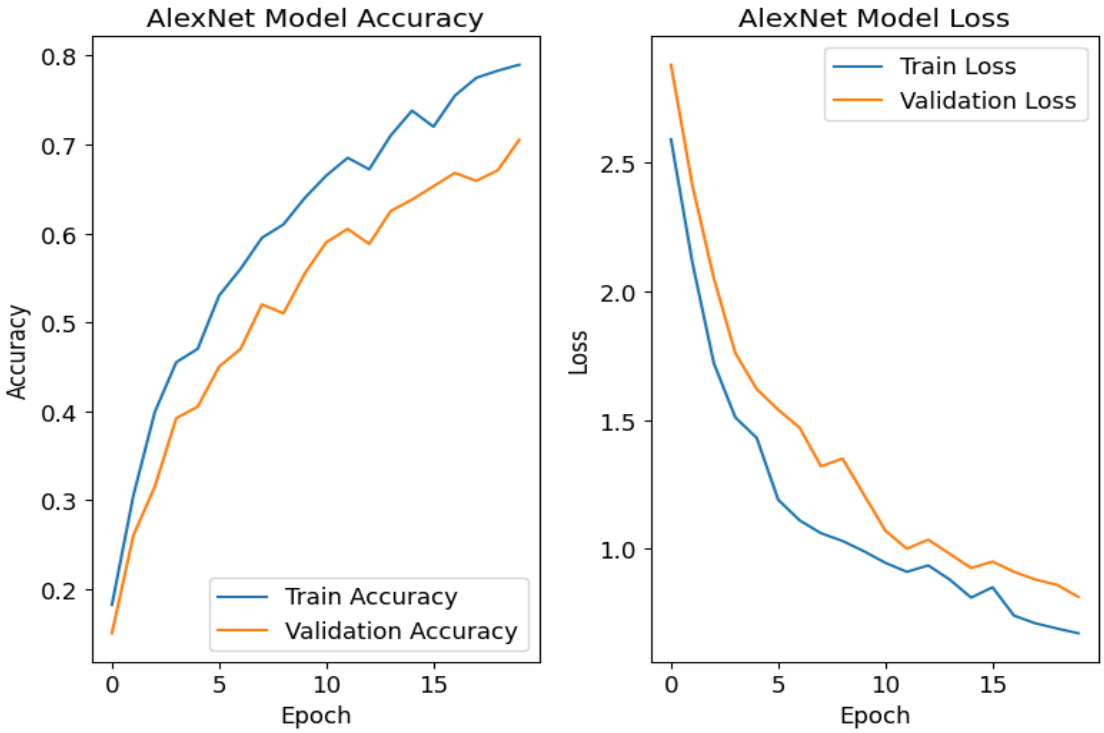
The trained models were tested on the testing set using standard performance metrics such as precision, recall, accuracy, and the F1-score. Table 2 displays the confusion matrix generated by each model for all bird species,

showcasing classification performance across multiple categories, while Table 3 shows the associated performance metrics.

