Bioacoustic monitoring to improve conservation efforts for endangered species.

Dr.S.Seema
HOD,
Dept. of MCA,
Ramaiah Institute of Technology,
Bangalore, India,
hod_mca@msrit.edu

Mohammed Wasim Changapur
Student,
Dept. of MCA,
Ramaiah Institute of Technology,
Bangalore, India,
thisiswasimnc@gmail.com

Abstract:- Bioacoustics is the study of sound as it relates to living organisms,particularly animals.This field encompasses a wide range of topics,including the production and reception of sounds by animals, the use of sound in communication, and the analysis of acoustic signals to better understand animal behaviour and ecology. Bioacoustic monitoring can be used to study animal populations and behaviour, track migration patterns,and detect the presence of endangered species.The conservation of endangered species has become a pressing concern due to the reduction in biodiversity and the growing threat of extinction of species. The use of bioacoustic monitoring, which provides real-time data on the existence, abundance, and distribution of endangered species, has the potential to enhance conservation efforts.

Index Terms: Bioacoustics, Audio Classifier,Convolutional Neural Network.

I. Introduction

Bioacoustics [1] is the study of animal vocalisations and sounds. By identifying target species from field recordings, bioacoustics assists wildlife monitoring, management, and conservation.
The inherent difficulties in processing bioacoustic data include target sounds that overlap, background and environmental noises, strength variations caused by different distances between the source and recorder, and variations in sound even within the same species. Lack of properly tagged datasets to train software is another problem.

For conservation efforts, deep audio classifiers have a number of benefits when used in bioacoustic monitoring. Deep audio classifiers can enable real-time monitoring of endangered species and can offer vital data on species existence, abundance, and distribution, helping conservationists to make wise management and protection decisions. Deep audio classifiers can also increase the effectiveness and precision of bioacoustic monitoring, obviating the necessity for manual analysis and broadening the scope and scale of monitoring operations.

Deep audio classifiers and bioacoustic monitoring have the ability to completely change how endangered species are conserved. Deep audio classifiers can assist in the development of targeted conservation strategies and help to the overall effectiveness of conservation initiatives by giving real-time data on species presence, abundance, and distribution. Bioacoustic monitoring has the potential to turn into an even more effective tool for conservation efforts in the future with continued development and improvement of deep audio classifiers.

II. Litrature Review

A comprehensive literature search was conducted using academic search engines, including Google Scholar and IEEE Xplore. The following keywords were used: Bioacoustics,Deep Audio classifier,Audio classification,Remote monitoring. Filters were applied to limit the search results to publications within the last five years and peer-reviewed articles.

Francisco J. Bravo Sanchez et.el [1] in his paper says that bioacoustics is a field of active research that supports wildlife monitoring, management and conservation through the identification of target species from field recordings.Autonomous recording units are now often used as a result of improvements in digital sound recording hardware and storage. These are digital sound recorders that can be placed in

the field for a short period of time (weeks to months) or permanently and collect a significant amount of acoustic data.

The paper also states that the major drawback of PAM(passive Acoustic monitoring) is that expert manual processing of audio recordings demands significant work. When millions of hours' worth of sound recordings are produced by replication at various spatial and temporal scales, it is not realistic to attempt it.

Fan Yang et.al [2] in his paper says that in order to create a data set of 264 different bird species, this article first gathers a significant amount of bird sound data. The sound data characteristic is then created using a single Mel spectrum. The classification outcome is then produced using a simple recognition model created to recognise the bird sound feature map.
In order to increase the accuracy of bird sound recognition, this paper designs a small, lightweight model. To improve the network's ability to scale the extraction of spatial information and channel information, the multi-scale feature fusion structure is first proposed, and then a PSA(pyramid split attention) module is implemented.

In contrast to human speech recognition, bird sound recognition in this study places greater emphasis on the form and function of the bird sound than on its content. Mel spectrum, which is commonly used in speech recognition systems, is chosen as the feature of the bird audio signal in order to lower the computational burden of the model and simplify the complexity of the feature fusion procedure.

