Animal Identification And Detection System Using Voice Recognition

[1]Amrutha Bhat, [2]Ankush Jana, [3]Jigme Shitro, [4]Nihal Shetty, [5]Sumukh KU
[1]Assistant Professor, [2,3,4,5]Student
[1,2,3,4,5]Computer Science and Engineering
[1,2,3,4,5]Mangalore Institute of Technology and Engineering, Moodabidri, Karnataka, India

Abstract - Animals tend to change their activities as well as their habitats due to the adverse effects on the environment or other natural or man-made calamities. Continuous monitoring of remote areas is essential to ensure the safety and protection of them. Audio recognition is the most effective way to monitor them and it also helps in studying the behavior and communication of different animals. Our audio datasets were carefully curated to include a diverse range of animal sounds. To ensure the accuracy and reliability of our data, we conducted extensive quality control measures, including manual review and verification of each audio clip. We utilized the convolutional neural network on the training data and validation data using various parameter and layer changes. The training and validation data for the model were split into 70-30 using the holdout split method. The model was then trained with features extracted by the MobileNetV2 model, in order to improve accuracy. Our future plans include expanding our dataset to include more exotic and endangered animals.

Index Terms – Convolutional Neural Network (CNN), Transfer Learning, Melspectrogram, Tensorflow

I. INTRODUCTION

We need to make sure the animals living in the forest are well protected for various reasons. 80% of biodiversity on land lives in the forests which signifies the importance it has in maintaining the ecological balance. It is also important to understand how the ecological footprint of human is affecting the animals and their ecosystem in the forest. This method not only would be helpful in finding how the animals are affected by our footprint but also can help us warn about a lot of natural calamities beforehand. As maintaining a proper balance between humans and other species, we need to make sure we understand their problems and act accordingly. To understand their problems, the only way is through audio and visual. In deep forests, due to many trees we cannot monitor a large area. By utilizing a tool that can recognize syllables of different species, it is possible to cover a large area with minimal equipment and resources. This will save a ton of money for the government as well as rescue animals when they are in emergency situation.

Our approach includes audio datasets with and without background noise, we can even take background noise into consideration and execute various methods in saving the environment as well. We can even record various background noise of forest fires, cutting down of trees with equipment. We can even take these into consideration and save environment. To get a clear audio file without any background noise, we need to have pre-recorded sounds of all the possible background noise and make sure we cancel these noises in the considered dataset and after many approaches, we got an accuracy of 87.6% in detecting animal sound.

II. LITERATURE SURVEY

We will focus specifically on Audio Recognition with the help of Tensor Flow in this section, rather than providing a broad overview of how machine learning is currently applied in various fields such as internet of things (Zeng, E- AUA: An Efficient Anonymous User Authentication Protocol for Mobile IoT. IEEE Internet of Things Journal., 2018) (Zheng, 2014), social networks (Wang, 2018), activity recognition (Bhandari, 2017) (Pan, 2018), and recommendation (Fu).

- "According to Sophia, Tensor Flow is used to implement complex DNN structures without getting complex mathematical details, and availability of large datasets. The machine learning model used in this paper is CNN consists of three hidden layers." (Thakare, 2017) This algorithm predicts the animal sounds and gives output as 1 if found the audio clip or 0 if it couldn't find any. Spectrogram features are learned from the audio signals to detect animal audio. This network is implemented in Keras.
- "Hassanali identifies and differentiates between bird's call and bird's song. These are separated with the help of syllable which is denoted as smallest part in audio." (Virani, 2017) Theauthor depicts that recognition based on class probabilities and recording level detection is more suitable than syllable level detection. The frequency calculating technique used herewas Mel Frequency Cepstral Coefficients (MFCC) which showsaccurate results for audio recognition.
- In her proposal, Arti V. Bang suggested several pattern techniques that can be used for sound classification, including "Mel Frequency Cepstral Coefficients (MFCCs), Gaussian Mixture Modeling (GMM), and Hidden Markov Model (HMM)" (Bang, 2018). The recordings obtained here are in different formats with different sampling rates and were casually recorded in a noisy environment. Various techniques such as Mean Computation,Principal Component Analysis along with K- Nearest Neighborsare used while testing the machine. The overall accuracy of animal's audio recognition kept increasing as k-folds are increased.

