Proceedings of 13th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2023 Month: April Volume: 13

Audio Classifier for Automatic Identification of Endangered Bird Species of Nepal

Rewan Gautam [a], Bhuwan Khatiwada [b], Bishwa Prakash Subedi [c], Niraj Duwal [d], Kiran Chandra Dahal [e]

[a, b, c, d, e] *Thapathali Campus, IOE, Tribhuvan University, Nepal*

[a] tha075bei032@tcioe.edu.np, [b] tha075bei009@tcioe.edu.np, [c] tha075bei011@tcioe.edu.np, [d] tha075bei028@tcioe.edu.np, [e] dahalkc@ioe.edu.np

Abstract

Home to around 11,000 species of fauna, Nepal is a country rich in biodiversity. Among them about 900 species are birds. Due to various reasons like encroachment of their natural habitats, and rampage killing many of these species are facing the threat of extinction. Around 38 endangered birds in Nepal need conservation. The growing advancement in machine learning can support the preservation of those species and monitoring the status of birds in the ecosystem can assist researchers of Nepal's biodiversity in planning different strategies for their preservation. We developed an endangered bird classification system to identify bird calls from the audio data set collected from Xeno-canto.org. This was achieved by extracting the spectral characteristics of the audio signal through Mel-Spectrogram and MFCC (Mel-Frequency Cepstral Coefficients) which generated the spectrogram. It was fed into the deep learning model architecture like efficientNet which is based on a convolutional neural network. A genetic algorithm was used for hyper-parameter optimization. Our best results showed an F1-Score of 79% for 41 species of birds (38 endangered and few other birds from Nepal). This has significant implications for the field of biodiversity conservation, as it allows researchers to monitor and protect threatened and endangered species.

Keywords

Audio, Birds, CNN, efficientNet, Genetic-Algorithm, Mel-Spectrogram, MFCC

1. Introduction

According to IUCN Red List of Threatened Species in 2021 [1], an estimated 159 bird species have gone extinct in the last 500 years. Also, out of 11,162 species 1,445 species are classified as threatened or endangered which is about 13 % of the total species on earth i.e. around 1 in 7 species are endangered. Brazil is home to nearly 172 of these endangered species and Indonesia has about 106 endemic species that are threatened [2]

Nepal is home to 875 species of birds out of 11,162 species found worldwide making it 27^{th} on the list of having the most endemic bird population. Out of these about 38 species are globally threatened i.e. about 5 % of the total bird species. Among the 38 threatened species, 9 are critically endangered, 8 are endangered and 21 species are vulnerable [3]. Some of the endangered birds of Nepal are Black-breasted Parrotbill, Black-necked Crane, Greysided Thrush, Rustic Bunting, Sarus Crane, Wood Snipe, Yellow-breasted Bunting, etc.

Birds are wildlife that can be found in the wild i.e. in deep and dense jungles and forests. Millions of years of evolution have made many creatures able to adapt themselves to their habitat making them practically invisible. Many birds have features that give them the ability to hide in plain sight. So visual recognition may not be the best approach for the classification of birds in their natural habitat. Thankfully birds are also well known for their melodious singing. If not visible it is possible that we might be able to hear the songs of the birds in the vicinity. So it is possible to apply audio recognition tools for the classification of various birds. For this, the data set of different birds singing can be used to recognize the birds using a machine learning algorithm that analyses the birds singing.

The following efforts were made for this research:

- Data collection, processing and augmentation
- Feature extraction using mel spectrogram and MFCC
- Classification of birds through efficientNet and custom CNN model

The organization of the paper is summarized as follows. The background information and a brief overview of related work are provided in Section 2. The CharaNet dataset used for preliminary analysis has been elaborated in Section 3. Section 4 describes the proposed approach in detail, including various alternative methods for feature extraction and audio classification. The preliminary findings from the CharaNet dataset and the results are demonstrated and discussed in Section 5. Finally, Section 6 concludes the paper by giving it a short conclusion.