Samruddhi Bhor et.al [3] in her research article suggests that the problem of recognising birds using an automated system with the usage of bird sounds can be defined as the challenge of differentiating several bird species from their recorded songs.Experts claim that unlike the bird noises used here, bird songs are more melodic and better for identifying species. The entire signal is pre-processed in order to identify the most pertinent portion of the signal and extract characteristics.

The main objective of the system is to automatically identify bird species from field recordings. In order to divide the audio picture into distinct Spectro-temporal segments, a sinusoidal detection method is required. A frequency track, which is a time series of the observed sinusoid's frequencies, is used to depict each segment. The HMM collection was utilised to represent each bird species during the experiment. These HMM sets are trained using unsupervised learning techniques. The likelihood ratio of the background model was compared to the probability ratio of the target model to make the detection.

Stefan Kahl et.al[4] in his research paper says that says that compared to less diversified systems, automated assessment of tropical soundscapes poses major hurdles, but it also offers greater potential rewards. More than three-quarters of all species and more than 90% of the world's terrestrial birds are found in the tropics, and if these ecosystems are not preserved, it will be impossible to satisfy international biodiversity targets.Despite their crucial importance, the tropics are largely ignored in research on biodiversity and ecosystem function, and conclusions from studies conducted in temperate regions are frequently misapplied to assume things about tropical systems. The abundance of bird species in places like the Amazon makes it difficult for most field observers to undertake credible avian surveys there, especially when combined with the logistical difficulties and poor viewing conditions typical of deep tropical woods.

Noumida A. et.al [5] says in his research that that in the work mentioned above, an iterative maximum likelihood technique was developed to train the individual HMMs for each species syllables. An HMM-based detector with a general model learned from all syllables was developed as a baseline system. Further it is tried to split the audio recording into time chunks, send it to CNN, and get output for each part.

In the above work a deep convolutional neural network structure called ResNet50 is used to identify different bird species with an accuracy of 60% to 72%. With a precision score of 0.686, the method in use predicts the most prevalent foreground bird species inside the audio environment using a deep learning-based neural network approach. In the proposed work, we use multiple transfer learning algorithms to identify different bird species from isolated recordings and analyse performance.

Emre Cakir et.al[6] says in his research article that higher level features that are resilient to regional spectral and temporal variations can be extracted by convolutional neural networks (CNN). Recurrent neural networks (RNNs) are effective at learning the audio signals' longer-term temporal context.In the present study, we merge these two methods into a convolutional recurrent neural network (CRNN) and use it to analyse spectral auditory data for the detection of birds.

Two phases are used in the model mentioned above. To create the sound representation, spectro-temporal features (spectrogram) are first retrieved from the raw audio recordings. The auditory features are mapped to a binary estimate of the existence of bird song in the second stage using a CRNN.By employing material consisting of acoustic features taken from a training database and the annotations of bird song activity, supervised learning is used to produce CRNN parameters.

Hasan Abdullah Jasim et.al[7] says the following in his research paper that different strategies have been used to approach the classification stage. Some methods extract features using traditional machine learning techniques before using any kind of classifier.The other approaches rely on DNNs that run from beginning to end and employ convolution to complete the feature extraction process.

In the above research paper there were various steps to the process. As part of this procedure, we produced a number of modules, each with a unique set of duties. This dissertation's practical application was carried out using Python programming. We take a deeper look at some of the intriguing possibilities that the CNN approach has provided for recognising bird species identification on both static and dynamic levels with regard to the sculpture that appears in the scene.

Agnes Incze et.al[8] in her research paper uses transfer learning to hone an already-built neural network's ability to identify bird noises. Many of these pre-trained networks have been trained to identify common features in photos. However, as images are represented as two-dimensional signals rather than one-dimensional sounds, a representational transformation is required for compatibility. In order to achieve this, spectrograms, a visual representation of the magnitude returned by the Short Time Fourier Transform, are used. (STFT). Instead of conducting a single DFT over a longer signal, the STFT is a variation of the DFT that divides the signal into partially overlapping chunks and applies the DFT to each using a sliding window.
The general ideas and approaches that went into developing the current bird sound recognition system are discussed in this research article.The offline training of the CNN with the required data and the online evaluation for a single sound are the two processes involved in system setup.

