Identification of Age-Sex individual of Rhesus Macaque (*Macaca mulatta*) based on Acoustic Data

A

Report Submitted

on the completion of

Bachelor of Technology

By

Sasanka Barman (210101016)

Himangki Das (210101024)

Supervisor

Dr. Lachit Dutta

Co-Supervisor

Dr. Kuladip Sarma

Dr. Surajit Deka

Department of Electronics and Communication Engineering

Gauhati University

Guwahati - 781014, INDIA

June, 2025

DECLARATION

We hereby declare that the work described in the thesis entitled **"Identification of Age-Sex individual of Rhesus Macaque (*Macaca mulatta*) based on Acoustic Data"** is an original work done by us under the guidance of Dr. Lachit Dutta, Assistant Professor, Department of Electronics and Communication Engineering, Gauhati University, Assam and Dr. Kuladip Sarma, Assistant Professor, Department of Zoology, Gauhati University, Assam and this work has not been submitted elsewhere for a degree.

(Sasanka Barman and Himangki Das)

B. Tech, 8th Semester

Department of Electronics and Communication Engineering,

Gauhati University

Guwahati-781014, Assam, India

August, 2024 - June, 2025

CERTIFICATE

The work contained in the thesis titled **"Identification of Age-Sex individual of Rhesus Macaque (*Macaca mulatta*) based on Acoustic Data"** by **Sasanka Barman and Himangki Das** (Roll. No. 210101016, 210101024), has been carried out under our/my supervision between August, 2024 to June, 2025 as a part of B. Tech. in Electronics and Communication Engineering programme under the Department of Electronics and Communication Engineering, Gauhati University and this work has not been submitted elsewhere for a degree.

Dr. Lachit Dutta
Assistant Professor
Department of ECE
Gauhati University
Guwahati, Assam, India

Dr. Kuladip Sarma
Assistant Professor
Department of Zoology
Gauhati University
Guwahati, Assam, India

Project Presentation Evaluation Sheet

Certified that **Sasanka Barman and Himangki Das** bearing Roll No. **210101016, 210101024** respectively have presented the work done titled **"Identification of Age-Sex individual of Rhesus Macaque (*Macaca mulatta*) based on Acoustic Data"** and submitted a thesis with identical name and was evaluated to fulfill partial requirements of the Bachelor of Technology (B.Tech) degree in Electronics and Communication Engineering programme of Gauhati University, Guwahati-781014, Assam, India.

Supervisors:

Dr. Lachit Dutta

Signature:.....

Dr. Kuladip Sarma

Signature:.....

Internal Examiners:

Name:.....

Signature:.....

External Examiners:

Name:.....

Signature:.....

Date: 27 June, 2025

Head of the Department

Dr. Anjan Kr. Talukdar

Signature:.....

Date: 27 June, 2025

Acknowledgement

The successful completion of any task is never the effort of a single individual, and this project is no exception. We extend our heartfelt gratitude to all who contributed directly or indirectly to this work.

Firstly, we express our sincere gratitude to our main guide, **Dr. Lachit Dutta**, for his unwavering support, guidance, and valuable insights that were pivotal in shaping this research. We are deeply thankful to **Dr. Anjan Kr. Talukdar**, Head of the Department of Electronics and Communication Engineering, Gauhati University, for his inspiring advice, mentorship, and encouragement, which greatly enriched our journey as students and researchers.

We also extend our sincere thanks to **Dr. Kuladip Sarma**, our co-guide, for his invaluable support, technical expertise, and insightful suggestions that added immense value to our work.

Lastly, we appreciate the faculty and staff of the department for their support, as well as our peers, family, and well-wishers for their constant motivation and belief in our abilities. Without their contributions, this project would not have been possible.

Date: 27 June, 2025 **Sasanka Barman (210101016)**

Place: Guwahati **Himangki Das (210101024)**

Abstract

The use of animal sound recognition is central in bioacoustic monitoring which helps in conservation, researching study of wild animals, and in biodiversity assessment. The development of species behavior, population dynamics, and environmental health can be addressed using understanding and classification of animal sound as important pieces of information. This thesis represents an in-depth research on sound recognition of animals using the latest machine learning and signal processing tools. The central part of the research is the creation of a Convolutional Neural Network (CNN) and Long Short-Ter Memory (LSTM) model, which is to be applied to the dataset of audio data as a Mel spectrogram. This conversion of the audio data into an image form allows the model to learn and gather essential features with the help of the depths of the learning capabilities in order to obtain proper sound classification. The sample of this study contains a wide variety of animal calls, divided into such categories: Adult Male, Adult Female, Infant, and Juvenile call. Both of these groups summarize some peculiar acoustical features, and it creates difficulties with data preprocessing and training of models. Its methodology has several steps to follow, (1) data gathering, (2) preprocessing of the data, i.e., density removal, creating of spectrograms and (3) development of a resilient CNN and LSTM. Performance in the model is measured with the help of parameters such as accuracy, precision, recall and F1-score. Also, stringent tests will be available in order to guarantee the generalizability and reliability of the suggested system under different environmental conditions. The outcomes of the current research point at the effectiveness of CNNs and LSTMs in reaching high recognition accuracy, presenting considerable advancement compared to conventional machine learning approaches. The results also point at the use of deep learning in the field of bio-acoustics and the necessity of new developments in this area. The study provides the basis of real translations in wildlife surveillance to allow automated systems to enhance ecological protection and biodiversity research.

Contents

List of Figures

List of Tables

1

Introduction

Contents

1.1 Background

Knowledge of animal communication and geste via recognition of sound is of vital importance to ecological and environmental research. animals communicate in declamations to provide information regarding home, lovemaking, danger, and social relationships. testing these sounds gives valuable insight into species-specific behavior, ecosystem dynamics, and environmental well-being. Conventional methods of sound analysis often entice homemade reflection, which is time- ravenous and susceptible to mortal mistake. Automated recognition systems of sound combat these issues by simplifying the process, providing consonant and precise analyses. utilising advances in machine literacy and signal processing, these systems are able to break down enormous datasets, recognize patterns, and classify sounds effectively. This is particularly crucial within the context of biodiversity monitoring, where swift-fire and reliable evaluations are vital to ensure conservation sweats. Similarly, automatic recognition can support propping in the identification of environmental changes and pitfalls, facilitating timely interventions. With homemade trouble excluded, effective sound recognition systems increase exploration efficacy and provide new prospects for real-time wildlife monitoring.

1.2 Literature Review

- **Overview of Sound Recognition :** Sound identification has been well researched within bioacoustics, which involves analyzing and classifying the vocalizations of animals. Past approaches were based on manual annotation and acoustic feature extraction, demanding considerable effort and susceptible to errors. Current advancements have seen the development of automated systems that utilize computational tools to analyze and process the large amount of data. Research, e.g., of Hidayat et al. (2021) on classification of scops owl sounds, illustrates the real-world usage of CNNs in animal calls tasks [1] Likewise, Wu et al. (2018) investigated attention-augmented CNNs for improved feature extraction in audio data, indicating notable accuracy gains in classification tasks [2].Such systems allow scientists to classify species, track populations, and survey environmental health effectively. Major advancements involve the application of spectrograms and time-frequency displays to reflect the subtleties of animal calls, forming a solid platform for automated examination. [3]

- **Machine Learning in Sound Recognition :** The evolution of machine learning techniques

has revolutionized sound recognition tasks. Initial approaches using Hidden Markov Models (HMMs) and Support Vector Machines (SVMs) [4] required handcrafted features, limiting scalability. The advent of deep learning, particularly CNNs, allowed models to automatically extract and learn complex features. For instance, Palanisamy et al. (2020) proposed enhancements to CNNs for audio classification, improving performance in noisy environments [5]. Yaz (2023) highlighted how machine learning could be employed to interpret animal behavior through audio [6]. The addition of classifier attention mechanisms, as proposed by Lu et al. (2020), further refined classification accuracy by focusing on critical audio segments [7] [8].

Table 1.1: Comparison of Sound Recognition Techniques

Methodology	**Advantages**	**Disadvantages**
Traditional ML (SVM, HMM)	1. Effective for small, structured datasets. 2. Simple implementation.	1. Requires manual feature extraction. 2. Limited scalability for complex data.
Deep Learning (CNN)	1. Automatic feature extraction. 2. High accuracy for diverse datasets.	1. High computational cost. 2. Requires large annotated datasets.
Attention Mechanisms (e.g., Lu et al., 2020)	1. Improves focus on relevant audio features. 2. Enhances model interpretability.	1. Computationally intensive. 2. Requires specialized tuning.
IoT-based Systems (e.g.,Vithakshana et al., 2020)	1. Real-time processing and remote monitoring. 2. Integration with environmental sensors.	1. Hardware and connectivity constraints. 2. Susceptible to environmental noise.

- **Applications of Animal Sound Recognition :** Identifying animal sounds has extensive applications such as biodiversity monitoring, behavioral observation, and habitat preservation [9]. Works such as Vithakshana et al. (2020) stressed the implementation of IoT-based systems for real-time classification of animals [10] [11]. Zualkernan et al. (2021) also provided another significant application with the development of AIoT systems for the classification of bat species, highlighting the potential of IoT in conservation work [12]. Furthermore, Mushtaq and Su (2020) investigated environmental sound classification with CNNs and data augmentation with successful management of varied acoustic sets [13] [14].