- Sven Koitka identified a new technique wherein the audio files are preprocessed before being trained. They are extracted according to frequencies and then are examined with the help of bandpass filtering technique. After obtaining the output datasets, the author applied a noise filtering and silent region removal technique to isolate the animal sounds and create a pure audio dataset. This dataset was then used to train a Transfer Learning algorithm (Koitka, 2017).
- Dorota Kamiska and Artur Gmerek presented fully automated algorithm (kaminska, 2012). In their paper, the authors selected and compared two classifiers, namely OM and k-NN, for analyzing sounds of several species downloaded from various web sources (Disjoint sets with 70% - for preparing, 30% - for testing). The order precision for various highlights demonstrated that spectral features are the best for Automatic Species Recognition task. Their best outcomes had mean Classificationprecision of 69.94% with k-NN classifier and 52.92% with SOM classifier.
- Chang-Hsing Lee et al. (2006) utilized frequency information to accurately extract syllables and employed averaged MFCCs within each syllable to distinguish between species based on their vocalizations. The experimental results demonstrated that AMFCC was far more effective than both HMM and ALPC in both training and testing stages. The average classification accuracy was up to 96.8% and 98.1% for frog calls and cricket calls, respectively.
- Their study aimed to determine the effectiveness of using audio recordings for bird species recognition and assess the suitability of the selected bird species for this task. Iosif Mporas et al. recorded the vocalizations of seven common bird species in the Hymettus Mountain region and analyzed the data to evaluate the accuracy of identifying each species based on their unique acoustic features. Two temporal and sixteen spectral audio descriptors computedusing the open SMILE acoustic parameterization tool were used. The boosting algorithm exhibited superior classification performance over all other algorithms under conditions of high SNR, whereas the bagging meta-classifier yielded marginally better results than the boosting algorithm under conditions of low SNR. Notably, the maximum classification accuracy achieved by the bagging meta-classifier was 92.89%.

III. RELATED WORK

In the subsequent sections, we are going to provide a criticalanalysis of related works in the field of audio recognition using machine learning. Let's take a look at them below:

- "Species Recognition Using Audio Processing Algorithm": In this paper, the author identifies and differentiates between animal's call and bird's call. These are separated with the help of syllable which is denoted as smallest part in audio.The author depicts that recognition based on classprobabilities and recording level detection is more suitable than syllable level detection. The frequency calculating technique used here was Mel Frequency Cepstral Coefficients (MFCC) which shows accurate results for audio recognition.
- "Deep Learning based Animals Audio Detection": In this paper, the author introduces a deep learning-based solution for detecting bird audio, leveraging the power of TensorFlow. "Tensor Flow is used to implement complex DNNstructures without getting complex mathematical details, andavailability of large datasets. The machine learning model used in this paper is CNN which consists of three hidden layers." This algorithm predicts the animals sounds and gives output as 1 if found the audio clip or 0 if it couldn't find any. Spectrogram features are learned from the audio signals to detect animal audio. Keras is the framework in which this network has been developed. This is inspired by human biological neurons. The processing neural signals along with deep learning is used in recognizing species audio. The preprocessed data is then mapped to the training data. By undergoing various normalizations through preprocessing functions, CNN is developed. Back-propagation through time with the Adam optimizer is utilized to train the network.
- "The author of this paper explores various pattern techniques for sound classification, including Mel Frequency Cepstral Coefficients (MFCCs), Gaussian Mixture Modeling (GMM), and Hidden Markov Model (HMM), and how they can be used to recognize animal species through data reduction methods. The recordings obtained here are in different formats with different sampling rates and were casually recorded in a noisy environment. Various techniques such as Mean Computation, Principal Component Analysis along with K-Nearest Neighbors are used while testing the machine. The overall accuracy of bird's audio recognition keptincreasing as k-folds are increased.

IV. RESEARCH METHODOLOGIES

The goal of this project is to enhance the ecosystem by teaching a machine to recognize various animal sounds, which can benefit both animals and humans. It is an arduous task for us to deploy human assets everywhere for the sole purpose of monitoring animals even after which it is susceptible to obvious errors. So, it is better to develop a machine which can do it all for us. This research will help start the whole process spoken about with detailed understanding of how to train the machine from scratch with very few exceptions. This can be used now during the preliminary times of development which will only improve themachine to be more effective as it learns further.

