# Bioacoustic Bird Monitoring: A Deep Learning Solution for Effective Biodiversity Conservation

1st Mansi Raval
*Department of Computer Science and Engineering*
*CSPIT, Charusat University*
Changa, India
mansi.raval24@gmail.com

2nd Pavan Chauhan
*U & PU. Patel Department of Computer Engineering*
*CSPIT, Charusat University*
Changa, India
21ce018@charusat.edu.in

3rd Mrugendrasinh Rahevar
*U & PU. Patel Department of Computer Engineering*
*CSPIT, Charusat University*
Changa, India
mrugendrarahevar.ce@charusat.ac.in

4th Amit Thakkar
*Department of Computer Science and Engineering*
*CSPIT, Charusat University*
Changa, India
amitthakkar.it@charusat.ac.in

***Abstract*—Accurate identification of bird species is crucial for effective biodiversity monitoring and conservation efforts. Traditional visual surveys for monitoring birds are labor-intensive and limited. This paper introduces a CNN-based approach for bird species classification using audio recordings from the BirdCLEF 2023 dataset, comprising 16,900 recordings of 264 species from Kenya. Raw audio was converted into Mel spectrograms and the EfficientNet-B1 model was utilized with data augmentation and mixup techniques. The model achieved a cmAP score of 0.84 on the validation set, surpassing existing models. This demonstrates the model's robustness and accuracy in classifying diverse bird species in complex soundscapes. These findings highlight the potential of this approach for real-world bioacoustic bird monitoring applications in ecological research and conservation efforts.**

***Index Terms*—BirdCLEF2023, LifeCLEF, audio, bioacoustics, EfficientNet-B1, bird recognition, Kaggle**

## I. Introduction

Birds play a vital role in ecology, and their populations are crucial indicators of ecosystem health and biodiversity trends. Traditional bird monitoring methods, such as visual surveys, are labor-intensive and limited in scope. Autonomous recording devices have revolutionized bird monitoring by capturing large amounts of audio data, with tens of terabytes frequently generated during data collection projects. However, manual monitoring of population changes based on audio data is nearly impossible, and machine learning-based analytical techniques are needed to process and analyze these large datasets. [1]

The BirdCLEF challenge is an initiative that aims to develop robust frameworks for avian vocalization detection and identification in continuous soundscape data. BirdCLEF 2023 is organized by the Cornell Lab of Ornithology's K. Lisa Yang Center for Conservation Bioacoustics and NATURAL STATE, with data sourced from the Xeno-Canto Foundation. Launched in 2014, BirdCLEF has evolved into one of the largest bird sound recognition competitions, featuring an extensive dataset covering tens of thousands of recordings representing up to 1,500 species. [2] [3] This edition focuses on detecting bird calls in long-term recordings in Kenya, a region known for its extraordinary biodiversity and diverse range of vocalizations. East African animals rely on auditory cues for socialization, communication, and survival. Still, African avian species are frequently underrepresented in sound databases like Xeno-canto[1], leading to a weak label problem for competitors. [4]

The rate at which bird populations are falling globally is concerning. There is a need for more efficient and automated methods for bioacoustic monitoring to overcome the challenges of manual annotation and processing of large datasets in bird sound recognition. [5] Effective monitoring of bird populations is critical for developing conservation solutions, but automated bird species classification models face challenges such as background noise, labeling ambiguity, and mismatch between training and testing data. These challenges limit the effectiveness of automated bird species identification systems, which are essential for ecological research and conservation initiatives. To overcome these challenges, this study presents a novel convolutional neural network (CNN)-based approach for bird species classification, leveraging prior studies and modifying the model architecture to tailor it specifically for bird species classification tasks.

The contribution of this research lies in addressing the shortcomings of current methods and proposing an efficient and automated approach for bioacoustic monitoring. This work aims to surpass existing models in performance and accuracy, advancing the state-of-the-art in bird audio classification and providing valuable tools for ecological research and conservation efforts.

## II. Related Work

Bioacoustic bird monitoring uses sound analysis and monitoring of bird diversity and abundance. Conventional tech-

[1]https://www.xeno-canto.org/

979-8-3503-7311-0/24/$31.00 ©2024 IEEE

niques for species identification depend on human specialists; however, these approaches can be laborious, time-consuming, subjective, and have limited scalability. Automated bird audio classification provides a viable alternative by using machine learning algorithms to analyze and categorize bird audio. Bird audio classification is an important area of research in bioacoustics, as it has the potential to transform wildlife monitoring by enhancing species identification accuracy, scalability, and efficiency.