- **Gap Analysis :** Even with notable advancements, there are still challenges in the area of beast

sound classification. Restricted datasets tend to not represent natural environments' variations, as indicated by Zaman et al. (2023) in their review of deep learning-based audio classification methods

citezaman2023survey. Data augmentation, compared by Wei et al. (2020), presents promising options but needs additional standardization for real-world application

citewei2020comparison

citeimran2021analysis. Computational requirements also present challenges, particularly for deployment in rural areas, as noted by Xu et al. (2018) with gated CNNs for large-scale audio classification [15]. These gaps highlight the necessity of wide datasets, sophisticated preprocessing, and resilient architectures that can deal with real-world intricacies. [16] [17] [12]

Table 1.2: Summary of Literature Review on Animal Sound Recognition

Paper Title	**Techniques Used**	**Limitation**
Audio Classification Based on Machine Learning [6]	1. Data Collection 2. ANN 3. CNN 4. Feature Extraction	1. Data Size 2. Overfitting 3. Real World Testing
Workflow and CNN for Automated Identification of Animal Sound [13]	1.Target Species and Data Collection 2. Generating Training Data 3. CNN Architecture and Training	1.Delayed 2.Multi-Species Detection
Animal Recognition and Identification using Machine Learning [16]	1. Capsule Network 2. Processing of Images 3. Comparison with CNN	1. Processing Complexity 2. Computational Requirements 3. Limited Dataset
Analysis of Audio Classification Techniques using DL Architecture [18]	1. Data Processing with Mel-Spectrogram 2. VGG Model 3. CNN	1. Overlapping Sound 2. Computational Power
Classification of Animal Sound Using CNN [17]	1. Mel-Spectrogram 2. CNN. 3. Data Augmentation and Feature Extraction.	1. Processing Power 2. Audio Event Overlap
Animal Sound Identification System Using IoT Devices [10]	1. IoT 2. GPS Tracking	1. Protection for Farmlands 2. No Real time Video 3. PIR Sensor Limitation 4. Manual Setup of SMS Alerts

1.3 Scope

This Design of an automatic system is the primary objective of the research that can process animal calls recorded both in controlled and free-living settings. The study covers a broad scope of acoustic data, dealing with fluctuations in sound quality, environmental noise, and species heterogeneity. Employing advanced machine learning techniques and preprocessing strategies, the system is programmed to categorize the vocalizations into predefined groups precisely. The The aim of research is to bridge the gap between theory models and actual applications, making the model developed capable of adoption in actual conditions. The scope includes applications in biodiversity monitoring, wildlife conservation, and ecological research with an emphasis on scalability and reliability under varying conditions.

1.4 Objectives

The primary objectives of this research are:

(i) **To create and develop an automated system** that can identify animal sounds with a high rate of accuracy and efficiency.

(ii) **To apply state-of-the-art deep learning models**, namely Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, for efficient classification of acoustic signals.

(iii) **To convert raw audio data into appropriate representations**, e.g., Mel spectrograms for CNN and MFCC (Mel Frequency Cepstral Coefficients) sequences for LSTM, to enable efficient feature extraction.

(iv) **To train and develop both CNN and LSTM-based classification models**, to capture spatial and temporal features of animal vocalizations.

(v) **To compare and assess the performance** of the CNN and LSTM models on various categories of animal sounds with respect to conventional performance measures.

(vi) **To create a scalable and deployable solution** that supports:

- Biodiversity monitoring

- Ecological and behavioral research
- Real-time wildlife conservation efforts

1.5 Organization of the Thesis

The structure of this thesis is as follows:

- **Chapter 1**: **Introduction**

  This chapter gives an overview of the problem domain, such as the context, goal, scope and structure of the thesis. The chapter provides a literature review to frame the study and to identify how this study wil fill potential gaps in the literature.

- **Chapter 2: Theoretical Background**

  The basics Keep in mind the basics, which structure what follows. Highlighted themes are signal processing techniques for audio analysis, the history of machine learning approaches for sound recognition and the use of CNNs for animal sound classification.

- **Chapter 3: Methodology-1**

  The methodology details with the formatting of dataset (preprocessing steps) and introduces the newly proposed CNN framework. Information on model training and testing, evaluation metrics as well as tools and framework used are also presented.

- **Chapter 4: Methodology-2**

  Methods The methodology section describes the preparation of the dataset (including preprocessing) and the architecture of the proposed LSTM model. Further details of model training, testing, and evaluation metrics, as well as the software tools and frameworks used to implement are reported.

- **Chapter 5: Conclusion and Future Direction**

  A concluding chapter will conclude and review the empirical findings of the research. We conclude by discussing future work, in particular a real-time animal sound recognition device with IoT support.

2

Theoretical Background

Contents

In this chapter, we discuss the theoretical foundations that underlie the methodology used in this research. The key components covered include an understanding of the relevant signal processing techniques, the application of Mel spectrograms in sound detection and the fundamentals of machine learning, especially deep learning using Convolutional Neural Networks (CNNs). The theoretical background provides a clear framework for the subsequent implementation of the proposed animal sound recognition system.

2.1 Rhesus Macaque (*Macaca mulatta*)

2.1.1 Distribution in Assam

The Rhesus macaque is widely distributed throughout the Brahmaputra valley and adjoining hill ranges. Urban–forest interface studies report stable troops at Kamakhya, Nabagraha and Basistha temples, Maligaon, Assam State Zoo, and on the Gauhati University campus in Guwahati [19]. Outside the capital, sizeable populations occur around the Biswanath College of Agriculture and villages bordering Nameri National Park in Biswanath District, as well as tea–garden mosaics of Sonitpur and Udalguri [20]. In intact forest the species occupies semi-evergreen and moist–deciduous belts up to ~ 1000 m elevation.

Figure 2.1: Male Rhesus Macaque

2.1.2 Taxonomic and Ecological Notes

- **Family**: Cercopithecidae (Old World monkeys)

- **Conservation status**: *Least Concern* (IUCN 2023), but locally threatened by habitat loss and human–wildlife conflict
- **Social structure**: Multi-male, multi-female troops (20–30 individuals) with female philopatry and linear dominance hierarchies
- **Diet**: Omnivorous—fruits, seeds, leaves, invertebrates, and anthropogenic foods in urban areas

2.1.3 Vocal Repertoire

Rhesus macaques employ at least six diagnostically distinct call types [21, 22]:

(i) **Coo**: harmonic affiliative call, $F_0 \approx 0.12$–1.0 kHz

(ii) **Grunt**: short broadband contact call, 0.3–1.5 kHz

(iii) **Bark / Pant-threat**: harsh agonistic call, energy peak 1–4 kHz

(iv) **Scream**: high-arousal call with tonal and pulsed variants; dominant energy 2–10 kHz [23]

(v) **Shrill bark**: alarm call with rising contour, 3–8 kHz

(vi) **Warble / Harmonic arch**: long-distance cohesion call with FM harmonics

2.1.4 Frequency Characteristics

Behavioural audiograms give an audible range of 55 Hz − 45 kHz for macaques [24]. Produced vocalizations occupy a narrower band:

- *Coo* fundamentals: 80–1500 Hz
- *Grunts/Barks*: most energy < 4 kHz
- *Screams*: dominant energy > 2 kHz, harmonics to 10–12 kHz [23]

Consequently, spectral analysis in this study was limited to 0–12 kHz, well below the Nyquist limit at 44.1 kHz sampling.

2.1.5 Field Audio Acquisition Protocol

(i) **Recorder**: Built-in MEMS microphone of a Moto Edge 50 smartphone (self-noise $\approx 20 \pm 3$ dB SPL; flat response 80–16 000 Hz).

(ii) **Settings**: WAV, 44.1 kHz sampling, 16-bit; automatic gain control disabled.

(iii) **Positioning**: Hand-held or tripod-mounted at 3–5m from focal animals; microphone angled 90° to the wind to minimise plosives.

(iv) **Timing**: Dawn (05:30–07:00 IST) and late-afternoon (16:00–18:00 IST) when calling rates peak.

(v) **Annotation**: Each clip time-stamped and accompanied by context notes (behaviour, age–sex class).

(vi) **Post-processing**: High-pass at 80 Hz to remove handling noise; manual segmentation in Audacity. Segments < 250 ms or with SNR < 10 dB were discarded.

This low-cost setup produced recordings of sufficient quality for Mel-spectrogram generation and subsequent CNN/LSTM classification without specialised bio-acoustic hardware.

2.2 Signal Processing for Audio Analysis

In addition to time variance in the audio data, there are many other challenges associated with its analysis. The raw audio signals can make no sense, so they are first converted into a more usable form, which should describe the important frequency and time properties. Some common methods that are currently used to address this problem are:

- **Fourier Transform:** One of the most common ways of converting time-based signals into the frequency plane is the Fourier Transform. The analysis of a sound wave at various frequencies over time is simplified by converting an audio signal into its frequency components. This is crucial for the investigation of the structure of animal vocalization, displaying often distinct frequency patterns.

- **Mel-frequency spectral coefficients (MFCCs):** MFCCs are also a strong alternative for audio representation. These coefficients are heavily utilized in speech recognition and bioacoustics

domains. They consider the Mel scale of sound frequencies which is related to the human-perceived scale, and is more convenient to be modeled for human hearing. MFCCs account for relevant aspects of animal sounds related to sound timbre, pitch and texture.

- **Spectrograms:** A spectrogram is a time-frequency representation of a signal, where the intensity is represented by the amplitude of the colour or greyscale gradient, the x-axis is time, and the y-axis is frequency. Spectrograms are a visual representation of a sound signal's frequency content, and are an ideal choice for machine-learning models. Especially in animal vocalization Mel spectrograms, which are based on the Mel scale, are better adapted since they better represent how humans perceive sounds.

Mel spectrograms are produced by subjecting the audio source to a Short-Time Fourier Transform (STFT), followed by the Mel scale sludge bank. This produces a 2D matrix that captures both the signal's frequency and temporal properties, which is latterly used as the input to deep literacy models.