JIE XIE et.al[9] in his research paper says that three modules make up the system used to classify bird sounds in the for mentioned research paper: preprocessing, feature extraction, model development, and ensemble.The research evaluates our suggested strategy using a public dataset (CLO-43DS) given by Salamon. 43 different North American wood warblers' flight sounds are included in this collection.Each audio clip is altered and cut during the preprocessing such that it only contains one flight call. The audio clips have a 22.05 kHz sampling rate.
The authors of the dataset used a frame size of 11.6 ms with an overlap of 1.25 ms and 40 Mel bands to create the Mel-spectrograms that they provide.

Dan Stowell et.al[10] explains in his research paper background noise should be taken into account, and even minimal noise reduction can aid in downstream processing. Although it is a first step in simple noise reduction, the assumption of temporally constant background noise levels is generally incorrect for outdoor sound recordings. Some methods allow for background noise that varies subtly. The major problem, though, is resilience to strongly fluctuating noise, particularly from wind and rain, as well as from other flora.
The research paper also says that how much human participation is actually required in practice, even with ostensibly autonomous systems, is a critical question for large-scale studies and for general applicability. The manual calibration of thresholds and/or templates for each species of interest required by commonly used programmes like Raven and SongScope before usage can have a significant impact on precision and sensitivity.

III. Methodology

By analysing the noises that animals make and using that information to track and monitor their behaviour, movements, and population levels, machine learning can be used to bioacoustic monitoring to enhance conservation efforts for endangered species.

Data collection: This involves the collection of audio recordings of the target species in their natural habitat using appropriate recording equipment and techniques. The audio recordings can be obtained from various sources, including remote acoustic sensors, audio traps, or hand-held recorders.

Data preprocessing: Once the audio recordings have been collected, they need to be preprocessed to remove background noise and prepare them for analysis. This may involve filtering, noise reduction, and segmentation of the recordings into individual vocalizations.

Audio feature extraction: In order to develop a deep audio classifier, relevant features need to be extracted from the audio recordings. This may involve time-frequency analysis, spectrogram analysis, and other signal processing techniques to identify unique features and patterns in the vocalizations.

Deep learning model development: Once the audio features have been extracted, a deep learning model, such as a convolutional neural network (CNN), can be developed to classify the vocalizations based on their acoustic features. The deep learning model can be trained using a dataset of labeled vocalizations.

Model evaluation: The performance of the deep audio classifier needs to be evaluated using various metrics, such as precision, recall, and accuracy. This can be done using a separate validation dataset of labeled vocalizations.

In general, the methodology for the study on "Bioacoustics monitoring to improve conservation efforts for endangered species using deep audio classifier" entails the gathering of audio recordings, feature extraction and preprocessing, development and evaluation of a deep audio classifier, real-time monitoring, and the use of monitoring data to inform conservation strategies.

IV.Implementation

Data Preparation
Audio files provided in the training and test sets are not only of different duration and quality, they also have different formats regarding sample rate, bit depth and number of channels. In the

first step, two data sets are formed with homogeneous file properties.
For the first set all files are resampled to 16 kHz followed by normalization to a maximum signal amplitude of -3 dB. The training set is augmented by additionally extracting 381 audio files using the time coded annotation of species in the metadata of the newly provided soundscape validation set.

Via segmentation, the audio content of each training file is separated in signal and noise parts. Segmentation is done in frequency domain applying image processing methods like median clipping [8] and further morphological operations on the spectrogram image to extract individual sound events.

Training Setup

Different models pre-trained on the Xeno-cento data set are fine-tuned with spectrogram images representing short audio chunks. For audio file reading and processing the PySoundFile and librosa python packages are utilized. The basic data loading pipeline can be summarized as follows:

- extract audio chunk from file with a duration of ca. 5 seconds
- apply short-time Fourier transform
- normalize and convert power spectrogram to decibel units (dB) via logarithm
- convert linear spectrogram to mel spectrogram
- remove low and high frequencies
- resize spectrogram to fit input dimension of the network
- convert grayscale image to RGB image

Feature Generation:

The success of the neural network depends on the creation of high-quality input characteristics. Three main phases can be seen. First, we determine which parts of the sound file are noise or silence and which parts are associated with a bird singing or calling (signal sections). The spectrogram is then computed for both the signal and noise parts. Third, we separate each part's spectrogram into equal-sized segments. Then, we may add a chunk from the noise spectrogram to each chunk from the signal spectrogram to make it a unique sample for training and testing.