BirdCLEF is an annual contest focused on developing and assessing automated analysis frameworks for bird audio signals in soundscape recordings. The competition aims to foster the development of analytical frameworks that can consistently identify and classify the vocalizations of rare avian species in continuous sonic landscapes, even in the absence of large training data. [6] It provides a standardized dataset and evaluation framework for automated bird species classification using audio recordings The competition aims to focus on advancing the automated detection of rare and endangered bird species and provides an opportunity to explore the application of deep learning models in bird audio classification[6]. This is particularly important because manual processing of a huge volume of soundscape data is undesirable, and automated attempts can help accelerate the process. [7]

The early stages of bird audio classification research focused on focus was on manual feature extraction from audio recordings, such as B. Gammaton filter bank energies and Mel frequency cepstral coefficients (MFCCs). After that, machine learning algorithms were used to identify the retrieved features, including random forests and support vector machines (SVMs). [8] In the present-day landscape of bioacoustic research, there is a growing interest in leveraging deep neural networks for the detection of sound events. [9], [10] Previously, attempts have been made to utilize CNN classifiers for distinguishing bird calls, utilizing visual representations of these auditory signals in spectrograms. [11]–[14] It has emerged as the most popular model in terms of accuracy and benefits, beating earlier classification approaches. Presently, combinations of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) like LSTM, GRU, and RMUs are used to enhance the spectral and temporal features of bird sounds. [15] Researchers are adopting pre-trained models such as YamNet, AlexNet, and ResNet-50 to improve accuracy.

The application of deep learning models, such as EfficientNets, has produced encouraging results in bird audio classification. EfficientNets are a class of efficient and scalable convolutional neural network models that have achieved cutting-edge performance in various image recognition tasks. Recent studies have explored the application of EfficientNets to the problem of bird species identification using audio data. Researchers have achieved high classification accuracy by fine-tuning pre-trained EfficientNet models on large-scale bird audio datasets, even in the presence of imbalanced or insufficient training data. [16]–[18]

Bird audio classification has become an essential tool for bioacoustic bird monitoring. The field has experienced significant progress with the adoption of deep learning techniques, particularly Convolutional Neural Networks (CNNs). Transformers are receiving significant attention due to their exceptional performance in several classification tasks, particularly in the field of bird audio classification. One prominent example is the STFT Transformer, which demonstrated competitive performance in the BirdCLEF 2021 competition and uses time slices of spectrograms as input fields. [19] While these developments have transformed the field of bioacoustic bird monitoring, there are still limitations to consider, such as the need for large-scale datasets and the potential for overfitting. Nevertheless, the advancements in bird audio classification have paved the way for more thorough and efficient monitoring of bird populations, ultimately contributing to a better understanding of avian ecology and conservation activities.

## III. Proposed Methodology

This section provides an extensive description of the bird audio detection technique. The proposed methodology is divided into three main sections: dataset description, data preparation, model architecture, and implementation details. Fig. 1 shows the model pipeline. It also shows the EfficientNet-B1 architecture. [20]

### A. Dataset Description

BirdCLEF2023[2] dataset is a large-scale audio dataset of bird vocalizations. The data consists of short and individual audio recordings of different bird species uploaded on the Xeno-canto network. Xeno-canto provides 16,900 audio recordings that can be used to train classifiers. The training data contains 264 species from Kenya, Africa. To match the audio in the test set, these files have been converted to the ogg format and downsampled to 32 kHz when necessary. The audio data are provided in the corresponding folders representing the bird names. The duration of each audio recording varies. Fig. 2 shows the number of training samples per species.

A wide range of metadata is provided for the training data with key fields like primary labels, coordinates for latitude and longitude, scientific name, common name, etc. This extensive metadata allows researchers to contextualize the audio recordings and explore factors like habitat and location that may influence the acoustic characteristics of the bird vocalizations.

### B. Data Preprocessing

The proposed methodology includes a comprehensive preprocessing pipeline that converts raw audio data into a machine learning model-compatible format. The first step in the preprocessing pipeline is to extract metadata, such as the labels of the audio files, from the provided '.CSV' file. To ensure an even distribution of labels between training and validation sets, a stratified split is performed based on the primary labels after using regular expressions to extract secondary labels. Key audio properties, including the number of frames, sample rate, and duration, are then extracted and appended to the dataset.

