2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT) | 978-1-6654-1005-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICICICT54557.2022.9917735

Classification of Bird Species using Audio processing and Deep Neural Network

Sachin Aggarwal,
Department of CSE, Amity School of Engineering & Technology,
Amity University,
Noida, India,
sachinaggarwal1296@gmail.com

Smriti Sehgal,
Department of CSE, Amity School of Engineering & Technology,
Amity University,
Noida, India,
smriti1486@gmail.com

***Abstract*— Ornithology deals with the methodological study of birds and it consists of knowledge of birds and all other things which relates to them. The people who work in this field are knowns as Birders. One of their main tasks is to find and track some rare species of birds. Classification of a bird species can be a very challenging task and especially based on the sound they produce. Classification of bird species based on audio can be very helpful to find some rare birds as it will tell the species of the bird before actually tracking it and this can save a lot of time and efforts. In order to reduce this time ands efforts a lot of researchers have proposed different methods and techniques like machine learning, deep learning and wavelet study to classify the species of a bird. Some of these techniques are VGG16 and ResnetV2, Multilayer Perceptron, Naïve Bayes, and J4.8 Decision tree, Convolution Neural Network and many more. Along with this they have used Audio feature extractors like spectrogram, Inverse Short Time Fourier Transform and Mel-frequency cepstral coefficient. These techniques are discussed in detail in the related work section of this paper. In order to classify the species of a bird in this paper we have used several preprocessing techniques like conversion of .mp3 audio file to .wav audio files, feature extraction, class balancing also we have used one approach to minimize the time to pre process these audio files and we have also used to the Mel-frequency cepstral coefficient to extract features from these audio files and finally we have used a Deep Learning based Classification model to classify and predict the species of a Bird by analyzing the sound produced by that bird.**

Keywords—LSTM (Long Short-Term Memory), IoT (Internet of Things), RNN (Recurrent Neural Network), Accuracy, Class, Dataset, Attribute Selection, TensorFlow

I. Introduction

Birders are people who work in the field of Ornithology in which they study about different birds, and everything related to them. This study requires a lot of efforts to find out the location and species of different bird around the world and it is done to track the changes in the ecosystem. This study of birds can help a lot to understand how and what changes are happening in our environment. One of the main task for Birders is to find the species and location of some rare birds which can be a very difficult task as it requires a lot of time and efforts to find these birds in places like forests and deserts. One of the possible solution for this problem was to perform image classification to identify the species of a bird by its image. But this concept was not efficient at all since most of the time these birds are hiding at some place due to which even a human finds it very difficult to locate these birds and since is very hard to locate them it is very difficult to take picture of these birds.

Another solution for this problem is that to use the sound of the birds to find out their species. This concept is based on the fact that same species of birds makes very similar sound which is different from other species. This makes it possible to classify a bird species by its sound. Also, when e are processing sound we don’t actually need to locate a bird as sound travels in all directions it is not possible for a bird to hide its sound. Birdwatching or Birding deals with finding and observation rare birds and it is done by a lot of people as a hobby they may or may not have some expert knowledge. This makes it very difficult for them to find rare birds but with the help of Audio Processing and Artificial Intelligence they can easily do this task with expert knowledge where the AI will serve as an expert.

A lot of researchers have created so many different models, techniques and methods like Machine Learning, Deep Learning and wavelet study to classify the species of a bird. Some of these techniques are VGG16 and ResnetV2 [1][2], Multilayer Perceptron [3], Naïve Bayes, and J4.8 Decision tree [4], Convolution Neural Network [5] and many more. Along with this they have used Audio feature extractors like spectrogram, Inverse Short Time Fourier Transform and Mel-frequency cepstral coefficient. These techniques are discussed in detail in the second section of this paper where we have discussed the methodology used by them and the corresponding results which they got from their approach.

In our work we have worked on creation of a Deep Learning model which can help to classify the species of a bird among 264 number of different species of bird. In our work we have used the Tensorflow with keras to create a sequential deep learning model to classify the species of a bird. This proposed model is discussed in detail in the fifth section of this paper. In our work we have also performed a lot of data pre-processing techniques like conversion of .mp3 audio file to .wav audio files, feature extraction, class balancing. Apart form this we have used some approach to minimize the time to pre-process these audio files and we have also used to the Mel-frequency cepstral coefficient to extract features from these audio files. This pre-processing task is very important as we are working on audio files a we cannot use the files directly to train our model. This Data pre-processing is discussed in the fourth section of this paper and in the third section we have given a description of the dataset which we have used in our work.