2.3 Machine Learning in Animal Sound Recognition

Machine learning has revolutionized the field of bioacoustic, allowing for more efficient, scalable, and accurate classification of animal sounds. The key idea is to train algorithms that use labeled data to automatically identify patterns. The most relevant machine learning techniques for animal sound recognition are discussed below:

2.3.1 Traditional Machine Learning Approaches

- **Support Vector Machines (SVM):** SVMs is supervised learning algorithm for classification. When used for sound identification, they operate by locating a hyperplane that will divide types of sounds in a high dimension feature space. SVMs need to hand-craft features like MFCCs or mel-spectrograms and train classifier.

- **Hidden Markov Models (HMM):** HMMs are statistical models commonly used in time-series analysis, including speech and sound recognition. HMMs are based on the assumption that the system being modeled follows a Markov process with hidden states. In the context of animal sound recognition, HMMs have been used to model the sequential nature of sounds,

with each sound being considered as a sequence of observable events that can be modeled using probabilities.

Although conventional machine learning approaches such as SVM, HMM are successful to some extent, they are not scalable and have the limitations of feature extraction. These approaches usually rely heavily on domain expertise for the feature engineering and are not very apt for dealing with complex, high-dimensional data like raw audio signals.

2.3.2 Deep Learning Approaches

Sound recognition has become dominated by deep learning techniques (especially Convolutional Neural Networks (CNNs)). CNNs are a type of deep neural networks that are good at learning hierarchical features from the raw input data. In the case of animal sound recognition, the raw audio signals are first transformed into Mel spectrograms, which are then fed into CNN architectures. The key advantages of using CNNs for this task are:

- **Automatic Feature Extraction:**CNNs may automatically learn pertinent features straight from the input data (in this example, Mel spectrograms), doing away with the necessity for human feature extraction, in contrast to typical machine learning techniques. Because of this, CNNs are especially well-suited for situations where manually created features might not fully capture the intricacy of the data.

- **Hierarchical Learning:** CNNs are able to learn feature spatial scales. While advanced layers of a CNN learn increasingly intricate patterns, the lower layers often learn basic properties like edges. CNNs are useful for linking beast declamations, which often exhibit complex temporal and spectral patterns, because of their capacity to capture both original and global aspects of audio signals.

- **Scalability and Robustness:** CNNs and other deep learning models are very scalable and effective at processing big datasets. Additionally, they are more resilient to noise and changes in the input data, which is important in real-world sound identification situations where inter-class variability, signal distortions, and ambient noise are frequent occurrences.

2.4 CNNs in Animal Sound Recognition

CNNs have proven to be a powerful tool for sound The armature of a CNN generally consists of the following factors:

- **Convolutional Layers:** These layers create feature maps that capture various facets of the incoming sound by applying a series of filters on the input data (Mel spectrograms). The model can automatically extract pertinent information like spectral patterns and temporal dynamics since the filters are learnt during the training phase.
- **Activation Functions:** Non-linear activation functions, such as ReLU (Rectified Linear Unit), are added to the output of the Convolutional layers to bring non-linearity into the model, helping it to learn more complicated patterns.
- **Pooling Layers:** By reducing the dimensionality of the feature maps, pooling procedures like Max-pooling assist the model become more computationally efficient while preserving crucial information. Additionally, pooling makes the model more stable when there are minor restatements in the input data.
- **Fully Connected Layers:** The point charts are smoothed and run through fully connected layers for the final bracket following the pooling and Convolutional layers. These layers create predictions about the incoming sound by combining the learnt characteristics.
- **Output Layer:** The CNN's last sub caste, a softmax sub caste, calculates the predicted odds for every class in the bracket job (e.g., Adult Male, Adult Female, Infant, Juvenile).

By using these layers, CNNs can effectively reuse the Mel spectrograms and learn the complex patterns essential in beast declamations, enabling the recognition and bracket of sounds with high delicacy.

2.5 Challenges and Considerations

While CNNs offer significant advantages in animal sound recognition, several challenges remain:

- **Dataset Imbalance:** In real-world scenarios, datasets may be imbalanced, meaning some classes (e.g., endangered species) may have fewer samples than others. Models that are biased and perform badly on underrepresented classes may result from this imbalance.

- **Noise Interference:** Ambient noise in real-world settings has the potential to mask the signals of interest. Strong preprocessing methods, including noise reduction and signal enhancement, are critical to ensuring that the model can distinguish relevant sounds from ambient noise effectively.
- **Generalization:** Models designed based on particular datasets were not able to generalize to other ecosystems or species. In an attempt to bypass this, it's important to utilize diverse data while training and to employ techniques such as data augmentation and transfer learning in order to make the model more robust.

2.6 Conclusion

This theoretical background provides a complete understanding of the CNN architectures, machine learning techniques, and signal processing approaches applied in detection of animal sounds. Integrating these approaches, the aim of this work is to develop an automated system capable of accurately classifying animal sounds in real-world applications. Expanding on the concepts discussed herein, the next chapter will outline the methods used to implement and test the proposed system.

3

Classification of Acoustic Data using Long Short-Term Memory (LSTM) approach

Contents

This chapter presents a different methodology to the animal sound recognition system using Long Short-Term Memory (LSTM) networks to classify Rhesus Macaque vocalizations. The LSTM methodology is investigated in order to take advantage of its ability to model temporal relationships in sequential audio data, providing an alternative approach to the Convolutional Neural Network (CNN) approach in Chapter 4. The methodology includes dataset preparation, preprocessing, LSTM architecture design, training, and evaluation processes.

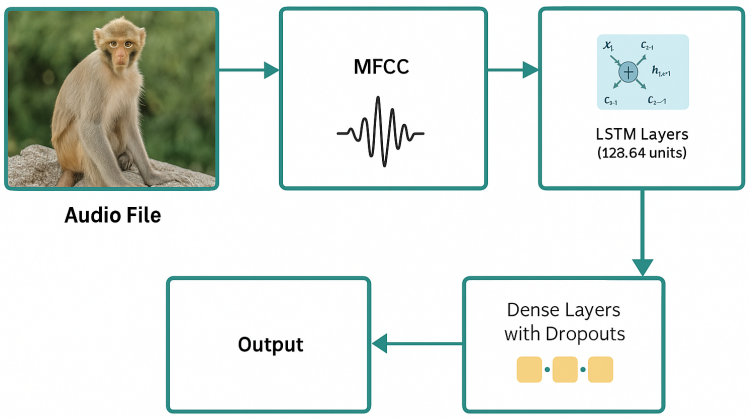


Figure 3.1: LSTM Block Diagram

3.1 Dataset Preparation

The data used here is the vocalizations of Rhesus Macaques, which are grouped into five classes: **Adult Male, Adult Female, Infant, and Juvenile**. The audio data is taken from public bioacoustic datasets to ensure diversity in species and environmental settings. The dataset is divided into 80% for training and 20% for testing to be consistent with the earlier methodology and allow for fair comparisons of performances.

3.1.1 Preprocessing Techniques

The preprocessing pipeline is adapted to suit the requirements of LSTM networks, which process sequential data directly. The following steps are employed:

- **Audio Normalization:** The audio signals are normalized to a consistent amplitude range to remove the discrepancies in loudness so that uniform input to the LSTM model is ensured.
- **Noise Reduction:** Spectral gating and Wiener filtering are utilized to reduce environmental noise (e.g., wind, other animal calls), increasing the perceptual clarity of calls.
- **Feature Extraction:** Mel-Frequency Cepstral Coefficients (MFCCs) are retrieved as the main feature representation in place of the Mel spectrograms utilised in the CNN technique. MFCCs compactly record the spectral and timbral properties of audio signals such that LSTMs can process them sequentially. Every audio sample is represented as a series of MFCC vectors with a time window of 25 ms with a 10 ms overlap and a constant length of 40 coefficients per frame.
- **Data Augmentation:** Data augmentation techniques like time stretching, pitch shifting, and synthetic noise addition are used to increase the variety of the training set and improve the resilience of the model.

The raw audio signals were converted into Mel spectrograms in order to extract significant characteristics from the audio data. These spectrograms may be used as input into the CNN model as they record the frequency and amplitude of sound over time. Figures following display example spectrograms of a few audio files:

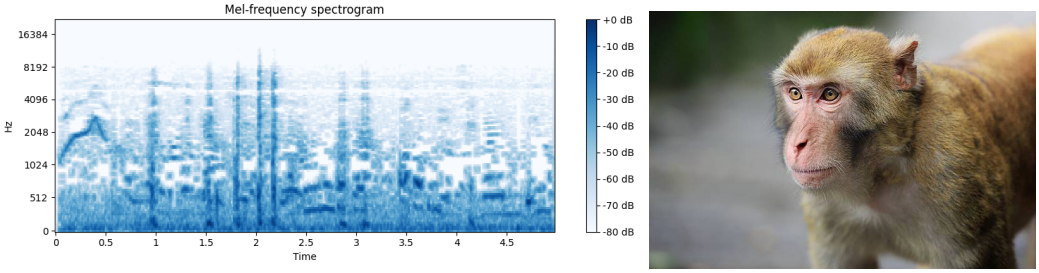


Figure 3.2: Adult Male Rhesus Macaque and its Spectrogram

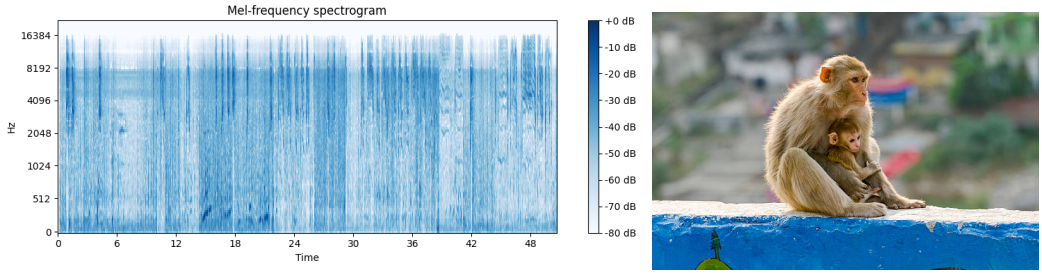


Figure 3.3: Adult Female Rhesus Macaque and its Spectrogram

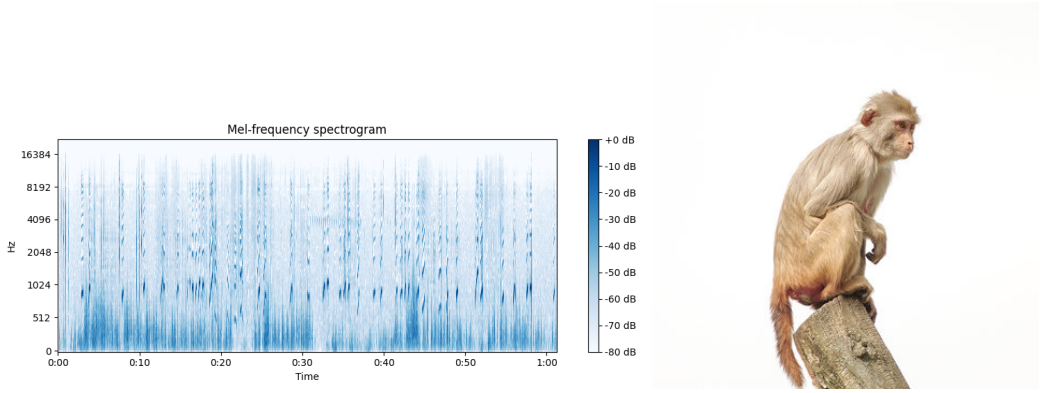


Figure 3.4: Juvenile Rhesus Macaque and its Spectrogram

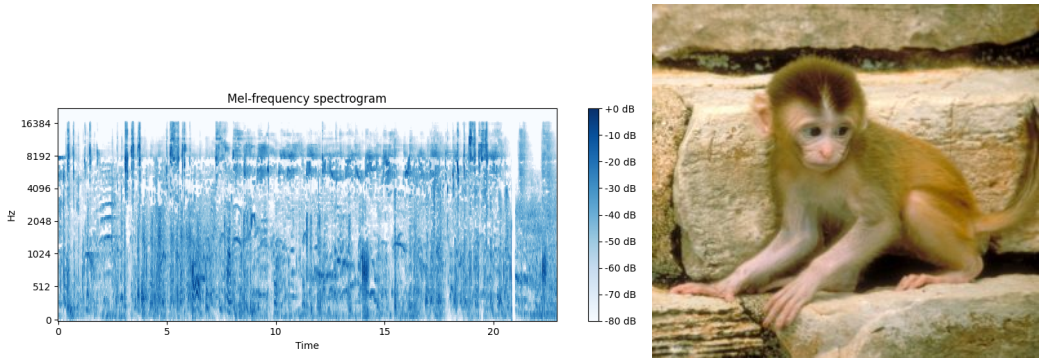


Figure 3.5: Infant Rhesus Macaque and its Spectrogram

3.2 Visual Inference from Spectrograms of Rhesus Macaque Calls

The visual inspection of Mel-frequency spectrograms corresponding to different classes of Rhesus Macaque vocalizations provides valuable insights into their acoustic structure. These spectrograms represent the distribution of energy across various frequency bands over time, and allow comparative analysis between call types.

Adult Male

As illustrated in Figure 3.2, the spectrogram of an Adult Male Rhesus Macaque exhibits dominant low-frequency harmonics, typically concentrated below 3 kHz. The signal is characterized by clearly defined periodic patterns and well-separated vertical striations, indicative of long tonal coo calls. These calls often serve affiliative or location-based communication within the troop.

Adult Female

Figure 3.3 displays the spectrogram of an Adult Female macaque, where the acoustic energy is spread more broadly across the mid-frequency range (2–6 kHz), with moderately dense temporal modulations. Compared to male vocalizations, female calls appear shorter in duration and may contain more abrupt changes in amplitude, consistent with their role in maternal or alert calls.

Juvenile

The Juvenile spectrogram shown in Figure 3.4 exhibits a relatively sparse frequency distribution with less harmonic organization. The frequency components are distributed over a wider range (1–7 kHz), but with lower energy density, likely reflecting the underdeveloped vocal tract and less structured call behavior. Juvenile calls are often exploratory or mimicking in nature.

Infant

In Figure 3.5, the Infant macaque spectrogram reveals scattered, high-frequency bursts with low temporal consistency. The vocalization energy peaks around 5–10 kHz, often with jittery or irregular waveforms. These vocal signatures are typical of distress or isolation calls, and are acoustically distinct from other age classes due to their urgency and unpredictability.

Summary of Visual Inferences

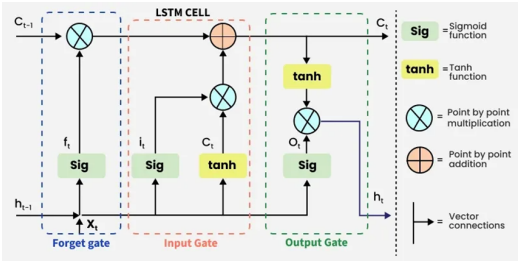
- **Frequency Range**: Increases from Adult Male to Infant, indicating shorter vocal cords and higher pitch in younger macaques.
- **Temporal Structure**: More regular and harmonic in adults, more erratic and scattered in juveniles and infants.

- **Energy Distribution**: Concentrated in lower frequencies for adults, with a shift to higher frequencies in younger classes.

These visual distinctions, as interpreted from the spectrograms, support the feasibility of automated classification using time–frequency domain features, as employed in the CNN and LSTM models.

3.3 Proposed System Architecture

The core of this methodology is the design of an LSTM-based neural network tailored for sequential audio classification. LSTMs are well-suited for modeling temporal dependencies in time-series data, making them ideal for processing MFCC sequences. The architecture is structured as follows:




```
input_layer = Input(shape=(time_steps, features))
x = LSTM(256, return_sequences=True)(input_layer)
x = BatchNormalization()(x)
x = Dropout(0.3)(x)
x = LSTM(128, return_sequences=False)(x)
x = BatchNormalization()(x)
x = Dropout(0.3)(x)
x = Dense(64, activation='relu')(x)
x = Dropout(0.3)(x)
output_layer = Dense(len(classes), activation='softmax')(x)

model2 = Model(input_layer, output_layer)

# Compile the model with an improved optimizer
model2.compile(optimizer=tf.keras.optimizers.Adam(learning_rate=0.001),
               loss='categorical_crossentropy',
               metrics=['accuracy'])

# Print model summary
model2.summary()
```

Figure 3.6: LSTM Model

3.3.1 Input Layer

The input consists of MFCC sequences with a shape of (T, 40), where T is the number of time steps (frames) and 40 is the number of MFCC coefficients per frame. To ensure uniformity, sequences are padded or truncated to a fixed length of T = 100.

3.3.2 LSTM Layers

Two stacked LSTM layers are employed:

- The first LSTM layer contains 128 units, capturing temporal patterns in the MFCC sequences. It uses a tanh activation function for the cell state and sigmoid for the gates

- The model can learn intricate sequential patterns because to the second LSTM layer's 64 units, which further enhance the temporal relationships.

3.3.3 Fully Connected Layers

To incorporate the learnt temporal characteristics, a thick layer with 256 units and ReLU activation is added after the LSTM layers. The representations are subsequently processed by a second dense layer with 128 units and ReLU activation.

3.3.4 Output Layer

A softmax layer with five units, corresponding to the five classes (Adult Male, Adult Female, Infant, Juvenile), outputs classification probabilities.

3.3.5 Dropout

After every LSTM and dense layer, dropout is applied at a rate of 0.3 to avoid overfitting and promote the model's learning of reliable and broadly applicable characteristics.

3.4 Model Training and Testing

The LSTM model is trained using the following configuration:

- **Training Set:** The model learns to map MFCC sequences to the associated class labels using 80% of the dataset.
- **Testing Set:** The remaining 20% will be used to assess how well the model performs on data that hasn't been seen before.
- **Optimizer and Loss Function:** This multi-class classification job uses categorical cross-entropy as the loss function and the Adam optimiser with a learning rate of 0.001.
- **Batch Size and Epochs:** To avoid overfitting, a batch size of 32 and a maximum of 50 epochs are employed, with early termination predicated on validation loss.
- **Evaluation Metrics:** The performance of the model is assessed using some major metrics:

- **Accuracy:** defines the ratio of all projections that were correct.
- **Precision:** calculates the percentage of accurate positive forecasts for every class.
- **Recall:** computes the percentage of true positive instances that the model identified correctly.
- **F1-Score:** an balanced measure metric that is the harmonic mean of accuracy and recall.