[2]https://www.kaggle.com/c/birdclef-2023/data

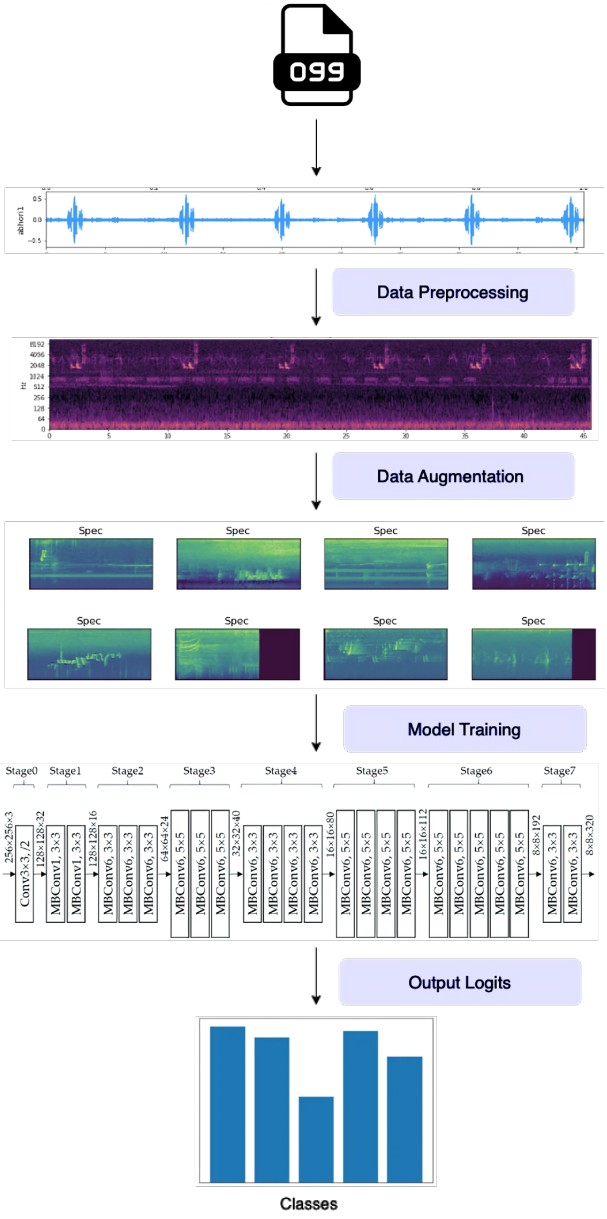


Fig. 1. Model Pipeline

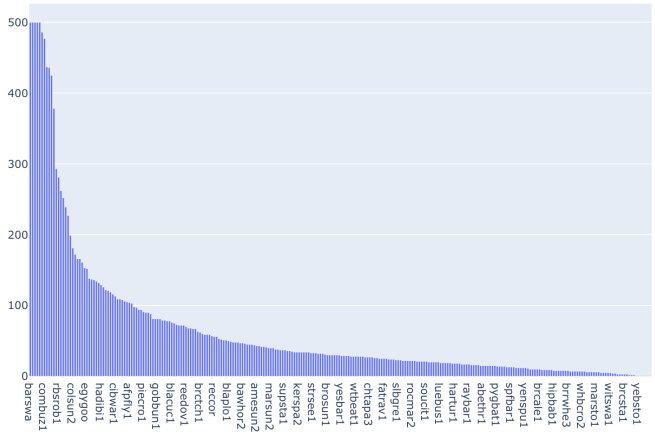

Fig. 2. Number of Training Samples Per Species

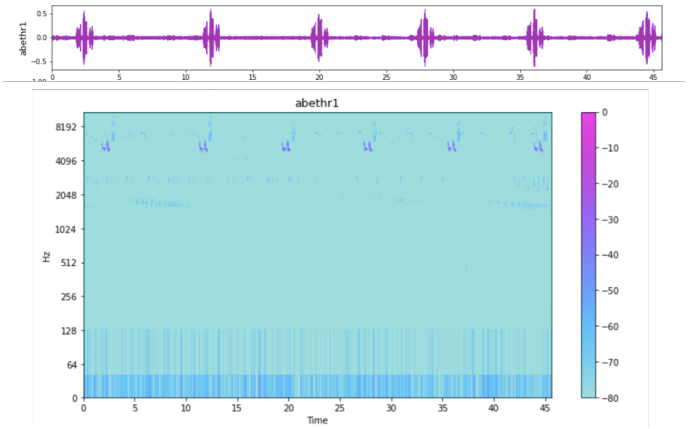


Fig. 3. Sound Wave and Mel Spectrogram of *abethr1*

From the recorded audio dataset, all audio signals were considered as part of the challenge. These audio signals were chopped into 5-second intervals with a 66.6% overlap (i.e., a step size of 3.33 seconds) and converted into Mel Spectrograms. The short-time Fourier transform (STFT) was applied to generate a spectrogram for the Mel spectrogram calculation. The spectrogram was then plotted on the Mel frequency scale, where equal distances on the Mel scale correspond to the same perceived distance because it is a logarithmic scale (1000 Mel = 1000 Hz). To create the Mel spectrograms, a sampling rate of 32k was used, and the number of Mel bands was set to 128. Fig. 3 shows the waveform and Mel spectrogram of one of the species in the dataset.

The Mel spectrograms were normalized and converted into image-like representations. Normalizing the data to the range of 0-255 ensures consistency and facilitates model training. Fixed-length audio segments were extracted using cropping and padding, and then converted into 224x224 pixel images. These images were saved in their respective output directories as Numpy files, with each file containing the Mel spectrogram images for a single audio recording. The preprocessing pipeline provides input for model training and assessment.

### C. Model Architecture

CNNs were first presented by LeCun in the late 1980s [21], it is composed of various layers, including the activation layer, pooling layer, fully connected layer, and convolutional layer. Unique features are extracted from input images using convolutional layers. Convolutional operations between the kernel and the input image are used to extract these features. The kernel generates the feature map by sliding over the image in a small, rectangular matrix measuring 5 × 5 and 3 × 3. The pooling layer preserves the most important information about the feature map, which is then used to make the feature map smaller. Average pooling and maximum pooling are two types of pooling layers used. Classifying images is done by fully

connected layers, and final result prediction is done by the output sigmoid layer. [22], [23]