A.*Data Collection*

Data collection is the process of collecting and measuring information on interesting variables in an established systematic manner that allows one to answer stated research questions, test hypotheses, and evaluate results. For our project, the dataset which comprises of the audio from different websites were collected. 3200 animal audio files were collected from various multimedia files.

B.*Data Description*

It involved the primary task of designing the framework for animal audio classification, involves description of the raw audio data collected from various websites by downloading huge numberof audio files and loading the same in python to understand the descriptive analysis.

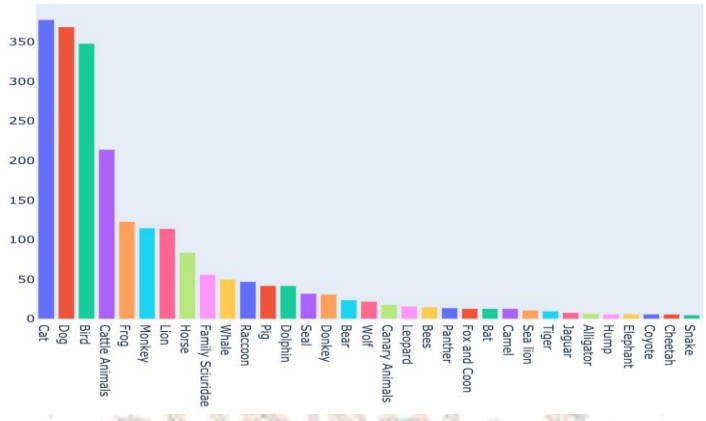

Fig.1 Distribution of Classes in Dataset

C.*Data Preprocessing*

Data preprocessing is a technique of data mining involving the transformation of raw data into a comprehensible format. Data mining employs this technique to convert raw data into a format that is both useful and efficient. When dealing with machine learning problems, a significant amount of time is spent ensuring the quality of the data, especially since it is often sourced from multiple unreliable sources in varying formats.

D.*Feature Extraction*

Melspectrogram uses a frequency-domain filter bank for time-windowed audio signals. As the second and third output arguments from melspectrogram, we can get the center frequencies of the filters and the time instants corresponding to the research frames. Mathematically, the mel-scale is the result of some non-linear transformation of the frequency scale.

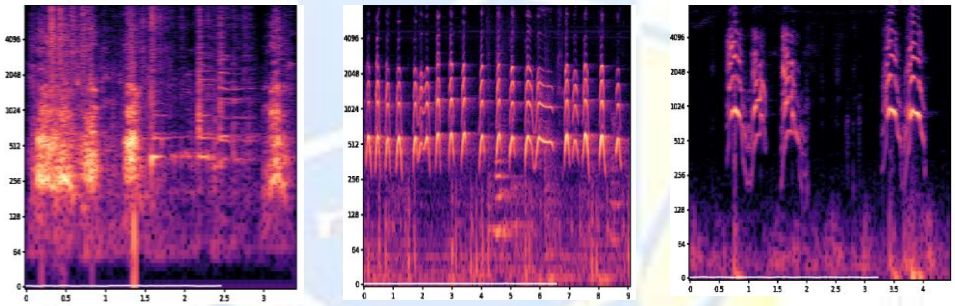

Fig.2 Melspectrogram of Dog, Cat and Cow

E.*Project Design*

The overall system of our work can be viewed as two entities decomposed into the identification and classification of animal audio. The Figure 3 provides a high-level work-flow overview modeling various steps in the research design and the components associated. In summary, the research follows an approach which involves reading the data, generating the spectrogram and analyzing the spectrogram to train the model and further make predictions on the trained model.