2. Background and Related Works

Some works have been done on bird classification using different approaches. Marini et al., 2013[4] classified bird species based on color features. Here, color segmentation was applied to remove the background elements and define candidate regions for the presence of the birds. The normalized color histograms were computed from there. The histogram bins were fed to the machine learning algorithm.

But it is true that it's easier to hear birds than to see them. Extensive bird classification research can be observed with audio data. The annual BirdCLEF challenge made this problem

more popular. The 2021 edition of this challenge was to predict bird species among 397 birds in every 5 seconds of test data. The training dataset was provided in the challenge from all over the regions but test recording was provided only from four different places: New York, California, Costa Rica, and Colombia. The 10th place holder of the 2021 BirdCLEF challenge, Conde et al. [5] discussed a classification solution to this problem using custom augmentations and not limited to only the audio dataset. The additional features provided in the competition like the appearance of other birds in the audio, the rarity of the bird, and (latitude and longitude) were considered.

In 2018, Incze et al.[6] used compact and performance-oriented MobileNet as their starting checkpoint. Spectrograms based on grayscale and jet color were used for the input to the neural network. In the grayscale, 0 represents white and 1 represents black while in jet color 0 represents blue, around 0.5 represents yellow and 1 represents red. The accuracy achieved with both approaches decreased with the increase in the number of classes.

"Automatic bird audio identification using convolutional neural networks and Mel-spectrogram features" [7] Fabian-Robert Stöter et al. (2018). In this study, the authors used Mel Spectrogram as a feature extractor in combination with a CNN model to classify bird audio recordings into different bird species. They considered a total of 10 bird species in their study. They achieved an accuracy of 92.3% in their classification task. One limitation of the study is that it only considered a small number of bird species, which may not be representative of the entire bird population. "Bird sound recognition using deep convolutional neural networks and Mel-frequency cepstral coefficients" [8] Fabian-Robert Stöter et al. (2017) - In this study, the authors used MFCC as a feature extractor in combination with a deep convolutional neural network (DCNN) for bird audio recognition. They found that the combination of MFCC and DCNN achieved good performance in the recognition task.

We propose our system to extract features using Mel Spectrogram, MFCC and use a genetic algorithm for parameter optimization.

X. Xiao et al. [9] proposed Genetic Algorithm to optimize the hyper-parameters in CNNs. In their work, they did not restrain the depth of the model. Experimental results show that they can find satisfactory hyper- parameter combinations efficiently with accuracy about 88.92% and within 24.55 hours which is relatively better than random search algorithm.

3. Dataset analysis

On the website xeno-canto.org, bird sounds from throughout the globe are shared. The recordings are uploaded on the website by the contributors who travel around.

We collected 2215 audio recordings from this website of 41 birds of which 38 of them are endangered species. The dataset was increased to 6733 audio recordings after the 10-second splitting and data augmentation through gaussian noise addition for birds having less than 30 files. A total of 5407 audio recordings were used for training, 639 for validation, and 687 for testing purposes.

4. Proposed Methodology

The audio dataset was collected from https://xeno-canto.org/ and augmented with noise to increase its size. Imbalanced class distribution was handled using data augmentation for minority classes instead of sampling techniques. Mel spectrograms and MFCCs were used as feature extractors due to their ability to capture the spectral content of audio signals over time. Custom CNN model and efficientNet model were built whose hyperparameters were optimized with genetic algorithm.

4.1 Mel Spectrogram

Often called "pictures of a sound", spectrograms are simply a graphical or a visual representation of a signal. Spectrograms give us information about the presence of various frequency components at different times and the strength or loudness of a certain frequency at a particular time. Mel spectrograms are those spectrograms in which the signal frequencies are transformed into the Mel scale. It is a representation of signals which is perceptually relevant in terms of frequency.