Convolutional Neural Networks (CNNs) are often developed with a fixed budget of resources, and when additional resources become available, they are scaled up to achieve higher accuracy. Tan and Le employed neural architecture search to create a new baseline network and scale it up, resulting in a family of models known as EfficientNets that outperform earlier CNNs in terms of accuracy and efficiency. [24] EfficientNets models are a set of image classification algorithms that achieve exceptional accuracy while having orders of magnitude smaller and faster than previous models. The authors propose a compound scaling technique that uses a predetermined set of coefficients to scale width, depth, and resolution consistently. There are eight models in the EfficientNet model group, ranging from B0 to B7. The model numbers that follow correspond to versions that have additional parameters and are more accurate.

The proposed model architecture for the bird species classification task on BirdCLEF2023 is based on the convolutional neural network EfficientNet-B1. [24] The EfficientNet-B1 architecture is a series of MBConv (Mobile Inverted Bottleneck Convolution). EfficientNet models are known for their high performance and efficiency, making them a suitable choice for this task. The backbone of the model is the pre-trained EfficientNet-B1 model. It is highly efficient in performing feature extraction from input images and is effective for image classification tasks, consisting of a series of convolutional layers, batch normalization, activation functions, and pooling layers. As the input passes through the EfficientNet-B1 backbone, the spatial dimensions of the feature maps are gradually reduced through pooling operations such as max-pooling or average pooling. In the proposed model, the last layer of the original EfficientNet-B1 model is modified according to the number of bird species classes in the dataset and it uses binary cross-entropy with logit loss. This model architecture leverages the powerful feature extraction capabilities of the pre-trained EfficientNet-B1 backbone while adapting it to the specific requirements of the bird species classification task through modifications to the final layer and the use of appropriate loss functions and optimization. Fig. 4 shows the model architecture for bioacoustic bird monitoring.

The model processes a series of Mel spectrograms derived from audio recordings of bird calls. These spectrograms are generated using a window size of 2048 samples, a hop length of 512 samples, and 128 Mel bins. Its output corresponds to the primary label of the dataset, representing the bird species in the audio recordings. The model calculates these probabilities using a sigmoid activation function, which may be interpreted as the probability of each species being in the input audio recording.

