**Bird Species Detection Through Sound**

1. **Introduction**

Detecting birds from their call helps scientists to identify population diversity and is considered important for the preservation of ecosystem health. With the growing impact of climate change, bird populations are expected to fluctuate in size and distribution. Therefore, it is crucial to survey bird species. For instance, Hawaii has lost 68% of its bird species over the years, the consequences of which can harm the food chain.

Manual data collection is inefficient as it was time-consuming and sometimes can be physically challenging for scientists to access the sound of wild birds whose habitats are on an isolated island of the highest altitude. With the physical monitoring complex, scientists have turned to sound recordings knowns as bioacoustics monitoring which provide low labor and cost-effective strategy for studying endangered bird species. I used the dataset provided by Kaggle and inserted the MFCC of the audio file into a different machine-learning model after some preprocessing in the dataset to get better accuracy.

1. **Proposed Model**

Extracting deep features based on image representation of sounds has been very effective while applying it to the audio classification of bird calls. Past research has found that CNN models are the most common approach in bird call detection as features can be effectively extracted from spectrogram and classified as images.

I experimented with different machine learning models, but I got 94% of accuracy with the convolutional neural network. I converted audio snippets into a spectrogram and then into Mel-frequency cepstral coefficient (MFCC). I save 13 coefficients of MFCC in JSON life along with its label and insert them into CNN. First, I built a training model from a baseline CNN architecture of just three convolutional layers, then improved the model with hyperparameter tuning in the learning rate, number of hidden layers and epochs, and dropout rates for regularization.

Datasets and Preprocessing

Datasets are in Kaggle for the competition which was provided by an association of different institutes The Cornell Lab of Ornithology’s K.Lisa Yang Center for Conservation Bioacoustics, Google Bioacoustics Group, LifeCLEF, Listening Observatory for Hawaiian Ecosystem, Bioacoustics lab at the University of Hawaii at Hilo and Xeno-Canto. All these institutes

Are working to find out the preservation method to stop the extinction of birds in Hawaii and experiment same preservation measures in other parts of the world to protect endangered birds. Datasets have recordings of 152 species found in Hawaii.

Initially, I took 4 classes on birds. In I have 960 audio files of thirty-second recordings. Audio\_files were in .ogg format. So, I converted the .ogg format into waveform (WAV) using the MediaHuman Audio Converter application. Again, after converting the waveforms into spectrograms, I converted the spectrograms into MFCC and saved 13 coefficients of MFCC into a JSON file along with its label and inserted it into a different machine learning model.

1. **Experiments**

Datasets have 960 audio files with 30-second-long recordings. Firstly, I segmented each audio file into 3 segments of 10 seconds and then converted them into MFCC and insert the array of MFCC into the different machine learning model.

Model 1: Simple neural network with 3 hidden layers. I kept 512, 256, and 64 neurons in

respective layer with 0.3 dropout probability and regularization of 0.001 in each layer.

Model 2: Convolutional neural network with 3 convolution layers with Kernel size 3 in 1st and 2nd

layer and size 2 in 3rd layer. I used max-pooling for downsampling and batch

normalization in each convolution layer. One fully connected dense layer in the

bottom with 64 neurons and a dropout probability of 0.3 to avoid overfitting. And

output dense layer with 4 neurons.

Model 3: Recurrent neural network with 2 long short-term memory layers. Each LSTM layer with

64 units. Then fully connected dense layer with 64 neurons and a dropout probability

of 30%. And output layer with 4 neurons.

In each model, I used Adam optimizers with a learning rate of 0.001 and sparse categorical cross entropy as a loss function. Each model performs best with 64 batch size giving high accuracy while running for 100 epochs. I used SoftMax activation in the output layer of each model and Relu activation in other layers.

4. Results and Discussion

Model 1 (Multi perceptron neural network)

Graphical user interface

Description automatically generated

Model 2 (Convolutional Neural Network)

Graphical user interface, application, Word

Description automatically generated

Model 3 (Recurrent Neural Network with LSTM)

Graphical user interface, application

Description automatically generated

I got 94% of test accuracy from model 2 where I used a convolutional neural network. Firstly, I used 5 seconds of audio in CNN which gave me 82% of test accuracy with a case of overfitting afterward applied dropout and regularization in the hidden layer which increase the accuracy to 90%.

Again, I tried the same model with 10 seconds of audio recording then I ended up getting 94% test accuracy with the optimal fitting of the model. Model 1 (Simple neural network) and Model 3 (RNN) got an accuracy of 73% and 86% respectively.

5. Conclusion/Future Work

After reading past research and working with Kaggle datasets, I concluded that the audio dataset performs well in Convolution Neural Networks. In the future, I am going add more classes of bird audio to my network and same time work on making the model architecture strong enough to handle 152 categories of birds also audio files in datasets are not enough for classification. Therefore, I am planning to work in data preprocessing like data augmentation.