II. Related Work

In [1] the authors have used two different techniques to classify the Bird audio. The first technique which they used was Deep Neural network and the other one was also Deep neural Network but in the second one they used Convolution layers. They used these techniques to create a pipeline for classifying bird speech on edge device. Apart from this they have discussed the advantages and disadvantages for both of these approaches [2]. The dataset which they used for their study consist of 24000 samples of bird audio for different species and it was taken from the Xeo-Canto. They have compared their work against some of the pretrained models like MobileNetV2, VGG16 and ResnetV2. As for the performance evaluation they have used five parameters which includes accuracy, precision and loss also it was found that the VGG19 performed best among all algorithms.

Similarly, in [3] also the authors have compared the performance of three different machine learning algorithms which are Multilayer Perceptron, Naïve Bayes, and J4.8 Decision tree algorithm. Along with these algorithms they have used Mel-frequency cepstral coefficient to improve the accuracy of these algorithms [4]. In their study they used five different parameters to evaluate the performance of these algorithms and these parameters are False Alarm Rate, Accuracy, and True Negative Rate. Their results shows that the J4.8 Decision Tree algorithm was able to perform best among all with 78.40% of accuracy.

In [5] a Convolution Neural Network based approach was used to create a model to classify bird audio. In their work the have used the Expanded Convolution Neural Network. This model was made to classify only four species of bird. The dataset used for the training for the training of this model was provided by Stanford Biology department which includes 10000 bird audio files for 14 different species of birds [6]. This work was named as CS230 project by the authors. The authors selected the F1 dev score as the evaluation parameter for the work and they were able to achieve 0.886 F1 score for their classification model.

Similarly, in [7] the authors have used a baseline Convolution Neural Network model which serves as a training model with two convolution layers in it. The dataset used in their work contains 4327 audio files for 101 species of birds which can be seen in United States. Each of this file was converted into a spectrogram and then it was used in the training process [8]. In this work the authors have also performed the hyperparameter tuning which includes learning rate epochs, drop rate and several other parameters. With the help of this approach, they were able to get 98% of training and 67% of testing accuracy.

In [9] the authors have used a very similar approach except the spectrogram in their work they have used a Inverse Short Time Fourier Transform also the CNN model using in this work make use of NaK architecture with a total of 6 layers. Along with this they have compared the performance of ResNet and AlexNet on five different performance evaluation parameter which include F-Score [10], Specificity Recall and Accuracy. In this study it was found they the ResNet was able to perform better as compared to the AlexNet and the accuracy for ResNet was 90%.

In [11] the authors have combined several techniques to create a robust model to classify different environmental sound. In their work the have used five Convolution Neural Network Models which are pre-trained and the have used four representation techniques to extract features from the audio files. Apart from this they have also used six data augmentation methods to increase the size of the dataset. The final model was trained and tested on 3 datasets which includes ESC-50 [12] dataset, bird sound and cat sound dataset. With the help of this approach the were able to achieve 90% accuracy for cat dataset, 97% for bird dataset and 88.65% accuracy for ESC-50 dataset.

III. Daset Description

The dataset which we have used to train our model is taken from “Kaggle.com”. This dataset contains 21375 number of audio file which is the audio sample of 264 different bird species. Along with this we have five .csv files as well and among these files the most important file is “train.csv” infarct we will be specifically using only “train.csv” file in our work. These 21375 audio files are divided into subfolders which are the “ebird-code” assigned to them each bird species is assigned one unique code and since we have 264 nomber of species the number of folders for ebird-code is also 264 which means one folder for each class. The train.csv file contains a lot of information regarding this dataset like ebird-code, channel (stereo or mono), file name, species, location, date and many more. Out of all these attributes we are only interested in three attributes which are ebird-code, file name, and species. The reason for using only these three attributes is that we can use ebird-code and file name can provide the appropriate path to extract the audio file from the storage (since these files are divided into subfolder with ebird-code as subfolder name we need these two attributes to get the file from correct path) and species attribute is needed to map the corresponding audio file and bird code to the corresponding species of that bird.