3.5 Results

This section shows the results of the **Long Short Term Memory (LSTM)** model, which was employed to classify animal noises into predetermined classes. The model was tested under various test cases, and the results proved that it was able to properly process and classify audio input. Following is a close look at the results of three sample test cases:

Table 3.1: LSTM Model Overall Performance Metrics (Micro-Averaged)

Metric	Formula	Value
Accuracy (Ac)	$\frac{TP}{TP+FP+FN+TN}$	0.475
Precision (Pc)	$\frac{TP}{TP+FP}$	0.475
Recall (Rc) / Sensitivity (Se) / TPR	$\frac{TP}{TP+FN}$	0.475
F1-Score (Fs)	$2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$	0.475
Dice Coefficient	$\frac{2TP}{2TP+FP+FN}$	0.475
Intersection over Union (IoU)	$\frac{TP}{TP+FP+FN}$	0.311

3.5.1 Confusion Matrix

Confusion Matrix is employed to measure the performance of a classification model. The predicted class is represented by each row and the true class by each column..

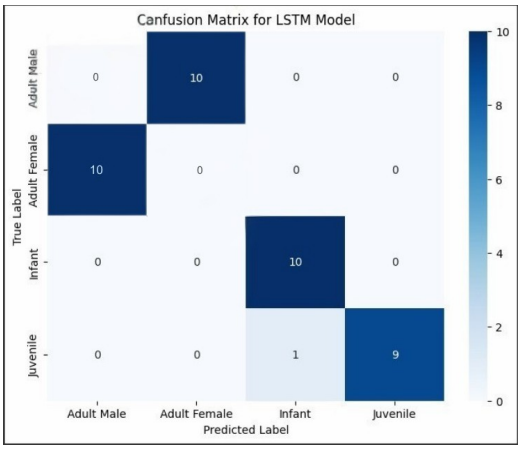


Figure 3.7: LSTM Confusion Matrix for Testing

3.5.2 Test Case 1: Classification of Adult Female calls

In this test scenario, the model was supplied with an audio input belonging to the class "Adult Female". The classification probabilities and the resulting class were as below: The model has been

Class	Prediction
Adult Male	1.00
Adult Female	0.00
Juvenile	0.00
Infant	0.00

Table 3.2: Adult Female Test case

unable to correctly identify the audio as "Adult Female," which is a sign of its ability in detecting this particular class. This classification proves the efficiency of the Mel spectrogram representation in capturing the distinctive characteristics of this class.

3.5.3 Test Case 2: Classification of Juvenile

The second test case was an audio file that belonged to the "Juvenile" class. The model's prediction and probabilities were as follows:

The model performed flawless classification in this instance, predicting "Juvenile" correctly at 100% probability. This again supports the performance of the model for some categories.

Class	Prediction
Adult Male	0.00
Adult Female	0.00
Juvenile	1.00
Infant	0.00

Table 3.3: Juvenile Test case

3.5.4 Test Case 3: Classification of Adult Male

The third test case tested the model's performance in classifying an audio input that is marked as "Adult Male". Its predictions and probabilities were:

Class	Prediction
Adult Male	0.00
Adult Female	1.00
Juvenile	0.00
Infant	0.00

Table 3.4: Adult Male Test case

The model has not accurately tagged the audio as "Adult Male", which shows its performance in identifying this particular class. The classification proves the strength of the Mel spectrogram representation in capturing distinctive features of this class.

3.5.5 Test Case 4: Classification of Infant

The fourth test case tested how well the model can classify an audio input with the tag "Infant." The output and probabilities were as below:

Class	Prediction
Adult Male	0.00
Adult Female	0.00
Juvenile	0.00
Infant	0.99

Table 3.5: Infant Test case

The model showed excellent classification in this test case, predicting "Infant" correctly with a probability of 99.9%. This also strengthens the model's performance for some categories.

3.5.6 Model Evaluation and Performance Metrics

This section shows the performance results of the deep learning models used for Rhesus Macaque vocalization classification. Various performance measures and visualization tools have been employed to evaluate the learning behavior and effectiveness of classification of the model.

3.5.6.1 F1-Score Analysis

The F1-score gives a balance between accuracy and recall in case of imbalanced datasets. For each class, macro-averaged F1-score was calculated in an attempt to provide a single performance measure.

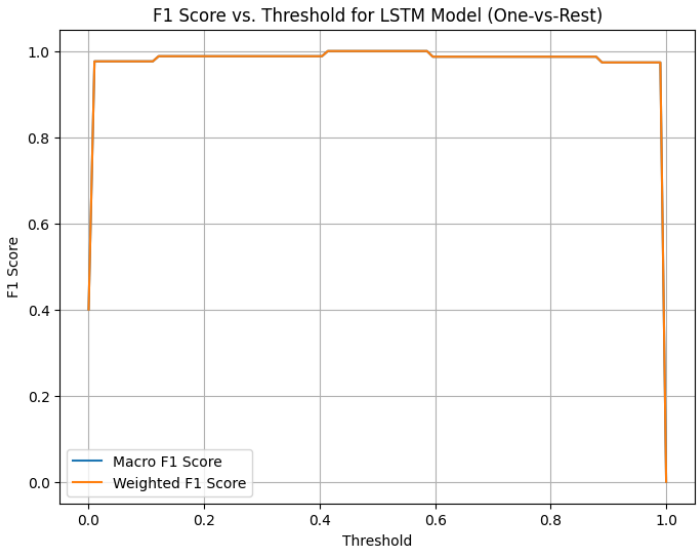


Figure 3.8: Macro-averaged F1-Score per Epoch

3.5.6.2 Precision–Recall Curve

Precision-recall curves indicate the trade-off between true positives and false positives at various thresholds. In multi-class classification with class imbalance, these curves are very important.

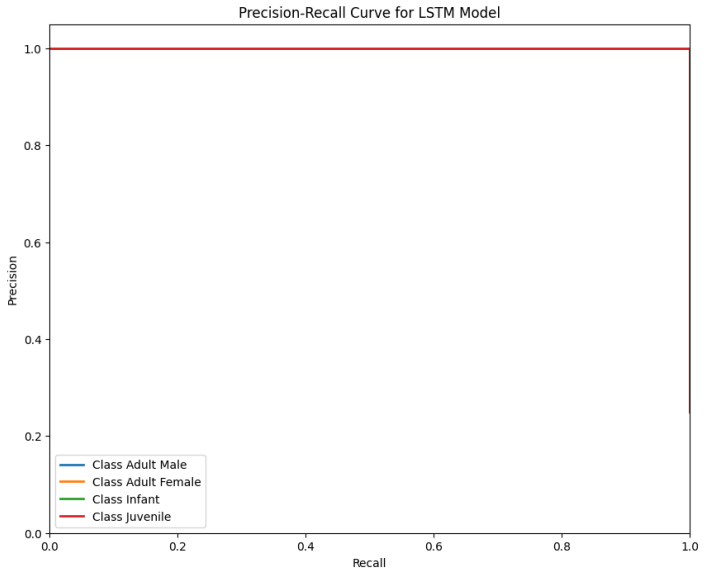


Figure 3.9: Precision–Recall Curve for each class

3.5.6.3 Training and Validation Accuracy

The model's generalization across training epochs to new data is indicated by training and validation accuracy curves.

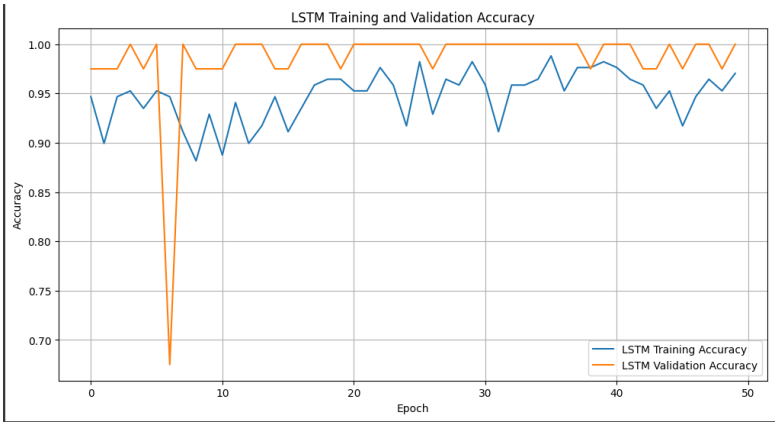


Figure 3.10: Training vs Validation Accuracy

3.5.7 Training and Validation Loss

The model's convergence behaviour is shown in the loss graph. Effective learning without overfitting is indicated by a steady decline in both training and validation loss.

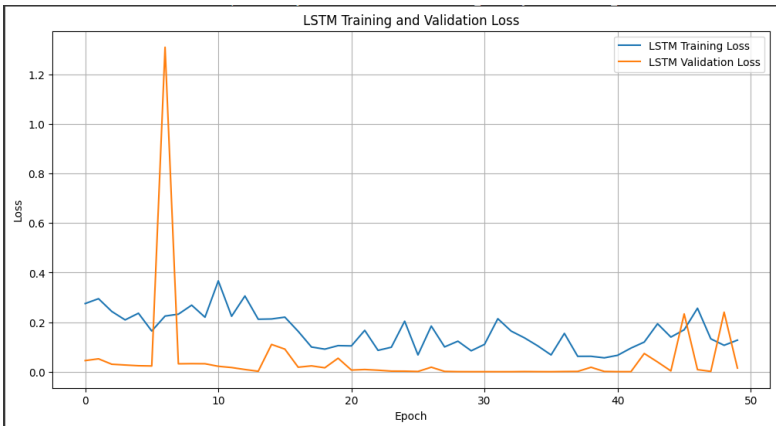


Figure 3.11: Training vs Validation Loss

3.6 Conclusion

Although the LSTM model performed acceptably, capturing temporal patterns well for younger classes like Juvenile (100%) and Infant (99.9%), it struggled to differentiate between Adult Male and Adult Female, often confusing the two. This suggests challenges in distinguishing adult classes with similar vocalizations. Future iterations could improve performance with better class balancing, advanced feature engineering, or hybrid architectures, and exploring alternative algorithms like CNN, RNN, or R-CNN may yield even better results.