## IV. Results

### A. Evaluation Parameters

BirdCLEF competitions used the cmAP metric before moving to Kaggle, which does not support it. Therefore, different versions of the F1 Score were used in the 2020, 2021 and 2022 competitions. The disadvantage of F1 is that a binary label for the species must be selected in each inference window. This has led to the development of threshold selection techniques that make it difficult to assess the quality of the base model. In practice, thresholds are selected based on end-user preferences, which is why threshold-free model quality assessment is preferred. [4]

The cmAP score (1) is a metric that is defined as the average of the per-class mean average precision (AP) for all classes.

$$cmAP := \frac{\sum_{c=1}^{C} AveP(c)}{C} \tag{1}$$

where $C$ is the number of target classes, and $AveP(c)$ (2) is the average precision for the $c$th species, computed as:

$$AveP(c) := \frac{\sum_{k=1}^{N} P(k) \times rel(k)}{n_{rel}(c)} \tag{2}$$

where k represents an item's rank in a list of examples with class $c$, $P(k)$ represents the precision at cut-off $k$ in the list, $rel(k)$ is an indicator function indicating whether class $c$ is present in the $k$th example, and $n_{rel}(k)$ is the total number of examples containing class $c$.

The cmAP score calculates the mean average precision per class and then averages it across all classes, weighting each type equally. However, for species with few samples in the collection, the MAP of each species may be noisy, affecting the overall cmAP score. A modified cmAP metric called padded cmAP (pcmAP), was proposed to address this issue. This method involves adding "free" examples at the beginning of each class, reducing the fluctuation of MAP values, and mitigating the impact of sparse labels. During the contest, the parameter was set to 5. [4] Table I shows the validation metrics at the end of the training.

### B. Experimental Setup

In the experimental setup, a Nvidia GeForce GTX TITAN X GPU alongside an Intel Xeon CPU, supported by 30 GB of DDR4 RAM running at 3200 MHz, was utilized.

The Adam optimizer with cosine annealing and warm restarts was used to schedule learning rates, enhancing convergence and generalization. Mixup data expansion was employed, linearly combining two input instances and their respective goals to produce a new training example. For training and evaluation, pre-processed Mel spectrogram images and labels were fed into a custom dataset class using PyTorch Lightning. For validation, binary cross-entropy loss was computed. The hyperparameters included a batch size of 64, an Adam optimizer, a 10e-4 learning rate, and 20 epochs with Early Stopping configuration. The dataset comprised 264 classes.

### C. Experimental Results

Experimental results show that the model correctly detects bird species in audio recordings, achieving a remarkable cmAP score of 84.242%. As shown in Table II, this performance

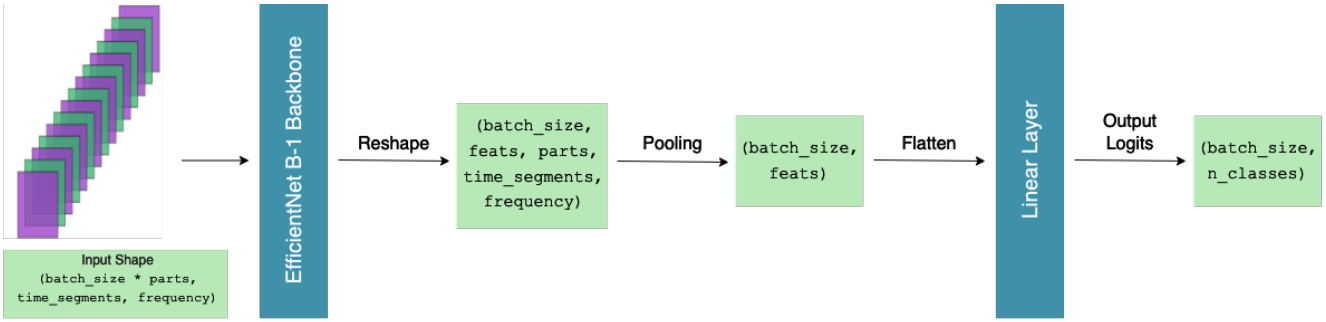


Fig. 4. Model Architecture

surpasses the top-performing models in the reference study, such as those leveraging EfficientNetV2s (M9) and ResNet50 (M8), which achieved cmAP scores of 83.386% and 83.288%, respectively. The model's performance was evaluated on the validation set during the training process. Fig. 5 shows training and validation accuracy over 20 epochs, indicating that the model learned to reduce the loss function and attained a stunning accuracy of more than 0.95. This excellent performance illustrates the model's capacity to correctly identify a diverse variety of bird species based on their distinctive vocalizations, as well as its potential application in ecological studies, conservation projects, and biodiversity monitoring programs.