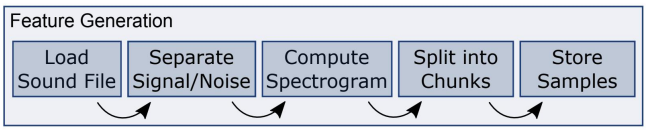

Figure 1 The Process of Feature Extraction

Separate Signal/Noise:
We compute the spectrogram of the entire sound file before splitting it into a signal and a noise part. Notably, the spectrograms in this paper were all calculated using the same method.
A Hanning window function is used to perform a short-time Fourier transform (STFT) on the signal first.The STFT's amplitude logarithm is then calculated. The signal/noise separation is an exception to this rule, though, since we divide each element by the highest value rather than taking the logarithm of the amplitude in order to bring all values into the range [0, 1]. Now that we have the spectrogram, we can search for the signal/noise intervals.
We adhere to the signal part very completely. First, we choose every spectrogram pixel that is three times larger than both the row median and the column median. Since a high amplitude typically indicates a bird singing or calling, this gives us all the crucial information from the spectrograms. All other pixels are set to 0, and these pixels are set to 1. To remove the noise and combine segments, we use a binary erosion and dilation filter.

For the noise component, we use the identical procedures, but instead of choosing pixels that are three times larger than the row and column median, we choose all pixels that are 2.5 times larger. After that, we carry on as indicated before but invert the outcome at the end.

Due to the fact that a single column should never belong to both the signal and noise parts, we constructed our method to choose all pixels that are 2.5 times greater r than the row and column median. On the other hand, because we employ different thresholds (3 versus 2.5), it is possible that a column is neither part of the noise nor the signal.

Dividing the Spectrograms into Chunks

As mentioned in the last section, we construct a spectrogram for the sound file's signal and noise components. After that, we divided both spectrograms

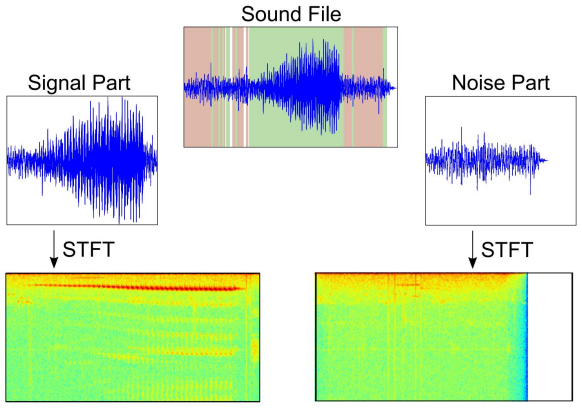

Figure 2 Separation of signal and noise part for the sound file

into equal-sized pieces.Three factors led to the divide. One of the requirements for our neural network architecture is a fixed sized input.We could add padding to the input, however given

the wide range in recording lengths, some samples would have over 99% padding. Although it would stretch or compress the signal in the temporal dimension, we could also try to employ different step sizes for our pooling layers. In contrast, chunks let us maintain a fixed step size while padding only the final portion.

Second, we can utilise each chunk as a distinct sample for training and testing because of the way our signal/noise separation method eliminates the problem of empty chunks (without a bird calling or singing). Third, we may allow the network to make numerous guesses per sound file (one forecast per chunk), and then we can average those predictions to produce a final prediction. As a result, our projections are more solid and trustworthy. A more sophisticated attempt at merging numerous predictions may be made as an extension, although no thorough testing has been done to date.

The next process is Data Augumentation.Through the use of a bigger and more varied set of data, data augmentation aims to enhance the performance of machine learning models.

Data Augumentation:

We need extra approaches to prevent over fitting because the number of sound files is very low when compared to the number of classes. Data augmentation was one of the most crucial components, along with drop-out, to enhance the system's generalisation performance. We use four different techniques for data augmentation.By examining Table 1, it is possible to comprehend the effects that each data augmentation technique has.

