Paper No: SC-14 **Smart Computing**

IoT based animal classification system using convolutional neural network

L. G. C. Vithakshana*
Department of Statistics & Computer Science
Faculty of Sciene, University of Kelaniya, Sri Lanka
chathuravithakshana@gmail.com

W. G. D. M. Samankula
Department of Statistics & Computer Science
Faculty of Sciene, University of Kelaniya, Sri Lanka
samankula@kln.ac.lk

***Abstract*: The kingdom "Animalia" is used to represent all living creatures on the planet earth, which is fallen into six categories. The language is the most common factor to divide humans and animals. Numerous classification techniques can be used for classification purposes, and the classification commonly can be done acoustically and visually. The classification systems are playing a considerable role, and bioacoustics monitoring was a significant field of study. Visual classification of animals is done by using either satellite images or established camera images. Nevertheless, due to some circumstances, image processing techniques cannot be applied. Then the acoustical classification techniques are taken place to encounter those problems. Even with acoustical methods, a remote observing method is required due to a few issues. Applying an IoT based acoustic classification system was designed using Convolutional Neural Networks (CNN), which is beneficial for those who are interested in monitoring ecosystems such as animal scientists, zoologists, and environmentalists. The hardware implementation was designed to collect the data from the place it was placed. The audio clips were preprocessed using the Mel-frequency Cepstral Coefficient (MFCC). A CNN architecture based on TensorFlow was used for the training process. To train and test the network, 400 sound clips of two seconds, such that 40 per each ten animal species, which were gathered from online libraries and formatted using Audacity, were used. The network was trained by changing the different gradient descent optimizers and eventually obtained the confusion matrices for each. The best result was gained by the AdaDelta, Gradient Descent, and RMSProp optimizers with 91.3% accuracy for each. Among them, AdaDelta had the most stable and increasing learning approach. As a future extension, to improve accuracy, a large number of data will be used.**

Keywords: *Animal, Convolutional Neural Network (CNN), Internet of Things (IoT), Mel-frequency Cepstral Coefficient (MFCC), TensorFlow*

I. INTRODUCTION

The word "animal" represents the massive varieties of species living on the planet earth that falls into the kingdom "Animalia" including the categories, mammals, birds, reptiles, fishes, amphibious, and arthropods [1] which can able to move, breath oxygen, consume organic materials, and sexually reproduce. The term "animal" is commonly used for all non-human species. Among the animals who live on the earth that is 8.7 million, there are only 1.2 million of them were identified, and there are more than seven million of them to be recognized as 2011 [2].

Since there is a difference between human calls and animal calls, language is the most common factor that separate human from other animals. In a human call, each word means subject, concept, or action. The vocal patterns can be used to distinguish each other in humans, but when it applied to the animals, it was quite a hard task to do so. However, there are some variations between the same animal species as well. Humans can express their emotions and moods using words as well as using the face. If a word was pronounced with bass, it could be classified as an adult man, and if the word was pronounced with a thin voice, it could be classified as a child. The animal calls also have information regarding emotions and mood. These calls consisted of the pitch, loudness, repetitions, and many more [3].

Classification systems are one of the most widely used technologies in the world to classify humans, animals, or objects. In the present, there are plenty of types of classification systems are available, including face classifications systems, vocal classification systems, fingerprint classification systems, iris classification systems, and so on. Those classifications and identifications were done by either visually or acoustically. For visual classification, image processing techniques are used, and for acoustical classification, audio (or vocal) processing techniques are used. In the sense of sound classification, environmental sound classification and bird sound classification are most popular.

As the name implies, the Internet of Things (IoT) is any "thing" that uses the internet to communicate and transfer data between each other without direct human interaction. Nowadays, IoT is touched by almost every field and industry, including business transactions, logistics, real-time transactions, healthcare, voice-enabling systems, and many more. Besides that, IoT can be used to observe animal behaviors and classify them in an efficient and useful way.

Convolutional Neural Networks (CNN or ConvNet) [4] is a well-known deep learning algorithm. It was invented based on the natural visual perception mechanism of the living creatures. CNN is commonly used for the classification. The most basic building blocks of neural networks are called neurons. Apart from that on CNN, there are convolutional layers, pooling layers, activation functions such as Rectified Linear Unit (ReLu), and fully connected layers (dense layers) available. There also have some variations of neural networks, including CNN, Recurrent Neural Networks (RNN), AutoEncoders, and so on.