During training, effective data augmentation methods like random cut squares were employed to increase the model's flexibility and capacity for generalization. To further improve the model's performance, mixup data augmentation with an alpha value of 0.6 was used. These strategies added variety to the training dataset and helped minimize overfitting, allowing the model to learn distinctive features and accurately classify species based on previously unseen samples.

A qualitative analysis of the model's predictions was conducted to gain further insights into its behavior. Fig. 7 illustrates a confusion matrix, providing a visual representation of the model's performance in correctly identifying various bird species. The diagonal elements of the matrix denote the count of correctly identified samples for each class, reflecting instances where the actual label matches the predicted label. Examination of the confusion matrix illustrates the model's general capacity to distinguish among the predominant bird species effectively. The model excels notably in detecting prevalent bird species, although some uncommon species may necessitate adjustments to the model's structure or supplementary training data.

TABLE I
VALIDATION METRICS AT THE END OF TRAINING.

| Metric | Value |
|---|---|
| Validation Loss | 0.11 |
| Validation C-MAP (Padding Factor 5) | 0.84 |
| Validation C-MAP (Padding Factor 3) | 0.79 |
| Validation AP Score | 0.83 |

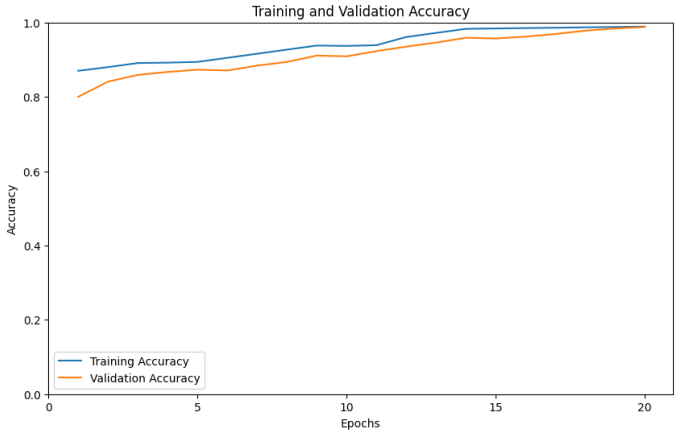


Fig. 5. Training and Validation Accuracy

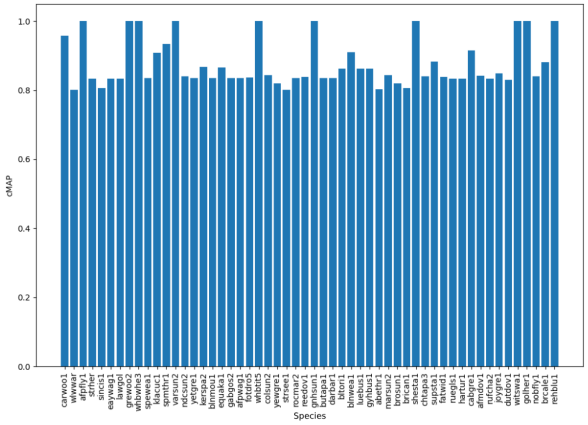


Fig. 6. cmAP Score Per Species

## V. Discussion

As shown in Table II [25], the proposed model achieves a higher cmAP of 84%, surpassing all the mentioned models. This cmAP score on the validation set outperforms the previously best-performing model (M9) by a margin of 0.856, showcasing the effectiveness of the proposed approach. The model's performance was specifically optimized for the cmAP score ensuring exceptional performance in ranking the predic-

TABLE II
COMPARISON WITH STATE-OF-ART METHODS

| Model ID | Model Architecture Description | cmAP[%] |
|---|---|---|
| M1 | Baseline (EfficientNetB0, chunk selection unweighted) | 80.949 |
| M2 | M1 w. chunk selection weighted 75 % RMS, 25 % pseudo label | 81.289 |
| M3 | M2 with EfficientNetV2s | 81.853 |
| M4 | M3 w. chunk selection weighted 45 % RMS, 15 % pseudo label | 82.299 |
| M5 | M4 w. EfficientNetB0 and extended 2023 Xeno-canto data | 82.808 |
| M6 | M5 with reverb augmentation | 83.164 |
| M7 | M6 w. original SED head using attention on time frames | 82.897 |
| M8 | M6 with ResNet50 | 83.288 |
| M9 | M6 with EfficientNetV2s | 83.386 |
| **M10** | **EfficientNet-B1 Backbone with modified layer** | **84.242** |