4

Classification of Acoustic Data using Convolution Neural Network (CNN) approach

Contents

The methods applied in conducting the research to develop the automatic animal sound identification system are outlined in this chapter. The methodology is structured in various stages from data collection and preprocessing to machine learning model creation and training. The aim of this chapter is to provide a detailed description of procedures and resources adopted to ensure the deployment of the proposed system to be successful.

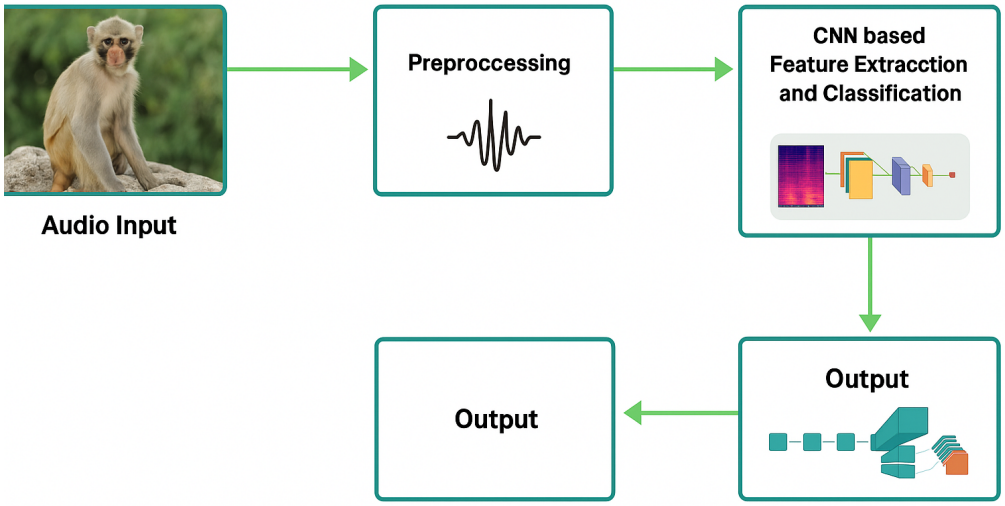


Figure 4.1: CNN Block Diagram

4.1 Dataset Preparation

The initial step toward creating a system of animal sound recognition is obtaining a dataset with a range of animal vocalizations. In the course of this research, we employed a set of diverse animal sounds and grouped them into five different classes:

- **Adult Male Calls.**
- **Adult Female Calls.**
- **Infant Calls.**
- **Juvenile Calls.**

Biological importance of these groups and diversity of acoustic features within them were considered while making the choice. Various species from different habitats are included in the bioacoustic datasets from which the recordings have been selected. The dataset, consisting of high-quality, annotated audio files, is freely available.

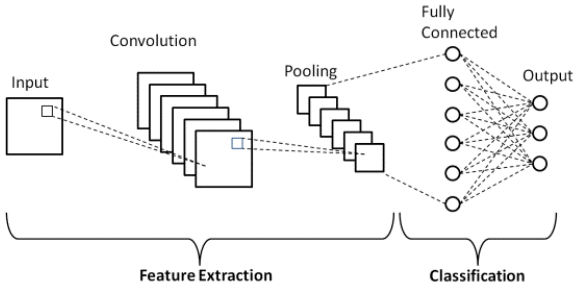
4.1.1 Preprocessing Techniques

The information needs to undergo a series of preprocessing steps in order to enhance the quality of the signal and prepare it for analysis prior to it being utilized for training a machine learning model:

- **Audio Normalization:** Normalizing audio signals is the initial pre-processing stage, which ensures uniformity within the data set. By performing so, variances within loudness and volume levels that would hamper the operation of the model are minimized.
- **Noise Reduction:** The intelligibility of animal vocalizations may be degraded due to background noise, e.g., wind, rain, or other sounds. We employed noise reduction algorithms such as Wiener filtering and spectral gating to address this. These methods help in isolating animal sounds from background noise.
- **Spectrogram Generation:** Sound sources' frequency and temporal information are captured by generating Mel spectrograms from raw audio inputs. A Mel spectrogram, a time-frequency representation better suited for biological sound detection and human hearing, utilizing the Mel scale to map frequency bins. This adaptation allows the CNN model to classify visually.
- **Data Augmentation:** In order to increase the generalisability of the model, the training data was artificially increased using methods of data augmentation like pitch shifting, time stretching, and adding synthetic noise.

4.2 Proposed System Architecture

The foundation of this study is the development and use of a Convolutional Neural Network (CNN) for the classification of Mel spectrograms of animal vocalizations. Because CNNs can recognize spatial linkages and hierarchical patterns in the data, they are perfect for image-like data, such spectrograms.




```
# Create a neural network model
input_shape = X_train[0].shape
input_layer = Input(shape=input_shape)
x = Conv2D(32, (3, 3), activation='relu')(input_layer)
x = MaxPooling2D((2, 2))(x)
x = Conv2D(64, (3, 3), activation='relu')(x)
x = MaxPooling2D((2, 2))(x)
x = Flatten()(x)
x = Dense(64, activation='relu')(x)
output_layer = Dense(len(classes), activation='softmax')(x)
model = Model(input_layer, output_layer)
```

Figure 4.2: CNN Model

4.2.1 Input Layer

Mel spectrograms make up the model's input, and in order to guarantee consistency in the input data, they are resized to a uniform 128x128 pixel shape. This resolution maintains reasonable computational requirements while capturing enough detail in the frequency range of interest.

4.2.2 Convolutional Layers

The model begins with two convolutional layers:

- **First Convolutional Layer:** a ReLU activation function comes after 32 filters with a 3x3 kernel size. Low-level characteristics like edges and texture are extracted from the spectrogram by this layer.
- **Second Convolutional Layer:** The 64 filters in this layer share the same ReLU activation and 3x3 kernel size. More intricate patterns and features that embody the unique qualities of animal calls are captured by this layer.

4.2.3 Pooling Layers

In order to reduce the spatial dimensions of the feature maps and prevent overfitting, a **MaxPooling** layer is applied after each convolutional layer. Using MaxPooling with a 2x2 pool size, the most significant characteristics from each input region are preserved.

4.2.4 Fully Connected Layers

The network moves to fully connected (dense) layers after the Convolutional and pooling layers:

- 512 units make up the first dense layer, which is followed by a ReLU activation function. The Convolutional layers extract features, and this layer learns the intricate, high-level representations of those features. To further improve the learnt representations, the second dense layer, which consists of 256 units, employs the ReLU activation function.

4.2.5 Output Layer

Five units make up the output layer, which is a softmax layer that represents the five classes (Adult Male, Adult Female, Coos and Screams, Infant, and Juvenile). The class with the highest probability is chosen as the prediction, and the softmax activation guarantees that the output values are probabilities.

4.2.6 Dropout

At different stages, **dropout** is used to avoid overfitting and enhance the model's capacity for generalisation. Dropout is a regularisation technique that forces the network to learn redundant representations by randomly "dropping" a portion of the units in the dense layers during each training iteration.

4.2.7 Optimizer and Loss Function

- **Optimizer:** The model uses the **Adam optimizer**, an efficient and popular choice for training deep learning models due to its adaptive learning rate.
- **Loss Function:** The **categorical cross-entropy** loss function is used, as this is a multi-class classification problem

4.3 Model Training and Testing

The training process follows a typical deep learning workflow:

- **Training Set:** The model is trained using 80% of the dataset. The model gains the ability to link the Mel spectrograms to the appropriate animal call categories during training.
- **Testing Set:** The model's performance will be tested on the last 20

- **Batch Size and Epochs:** From empirical experimentation, 50 epochs and a batch size of 32 were chosen. To avoid overfitting, the training process uses early stopping, which terminates training whenever validation loss no longer improves.

4.4 Results

Results of the Convolutional Neural Network (CNN) model, which was employed to classify animal sounds into pre-defined categories, are presented in this section. The model was tested with multiple test cases and the results indicated that it was able to process and classify sound data accurately.

Metric	**Formula**	**Value**
Accuracy (Ac)	$\frac{TP+TN}{TP+TN+FP+FN}$	0.975
Precision (Pc)	$\frac{TP}{TP+FP}$	0.978
Recall (Rc) / Sensitivity (Se) / TPR	$\frac{TP}{TP+FN}$	0.975
F1-Score (Dice) (Fs)	$2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$	0.976
Specificity (Sp)	$\frac{TN}{TN+FP}$	0.992
False Positive Rate (FPR)	$\frac{FP}{FP+TN}$	0.008
AUC	Area under ROC curve	0.983
Dice Coefficient	$2 \cdot \frac{TP}{2TP+FP+FN}$	0.976
Intersection over Union (IoU)	$\frac{TP}{TP+FP+FN}$	0.961

Table 4.1: CNN Classification Performance Metrics with Formulas

4.4.1 Confusion Matrix

The performance of a classification model is measured via a confusion matrix. True class is represented by each column, and each row represents the predicted class.