The different Augumentation methods are:

- Time Shift
- Pitch Shift
- Combining Same Class Audio Files
- Adding Noise

Time Shift:

We move the neural network's position in time by a random amount each time we give it a training example. This implies that we divide the spectrogram into two pieces and arrange the second piece in front of the first (wrap around shifts). As a result, there is a sharp corner where the first part's beginning and second part's finish meet, but all the information is still there. With this addition, we require the network to deal with spectrogram anomalies and, more crucially, we teach the network that bird songs and calls can appear at any time, regardless of the species of bird.

Pitch Shift:

Pitch alterations (vertical shifts), according to an analysis of various augmentation techniques , also contributed to lowering the categorization error. While a little change (approximately 5%) appeared to be helpful, we discovered that a bigger shift had no positive effects.Again, we preserved the entire piece of information by using a wrap-around technique.

Combining Same Class Audio Files:

We include audio tracks that belong to the same category. Since each sound file may be represented by a single vector, adding is an easy operation. We replay the shorter sound file as many times as necessary if one is shorter than the other. We re-normalize the outcome once two sound files are added in order to maintain the sound recordings' original maximum loudness. The operation depicts what happens when several birds (of the same species) sing simultaneously. Adding files enhances convergence since the neural network can observe more crucial patterns at once. We also discovered a minor improvement in the system's accuracy.

Adding Noise:

Background noise enhancement is one of the most crucial steps. The process of separating each file into a signal and noise element was previously covered.

Since background noise shouldn't depend on the class label, we can pick any random noise sample for every signal sample and layer it on top of the existing training sample. When combining audio files of the same class, this process should be carried out in the time domain by adding both sound files and repeatedly adding the smaller one. Even more noise samples can be added. In our experiment, we discovered that the greatest results are obtained when three noise samples are put on top of the signal, each with a dampening value of 0.4.This indicates that, given enough training time, we eventually add every type of background noise to a single training sample, reducing the generalisation error.

	MAP (FG only)	MAP (FG & BG)
Baseline	0.842	0.728
w/o Noise	0.831	0.731
w/o Same Class	0.839	0.730
w/o Time Shift	0.801	0.701
w/o Pitch Shift	0.828	0.725
w/o Noise and Same Class	0.768	0.661

Table 1

Network Architecture:

Five convolutional layers and a max-pooling layer make up the network. Before the final soft-max layer, we add one dense layer. A probability is produced for each class using the dense layer, which has 1024 units, and the soft-max layer, which has 1000 units.Figure 3 displays the architecture of the network.

Before each convolutional step as well as the dense layer, batch normalisation is used. A rectified activation function is used in the convolutional layers. The input layer, the dense layer, and the soft-max layer all use drop-out. We employ the single label categorical cross entropy function as a cost function.

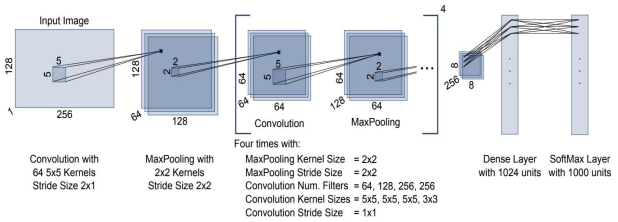

Figure 3 Architecture of the Network

Batches of 8 or 16 training examples are available here. Although we discovered that utilising 16 training samples per batch yielded somewhat better results, some models had to be trained with only 8 samples per batch due to GPU memory restrictions.

The batch normalisation algorithm performs much worse when several samples from the same sound file are included in a single batch. As a result, we randomly choose the samples for each batch without replacing them.

The updates for the weights are calculated using the Nesterov momentum method.

The starting learning rate is set to 0.1 and the momentum is set at 0.9.

V.Results

Due to the exceptionally high vocal frequency of many bird species, audio recordings have emerged as one of the most useful methods for studying birds.