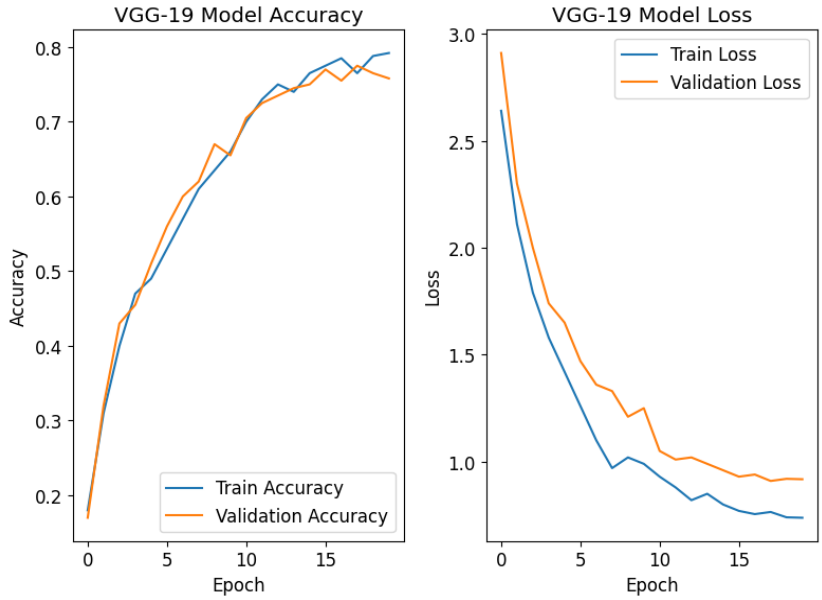
The result reveals that EfficientNet-B0 outperformed every other model with an accuracy of 0.93, precision of 0.94, recall of 0.93, and F1-score of 0.93. EfficientNet-B0's architecture, which balances model depth, width, and resolution, allows it to extract significant features more effectively from spectrogram images, even with a smaller dataset. Models such as CNN and AlexNet performed ineffectively, illustrating the difficulties of generalizing with insufficient training data. ResNet-50 and VGG-19 performed rather well, but they were unable to match EfficientNet-B0's consistency and accuracy. It demonstrates EfficientNet-B0's greater capacity to capture small audio patterns required for accurate bird call classification under data-constrained environments.



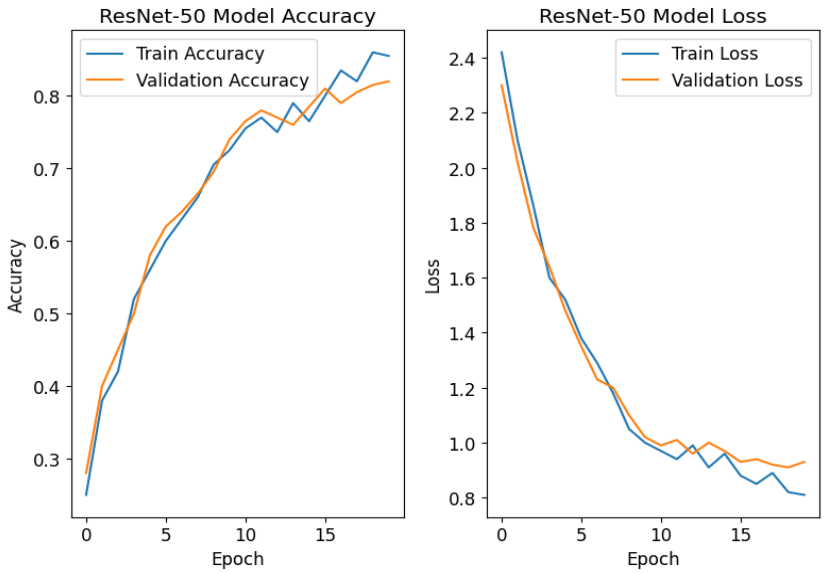
Convolutional Neural Networks (CNN)