IV. Data Pre-Processiong

Before we start with the actual training of our model, we need to perform a lot of data preprocessing task which can help a lot to increase the efficiency of the classification model. The first task in data preprocessing is to convert all the audio files into .wav format. The reason to this task is very important as the original audio files is of .mp3 format which is not supportable with scipy and if we use librosa then we can process it but the accuracy will be much less as compared to the .wav file. SO, in order to solve these problems in initial stage only we have converted whole data into .wav format. The implementation of our work was carried out on Google colab so in order to use we had to upload whole data on google drive and later we mounted the drive to get the dataset on Colab notebook. But the main problem here is that the size of the dataset was 23.5 GB which is much higher then the storage limit of Google Drive and we can mount only one Google Drives in one Colab notebook. To solve this problem we have simply divide the dataset into two parts and then we have saved it in two different drives in dataset folder. After this we have created a shortcut of both of these folders in a third drive. By creating shortcut of those folders in other drive helps use to access the content of those two folders without actually mounting those two drives.

Since we have divided the dataset into two parts, we have to process them separately. The first file will contain the result of processing of 9894 audio files and the second file will contain the result of remaining audio files which is 11481 records. These output for the processing of these audio file is an array of size 40 which represents the 40 features which we have extracted from these audio files. This processing of audio files takes a huge amount of time so is the reason it s not possible to perform feature extraction again and again. To solve this problem we have processed these audio files one time and the resultant arrays were saves in a list which was combined with the corresponding class label and later this list was converted into a data frame of 41 attributes and then that data frame was finally saved into a .csv file. By using this approach, we will not need to perform data preprocessing again and again we will simply perform it once and save the results into a .csv file and late we can simply import that .csv file whenever and wherever we need it. In figure 1 we have shown two graphs in which the first one tells us the number of records for which the bird was seen or not scene. In the second graph we have shown the number of records in the mono and stereo categories which represents the channel type and we can clearly see the difference here i.e. 3364 number of time it happened that the sound of the bird was recorded but it was not seen.

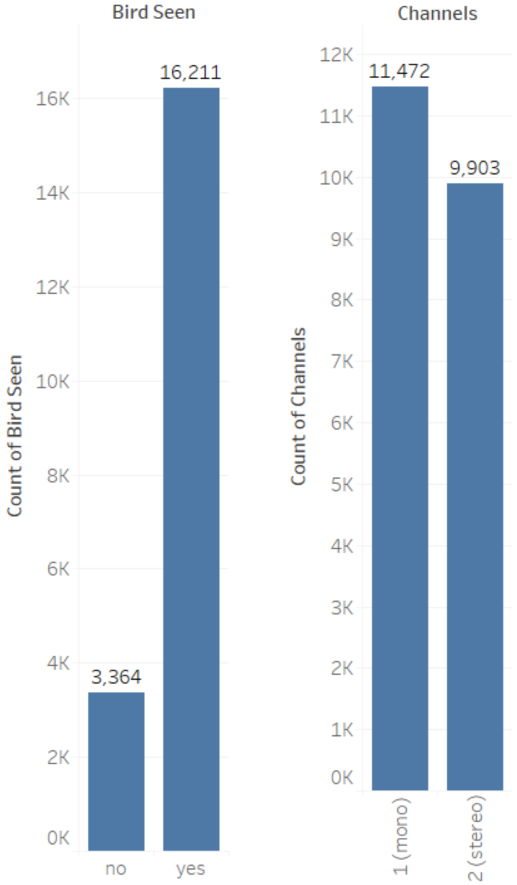

Fig. 1. NO. of records for both channels and NO. of records where the birds were scene or not.

In our work we have used the librosa library to process these audio files. The reason to use librosa over scipy is that the dataset which we have it contains a mixture of mono channel (single source of sound) and stereo channel (multiple sound source). Scipy is not as efficient as compared to librosa when it comes to processing of audio file with stereo channel and on top of that librosa provides higher bitrate as compared to scipy. Higher btrate helps to represent audio signal more accurately. So given all the points related to channel type and bitrate we decided to use librosa over scipy to process these audio files. The feature extractor which we have used in this work is MFCC which is Mel-Frequency Cepstral Coefficients [13] which is one of the most efficient feature extractors from any audio signals. Since we divided our data into two parts we got 2 .csv file which was later combined to get one file with all the processed audio file.