The mel spectrogram of audio signals are extracted following the given series of steps:

1. The Short-Time Fourier Transform (STFT) is applied on the given audio signal to generate a standard spectrogram.

2. The amplitudes of different frequencies are then converted to decibel scale.

3. The various frequency components found in the audio signal are converted to the Mel scale. The conversion of frequencies to Mel scale involves the following procedure:
   (a) First, the number of mel bands or the number of filters in the filter bank is chosen.
   (b) The next step is the construction of the mel filter banks, which require the following steps:
      i. The lowest and highest frequency of the extracted spectrogram are converted to the Mel scale using the following equation.
      $$m = 2595 \cdot \log\left(1 + \frac{f}{700}\right) \quad (1)$$
      where,
      m = frequency in Mel scale,
      f = frequency in Hertz scale.
      ii. The mel scale frequency range is divided into equally spaced intervals called Mel bands. The center frequency of each filter must be at the center of each interval.
      iii. The center frequencies of each filter are converted back to Hertz scale and rounded to nearest frequency bins. The conversion back to the hertz scale is performed using 1.
      iv. For each filter a band pass filter (triangular filter) is designed that passes frequencies of certain range around the center frequency.
   (c) The mel filter banks are applied to the spectrogram resulting from the short-time fourier transform to extract mel spectrogram.