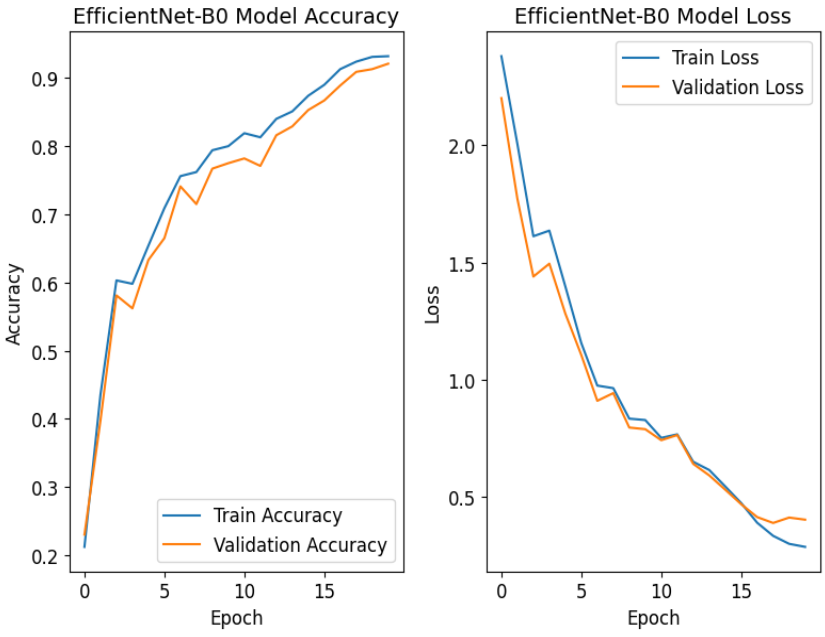
AlexNet



VGG-19



ResNet-50



EfficientNet-B0

Figure 3. Training, validation accuracy, and loss curves for different models

Table 2. Model-wise confusion matrix analysis for test image classification

	Alcippe Cinerea	Arachnothera Magna	Arborophila Torqueola	Cuculus Micropterus	Dicrurus Andamanensis	Glaucidium Cuculoides	Todiramphus Chloris	Liocichla Phoenicea	Pellorneum Ruficeps	Polyplectron Bicalcara	
Alcippe Cinerea	4	1	1	0	0	0	0	0	0	0	
Arachnothera Magna	1	3	0	0	0	0	0	0	0	0	
Arborophila Torqueola	0	1	5	1	0	1	0	0	0	0	
Cuculus Micropterus	1	0	0	10	0	1	0	2	0	0	
Dicrurus Andamanensis	1	0	0	0	3	0	0	0	0	0	**Convolutional Neural Networks (CNN)**
Glaucidium Cuculoides	0	1	2	0	0	7	0	0	0	0	
Todiramphus Chloris	0	0	0	0	0	0	5	1	0	0	
Liocichla Phoenicea	0	0	1	0	0	0	1	4	0	0	
Pellorneum Ruficeps	0	0	1	0	0	1	3	1	13	1	
Polyplectron Bicalcara	0	0	0	0	0	0	0	0	1	2	
Alcippe Cinerea	5	1	0	0	0	0	0	0	0	0	
Arachnothera Magna	1	3	0	0	0	0	0	0	0	0	
Arborophila Torqueola	0	1	5	1	1	0	0	0	0	0	
Cuculus Micropterus	1	0	0	10	0	1	1	1	0	0	
Dicrurus Andamanensis	1	0	0	0	3	0	0	0	0	0	**AlexNet**
Glaucidium Cuculoides	0	1	1	0	0	8	0	0	0	0	
Todiramphus Chloris	0	0	0	0	0	0	5	1	0	0	
Liocichla Phoenicea	0	0	0	0	0	0	1	5	0	0	
Pellorneum Ruficeps	0	0	1	0	1	3	1	0	12	2	
Polyplectron Bicalcara	0	0	0	0	0	0	0	0	1	2	
Alcippe Cinerea	6	0	0	0	0	0	0	0	0	0	
Arachnothera Magna	0	3	1	0	0	0	0	0	0	0	
Arborophila Torqueola	0	1	5	1	0	1	0	0	0	0	
Cuculus Micropterus	0	0	0	10	1	2	1	0	0	0	
Dicrurus Andamanensis	0	1	0	0	3	0	0	0	0	0	**VGG-19**
Glaucidium Cuculoides	0	1	1	0	0	8	0	0	0	0	
Todiramphus Chloris	0	0	0	0	0	0	5	1	0	0	
Liocichla Phoenicea	0	0	0	0	0	0	1	5	0	0	
Pellorneum Ruficeps	0	1	0	0	0	2	1	0	15	1	
Polyplectron Bicalcara	0	0	0	0	0	0	0	0	1	2	
Alcippe Cinerea	6	0	0	0	0	0	0	0	0	0	
Arachnothera Magna	0	3	0	1	0	0	0	0	0	0	
Arborophila Torqueola	0	0	6	1	1	0	0	0	0	0	
Cuculus Micropterus	0	0	0	11	1	2	0	0	0	0	
Dicrurus Andamanensis	0	0	0	0	4	0	0	0	0	0	**ResNet-50**
Glaucidium Cuculoides	0	1	0	1	0	8	0	0	0	0	
Todiramphus Chloris	0	0	0	0	0	0	5	1	0	0	
Liocichla Phoenicea	0	0	0	0	0	0	0	5	1	0	
Pellorneum Ruficeps	0	1	0	0	0	2	0	0	16	1	
Polyplectron Bicalcara	0	0	0	0	0	0	0	0	0	3	
Alcippe Cinerea	5	1	0	0	0	0	0	0	0	0	
Arachnothera Magna	0	4	0	0	0	0	0	0	0	0	
Arborophila Torqueola	0	0	8	0	0	0	0	0	0	0	
Cuculus Micropterus	0	0	0	14	0	0	0	0	0	0	
Dicrurus Andamanensis	1	0	0	0	3	0	0	0	0	0	**EfficientNet-B0**
Glaucidium Cuculoides	0	0	0	0	0	10	0	0	0	0	
Todiramphus Chloris	0	0	0	0	0	0	5	1	0	0	
Liocichla Phoenicea	0	0	0	0	0	0	1	5	0	0	
Pellorneum Ruficeps	0	0	0	0	0	0	1	1	15	0	
Polyplectron Bicalcara	0	0	0	0	0	0	0	0	0	3	