Feature extraction is used for dimensionality reduction of raw data into more manageable data called the features (or variations). To extract the features from given audio or image data, the number of techniques are used. Mel-frequency Cepstral Coefficient (MFCC) is one of the best algorithms to

do so. Apart from that, there are other speech feature extraction methods available, including Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC), Line Spectral Frequencies (LSF), Discrete Wavelet Transform (DWT), and Perceptual Linear Prediction (PLP) [5].

Machine learning (ML) is one of the most influential and powerful technologies in today's world. ML is a subfield of well-known and accessible technology, and ML can be divided into four main subfields, including supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. The main goal of ML techniques to convert information into knowledge. To classify the images as well as to classify the audio, ML techniques are used.

The same classification techniques applied to humans and objects can be applied to animals as well. Animal classification can be done either visually or acoustically. Visual classification of animals can be done using satellite images or images from an established camera in the forest. However, due to some circumstances, visual classification mechanisms cannot be used with animals. Satellite images cannot be obtained in a deep well-grown forest is one of the reasons. It is hard to detect animals who live in a well-grown forest, even by using a camera, without cannot spot them to a camera. That is why the acoustic classification methods take place to encounter those problems. Even using acoustic techniques, a remote observing method is required due to a few reasons. The in-situ methods can be expensive, and time-consuming is one of the reasons.

Bioacoustics monitoring is a significant field of study in the world. There are several techniques available for bioacoustics monitoring ranging from traditional in-situ methods to semi-automated bioacoustics monitoring systems [6] on selected ecosystems. In most of the previous studies, bird sound classification was done.

This type of designed system has benefited the people who are interested in bioacoustics monitoring within a specific ecosystem, including animal scientists and environmentalists.

To address the problems mentioned above for the targeted people, this study was done using the deep learning perspective with supervised learning. During this study, a well-known feature extraction technique called Mel-frequency Cepstral Coefficient (MFCC) was used to extract the features from the given audio clips, fed those to the CNN architecture, trained, and tested the model using the collected dataset. The used CNN architecture consisted of convolutional, ReLu, max-pool, and fully connected layers.

Usually, deep learning techniques are required for a large number of data to achieve the best results for a selected model. However, unfortunately, since there is no ready-made animal sound dataset available on the internet, the audio clips were collected from different online libraries [7], [8], [9], [10]. The collected dataset consisted of 400 audio clips with two seconds of each of ten animal species including bat, elephant, hornbill, junglefowl, macaque, myna, peafowl, pig, squirrel, and toad. Background noise clips were also added to conduct the training process more accurately.

The hardware module was designed using Arduino UNO, NodeMCU for wireless communication, and a microphone sound sensor. This module collects the sound and sent those to the cloud server for easy access to the data in the implemented Android application. An android-based mobile application was implemented, and it was used to collect those data on demand and finally display the results on the Android interface when an animal was called.

Section II will describe the related works done in this area of study, Section III will describe the methodology used to implement the system, Section IV will discuss on the experimental process, and the final section, Section V will discuss on the conclusion and future work.

II. Related Works

There were several related work done in this area of interest by using numerous technologies include CNN, joint algorithms using Zero Cross Rate (ZCR), MFCC, and Dynamic Time Warping (DTW), Hidden Markov Model (HMM), and using most common libraries and frameworks including LibROSA, TensorFlow, and Keras. Furthermore, most of the studies were done to classify bird sounds, and there were few studies done in the field of farming also. Below are some of the previous studies done in this area.

Yeo, Al-Haddad, and Ng [3] researched to develop an identification detection system using animal voice recognition. In their system, they used the Zero Cross Rate (ZCR), MFCC, and Dynamic Time Warping (DTW) joint algorithm for animal voice recognition. ZCR was used to detect the endpoint of given input voice data and remove the silenced voice. MFCC was used for feature extraction of the given audio data files. Finally, voice pattern recognition was done using the DTW algorithm. As their obtained results, the developed system has worked as expected.

K.H. Frommolt and K.H. Tauchert [6] researched on birds using the bioacoustics monitoring system. Cardioid microphones and four four-channel recorders were used to record the audio files during entire nights for five years from 2008 to 2012. During the study period, the changes in the number of birds, as well as their spatial distribution with changes in habit structure of birds, were monitored.