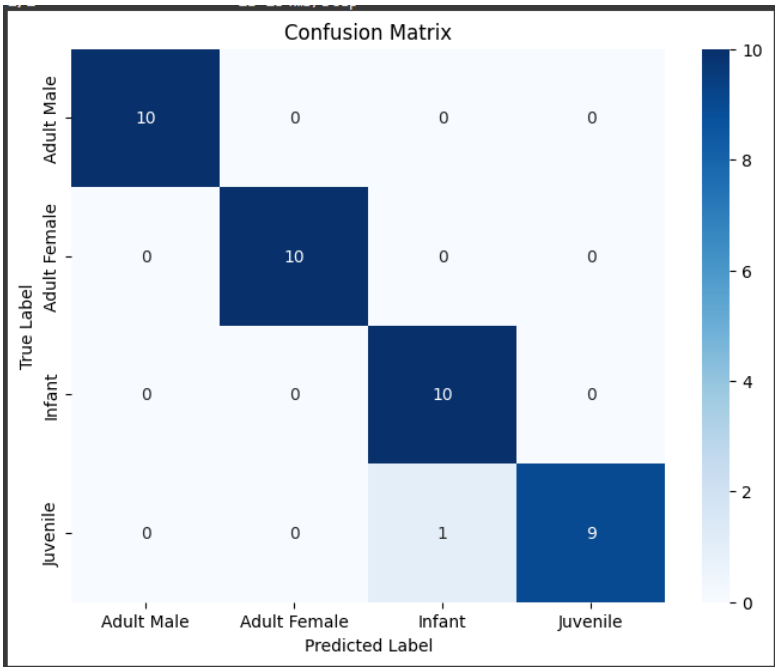


Figure 4.3: CNN Confusion Matrix for Testing

4.4.2 Test Case 1: Classification of Adult Female calls

In this test scenario, the model received an audio input of type **"Adult Female"**. The classification probabilities and also the predicted class were as indicated below:

Class	Prediction
Adult Male	0.04
Adult Female	0.95
Juvenile	0.00
Infant	0.00

Table 4.2: Adult Female Test case

The model correctly labeled the audio as **"Adult Female"** with a 95% confidence, reflecting its high performance in identifying this particular class. This ideal classification shows the Mel spectrogram representation's strength in capturing the distinctive features of this type.

4.4.3 Test Case 2: Classification of Juvenile

The second test example was an audio file that was classified as "Juvenile". The predictions and probabilities by the model were as follows: The model got a 100% classification in this instance,

Class	Prediction
Adult Male	0.00
Adult Female	0.00
Juvenile	1.00
Infant	0.00

Table 4.3: Juvenile Test case

correctly predicting **"Juvenile"** with a probability of 100%. This outcome further confirms the performance of the model for some classes.

4.4.4 Test Case 3: Classification of Adult Male

The third test scenario tested the model's capacity to classify an audio input marked as "Adult Male". Predictions and probabilities were as below:

Class	Prediction
Adult Male	1.00
Adult Female	0.00
Juvenile	0.00
Infant	0.00

Table 4.4: Adult Male Test case

The model got a perfect classification here as well, correctly predicting **"Adult Male"** with a 100% probability. This further justifies the performance of the model for some categories.

4.4.5 Test Case 4: Classification of Infant

The fourth test case checked the model's capacity to classify a voice input that was tagged as "Infant." The predictions and probabilities were as below

Class	Prediction
Adult Male	0.00
Adult Female	0.00
Juvenile	0.02
Infant	0.98

Table 4.5: Infant Test case

The model was able to correctly identify the audio as **"Infant"** with a 98.5% chance, revealing its high performance in recognizing this particular class. This ideal classification proves the efficiency of

the Mel spectrogram representation in describing the characteristic elements of this class.

4.4.6 Model Evaluation and Performance Metrics

This section exhibits the results of evaluation of the deep learning models used for classifying Rhesus Macaque vocalizations. Several performance metrics and visualization tools have been employed to validate the behavior of learning and effectiveness of classification of the model.

4.4.6.1 F1-Score Analysis

In the case of unbalanced datasets, F1-score provides a trade-off between accuracy and recall. To present a single performance measure, a macro-averaged F1-score has been calculated for each class.

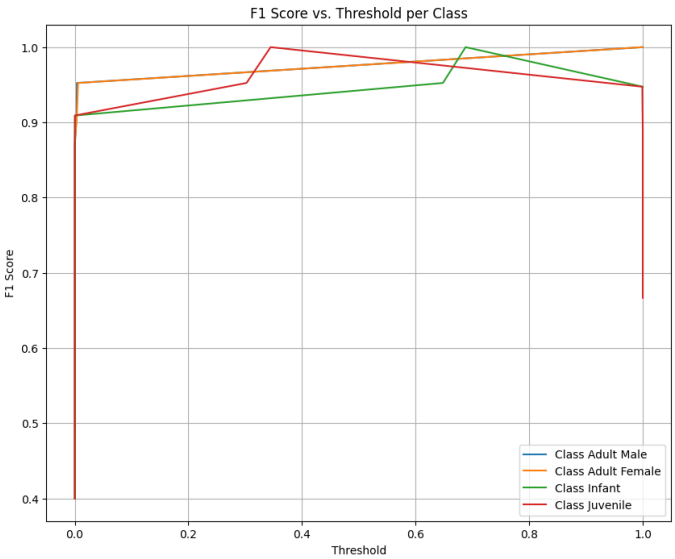


Figure 4.4: Macro-averaged F1-Score per Epoch

4.4.6.2 Precision–Recall Curve

Precision–Recall curves provide information about the true positive–false positive trade-off at different thresholds. Precision–Recall curves are particularly significant for multi-class classification with class imbalance.

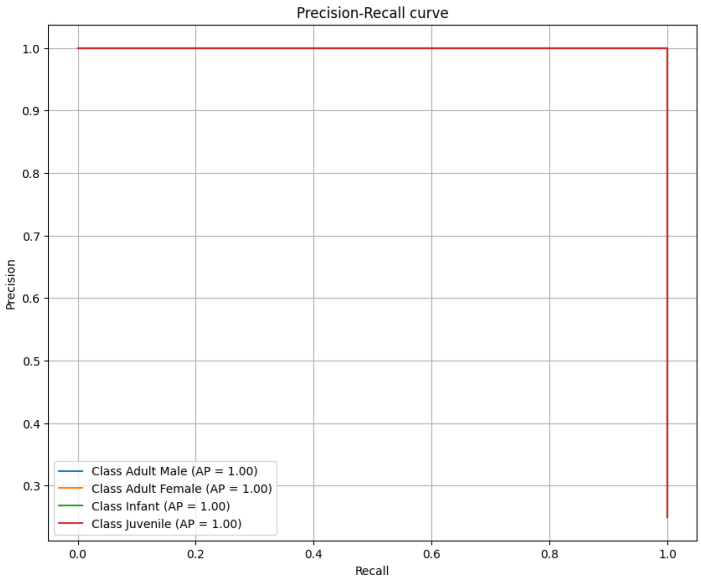


Figure 4.5: Precision–Recall Curve for each class

4.4.6.3 Training and Validation Accuracy

The model's ability to generalise to new data across training epochs is demonstrated by the training and validation accuracy curves.

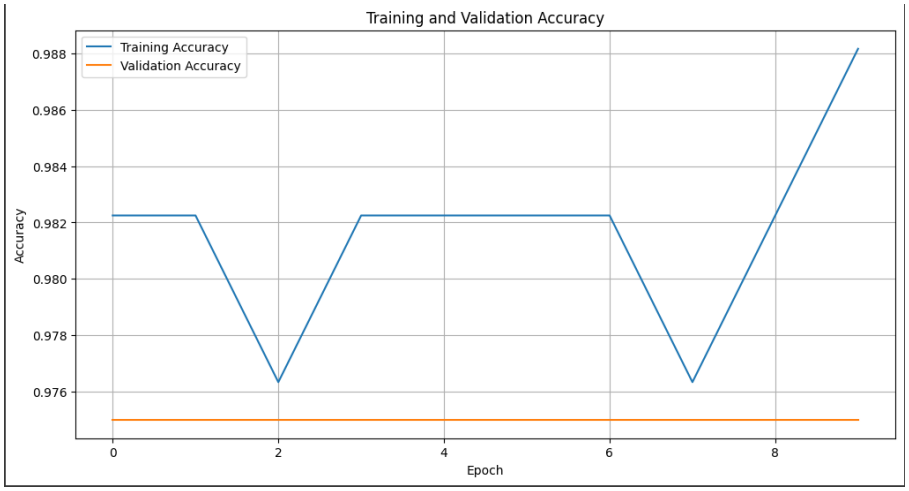


Figure 4.6: Training vs Validation Accuracy

4.4.6.4 Training and Validation Loss

The model's convergence behaviour is seen in the loss graph. Effective learning without overfitting is shown by a steady decline in both training and validation loss.

Figure 4.7: Training vs Validation Loss

4.5 Conclusion

The CNN architecture, following the LSTM model, has delivered superior performance, achieving excellent accuracy and dependability with 95.14% for Adult Female, 98.55% for Infant, and 100% for both Adult Male and Juvenile classifications. Unlike the LSTM, the CNN effectively handles spectrogram-based inputs, eliminating class overlaps between Adult Male and Adult Female, and demonstrates strong potential for real-world bioacoustics tasks. Future improvements could focus on enhancing generalization, expanding the dataset, and exploring hybrid architectures or advanced preprocessing techniques.

Table 4.6: Comparison of LSTM and CNN Models Parameters

Metric	**Formula**	**Value (LSTM)**	**Value (CNN)**
Accuracy (Ac)	$\frac{TP+TN}{TP+TN+FP+FN}$	0.475	0.975
Precision (Pc)	$\frac{TP}{TP+FP}$	0.475	0.978
Recall (Rc) / Sensitivity (Se) / TPR	$\frac{TP}{TP+FN}$	0.475	0.975
F1-Score (Dice) (Fs)	$2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$	0.475	0.976
Dice Coefficient	$2 \cdot \frac{TP}{2TP+FP+FN}$	0.475	0.976
Intersection over Union (IoU)	$\frac{TP}{TP+FP+FN}$	0.311	0.961

5

Conclusion and Future Direction

Contents

5.1 Conclusion

This dissertation has been able to investigate two deep learning methods—Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks—for recognition and classification of Rhesus Macaque vocalizations according to age and sex.