Birds' natural noises can provide precise and consistent information about the dynamics and distribution of animal habitats. Due to the fact that many bird species emit distinctive and dependable sounds, audio research and surveys are a useful technique for determining the species' density, abundance, and occupancy. Additionally, picture-based monitoring is troublesome for many delicate and vulnerable birds, enigmatic species, and species that live in locations that environmentalists find challenging to access. Other environmental operations, such as evaluating the effects of wildfires and figuring out how much forest regeneration has occurred, can also benefit from bird audio tracking.

In this project we are dividing the initial training set into a training and validation set, we may locally assess our outcomes. By class id (species), we grouped the files and used 10% of each group for validation and the remaining 90% for training in order to preserve the original label distribution.The neural network must be trained extensively. In order to fine-tune parameters, here we select a portion of the training set that contains 50 different species.

We were able to test more than 500 alternative network configurations using this (20 times smaller) dataset.

After training the final configuration on the entire training set (taking into account all 999 species), it achieved accuracy scores of 0.59 and mean average precision (MAP) scores of 0.67 on the local validation set (999 species).

Other methods fared nearly as well as ours when background species were taken into account. When there were no foreground species, another strategy was able to outperform us. Given our strategy for data augmentation and preparation, this should not come as a surprise to us. First, we were removing the noise and concentrating just on the signal. Theoretically, this should assist our network in concentrating on the crucial components, but in practise, we might ignore less audible background species. In order to improve our data, we also add background noise from additional files on top of the signal portion.

VI .Conclusion

In conclusion, our study showed how deep audio classifiers can be used to monitor endangered species' bioacoustics and boost conservation efforts. Using passive acoustic monitoring, we created and evaluated a deep convolutional neural network that successfully identified the target species and offered correct estimates of population size and dispersion. According to our findings, this technology has the potential to deliver real-time data on the existence, abundance, and distribution of endangered species. This information can help inform focused conservation strategies and boost the overall effectiveness of conservation efforts.

Additionally, we discovered that the deep audio classifier could recognise previously undiscovered vocalisations and tell them apart from vocalisations belonging to other species, highlighting the capability of this system to learn new knowledge on species vocal behaviour. We encourage more study in this field to improve and hone this technology for conservation reasons. Deep audio classifiers have the potential to revolutionise the way we monitor endangered species through bioacoustic monitoring. Ultimately, by strengthening our capacity to monitor threatened and endangered species, we can better comprehend their physiology, behaviour, and conservation status and create conservation plans that will increase the likelihood of their survival.

Comparative Analysis:

There are various solutions for the problem of Bird species identification in the present time but the current approach is the most efficent and effective one.

Some of the various others approaches for the problem are listed below:

Approach 1:

To decrease the size of the training data, each audio file is first downsampled to a frequency of 16 kHz and then applied with a low-pass filter having a cutoff frequency of 6250 Hz. The method then seeks to remove the cells with little information (according to the mean and variance) by dividing the

spectograms into cells of 0.5 seconds x 10 bands of frequency. Following these preprocessing processes, they reassembled the remaining spectrogram segments into five-second segments, produced arrays of 200310 (where 310 samples equal five seconds), and utilised these arrays as the input for the CNN.
Here, two different CNN designs have been used: the well-known AlexNet with the inclusion of batch normalisation, and a CNN that is more influenced by audio recognition systems and is built using four convolutional layers, one fully connected layer, ReLU activation functions, and batch normalisation.However it obtained mean average precision (MAP) ratings of 0.53 and accuracy values of 0.42.

Approach 2:
The classifier based on pairs of spectrogram peaks discussed in the context of audio fingerprinting was improved in a late fusion scheme using the technique presented in this approach, which was initially employed on smaller datasets. The technique is based on the bag-of-words strategy. First, the 44.1 kHz audio files were divided into 0.2s segments with 50% overlap, and only the segments with energy values greater than a relative (to the whole audiofile) value and spectral flatness values less than an absolute threshold were kept for Mel Frequency Cepstral Coefficient computation (MFCC). All of the MFCC and its derivatives were subjected to a k-means clustering with k=500 in order to extract for each file the normalised histogram of MFCC-based words (i.e., the 500 clusters), using just segments.A random forest classifier was then fed the generated feature vectors.Nevertheless, it received accuracy scores of 0.18 and mean average precision (MAP) ratings of 0.14.