After receiving the final output .csv file we have to perform a very important step which is to perform class balancing [14]. In our dataset we have 134 classes whose number of records was found to be having 100 records and remaining 130 classes having records from 9 to 99. Clearly, we can see that there is a huge class imbalance present in our dataset and if we use it directly to train our model the that training model will be definitely skewed. One way to solve this problem is that to take only those classes whose number of records is equal to maximum number of records in a class which is 100 and since there are 134 such classes so the number of records in final dataset will be equal 13400. But there is one problem in this approach that if we use this then we will not be able to classify the remaining 130 class which is almost equal to the 50% of the total number of classes in the original dataset. This is happening because we are simply discarding all the class whose number of records is not equal to 100. This makes this approach very inefficient and so is the reason we will not be using this I our work.

In order to solve this problem, we have used a very simple approach in which the number of records for each class was reduced to 9. The number 9 is the minimum number of records in a class so with this we can make sure we have equal number for records for each class and this will make sure that there is no class imbalance, and this will provide us the surety that the training model is not skewed. Next we have used the first come first serve approach to decide which record will be selected in the final dataset for class balancing process. The reason to choose this approach is that this approach will allow us to make sure that the dataset which we are using not biased in any way. Finally, after performing class balancing, we receive a new data frame with 2376 number of records in it and this data frame is converted into a .csv file. Also, this is the final .csv file which will be used for the training of our classification model. This whole data preprocessing task was performed once and now we can use this .csv file whenever we can simply import this csv file to train any kind of classification model on it. In figure 2 we have a graph which shows the number of records in each species of bird before performing class balancing. Here in the graph we can clearly see that the difference in number of records in these classes.

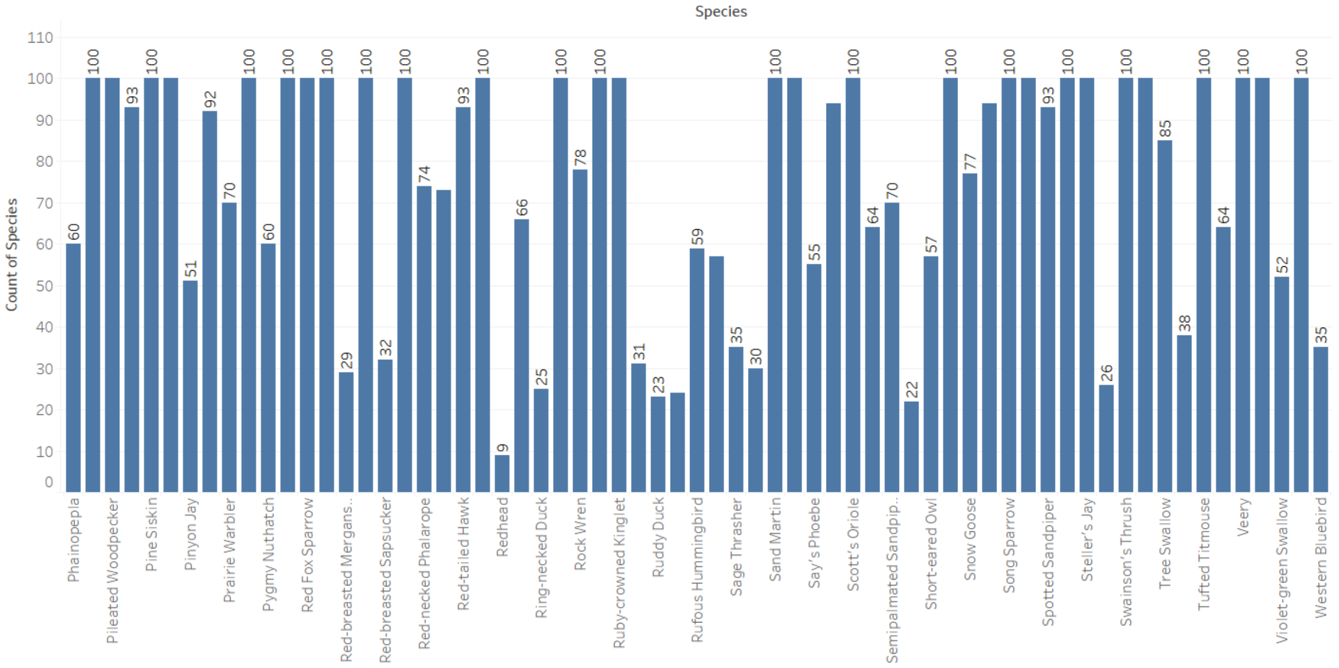

Fig. 2. The above image contains the block diagram of data pre-processing which shows different stage and output of those stages.