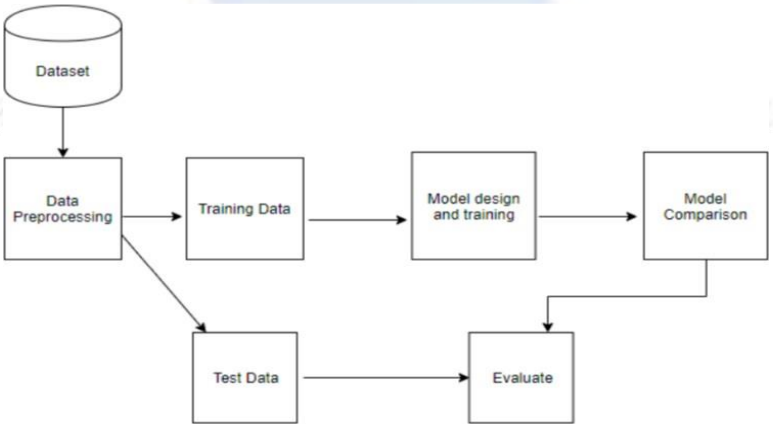

Fig.3 Workflow Diagram

V. IMPLEMENTATION AND RESULT

This section outlines the execution of this study's experiment followed by a technical assessment and evaluation of the methodology.

A.*Software*

The experiment for this research work was done primarily in python using various python libraries. Most frequently used libraries in this research work are Pandas, numpy, Librosa, glob, sklearn. For data exploration plotly. express and extended library for visualizations was used while for data modeling, Conv2D used keras with tensor flow backend.

B.*Data Modelling*

Convolutional neural network was applied on the training data and validation data using various parameter and layer changes. The training and validation data for first model was split in 70-30 using holdout split method. It includes several building blocks, such as layers of convolution, pooling layers and layers that are completely connected. The first layer is followed by activation layer, ReLU activation is applied on the input layer post convolution. Followed by second convolution layer Conv2D that learns from 64 filters and a kernel size of (3,3). The pooling layer is followed by a dropout layer to reduce the overfitting in the neural network as by preventing complex co-adaptations on training data. The architecture than has two sequence of convolution, activation and pooling layers as feature extractors followed by flatten operation and fully connect dense network to interpret the features and output layer for prediction.

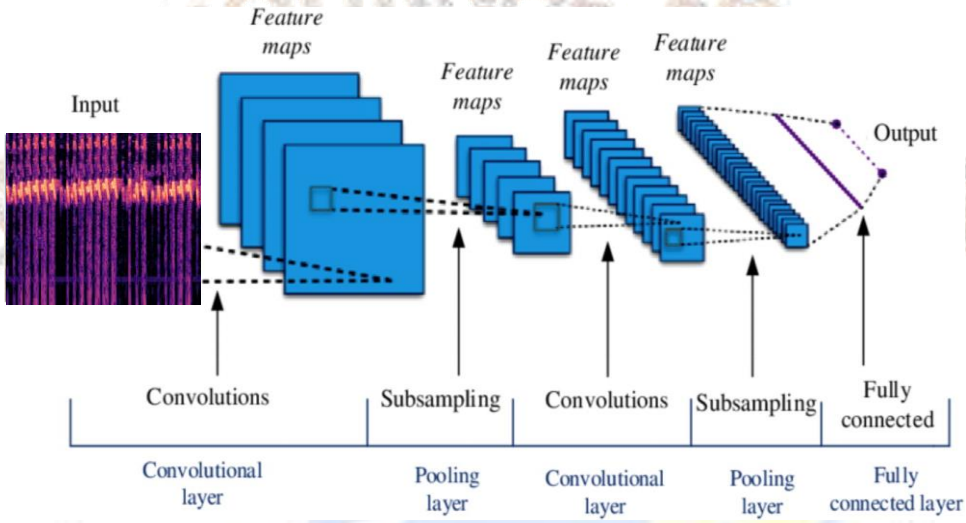

Fig. 4 CNN Architecture.

C.*Tranfser Learning*

Transfer learning is a machine learning technique where a pre-trained model is used as a starting point to solve a new, related task. Instead of training a model from scratch on a new dataset, transfer learning allows us to leverage the knowledge learned by an existing model that has been trained on a large dataset. The idea is to use the pre-trained model as a feature extractor and then fine-tune it on the new dataset. By doing this, we can use a smaller dataset to achieve better results than we would get by training a model from scratch. Transfer learning has been applied successfully in many domains, including computer vision, natural language processing, and speech recognition. MobileNetV2 is a deep neural network architecture designed for efficient computation on mobile and embedded devices. It was introduced by Google in 2018 as an improvement over the original MobileNet architecture. The main goal of MobileNetV2 is to achieve high accuracy on image classification tasks while using fewer parameters and lower computational resources.