Table 3. Performance evaluation metrics for test image classification

Model	Accuracy	Precision	Recall	F1-Score
CNN [30]	0.69	0.74	0.69	0.70
AlexNet [31]	0.72	0.76	0.72	0.72
VGG-19 [32]	0.77	0.80	0.77	0.77
ResNet-50 [33]	0.83	0.85	0.83	0.83
EfficientNet-B0	0.93	0.94	0.93	0.93

5. CONCLUSIONS

Identifying bird species based on their sounds is a difficult task, especially in environments such as forests or preserves for wildlife where visual confirmation is not always attainable. Manual identification needs professional knowledge of ornithology and may not be beneficial in situations that are immediate. Furthermore, the absence of efficient, automated

systems for classifying bird calls restricts biodiversity monitoring, species conservation efforts, and public connection with the environment. An automated approach based on deep learning models and spectrogram representations of bird calls may be beneficial to scholars, travelers, and conservationists. The technology not only improves species identification accuracy, but it also raises awareness and assists ecological study by making species data available and interpretable in real time. This study describes a robust and automated approach to distinguishing bird species using auditory inputs. The strategy achieves successful feature extraction from minimal data by transforming audio files into spectrograms and utilizing preprocessing techniques such as clipping, augmentation, and normalization. Several deep learning models have been evaluated, and EfficientNet-B0 was revealed as the most reliable and precise for bird call classification. This model's incorporation into a real-time assist system is beneficial to tourists and nature lovers, as it facilitates immediate recognition of bird species. GPT-4 Omni additionally improves users' knowledge, including species rarity and population information, making the system both instructive and user-friendly for field applications. Future study might involve widening the dataset by adding more bird species and geographic locations in order to boost model generality. Integrating methods for noise reduction and customizing the model to real-world environmental sounds can improve its performance in natural environments. Furthermore, integrating the system with conservation databases may assist in keeping track of biodiversity and endangered species tracking, thus supporting ecological preservation efforts.

REFERENCES

[1] Shuai, H., Hu, J., Zheng, S., Ma, Z., Liu, J. (2024). Most bird species remain poorly studied but threatened status promotes research effort. Avian Research, 15: 100215. https://doi.org/10.1016/j.avrs.2024.100215

[2] Pollock, H.S., Toms, J.D., Tarwater, C.E., Benson, T.J., Karr, J.R., Brawn, J.D. (2022). Long-term monitoring reveals widespread and severe declines of understory birds in a protected Neotropical Forest. Proceedings of the National Academy of Sciences of the United States of America, 119(16): e2108731119. https://doi.org/10.1073/pnas.2108731119

[3] Şekercioğlu, Ç.H., Daily, G.C., Ehrlich, P.R. (2004). Ecosystem consequences of bird declines. Proceedings of the National Academy of Sciences of the United States of America, 101(52): 18042-18047. https://doi.org/10.1073/pnas.0408049101