Finally in figure 3 we have shown a block diagram for the whole data-preprocessing task which was discussed above. In this we can see the flow of data i.e. how the original data will be transformed in the final.csv file for training. In this block diagram we have taken the data or the output of each preprocessing technique used in this work and on arrows we have mentioned the preprocessing technique used.

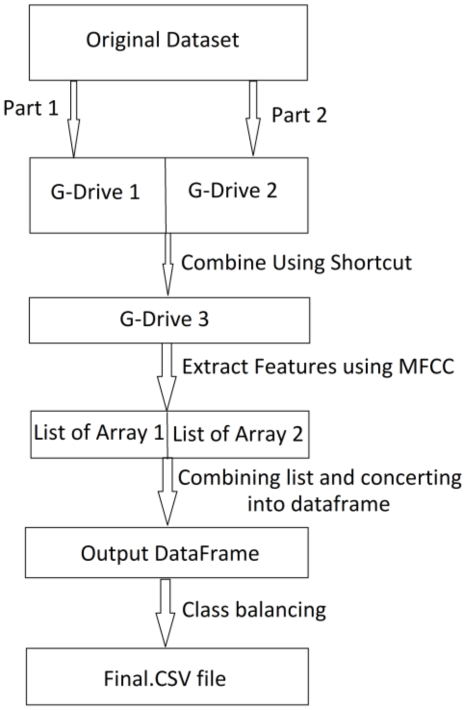

Fig. 3. Block diagram of Data Preprocessing which shows different stage and output of those stages.

V. Proposed Work

After performing the data preprocessing task, we will be left with a .csv file which is just like any other csv file containing classification data with independent variables and discrete dependent variable. In our csv file we have 41 attributes and 2,376 records. In this dataset we have 40 independent attributes and 1 dependent attribute which contains 264 number of classes, and each class has 9 records for it. This work was implemented on Google Colab where we used librosa matplotlib tensorflow, pandas, numpy, scipy, and several other libraries [15]. Like any normal classification task pandas was uses to import the data from csv file and numpy was used to divide it in to array (one array with 40 number attribute consisting of dependent variable and the other array with 1 attribute of class label or discrete dependent variable). Before performing any training on this data label encoder was used to convert the non-numerical data into numerical data in which each class or species value was converted into an array of 264 values where each array contains only single value as 1 and rest 263 values as 0. Later with the help of same library these arrays were converted in their respective class labels so that it will be easy to understand the predictions made by the training model.

Finally, we have used the tensorflow and keras library to create our classifier in which we have used the sequential model of Keras library [16]. The total number of layers in this model is 8 and first layer is our input layer with 40 neurons (equal to the size of input vector) and we have used “relu” activation function with input shape as 40 (equal to the number of independent variables in our dataset). After this we have six hidden/dense layers which contains 792 number of neurons (three time the number of classes in the dataset) with “relu” activation function. The drop rate for all of the layers was taken as 0.15 to avoid any kind of overfitting. Finally in the last output layer we have taken 264 neurons (equal to the number of classes in the dataset) with “softmax” activation function (for multi-class classification). In the end for the compilation of this model we have use the “sparse categorical crossentropy” as the loss function, “Adam” as the optimizer for this model and “sparse categorical accuracy” as the metrics. In this compilation layer it is necessary to select “sparse categorical crossentropy” as loss function and “sparse categorical accuracy” as metrics to be able to perform multi-class classification. Finally, we have used 1430 number of epochs and batch size as 32. In the fig. 4 below we have shown a block diagram of the overall training process discussed above. In this block diagram we have mentioned the data or the output of each process in the block and the corresponding technique applied on it is mentioned the directed edges.