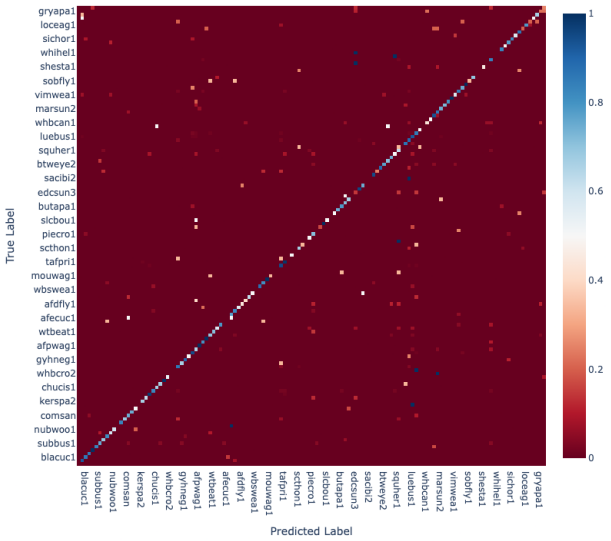


Fig. 7. Confusion Matrix

tions for every species of bird, which is a crucial component of the challenge. Fig. 6 shows cmAP score per species for a few selected species. The model can handle a wider range of variations in the data, making it more robust to unseen samples. Fig. 5 shows an accuracy of about 0.989. This proves that the model performs better than state-of-the-art models in accurately detecting bird species using audio recordings.

The BirdCLEF 2023 task's focus on cmAP and inference time constraints promoted the development of efficient models that balance accuracy and speed. The modified Sound Event Detection (SED) architecture was found to be effective in detecting birds in soundscapes where multiple species vocalize in different frequency ranges. The main strengths of this study include the use of a large dataset, which enabled the development of robust models that translate well to unseen data. The combination of a modified EfficientNet B1 backbone and effective data expansion techniques such as B. random intersecting squares and confusions led to improved generalizability and accuracy. The results highlight the importance of careful model selection and hyperparameter tuning to achieve state-of-the-art performance. While the approach achieved state-of-the-art performance, we also explored other methods, including various loss functions, model soups, and knowledge distillation, but did not produce satisfactory results. The proposed model occasionally had problems with rare species and suggested areas for future improvements, such as structural adjustments or additional training data. The performance of the model depends heavily on the quantity and quality of the training data. Future research should explore the applicability of this approach to other datasets and architectures, as well as investigate transfer learning and domain adaptation techniques to improve model performance.

## VI. CONCLUSION

In conclusion, the proposed approach achieved state-of-the-art performance on the BirdCLEF 2023 task, highlighting the potential of deep-learning solutions for bioacoustic bird monitoring. By leveraging a modified EfficientNet-B1 backbone and effective data augmentation techniques, the model achieved a cmAP score of 84.242% and showed strong generalization capabilities. This outstanding performance surpasses the top-performing models in the reference study. We envision a future where automated bird species identification systems play a critical role in biodiversity conservation, enabling scalable and cost-effective solutions for monitoring bird populations informing conservation efforts, tracking habitat health, and informing conservation efforts. This approach can serve as a baseline for future research in this area, aiming to inspire further advancements in deep learning-based bird classification. Ultimately, the goal is to contribute to a world where birds and their habitats are protected and preserved for future generations, and we believe that bioacoustic bird monitoring using deep learning solutions is a crucial step towards achieving this vision.

## REFERENCES


[1] D. Tuia, B. Kellenberger, S. Beery, B. R. Costelloe, S. Zuffi, B. Risse, A. Mathis, M. W. Mathis, F. Van Langevelde, T. Burghardt *et al.*, "Perspectives in machine learning for wildlife conservation," *Nature communications*, vol. 13, no. 1, pp. 1–15, 2022.

[2] A. Joly, H. Goëau, S. Kahl, B. Deneu, M. Servajean, E. Cole, L. Picek, R. Ruiz de Castañeda, I. Bolon, A. Durso *et al.*, "Overview of lifeclef 2020: a system-oriented evaluation of automated species identification and species distribution prediction," in *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 2020, pp. 342–363.

[3] S. Kahl, M. Clapp, W. A. Hopping, H. Goëau, H. Glotin, R. Planqué, W.-P. Vellinga, and A. Joly, "Overview of birdclef 2020: Bird sound recognition in complex acoustic environments," in *CLEF 2020-Conference and Labs of the Evaluation Forum*, vol. 2696, no. 262, 2020.