D.*Evaluation of Result*

Convolutional Neural Network, the model was trained on 6879 images and validated on 2293 images. The model was trained with 10 epochs after which the accuracy was found to be 74.24%.

```
Epoch 1/10
38/38 [==============================] - 46s 1s/step - loss: 1.7171 - accuracy: 0.3935 - val_loss: 1.4495 - val_accuracy: 0.5562
Epoch 2/10
38/38 [==============================] - 46s 1s/step - loss: 1.0374 - accuracy: 0.6334 - val_loss: 1.0777 - val_accuracy: 0.6313
Epoch 3/10
38/38 [==============================] - 40s 1s/step - loss: 0.8057 - accuracy: 0.7412 - val_loss: 0.8713 - val_accuracy: 0.7312
Epoch 4/10
38/38 [==============================] - 38s 1s/step - loss: 0.5737 - accuracy: 0.8005 - val_loss: 1.0110 - val_accuracy: 0.6812
Epoch 5/10
38/38 [==============================] - 38s 1s/step - loss: 0.4180 - accuracy: 0.8410 - val_loss: 0.6392 - val_accuracy: 0.7875
Epoch 6/10
38/38 [==============================] - 38s 1s/step - loss: 0.2094 - accuracy: 0.9272 - val_loss: 0.8548 - val_accuracy: 0.7812
Epoch 7/10
38/38 [==============================] - 39s 1s/step - loss: 0.1289 - accuracy: 0.9515 - val_loss: 0.8676 - val_accuracy: 0.8250
Epoch 8/10
38/38 [==============================] - 41s 1s/step - loss: 0.0731 - accuracy: 0.9677 - val_loss: 0.9152 - val_accuracy: 0.7875
Epoch 9/10
38/38 [==============================] - 41s 1s/step - loss: 0.1416 - accuracy: 0.9569 - val_loss: 0.7231 - val_accuracy: 0.8188
Epoch 10/10
38/38 [==============================] - 41s 1s/step - loss: 0.0283 - accuracy: 0.9919 - val_loss: 1.0746 - val_accuracy: 0.8250
```

Fig.5 CNN Model Performance with accuracy 74.24%.

Now the model was trained with features extracted by `MobileNetV2` by the concept of transfer learning and as a result there was an improvement in the accuracy which is 83.25 %.

```
38/38 [==============================] - 19s 505ms/step - loss: 25.3198 - accuracy: 0.6146 - val_loss: 10.2536 - val_accuracy: 0.6750
Epoch 2/10
38/38 [==============================] - 19s 503ms/step - loss: 2.5250 - accuracy: 0.8841 - val_loss: 3.0206 - val_accuracy: 0.7875
Epoch 3/10
38/38 [==============================] - 19s 491ms/step - loss: 0.5313 - accuracy: 0.9434 - val_loss: 4.3236 - val_accuracy: 0.7937
Epoch 4/10
38/38 [==============================] - 18s 475ms/step - loss: 0.3100 - accuracy: 0.9811 - val_loss: 4.5134 - val_accuracy: 0.7937
Epoch 5/10
38/38 [==============================] - 18s 477ms/step - loss: 0.0736 - accuracy: 0.9892 - val_loss: 3.3991 - val_accuracy: 0.8250
Epoch 6/10
38/38 [==============================] - 18s 479ms/step - loss: 0.0780 - accuracy: 0.9919 - val_loss: 2.5860 - val_accuracy: 0.8250
Epoch 7/10
38/38 [==============================] - 18s 473ms/step - loss: 0.0052 - accuracy: 0.9973 - val_loss: 3.0224 - val_accuracy: 0.8188
Epoch 8/10
38/38 [==============================] - 18s 471ms/step - loss: 8.6119e-04 - accuracy: 1.0000 - val_loss: 2.8453 - val_accuracy: 0.8438
Epoch 9/10
38/38 [==============================] - 18s 479ms/step - loss: 1.8758e-06 - accuracy: 1.0000 - val_loss: 2.8216 - val_accuracy: 0.8438
Epoch 10/10
38/38 [==============================] - 18s 480ms/step - loss: 9.1730e-07 - accuracy: 1.0000 - val_loss: 2.8206 - val_accuracy: 0.8438
```

Fig.6 CNN Model (using features extracted by MobileNetV2) Performance with accuracy 83.25%.