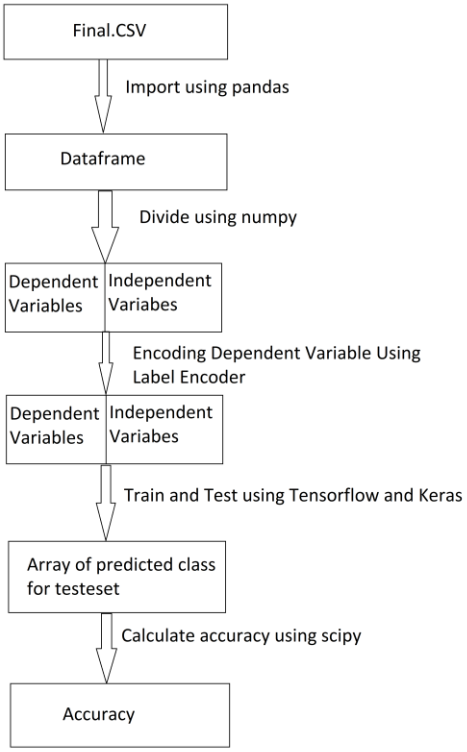
Final.CSV
Import using pandas
Dataframe
Divide using numpy
Dependent Variables
Independent Variabes
Encoding Dependent Variable Using Label Encoder
Dependent Variables
Independent Variabes
Train and Test using Tensorflow and Keras
Array of predicted class for testeset
Calculate accuracy using scipy
Accuracy

Fig. 4. Block diagram of training process which shows different stages involved and the output of those stages.

VI. RESULTS

In order to test the performance of this work three performance evaluation parameter was used which are Accuracy F1-Score and Specificity. Also, we have taken two set of data in which first one has five number of classes and the second one has all 264 classes. This is done because while dealing with a multiclass classification we cannot calculate the overall F1-score and Precision they are calculated for every class separately and it is normally not possible to show the results of all 264 classes so for this reason we have done validation on 2 dataset in which the dataset with 5 classes will have all detailed results and for the other one we have calculated the average of all the parameters to show their results.

In the table 1 we have shown the confusion matrix which is obtained for the dataset with 5 classes. In this validation set we have a total of 135 number of records. In this we can note that the class Pine Siskin is having maximum number of correct predictions whereas maximum number of wrong predictions is observed in Song Sparrow class. Also, it was observed that the overall accuracy for this dataset was found to be 0.9407407%.

TABLE I. RESULTANT CONFUSION MATRIX FOR THE FIRST DATASET

	Veery	Song Sparrow	Sand Martin	Scott’s Oriole	Pine Siskin
Veery	25	0	0	0	0
Song Sparrow	0	22	0	2	0
Sand Martin	0	2	27	0	0
Scott’s Oriole	2	1	0	25	0
Pine Siskin	0	1	0	0	28

After this in table 2 we have shown the accuracy, F1 score and the Precision for the first dataset. As discussed earlier that for a multiclass classification we need to calculate these parameters for each class separately and same information can be observed in the Table 2 as well.

TABLE II. PERFORMANCE EVALUATION RESULTS FOR FIRST DATASET.

Class	Accuracy	F1 Score	Precision
Veery	98.52%	0.96	1.0
Song Sparrow	95.56	0.88	0.92
Sand Martin	98.52	1.0	0.93
Scott’s Oriole	0.89	0.93	0.89
Pine Siskin	0.97	1.0	0.97

From the above table we can note some points in which first we can note highest accuracy is obtained for the class Veery and Sand Martin. Also, we cannot that the class Pine Siskin and Sand Martin got maximum F1 score and class Veery got maximum precision among all. After this we have calculated the Average F1 score and Precision for both datasets. These values for the small dataset was found to be 0.954 average F1-Score and 0.942 average precision for the full dataset the accuracy was 0.92592, average F1-Score was 0.93621, and the Average Precision was 0.92375.

VII. CONCLUSION AND FUTURE DIRECTION

The identification of bird species can help a lot for birders and birdwatchers to find rare birds and not only that they can also examine the environmental conditions with this data. We have seen how Artificial intelligence and Machine Learning can help to perform this task very efficiently and even a person with very less expertise can identify the species of a bird just by its sound. In out study we found that the proposed model was efficient for the classification of 264 different bird species. The accuracy for this proposed model was found to be 92% which is very acceptable given the number of species.