[4] Thulin, C.G., Röcklinsberg, H. (2020). Ethical considerations for wildlife reintroductions and rewilding. Frontiers in Veterinary Science, 7: 163. https://doi.org/10.3389/fvets.2020.00163

[5] Lotfian, M., Ingensand, J., Brovelli, M.A. (2021). The partnership of citizen science and machine learning: Benefits, risks, and future challenges for engagement, data collection, and data quality. Sustainability, 13(14): 8087. https://doi.org/10.3390/su13148087

[6] Wilson, J.P., Amano, T., Fuller, R.A. (2023). Drone induced flight initiation distances for shorebirds in mixed species flocks. Journal of Applied Ecology, 60(9): 1816-1827. https://doi.org/10.1111/1365-2664.14467

[7] Cantu de Leija, A., Mirzadi, R.E., Randall, J.M., Portmann, M.D., Mueller, E.J., Gawlik, D.E. (2023). A meta-analysis of disturbance caused by drones on nesting birds. Journal of Field Ornithology, 94(2): 3. https://doi.org/10.5751/JFO-00259-940203

[8] Husby, M. (2025). Recommendations on how to use flight initiation distance data in birds. Biology (Basel), 14(4): 329. https://doi.org/10.3390/biology14040329

[9] Indumathi, C.P., Diviyalakshmi, K.R., Mahalakshmi, R. (2024). Bird sound identification system using deep learning. Procedia Computer Science, 233: 597-603. https://doi.org/10.1016/j.procs.2024.03.249

[10] Yang, Y., Zhang, Y., Liu, Z., Wang, H. (2024). Automatic bird species recognition from images with feature enhancement and contrastive learning. Ecological Informatics, 78: 102365. https://doi.org/10.1016/j.ecoinf.2023.102365

[11] Wang, K., Yang, F., Chen, Z., Chen, Y., Zhang, Y. (2023). A fine-grained bird classification method based on attention and decoupled knowledge distillation. Animals, 13(2): 264. https://doi.org/10.3390/ani13020264

[12] Xie, J., Zhu, M. (2023). Acoustic classification of bird species using an early fusion of deep features. Birds, 4(1): 138-147. https://doi.org/10.3390/birds4010011

[13] Wang, H., Xu, Y., Yu, Y., Lin, Y., Ran, J. (2022). An efficient model for a vast number of bird species identification based on acoustic features. Animals, 12(18): 2434. https://doi.org/10.3390/ani12182434

[14] Revadekar, S., Panchal, V., Kanani, P., Shah, K., Vasoya, A., Pandey, R. (2023). Bird sound classification using deep neural networks: A comparative analysis of State-of-the-Art models and custom architectures. International Journal of Intelligent Systems and Applications in Engineering, 11(4): 614-622. https://ijisae.org/index.php/IJISAE/article/view/3596

[15] Noumida, A., Rajan, R. (2022). Multi-label bird species classification from audio recordings using attention framework. Applied Acoustics, 197: 108901. https://doi.org/10.1016/j.apacoust.2022.108901

[16] Xiao, H., Liu, D., Chen, K., Zhu, M. (2022). AMResNet: An automatic recognition model of bird sounds in real environment. Applied Acoustics, 201: 109121. https://doi.org/10.1016/j.apacoust.2022.109121

[17] Sahoo, A. (2024). iBC53 – Indian Bird Call Dataset. Kaggle. https://www.kaggle.com/datasets/arghyasahoo/ibc53-indian-bird-call-dataset, accessed on Jun. 02, 2025.

[18] Xeno-Canto Foundation. (n.d.). Xeno-Canto: Bird sounds from around the world. Xeno-Canto. https://xeno-canto.org/, accessed on Jun. 02, 2025.