A confusion matrix or error matrix is a tabular visualization of statistical classification per class. A confusion matrix for the CNN model is shown in Figure 7.

Fig. 7 Confusion Matrix.

VI. STRENGTH AND LIMITATION

Convolutional neural network and hybrid models were designed to classify animal audio using melspectogram generated from the audio waveforms. The study's key strength was its ability to accurately identify and classify the animal sound in the audio files. The designed framework though new for audio sound is actually similar to existing AI architecture like one discussed in the paper Takahashi et al. (2016). Audio waveforms were preprocessed, and melspectrogram were generated from the waveform to train the algorithm. One of the key strengths of this model is that it can be used in future to classify different audio classes.

The limitation of the model is it requires huge computational power when the dataset is large. For this research work, the amount of data was limited making it simpler to train the model. If the model is extended in future for processing huge number of records, machine with high computational power will be required. Also, the preprocessing done on the data was specific to the format of the files downloaded. Additional mechanisms need to be implemented to make the audio generic in nature. The noise in the audio also needs to be filtered to maintain classification accuracy. The model is train on melspectogram which are generated from the audio waveform which will be a limiting factor.

VII. CONCLUSION

In this report, we have given a brief overview about the audio recognition of animals and how they are helpful in real lives. We also provided the usage and working of commonly used CNN Model. We used the MobileNetV2 model for feature extraction, which is a popular pre-trained convolutional neural network (CNN) architecture known for its high accuracy and efficiency. By leveraging the pre-trained weights of the MobileNetV2 model, we were able to train our own CNN model with a smaller dataset and achieve better performance than training from scratch. Overall, our study provides a valuable contribution to the field of audio recognition and machine learning. It opens up new opportunities for future research and innovation, and has the potential to make a positive impact on our environment and society.

VIII. FUTURE DIRECTIONS

Even with the extensive research on the approaches to recognize animal voice, there are many ways in which machines fail to recognize them accurately. Every proposed model so far just had an accuracy of about 66%. The challenges faced by audio recognition is more and by developing an accurate algorithm, we can detect species and help them from being endangered. We can even add various background sounds such as forest fires, breaking down of trees which helps to easily recognize the forest fires before it causes serious damage. Audio is the main weapon which can cover a large area than that of video where the vision of the camera will be interrupted by the trees.

IX. REFERENCES

[1].P. Tivarekar, R. (2017). Species Recognition Using Audio Processing Algorithm.
[2].V. Bang, A. (2018). Recognition of Animal Species from their Sounds using Data Reduction Techniques.
[3].Thakare, D. (2017). Extracting Frequency Domain Representation.
[4].G. Virani, H. (2017). Species Recognition Using Audio Processing Algorithm.
[5].Koitka, S. (2017). Recognizing Animal Species in Audio Files Using Transfer Learning.
[6].Kaminska, D. (2012). Automatic identification of Animal species: A comparison between KNN and SOM classifiers.
[7].Lee, H. (2006). Automatic recognition of animal vocalizations using averaged MFCC & linear discriminant analysis.
[8].Somervu, P. (2006). Parametric representations of Bird Sounds for Automatic Species Recognition.
[9].Mporas, I. (2012). Automated acoustic classification of bird species from real-field recordings.
[10].E. Sophia (2018) Deep Learning based Animal Audio Detection.
[11].Wang, Z. (2018). Deep Learning Based Intrusion Detection with Adversaries. IEEE Access, pp.1-1.
[12].Identification & Detection System for Animals from their Vocalization. (2013).
[13].Duan, S. (2014). Automated Species Recognition in Environmental Recordings.
[14].Stowell, D. (2017). On-Bird Sound Recordings: Automatic Acoustic Recognition of Activities and Contexts.
[15].Fritzler, A. (2017). Recognizing Bird Species in Audio Files Using Transfer Learning.