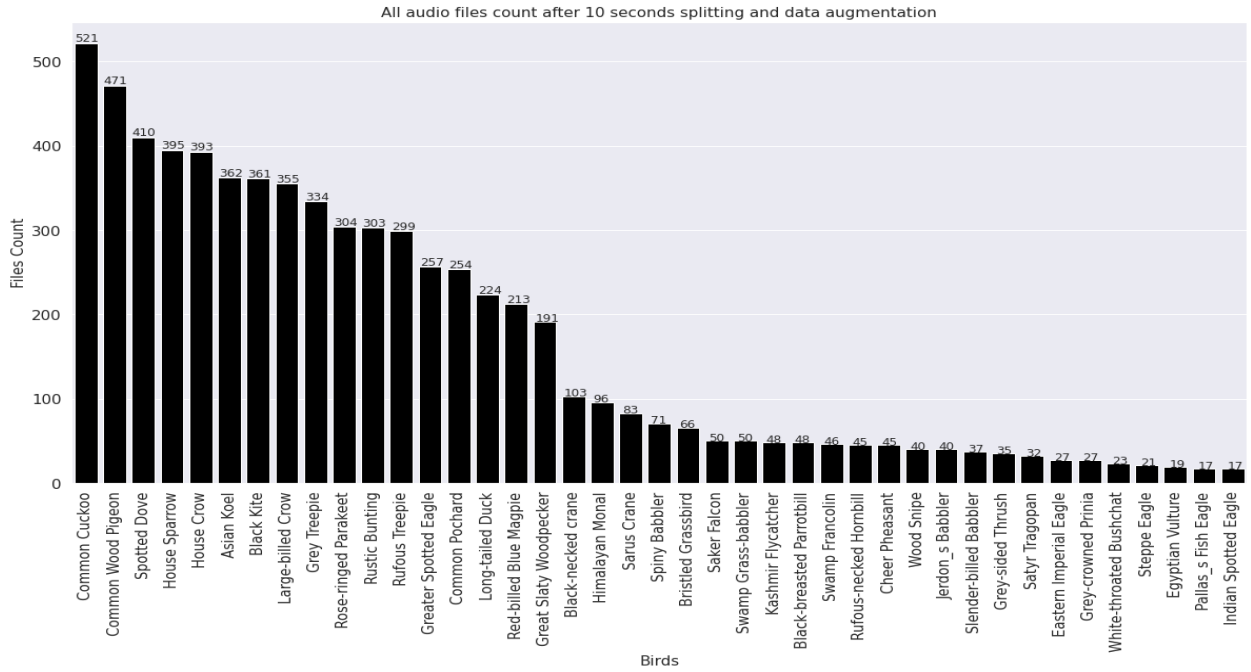

Figure 1: Audio Files Count for each Bird

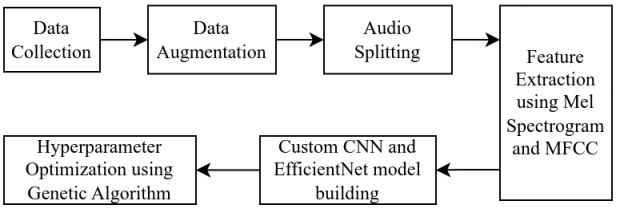

Figure 2: Research Methodology

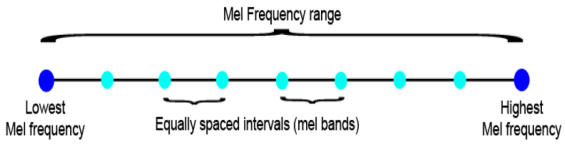

Figure 3: Equally spaced intervals (Mel bands)

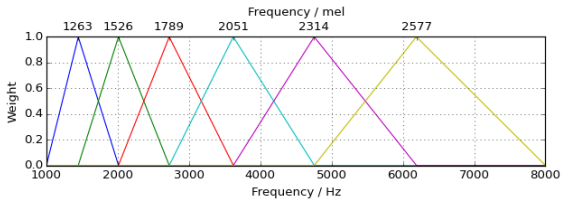

Figure 4: Mel Filter Banks

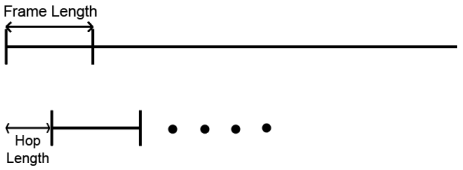

Figure 5: Framing of signal into short frames

4.2 MFCC

Mel-Frequency Cepstral Coefficients (MFCC) is a technique for feature extraction of audio signals. The coefficients of MFCC represent the spectral characteristics of an audio signal in a compact way. These coefficient values contain information about rate of change of different spectrum bands. MFCC's are relatively low dimensional and are able to describe large structure of spectrums. They ignore fine spectral details and hence are robust to background noises[10]

The majority of spectral energy is concentrated in the lower frequencies if the MFCC's have a positive value and if the MFCC's have negative value then, the majority of spectral energy is concentrated in the higher frequencies.

The extraction of MFCC's of a signal involves following steps:

1. The given signal is framed into short frames.
2. A window function is applied to the signal to nullify the effect of signal occurring outside the frame.
3. Discrete Fourier Transform (DFT) is applied to the windowed signal to generate the frequency spectrum of each frame.
4. Logarithm is applied to the spectrum to generate log amplitude spectrum.
5. Perform Mel scaling on the frequencies of the spectrum by applying necessary filter banks.This results in a Mel spectrogram.
6. Discrete cosine transform (DCT) is applied to the spectrogram , which gives a number of real valued coefficients. These coefficients are MFCC's.

4.3 EfficientNet

EfficientNet is a CNN architecture for image classification that uses compound scaling, which balances scaling of network depth, width, and resolution. As mentioned by Mingxing Tan

and Quoc V. Le in their research "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks"[11], performing scaling in any of the dimension of network i.e. depth, width and resolution increases the accuracy of the model but as the model gets bigger the accuracy decreases and the model faces the problem of vanishing gradient. EfficientNet achieves better accuracy and efficiency by increasing all dimensions in a balanced way. EfficientNet achieves high accuracy with fewer parameters than other networks.

We discussed that efficientNet uses a technique called compound scaling. Compound scaling is the balanced scaling of all three dimension depth, width and resolution in a constant ratio. The ratio for different scaling is given by the formula:

$$f = \alpha \cdot \beta^{\phi} \cdot \gamma^{\phi} \tag{2}$$

where, α is depth scaling factor, β is width scaling factor, γ is resolution scaling factor and f is network scaling factor. Here, α, β, γ are constant values and ϕ is a variable.