Grill and Schluter [11] conducted a study to detect the presence of birds on Mel-scaled log-magnitude spectrograms in a given audio signal by using two CNN architectures, namely "bulbul" and "sparrow". Both architectures were performed very similarly on test datasets. The "bulbul" system was performed slightly better in the development set. Eventually, they obtained an Area Under Curve (AUC) measure of 89% for the hidden test set, which was higher than any other previous contestant. They concluded that any other further improvement might not need for this study.

Tom, Antoni, and Andy [12] researched music pattern feature extraction of audio music using a CNN model for ten different music genres by using evenly distributed 1000 songs among those ten genres as the dataset for that experiment. As they mentioned that it was required minimum prior knowledge to construct their feature extractor compared to the previous works. In conclusion, they stated that the developed model was not robust enough to generalize the training result in unseen music data.

In the study of Sainath and Parada [13], they introduced the "cnn-trad-fpool3" CNN architecture for Small Keyword Spotting (KWS) which was roughly similar to the CNN architecture that used in this paper to conduct this project along with few other architectures. "cnn-trad-fpool3" was contained, two convolutional, one linear low-rank, and one Deep Neural Network (DNN) layer.

In the Khamparia, Gupta, Nhu, Khanna, Pandey, and Tiwari [14] study, they conducted a sound classification using CNN and Tensor Deep Stacking Network (TDSN). CNN architecture was built with two convolutional layers that activate with max-pooling layers, and a final fully connected layer. They used ESC-10 and ESC-15 datasets to train and achieved 77% and 49% for CNN and 56% in TDSN for both datasets.

Nordby [15] study was about the environmental sound classification on microcontrollers using CNN. By using the Urbansound8k dataset, he experimented on the STM32L476 low-power microcontroller by using Keras deep-learning framework. The noise classification was done on the sensor node. Furthermore, for future works, he proposed to build a wireless sensor network that is powered using this technology.

Ajibola Alim, S., & Khair Alang Rashid, N. [5] study, investigated on different feature extraction algorithms for speech recognition, including Mel-frequency Cepstral Coefficient (MFCC), Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC), Line Spectral Frequencies (LSF), Discrete Wavelet Transform (DWT), and Perceptual Linear Prediction (PLP). Moreover, they concluded that none of the methods were superior to the other; the area of application would determine which method to select.

Emre and Tek [16] conducted research using a CNN model with three convolutional layers with a max-pooling layer and three fully connected layers. Three convolutional layers and first two fully connected layers were used Rectified Linear Units (ReLu) as the activation function since the final fully connected layer was used softmax as the activation function. They were preprocessed the audio files using MFCC, and LibROSA and features were extracted. They have experimented by changing seven gradient descent optimizers to verify what was the most appropriate optimizer for their dataset of 875 audio clips, which belongs to ten different animal classes. They observed that Nesterov-accelerated Adaptive Moment Estimation (Nadam) was the best optimizer for their experiment with 75% final accuracy.

Like the previous studies, there was not any IoT based system that was proposed and implemented the domain of animal classification using vocal analysis. Some studies were obtained not much higher level of accuracies for the models. Some studies were observed the bias results due to the class imbalance of the used dataset. Some researches used the relevant technologies and achieved higher accuracies, but not in the domain of this study.

III. METHODOLOGY

This study was done by considering the following methodology. The first step was to collect the dataset and prepare it. Feature extraction was done using a Mel-frequency Cepstral Coefficient (MFCC) algorithm. Then the extracted features were fed to the layers for classification of Convolutional Neural Network (CNN). The hardware module and the mobile application were designed.

A. Dataset

The dataset was about forest animals. Since there was no ready-made dataset available, the sound clips were collected on different online libraries [7], [8], [9], [10], and prepared those as TABLE I using Audacity. In this process, a considerable number of audio clips had been removed due to low animal sound levels and high background sound levels. The background noise dataset was collected from an online library [17] and converted to WAV files with a 16000 Hz sample rate and a bit rate of 32 bit. The complete sound dataset of 417 audio files with 40 from each of ten animal classes and 17 background noise clips were prepared.

TABLE I. SPECIFICATIONS OF PREPARED AUDIO CLIPS

File Type	WAV
File Size	62.5 KB
Sample Rate	16 KHZ
Bit Rate	32
Channel	MONO
Clip Duration	2 seconds

B. Feature extraction

In previous studies, there were many algorithms used for feature extraction, including MFCC, LPC, and DTW. The used CNN was not supposed to feed the raw audio files as it was. In this study, a widely used feature extraction algorithm called MFCC was used. The feature extraction was not done separately. Instead, it was done at the beginning of the training using the same script.