[19] Foggia, P., Greco, A., Roberto, A., Saggese, A., Vento, M. (2023). Degramnet: Effective audio analysis based on a fully learnable time–frequency representation. Neural Computing and Applications, 35(27): 20207-20219. https://doi.org/10.1007/s00521-023-08849-7

[20] Takahashi, K., Shiraishi, T. (2025). Speech separation in time–frequency domain by deep learning with high performance and reducing parameters. Journal of Vibration Engineering & Technologies, 13: 57. https://doi.org/10.1007/s42417-024-01565-z

[21] Korade, N.B., Salunke, M.B., Bhosle, A.A., Sangve, S.M., Joshi, D.M., Asalkar, G.G., Kadu, S.R., Sarwade, J.M. (2025). Intelligent guitar chord recognition using

spectrogram-based feature extraction and AlexNet architecture for categorization. International Journal of Advanced Computer Science & Applications, 16(4): 758-766. https://doi.org/10.14569/IJACSA.2025.0160475

[22] Nadar, S., Gandhi, D., Jawale, A., Pawar, S., Prabhu, R. (2025). Multimodal audio violence detection: Fusion of acoustic signals and semantics. Acadlore Transactions on AI and Machine Learning, 4(4): 301-311. https://doi.org/10.56578/ataiml040405

[23] Levy, J., Naitsat, A., Zeevi, Y.Y. (2022). Classification of audio signals using spectrogram surfaces and extrinsic distortion measures. EURASIP Journal on Advances in Signal Processing, 2022(100): 1-23. https://doi.org/10.1186/s13634-022-00933-9

[24] Bai, K., Zhang, Z., Jin, S., Dai, S. (2025). Rock image classification based on improved EfficientNet. Scientific Reports, 15: 18683. https://doi.org/10.1038/s41598-025-03706-0

[25] Saadoon, Y.A., Khalil, M., Battikh, D. (2025). Predicting epileptic seizures using EfficientNet-B0 and SVMs: A deep learning methodology for EEG analysis. Bioengineering, 12(2): 109. https://doi.org/10.3390/bioengineering12020109

[26] Alruwaili, M., Mohamed, M. (2025). An integrated deep learning model with EfficientNet and ResNet for accurate multi-class skin disease classification. Diagnostics, 15(5): 551. https://doi.org/10.3390/diagnostics15050551

[27] Lilhore, U.K., Sharma, Y.K., Shukla, B.K., Vadlamudi, M.N., Simaiya, S., Alroobaea, R., Alsafyani, M., Baqasah, A.M. (2025). Hybrid convolutional neural network and Bi-LSTM model with EfficientNet-B0 for high-accuracy breast cancer detection and classification. Scientific Reports, 15: 12082. https://doi.org/10.1038/s41598-025-95311-4

[28] Haq, H. B. U., Akram, W., Irshad, M. N., Kosar, A., Abid, M. (2024). Enhanced real-time facial expression recognition using deep learning. Acadlore Transactions on AI and Machine Learning, 3(1): 24-35. https://doi.org/10.56578/ataiml030103

[29] Abd El-Ghany, S., Mahmood, M.A., Abd El-Aziz, A.A. (2024). Adaptive dynamic learning rate optimization technique for colorectal cancer diagnosis based on histopathological image using EfficientNet-B0 deep learning model. Electronics, 13(16): 3126. https://doi.org/10.3390/electronics13163126

[30] Zhang, F., Zhang, L., Chen, H., Xie, J. (2021). Bird species identification using spectrogram based on multi-channel fusion of DCNNs. Entropy, 23(11): 1507. https://doi.org/10.3390/e23111507

[31] Knight, E.C., Hernandez, S.P., Bayne, E.M., Bulitko, V., Tucker, B.V. (2019). Pre-processing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks. Bioacoustics, 29(3): 337-355. https://doi.org/10.1080/09524622.2019.1606734

[32] Permana, S.D.H., Rahman, T.K.A. (2024). Sound classification for Javanese eagle based on improved mel-frequency cepstral coefficients and deep convolutional neural network. International Journal of Advanced Computer Science and Applications, 15(2): 204-216. https://doi.org/10.14569/IJACSA.2024.0150222

[33] Zhang, S., Gao, Y., Cai, J., Yang, H., Zhao, Q., Pan, F. (2023). A novel bird sound recognition method based on multifeature fusion and a transformer encoder. Sensors, 23(19): 8099. https://doi.org/10.3390/s23198099