EfficientNet B0 is the base model used for scaling the depth, width, and resolution. All other models, such as B1, B2, etc., are scaled versions of B0. The base model is fixed and developed using Neural Architecture Search, which involves combining other neural networks. For efficientNet B0 or our baseline model, the value for α is 1.2, β is 1.1, γ is 1.1 and ϕ is 1. These values are determined by performing grid search.

4.4 Audio Splitting Algorithm

The audio dataset may have files of different durations. To maintain uniformity and increase the dataset size, each file is split into 10-second duration files. However, some splits may only contain silence or environmental noise. To overcome this issue, the validity of each audio split is checked by verifying if it contains at least half of the maximum amplitude of the original file. Additionally, if a split is too small, it is appended to the previous split instead of creating a separate audio file.

The algorithm to split the larger audio file into smaller samples is:

Step-1: Load audio sample into memory.

Step-2: If (AudioDuration) $>$ 10 then,
Calculate how many subsets of audio files can be created and length of last sample as:
nSplits = Audioduration / 10
rem = AudioDuration % 10
Else,
GOTO Step-5.

Step-3: Split the whole audio into nSplits subsets using the function.
newAudio = originalAudio[splits[n]:splits[n+1]]
GOTO Step-5.

Step-4: If rem $<$ 5 then,
Don't make separate audio file.
Else,
Make a separate audio file

Step-5: If the audio file is valid, then export the file.

4.5 Genetic Algorithm

The audio features are obtained using Mel Spectrogram and MFCC, resulting in an image input for the CNN model. For the optimization of the hyperparameters, an evolutionary algorithm like a genetic algorithm is used. Initially, a population of initial random parameters is generated, and the model is trained on the preprocessed dataset using a CNN architecture. Various evaluation metrics can be utilized to assess model performance, and if stopping criteria are met, the optimized parameters are saved for the best model. Otherwise, the selection is based on the fittest individuals, with their genes interchanged through the crossover. The mutation is applied to introduce randomness and maintain diversity in new offspring. The genetic algorithm follows the given steps:

Step-1: Randomly initialize populations 'p'

Step-2: Determine fitness scores of population.

Step-3: Until convergence repeat:
a) Select parents from population
b) Crossover and generate new population
c) Perform mutation on new population
d) Calculate fitness for new population

There are several reasons why one might choose to use a genetic algorithm (GA) for hyperparameter tuning instead of other techniques such as Bayesian search, grid search, or random search. Deep learning models can have a large number of hyperparameters, which can make exhaustive search methods like grid search computationally expensive or even infeasible. In contrast, the genetic algorithm can handle a large number of parameters with relatively fewer evaluations. Also, genetic algorithm is particularly well-suited for exploring a wide range of hyperparameters. It starts with a diverse population of individuals, and then it evolves them over multiple generations to find the optimal set of hyperparameters.

5. Results and Analysis

5.1 Feature Extraction

Mel spectrograms were extracted from each 10-second audio sample (long enough to capture meaningful information about the bird calls while still being short enough to be computationally efficient.) with a fixed shape of (48, 128), where the width determines the number of frames and the hop size. Hann (Hanning) window of size 1024 was used with a sampling rate of 32000 because of its capability of reducing spectral leakage and improving the resolution of the spectrogram. Mel spectrogram with 48 mel bands was generated with a frequency range of 500 Hz to 12500 Hz. For MFCC, the number of MFCCs was set as 13, while the sampling rate used was 22050.

5.2 Model Training

Three models were utilized for training in this research. The first model used the feature extracted through mel spectrogram as the input and EfficientNet as the CNN architecture. The second model used mel spectrogram in combination with a custom CNN, while the third model used MFCC with EfficientNet.