C. Convolutional Neural Network

Many CNN architectures were there, and one of them was selected for this study, which consisted of two convolutional 2D layers and a max-pooling layer. Each layer was activated with Rectified Linear Units (ReLu). The final fully connected layer was activated with softmax. A 1x1 stride size was used in convolutional layers, and a 2x2 stride size was used in max-pooling layers. The full architecture is shown in Fig. 1.

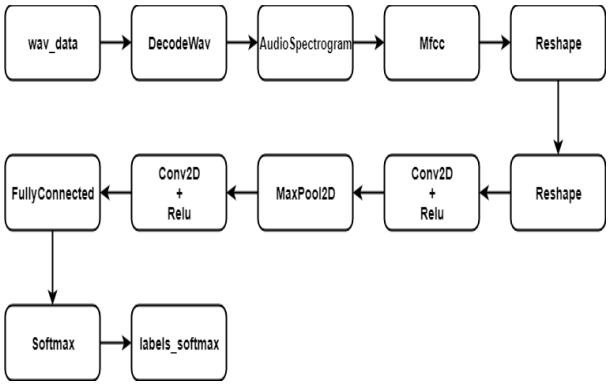

Fig. 1. CNN architecture

D. Hardware module

The hardware module was designed to collect the audio data from a faraway site that the device was placed. This module consisted of Arduino UNO powered by an ATmega328P microcontroller for getting connected the all devices, NodeMCU powered by ESP8266 (LX106) CPU for

wireless communication, and microphone sound sensor to collect the audio data. Those collected data were sent to the server for easy access later using implemented mobile applications.

E. Mobile application

The implemented mobile application based on the Android platform was designed to retrieve the audio data from the cloud server and eventually identify the animal with the help of the CNN model. The mobile application was implemented using Android Studio for native Android development. The application was built for API level 26, and above and Gradle version 3.5.1 was used. Fig. 2 shows some implemented interfaces.

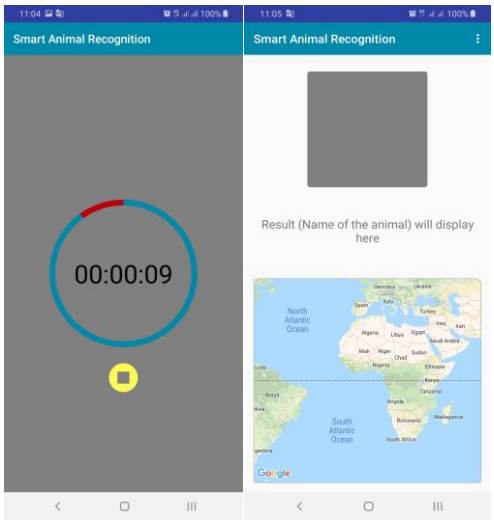

Fig. 2. Implemented Android mobile application interfaces

IV. Experimental Process

The experimental process was done using TensorFlow, and mainly it has consisted of three steps, including feature extraction, training, and testing. The experiment was done using Anaconda Prompt using TensorFlow GPU under the Windows 10 operating system. The dataset was split into three parts as training, testing, and validating. The 90% of data was split as training, the rest of the 10% was split as the testing set. A randomly selected another 10% was split as a validation dataset. For the optimization, six selected optimizers were considered before starting the training.

A. Training

During the training process, for the ten animal classes, bat, elephant, hornbill, junglefowl, macaque, myna, peafowl, pig, squirrel, and toad were used. For the background noise class, _background_noise_ label was used. The rest of the class was fell under the unknown label. The training process was conducted using the hyper-parameters, as illustrated in TABLE II. Features were extracted using the MFCC algorithm for the given dataset. The following Fig. 3 and Fig.4 show a raw audio file and a related spectrogram image for a given audio clip of the hornbill class.