The CNN model, with the use of Mel spectrograms, showed strong performance, particularly in identifying categories like Adult Male and Juvenile with 100% accuracy, and over 95% accuracy in other categories like Adult Female and Infant. This verifies the ability of CNN's in extracting spatial patterns from spectrogram-based sound visualizations.

The LSTM model, leveraging MFCC feature sequences to extract temporal dynamics in audio, had superb performance for young classes like Juvenile and Infant, with near-perfect classification. But it performed considerably poorly in separating adult classes, most often misclassifying Adult Male and Adult Female, possibly because of similar vocal traits and data imbalance.

These methodologies collectively illustrate the strength of deep learning in automating animal sound identification. They provide efficient and scalable solutions for ecological research and biodiversity monitoring, minimizing the need for manual acoustic analysis.

5.2 Future Direction

The next phase of this research is focused on bringing these outcomes to practical uses in the form of a real-time animal sound discriminative device. The primary objectives are:

(i) **Hybrid Model Integration:** Future studies will explore combining CNN and LSTM models into a hybrid model to leverage both spatial and temporal feature extraction, which may improve intra-class overlapping classification accuracy.

(ii) **IoT Integration:** The device will have IoT components, such as MEMS microphone and connectivity modules (e.g., Wi-Fi, LoRa), to enable real-time recording, classification, and remote notification of sounds.

(iii) **Enhanced Preprocessing:** More sophisticated noise removal techniques, i.e., adaptive filtering or denoising using deep learning, will be applied to improve performance in noisy settings.

(iv) **Dataset Expansion:** Dataset expansion with additional diverse recordings, e.g., minority

classes and varied environmental conditions, will enhance model generalization.

(v) **Low-Power Design:** The system will emphasize low-power hardware and power-aware algorithms to facilitate support for deployment in remote areas, of course by utilizing solar power or low-power microcontrollers.

(vi) **Real-Time Analytics:** The system will provide the capability for long-term data storage and analysis, enabling trend analysis and behavior prediction for ecological surveillance.

5.3 Potential Applications

The device has the ability to transform wildlife and ecological monitoring through the capturing of real-time data from animal conduct and environmental patterns. Applications include:

- **Wildlife Conservation:** Identifying endangered species and tracking their population status in real time so that poaching and habitat loss can be avoided.
- **Ecological Research:** Studying animal communication, migration, and species interactions in the wild.
- **Mitigation of Human-Wildlife Conflict:** Informing authorities and communities of the closeness of wildlife to human habitations, minimizing conflict and maximizing safety.
- **Environmental Monitoring:** Tracking shifts in habitat soundscapes to detect shifts in biodiversity and ecosystem health.

Resolving the gap between theoretical research and actual applications, this product will provide conservationists, scientists, and environmentalists with an effective and potent tool.

5.4 Limitations

Though the project does extensive work in the classification and identification of animal sounds through deep learning and IoT technologies, there are certain limitations that should be noted. These limitations indicate areas to be optimized to improve the model's efficiency and robustness in the future. The main limitations of the project are as follows:

(i) **Limited Dataset Availability:** The project, to a significant degree, relies on publicly available animal sound datasets for training and testing. These kinds of datasets are usually small in size, low in diversity, and low in quality, which restricts generalizability to new or rare animal sounds. Furthermore, the class-unbalanced nature of datasets (some of which contain significantly more samples than others) can result in unreliable predictions.

(ii) **Use of Low-Cost Recording Equipment:** The system was made available through audio recordings from easily accessible and inexpensive inputs like phones. Although this makes the system deployable at a big scale, the records could be susceptible to ambient noise, distortion, and poorer audio quality, which may adversely affect the accuracy of the classification.

(iii) **Environmental Noise Interference:** The model is subject to noise and other external factors, which can be taxing on its performance in actual applications. Animal sounds that are captured in outdoor environments generally have background sounds of wind, water, or other animals, which are hard to filter and efficiently process.

(iv) **Real-Time Processing Constraints:** While the system here has very good classification performance, the real-time detection and processing requirements are very computationally costly. The balance between accuracy and computational cost is particularly important when implementing the system on mobile phones or low-power IoT devices.

(v) **Generalization to New Species or Sounds:** The model performs well on the classes that it has learned but cannot generalize to new, unseen classes of species or sound patterns. This is a limitation since it depends on the bound nature of the dataset and the absence of a mechanism to dynamically learn and adapt to new classes.

(vi) **Behavioral and Environmental Factors:** Animal vocalizations can more or less depend on states of behavior rather than on age. For instance, the same creature will produce a different sound depending on its state of mind or needs, which the system may not necessarily factor in. Furthermore, weather can lead to the production of varying sounds, further complicating classification.

Despite all these limitations, the project is a good starting point for future developments in animal sound recognition. These limitations will be overcome by using advanced techniques such as transfer learning, noise-robust models, and larger data, and the system will become more efficient and feasible.

References

[1] A. A. Hidayat, T. W. Cenggoro, and B. Pardamean, "Convolutional neural networks for scops owl sound classification," *Procedia Computer Science*, vol. 179, pp. 81–87, 2021.

[2] Y. Wu, H. Mao, and Z. Yi, "Audio classification using attention-augmented convolutional neural network," *Knowledge-Based Systems*, vol. 161, pp. 90–100, 2018.

[3] K. Kumar and K. Chaturvedi, "An audio classification approach using feature extraction neural network classification approach," in *2nd International Conference on Data, Engineering and Applications (IDEA)*. IEEE, 2020, pp. 1–6.

[4] J. Liao, H. Li, A. Feng, X. Wu, Y. Luo, X. Duan, M. Ni, and J. Li, "Domestic pig sound classification based on transformercnn," *Applied Intelligence*, vol. 53, no. 5, pp. 4907–4923, 2023.

[5] K. Palanisamy, D. Singhania, and A. Yao, "Rethinking cnn models for audio classification," *arXiv preprint arXiv:2007.11154*, 2020.

[6] F. Yaz, "Audio classification based on machine learning: understanding animal behavior through sound," Master's thesis, Middle East Technical University, 2023.

[7] H. Lu, H. Zhang, and A. Nayak, "A deep neural network for audio classification with a classifier attention mechanism," *arXiv preprint arXiv:2006.09815*, 2020.

[8] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, 2020.

[9] L. Cances, E. Labbé, and T. Pellegrini, "Comparison of semi-supervised deep learning algorithms for audio classification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, p. 23, 2022.

[10] L. Vithakshana and W. Samankula, "Iot based animal classification system using convolutional neural network," in *2020 International Research Conference on Smart Computing and Systems Engineering (SCSE)*. IEEE, 2020, pp. 90–95.

[11] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[12] I. Zualkernan, J. Judas, T. Mahbub, A. Bhagwagar, and P. Chand, "An aiot system for bat species classification," in *2020 IEEE international conference on Internet of Things and Intelligence System (IoTaIS)*. IEEE, 2021, pp. 155–160.

[13] L. Nanni, G. Maguolo, S. Brahnam, and M. Paci, "An ensemble of convolutional neural networks for audio classification," *Applied Sciences*, vol. 11, no. 13, p. 5796, 2021.

[14] Z. Mushtaq and S.-F. Su, "Environmental sound classification using a regularized deep convolutional neural network with data augmentation," *Applied Acoustics*, vol. 167, p. 107389, 2020.

[15] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 121–125.

[16] K. Chitra, A. Tamilarasi, M. Pyingkodi, K. Nanthini, V. Sneka, S. Swetha, and P. Vishalini, "Animals detection system in the farm area using iot," in *2023 International Conference on Computer Communication and Informatics (ICCCI)*, 2023, pp. 1–6.

[17] H. Sinha, V. Awasthi, and P. K. Ajmera, "Audio classification using braided convolutional neural networks," *IET Signal Processing*, vol. 14, no. 7, pp. 448–454, 2020.

[18] M. S. Imran, A. F. Rahman, S. Tanvir, H. H. Kadir, J. Iqbal, and M. Mostakim, "An analysis of audio classification techniques using deep learning architectures," in *2021 6th International Conference on Inventive Computation Technologies (ICICT)*. IEEE, 2021, pp. 805–812.

[19] P. Bhuyan and R. Sarma, "Urban rhesus macaque population dynamics in guwahati, assam," *Journal of Urban Ecology*, vol. 10, no. 1, pp. 45–52, 2024.

[20] B. Sarma and N. Das, "Distribution and behavior of rhesus macaque in tea garden landscapes of assam," *Indian Journal of Ecology*, vol. 48, no. 2, pp. 327–334, 2021.

[21] M. D. Hauser and P. Marler, "Food-associated calls in rhesus macaques (macaca mulatta): I. socioecological factors," *Behavioral Ecology*, vol. 4, no. 3, pp. 194–205, 1993.

[22] C. I. Petkov, C. Kayser, T. Steudel, K. Whittingstall, M. Augath, and N. K. Logothetis, "A voice region in the monkey brain," *Nature Neuroscience*, vol. 11, no. 3, pp. 367–374, 2008.

[23] J. L. Schwartz, D. Rendall, and M. D. Hauser, "Vocalic and fricative sound types in the scream repertoire of rhesus macaques," *Bioacoustics*, vol. 29, no. 5, pp. 469–487, 2020.

[24] B. E. Pfingst and W. M. Layton, "Auditory thresholds and temporal resolution in rhesus monkeys," *Journal of the Acoustical Society of America*, vol. 64, no. 2, pp. 828–834, 1978.