Table 1: Parameters used for all three models

Parameters	Model-I	Model-II	Model-III
Batch Size	2	2	2
Epochs	30	30	30
Image Shape	(128, 128, 3)	(128, 128, 3)	(224, 224, 3)
Model	EfficientNetB3	-	EfficientNetB0
Weights	ImageNet	-	ImageNet
Filters	-	16, 32, 128, 256, 512	-
Kernel Size	-	(3,3)	-
Max Pooling Pool Size	-	(2,2)	-
Number of dense layer neurons	-	1024, 256, 41	-
Classification Classes	41	41	41
Learning Rate	0.0001	0.0001	0.0001
Loss Function	Categorical Cross Entropy	Categorical Cross Entropy	Categorical Cross Entropy
Optimizer	Adam	Adam	Adam
Fraction of the input units to drop (dropout)	0.2	-	0.2
Activation Function	ReLU and Softmax	ReLU and Softmax	ReLU and Softmax
Total parameters	11,791,192	3,892,297	67,037,364
Trainable parameters	11,703,889	3,890,409	66,995,341
Non-trainable parameters	87,303	1,888	42,023

The parameters used for model training on all three models are shown in the Table1. For a small dataset, 30 epochs were enough for the models to capture relevant patterns as the early stopping criteria were met in 15 epochs.

5.3 Hyperparameter Optimization

Before applying the genetic algorithm, several parameters needed to be considered such as population size, number of generations, crossover probability, and mutation probability. Larger population sizes can lead to more diverse solutions but require more computational resources and time. Smaller population sizes are faster but may have a harder time finding good solutions. In this project, a population size of 10 was used due to limited computational resources. The number of generations was set to 30. Crossover probabilities between 0.6 and 1.0 can be effective, with values around 0.8 often working well. Mutation probabilities are typically smaller, ranging from 0.001 to 0.01. For this project, a crossover probability of 0.8 and a mutation probability of 0.005 were used.

Table 2: Hyperprameter Search Space

Parameters	Values
Learning rate	0.00001 - 0.1
Dropout	0.1, 0.2, 0.3
Number of filters	32, 64, 128
Filter size	3, 5, 7
Activation function in hidden layers	ReLU, Sigmoid, Tanh

The hyperparameters search space used are shown in the Table 2. The best hyperparameters were found to be learning rate = 0.0001, dropout = 0.1, filter size = 3 and activation function = ReLU.

5.4 Results

The Table 3 shows a comparison of three different models used.

Table 3: Model Comparison

Models	Precision	Recall	F1-score
Model-I	0.81	0.79	0.79
Model-II	0.66	0.66	0.64
Model-III	0.72	0.73	0.71

From here onwards, the discussion will be on the best model i.e the first model. The Figure 6 illustrates the confusion matrix of the model.

The model had a high false negative rate for Grey Treepie, with a recall of 59%, where 12 actual Grey Treepies were mistakenly classified as Rufous Treepies. One possible reason for this misclassification is that Grey Treepies and Rufous Treepies belong to the same family, making it challenging for the model to differentiate between them. Additionally, Grey Treepies are known for their varied and complex calls, which may include mimicked sounds, and it's possible that the data was collected when Grey Treepies mimicked the vocalizations of Rufous Treepies, leading to misclassification.

The Figure 7 shows the graphical plot of the model's training and validation accuracy up to 30 epochs. While training, due to early stopping the training stopped after 15 epochs. Here, the accuracy seems to be increasing with the increment of epochs and remains stagnant afterward. Hence, accuracy of nearly 85% for the training dataset and 79% for the validation dataset was acheived .

Similarly, Figure 8 illustrates the Model loss which is Categorical Cross Entropy Loss and it was observed that the loss was reduced to as low as 0.2 for the training dataset and 0.8 for the validation dataset.