TABLE II. Hyper-parameters of the training process

Key	Value
batch_size	40
preprocessor	MFCC
how_many_training_steps	15000, 3000
learning_rate	0.001, 0.0001
model_architecture	Conv
optimizer	Gradient Descent, Momentum, Adam, AdaGrad, AdaDelta, RMSProp

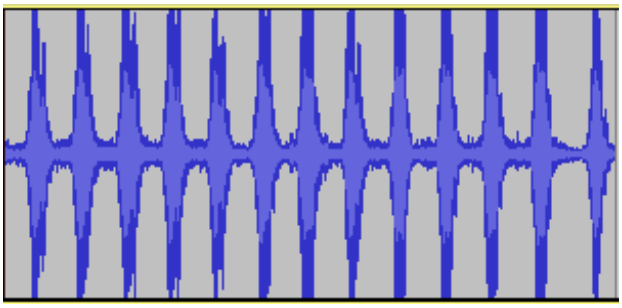

Fig 3. A raw audio file for hornbill class

At the end of the training, the trained model was saved as ". pb" file for future usage. Training accuracy was increased with the training steps. The model was trained up to 18000 steps of each by change one of the hyper-parameters called optimizer to verify what was the most appropriate hyper-parameter fitted for the model. Since the low learning rate was led to the highest accuracy, the first 15000 steps were trained using a 0.001 learning rate while the last 3000 steps were trained using a 0.0001 learning rate.

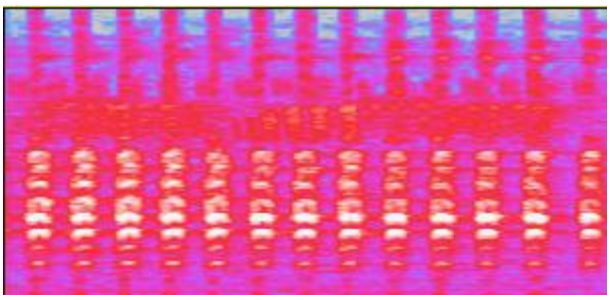

Fig 4. Related spectrogram

The training accuracy was not changed that much after 2000 steps in most of the cases except AdaDelta. Some of the observed graphs shown in Fig. 5 and Fig. 6. At the end of the training, confusion matrices for both test and validation were generated.

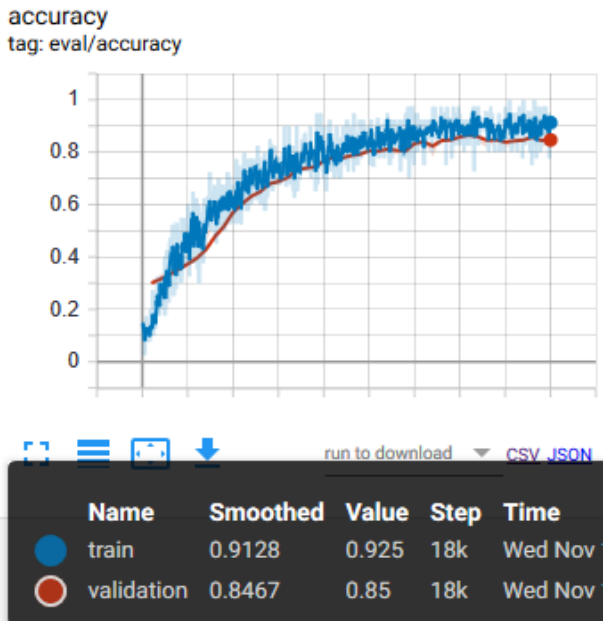

Fig. 5AdaDelta accuracy graph using TensorBoard

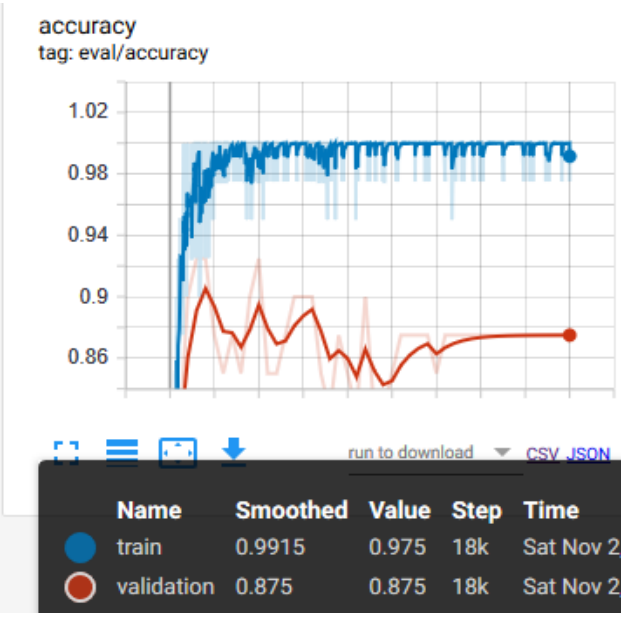

Fig 6. Gradient Descent accuracy graph using TensorBoard

B. Test and results

In this study, the trained CNN model was used to identify an unknown sound file as one of the labeled animal classes. For the given dataset, a confusion matrix was generated, and the visualized confusion matrix generated for AdaDelta optimizer is shown in Fig 7.