As for the future work we can a better approach to perform class balancing on our data as the current class balancing approach in not efficient enough. In the current approach the minimum number of records for a particular species was found to be 9 and just because of this the number of records for each class was reduced to 9 even though there were a lot of classes with 100 record du to this reason we lost a lot of data and to solve this problem we can use data augmentation to increase the number of records for those classes whole total records are very less. This will provide more data to our training model, and it will help to boost the performance. The other technique for class balancing which can be used outlier detection through clustering. The current approach follows the first come first serve methodology to select records for class balancing but with this approach a lot of outliers can enter in our training data which can lead to poor accuracy of training model. If we use clustering approach in which we make a cluster for each class then we can remove the records which the far from the centroid of the cluster which will remove all the outliers from our training data and hence provide better accuracy. Apart from this we can also use some pretrained models with transfer learning to perform classification task and if we combine this with other approaches discussed above then we can definitely get a very highly efficient model to predict the species of a bird with its sound.

REFERENCES

[1] Behr, D., wa Maina, C., & Marivate, V. (2021, September). An empirical investigation into audio pipeline approaches for classifying bird species. In 2021 IEEE AFRICON (pp. 1-6). IEEE.

[2] Ghani, B., & Hallerberg, S. (2021). A Randomized Bag-of-Birds Approach to Study Robustness of Automated Audio Based Bird Species Classification. Applied Sciences, 11(19), 9226.

[3] Mehyadin, A. E., Abdulazeez, A. M., Hasan, D. A., & Saeed, J. N. (2021). Birds Sound Classification Based on Machine Learning Algorithms. Asian Journal of Research in Computer Science, 1-11.

[4] Rajan, R., & Noumida, A. (2021, June). Multi-label Bird Species Classification Using Transfer Learning. In 2021 International Conference on Communication, Control and Information Sciences (ICCISc) (Vol. 1, pp. 1-5). IEEE.

[5] Ablikim, U., Ngoi, J., & Liu, P. Using CNNs to Recognize Bird Species by Song.

[6] Gunawan, K. W., Hidayat, A. A., Cenggoro, T. W., & Pardamean, B. (2021). A Transfer Learning Strategy for Owl Sound Classification by Using Image Classification Model with Audio Spectrogram. International Journal on Electrical Engineering and Informatics, 13(3), 546-553.

[7] Cheng, S., & Wang, J. Detection of Bird Species Through Sounds.

[8] Mohanty, R., Mallik, B. K., & Solanki, S. S. (2020). Automatic bird species recognition system using neural Network based on spike. Applied Acoustics, 161, 107177.

[9] Anand, R., Shanthi, T., Dinesh, C., Karthikeyan, S., Gowtham, M., & Veni, S. (2021, June). AI based Birds Sound Classification Using Convolutional Neural Networks. In IOP Conference Series: Earth and Environmental Science (Vol. 785, No. 1, p. 012015). IOP Publishing.

[10] Pahuja, R., & Kumar, A. (2021). Sound-spectrogram based automatic bird species recognition using MLP classifier. Applied Acoustics, 180, 108077.

[11] Nanni, L., Maguolo, G., Brahnam, S., & Paci, M. (2021). An ensemble of convolutional neural networks for audio classification. Applied Sciences, 11(13), 5796.

[12] Gupta, G., Kshirsagar, M., Zhong, M., Gholami, S., & Ferres, J. L. (2021). Recurrent Convolutional Neural Networks for Large Scale Bird Species Classification.

[13] Jung, S. Y., Liao, C. H., Wu, Y. S., Yuan, S. M., & Sun, C. T. (2021). Efficiently classifying lung sounds through depthwise separable cnn models with fused stft and mfcc features. Diagnostics, 11(4), 732.

[14] Sielenou, P. D., Viallon-Galinier, L., Hagenmuller, P., Naveau, P., Morin, S., Dumont, M., ... & Eckert, N. (2021). Combining random forests and class-balancing to discriminate between three classes of avalanche activity in the French Alps. Cold Regions Science and Technology, 187, 103276.

[15] Othmani, A., Kadoch, D., Bentounes, K., Rejaibi, E., Alfred, R., & Hadid, A. (2021, January). Towards robust deep neural networks for affect and depression recognition from speech. In International Conference on Pattern Recognition (pp. 5-19). Springer, Cham.

[16] Mustika, I. W., Adi, H. N., & Najib, F. (2021, August). Comparison of Keras Optimizers for Earthquake Signal Classification Based on Deep Neural Networks. In 2021 4th International Conference on Information and Communications Technology (ICOIACT) (pp. 304-308). IEEE.