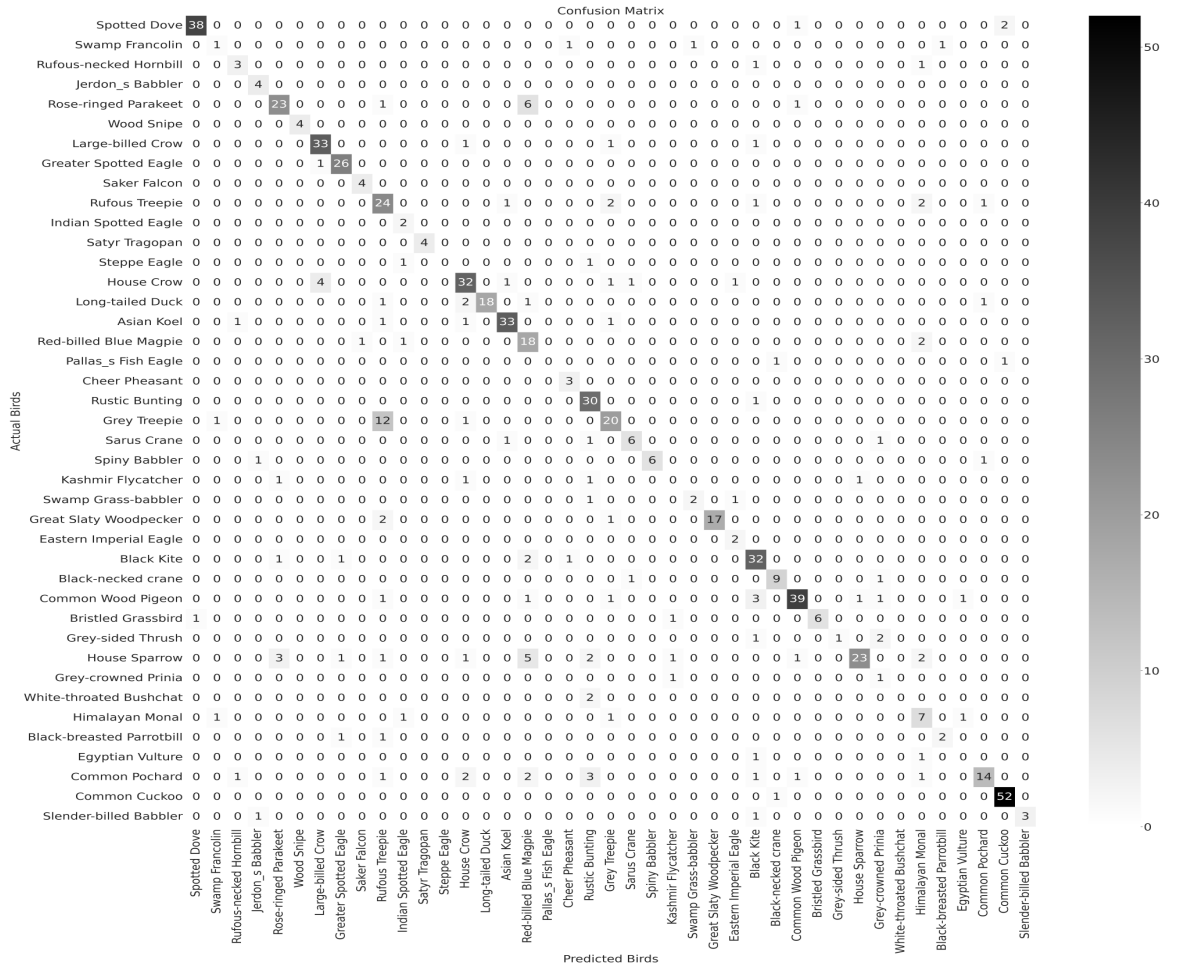

Figure 6: Confusion Matrix of Model-I

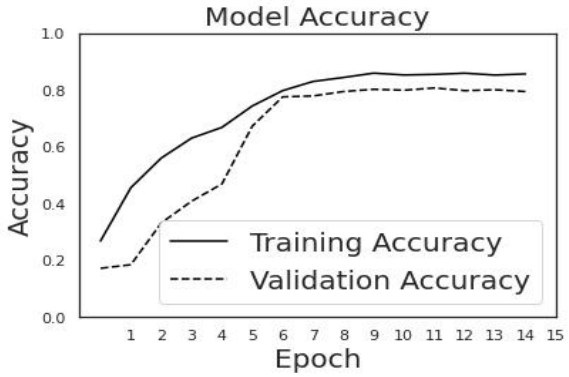

Figure 7: Accuracy plot of best model

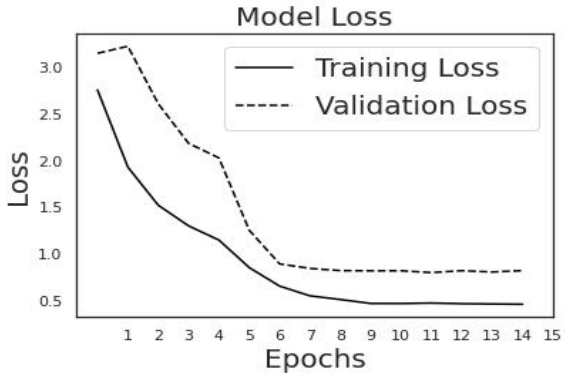

Figure 8: Categorical Cross Entropy loss of best model

6. Conclusion

The loss of a single species can disrupt the ecosystem's balance and biodiversity, emphasizing the importance of conservation efforts. This project proposes a machine learning-based system to assist in identifying endangered bird species through audio input which is useful in protected areas such as wildlife reserves and national parks. The system consists of three machine learning algorithms that achieved good accuracy in classifying 41 different bird species. Model I, using Mel Spectrogram and EfficientNet, achieved an F1-score of 79%, while Model II, using Mel Spectrogram and Custom CNN, achieved 64%, and Model III, using MFCC and EfficientNet, achieved 72%. Although improvements can still be made, these results demonstrate the potential of machine learning in supporting conservation efforts to preserve endangered bird species.

Acknowledgments

The authors would like to express there deepest gratitude to the Department of Electronics and Computer Engineering for all the inputs that were provided to carry out this study

References

[1] Iucn red list of threatened species.