Rather than mention the accuracy, a matrix can summarize the actual accuracy against the predicted values of the model for the provided test dataset. The columns show the predicted values while the rows show the actual ground truth values. The diagonal entries show the correct classification, while other entries (off-diagonals) show the incorrect classification of the given labels.

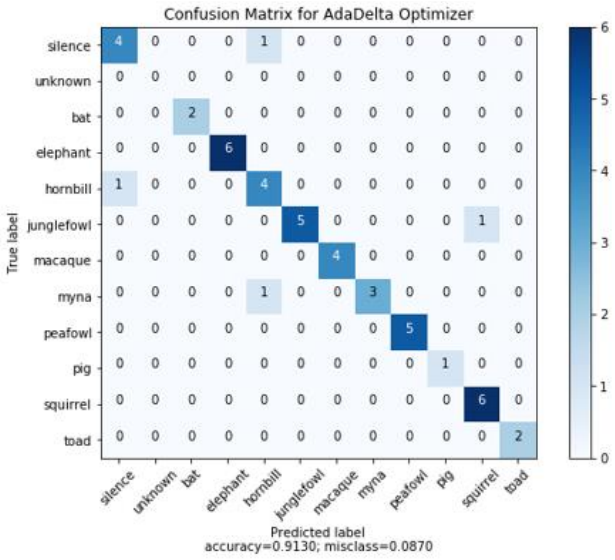

Fig. 7. AdaDelta optimizer visualized confusion matrix for test set

As an example, from the given 10% of dataset as the test dataset, among six elephant sound clips, the model used AdaDelta optimizer identified all six as elephants. Among junglefowl sounds, five clips were identified as junglefowl, and another one was incorrectly identified as a squirrel. Moreover, there were no unknown classes available within the dataset. The matrix was shown; the accuracy was divided among all classes almost the same way because a balanced dataset was used. There was no bias in any class.

Fig. 8 shows one of the observed results for a given audio clip. It was tested using a model trained by AdaDelta optimizer for one of the audio clips, which were included the sound of a peafowl. As observed, the accuracy score of 0.97427 out of 1.0000 was achieved successfully. Also, some inaccurate percentages were archived. Therefore, the given audio clip was successfully identified as a peafowl.

```
E:\train6\dataset\peafowl\peafowl_21.wav
peafowl (score = 0.97427)
hornbill (score = 0.01499)
elephant (score = 0.00624)
```

Fig 8. Observed results for given peafowl audio clip

The train was conducted by changing six optimizers, and as observed, the accuracies were not changed after 2000 steps in most of the cases. Once the test dataset was run on models, the best accuracy was obtained by AdaDelta, Gradient Descent, and RMSProp optimizers, which were 91.3%, and the worst accuracy was obtained by Momentum optimizer which was 82.6% as listed in Table III.

TABLE II. FINAL ACCURACY WITH RESPECT OPTIMIZERS

Optimizer	Accuracy (%)
AdaDelta	91.3
Gradient Descent	91.3
RMSProp	91.3
AdaGrad	89.1
Adam	87.0
Momentum	82.6

The IoT module and the mobile application were tested by placing it in a nearby place and observed the audio which was gathered around that place was saved in the database and successfully retrieved by the mobile application when request.

Then using the already included pre-trained model, the audio was recognized as one of the species that observed. Hence the both IoT module and mobile application were functioning well in the given conditions.

V. Conclusion and Future work

In this study, a system was proposed to classify a given animal sound clip for one of the pre-known ten animal classes. As described during this paper, the collected sound clips were preprocessed using MFCC and fed those to the created CNN architecture. The training and test were done using 417 audio clips, which were belonged to ten different animal species. The confusion matrices were generated for both test and validation datasets. One of the previous study [9] was shown that the accuracies were affected by class imbalance. To avoid this issue, the balanced classes were used. As they achieved 75% accuracy, this study was achieved 91.3% accuracy. The most suitable model was the model that used AdaDelta optimizer with 91.3% accuracy. Furthermore, a proper system was proposed to classify animals who lived in faraway locations by voice.