Approach 3:
Convolutional neural network learning is the framework used in this method. With the use of percentile thresholding, silent regions of de-noised spectrograms were eliminated, resulting in around 86,000 training segments, each of different length and linked to a single primary species. Segments were modified by cutting or padding as a data augmentation approach and to match the CNN's fixed 5 second input size. The first three runs were produced using filters that were broader, deeper, or both at the same time. The final run uses an ensemble of neural networks to average the results of the first three runs' predictions.Still, it obtained mean average precision (MAP) ratings of 0.52 and accuracy ratings of 0.41.

Approach Name	Mean Average Precision	Accuracy Rating
Approach 1	0.53	0.42
Approach 2	0.18	0.14
Approach 3	0.52	0.41
Proposed Approach	0.59	0.67

This means that in the category where background species were disregarded, our strategy outperformed the next best approach, that is approach 3, by 17%.This is because our CNN design is made up of five convolutional layers, a rectification activation function, and a max-pooling layer.The call and song components are first identified and isolated from the silent and noisy parts using spectrogram analysis and morphological techniques. Following multiple data augmentation approaches, spectrograms are subsequently divided into 3 second chunks and used as inputs for the CNN.First, three randomly chosen bits of background noise are concatenated with each chunk that was identified as a singing bird. The following data augmentation techniques included time shift, pitch sift, and mixes of audio files from the same species. was attained after just one training day.

REFERENCES

[1]. Francisco J. Bravo Sanchez1, Md Rahat Hossain1, Nathan B. English2 & Steven T.Moore,"Bioacoustic classification of avian calls from raw sound waveforms with an open-source deep learning architecture.",Nov 2021,DOI:10.1038/s41598-021-95076-6

[2].Fan Yang, Ying Jiang, Yue Xu"Design of Bird Sound Recognition Model Based on Lightweight",vol.10, 2022DOI:0.1109/ACCESS.2022.3198104

[3].Samruddhi Bhor,Rutuja Ganage,Omkar Domb,Hrushikesh Pathade,Shilpa Khedkar"Automated Bird Species Identification using Audio Signal Processing and Neural Network",2022,CFP22AV8-ART

[4].Stefan Kahl, Mary Clapp, W. Alexander Hopping, Herv´e Go ¨eau, Herv´e Glotin,Robert Planqu´e , Willem-Pier Vellinga, and Alexis Joly"Bird Sound Recognition in Complex Acoustic Environments", vol. 2696,2020,DOI:CEUR-wS.org/Vol-2696/paper_262.pdf

[5].Noumida A., Rajeev Rajan,"Deep Learning-based Automatic Bird Species Identification from Isolated Recordings",2021,DOI:10.1109/ICSCC51209.2021.9528234

[6].Emre Cakir,Sharath Adavanne,Giambattista Parascandolo,Konstantinos,Drossos,TuomasVirtanen"Convolutional Recurrent Neural Networks for Bird Audio Detection",2017,DOI:978-0-9928626-7-1

[7].Hasan Abdullah Jasim,Saadaldeen R. Ahmed ,Abdullahi Abdu Ibrahim,Adil Deniz Duru,"CLASSIFY BIRD SPECIES AUDIO BY AUGMENT CONVOLUTIONAL NEURAL NETWORK",2022,DOI:10.1109/HORA55278.2022.9799968

[8].Agnes Incze, Henrietta-Bernadett Jancso, Zoltan Szil,Attila Farkas and Csaba Sulyok,"Bird Sound Recognition Using a

Convolutional Neural Network",September 13-15, 2018,DOI:978-1-5386-6841-2/18/$31.00

[9].JIE XIE(Member, IEEE), KAI HU , MINGYING ZHU , JINGHU YU, AND QIBING ZHU,"Investigation of Different CNN-Based Models for Improved Bird Sound Classification", December 4, 2019,DOI:10.1109/ACCESS.2019.2957572

[10].Dan Stowell,Mike Wood ,Yannis Stylianou,Herve Glotin,"BIRD DETECTION IN AUDIO: A SURVEY AND A CHALLENGE",SEPT 13 2016,DOI:978-1-5090-0746-2/16/$31.00