[2] Hannah Ritchie and Max Roser. Biodiversity. *Our World in Data*, 2021. https://ourworldindata.org/biodiversity.

[3] Birdlife data zone. *Datazone.birdlife.org*, 2022. http://datazone.birdlife.org/country/nepal.

[4] Andréia Marini, Jacques Facon, and Alessandro L Koerich. Bird species classification based on color features. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pages 4336–4341. IEEE, 2013.

[5] Marcos V Conde, Kumar Shubham, Prateek Agnihotri, Nitin D Movva, and Szilard Bessenyei. Weakly-supervised classification and detection of bird sounds in the wild. a birdclef 2021 solution. *arXiv preprint arXiv:2107.04878*, 2021.

[6] Agnes Incze, Henrietta-Bernadett Jancso, Zoltan Szilagyi, Attila Farkas, and Csaba Sulyok. Bird sound recognition using a convolutional neural network. *2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)*, 2018.

[7] Fabian Robert Stöter. Automatic bird audio identification using convolutional neural networks and mel-spectrogram features. 2018.

[8] Fabian Robert Stöter. Bird sound recognition using deep convolutional neural networks and mel-frequency cepstral coefficients. 2017.

[9] Xueli Xiao, Ming Yan, Sunitha Basodi, Chunyan Ji, and Yi Pan. Efficient hyperparameter optimization in deep learning using a variable length genetic algorithm. *arXiv preprint arXiv:2006.12703*, 2020.

[10] Tanveer Singh. Mfcc's made easy, Jun 2019.

[11] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.