It was a well-known fact that deep learning strategies were required a large number of a labeled dataset to obtain much better results. Hence, the dataset will be expanded to have more than 417 audio clips in future studies to achieve more accurate and better results. Also, it will be better to consider changing other hyper-parameters rather than optimizers and specifications of collected audio clips.

The designed hardware module was supposed to send a gathered audio file to a cloud server and retrieve it by a mobile application on demand and it was functioning well. Rather than using a dataset that collected from online libraries, in the future extension, the data collected from this hardware module will be used to train the CNN model and to identify the animals, because it takes about months to collect a proper sample by placing a single hardware module in different forestry places.

In this study, the TensorFlow model was used. In a future study, to build a CNN model from scratch and conduct the CNN training for achieving a much better result is recommended.

In a future study, this same system can be improved to classify animals by their gender as well because there is a slight difference of voice between male and female animals of the same species. Also, this system can be improved to classify animals by their age groups, because cubs and elder ones of the same species have two different vocal patterns.

Furthermore, it is recommended that this system be developed to a more advanced level to recognize the mood of the same animal in the different stages of the day that they live.

References

[1] BYJUS. (2019). Classification Of Animal Kingdom - Non-chordates And Chordates. [online] Available at: https://byjus.com/biology/classification-of-animal-kingdom/ [Accessed 23 Nov. 2019].

[2] C. Mora, D. P. Tittensor, S. Adl, A. G. B. Simpson, and B. Worm, "How many species are there on Earth and in the ocean?," *PLoS Biol.*, vol. 9, no. 8, p. e1001127, Aug. 2011.

[3] Che Yong Yeo, S. A. R. Al-Haddad, and C. K. Ng, "Animal voice recognition for identification (ID) detection system*," in 2011 IEEE 7th International Colloquium on Signal Processing and its Applications*, 2011, pp. 198–201.

[4] M. C. Munteanu, A. Caliman, C. Zaharia, and D. Dinu, "Convolutional neural network," 10497089, 03-Dec-2019.

[5] S. Ajibola Alim and N. Khair Alang Rashid, "Some Commonly Used Speech Feature Extraction Algorithms," *in From Natural to Artificial Intelligence - Algorithms and Applications*, R. Lopez-Ruiz, Ed. IntechOpen, 2018.

[6] K.-H. Frommolt and K.-H. Tauchert, "Applying bioacoustic methods for long-term monitoring of a nocturnal wetland bird," *Ecol. Inform.*, vol. 21, pp. 4–12, May 2014.

[7] "Macaulay Library." [Online]. Available: https://www.macaulaylibrary.org/. [Accessed: 14-Dec-2019].

[8] "xeno-canto: Sharing bird sounds from around the world." [Online]. Available: https://www.xeno-canto.org/. [Accessed: 14-Dec-2019].

[9] "the Internet Bird Collection | HBW Alive." [Online]. Available: https://www.hbw.com/ibc. [Accessed: 14-Dec-2019].

[10] "Free Sound Effects." [Online]. Available: https://www.freesoundeffects.com/. [Accessed: 14-Dec-2019].

[11] T. Grill and J. Schluter, "Two convolutional neural networks for bird detection in audio signals," 2017 *25th European Signal Processing Conference (EUSIPCO).* 2017.

[12] T. L. H. Li, A. B. Chan, and A. H. W. Chun, "Automatic musical pattern feature extraction using convolutional neural network," *in Proceedings of the International MultiConference of Engineers and Computer Scientists* 2010, IMECS 2010, 2010, pp. 546–550.

[13] T. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," *INTERSPEECH*, 2015.

[14] A. Khamparia, D. Gupta, N. G. Nguyen, A. Khanna, B. Pandey, and P. Tiwari, "Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network," *IEEE Access*, vol. 7, pp. 7717–7727, 2019.

[15] J. O. Nordby, "Environmental sound classification on microcontrollers using Convolutional Neural Networks," *Norwegian University of Life Sciences*, 2019.

[16] E. Sasmaz and F. Boray Tek, "Animal Sound Classification Using A Convolutional Neural Network," *2018 3rd International Conference on Computer Science and Engineering (UBMK).* 2018.

[17] "Nature Sounds - Rain Sounds: Free Download, Borrow, and Streaming: Internet Archive." [Online]. Available: https://archive.org/details/relaxingrainsounds. [Accessed: 14-Dec-2019].