[4] S. Kahl, T. Denton, H. Klinck, H. Reers, F. Cherutich, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, and A. Joly, "Overview of birdclef 2023: Automated bird species identification in eastern africa," *Working Notes of CLEF*, 2023.

[5] C. K. Nagesh and A. Purushothama, "The birds need attention too: Analysing usage of self attention in identifying bird calls in soundscapes," *arXiv preprint arXiv:2211.07722*, 2022.

[6] S. Sharma, K. Sato, and B. P. Gautam, "A methodological literature review of acoustic wildlife monitoring using artificial intelligence tools and techniques," *Sustainability*, vol. 15, no. 9, p. 7128, 2023.

[7] C. M. Wood, S. Kahl, P. Chaon, M. Z. Peery, and H. Klinck, "Survey coverage, recording duration and community composition affect observed species richness in passive acoustic surveys," *Methods in Ecology and Evolution*, vol. 12, no. 5, pp. 885–896, 2021.

[8] S. A. Brooker, P. A. Stephens, M. J. Whittingham, and S. G. Willis, "Automated detection and classification of birdsong: An ensemble approach," *Ecological Indicators*, vol. 117, p. 106609, 2020.

[9] 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.

[10] Y. Shiu, K. Palmer, M. A. Roch, E. Fleishman, X. Liu, E.-M. Nosal, T. Helble, D. Cholewiak, D. Gillespie, and H. Klinck, "Deep neural networks for automated detection of marine mammal species," *Scientific reports*, vol. 10, no. 1, p. 607, 2020.

[11] J. Schlüter, "Bird identification from timestamped, geotagged audio recordings." *CLEF (Working Notes)*, vol. 2125, 2018.

[12] M. Lasseck, "Bird species identification in soundscapes." *CLEF (Working Notes)*, vol. 2380, 2019.

[13] M. Mühling, J. Franz, N. Korfhage, and B. Freisleben, "Bird species recognition via neural architecture search." in *CLEF (Working Notes)*, 2020, pp. 1–13.

[14] S. Kahl, A. Navine, T. Denton, H. Klinck, P. Hart, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, and A. Joly, "Overview of birdclef 2022: Endangered bird species recognition in soundscape recordings." in *CLEF (Working Notes)*, 2022, pp. 1929–1939.

[15] G. Gupta, M. Kshirsagar, M. Zhong, S. Gholami, and J. L. Ferres, "Comparing recurrent convolutional neural networks for large scale bird species classification," *Scientific reports*, vol. 11, no. 1, p. 17085, 2021.

[16] Y. Yang, K. Zhou, N. Trigoni, and A. Markham, "Ssl-net: A synergistic spectral and learning-based network for efficient bird sound classification," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 926–930.

[17] S. G. Gundabatini, S. Sai, S. Vinay, Reddy, T. Chandrika, S. Kaarthikeya, P. Kumaar, S. Siddik, and T. S. Akhil, "Bird species identification using audio processing and alexnet neural network," *international journal of food and nutritional sciences*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258034345

[18] A. Noumida, R. Mukund, N. M. Nair, and R. Rajan, "Stacked res2net-cbam with grouped channel attention for multi-label bird species classification," in *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE, 2023, pp. 446–450.

[19] J.-F. Puget, "Stft transformers for bird song recognition." in *CLEF (Working Notes)*, 2021, pp. 1609–1616.

[20] Y. Jie, X. Ji, A. Yue, J. Chen, Y. Deng, J. Chen, and Y. Zhang, "Combined multi-layer feature fusion and edge detection method for distributed photovoltaic power station identification," *Energies*, vol. 13, no. 24, p. 6742, 2020.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[22] J. Ding, B. Chen, H. Liu, and M. Huang, "Convolutional neural network with data augmentation for sar target recognition," *IEEE Geoscience and remote sensing letters*, vol. 13, no. 3, pp. 364–368, 2016.

[23] A. Hussain, S. Ul Amin, M. Fayaz, and S. Seo, "An efficient and robust hand gesture recognition system of sign language employing finetuned inception-v3 and efficientnet-b0 network." *Computer Systems Science & Engineering*, vol. 46, no. 3, 2023.

[24] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[25] M. Lasseck, "Bird species recognition using convolutional neural networks with attention on frequency bands," *CLEF Working Notes*, 2023.