KANTIPUR ENGINEERING COLLEGE

(Affiliated to Tribhuvan University)

Dhapakhel, Lalitpur

[Subject Code: CT654]

A MAJOR PROJECT PROPOSAL ON

FEATHERFIND : BIRD SPECIES IDENTIFICATION FROM AUDIO AND IMAGE

Submitted by:

Gaurav Giri	**[Kan077bct034]**
Iza K.C.	**[Kan077bct039]**
Prajwal Khatiwada	**[Kan077bct056]**
Samrat Kumar Adhikari	**[Kan077bct074]**

A MAJOR PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF BACHELOR IN COMPUTER ENGINEERING

Submitted to:

Department of Computer and Electronics Engineering

5 June, 2024

FEATHERFIND : BIRD SPECIES IDENTIFICATION FROM AUDIO AND IMAGE

Submitted by:

Gaurav Giri **[Kan077bct034]**

Iza K.C. **[Kan077bct039]**

Prajwal Khatiwada **[Kan077bct056]**

Samrat Kumar Adhikari **[Kan077bct074]**

A MAJOR PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF BACHELOR IN COMPUTER ENGINEERING

Submitted to:

Department of Computer and Electronics Engineering

Kantipur Engineering College

Dhapakhel, Lalitpur

5 June, 2024

ABSTRACT

SnapTag introduces an approach for efficiently analyzing handwritten documents by leveraging image processing techniques and Named Entity Recognition (NER). The primary objective is to develop a system capable of extracting meaningful information from handwritten content provided by users and subsequently generating relevant tags for improved document organization and categorization.

SnapTag employs image processing methods such as image binarization, thresholding, denoising to enhance the quality of scanned handwritten documents. Through these techniques, the system effectively preprocesses images, mitigating noise and improving the clarity of handwritten text. Then, the methodology involves the integration of Canny edge detection and Hough line transformation, coupled with K-means clustering, to accurately detect document boundaries. Subsequent stages of the process incorporate image segmentation to isolate words and characters, followed by a classification model that identifies each character within the document. The character recognition phase utilizes a trained classification CNN model, to accurately classify individual characters into predefined classes. This step is crucial for deciphering the handwritten content and preparing it for further analysis.In the final stage, NER is employed to extract meaningful tags from the processed document providing valuable metadata that enhances the document's categorization and searchability.

Keywords−Optical Character Recognition, Binarization, Thresholding, Denoising, Boundary Detection, Hough Line Transformation, K-Means Clustering, Convolutional Neural Network, Named Entity Recognition

TABLE OF CONTENTS

Abstract i

List Of Figures iv

Abbreviations v

1 Introduction 1

1.1 Background . . . 1

1.2 Problem Statement . . . 2

1.3 Objectives . . . 2

1.4 Application Scope . . . 3

1.5 Features . . . 3

1.6 Feasibility Study . . . 4

1.6.1 Economic Feasibility . . . 4

1.6.2 Technical Feasibility . . . 5

1.6.3 Operational Feasibility . . . 5

1.6.4 Schedule Feasibility . . . 5

1.7 System Requirements . . . 6

1.7.1 Development Requirements . . . 6

1.7.2 Deployment Requirements . . . 6

2 Literature Review 7

2.1 Related Works . . . 7

2.2 Related Research . . . 8

3 Methodology 15

3.1 Working Mechanism for Identification using Audio . . . 15

3.1.1 Dataset . . . 16

3.1.2 Dataset Overview for Audio . . . 17

3.1.3 Data Preprocessing . . . 19

3.1.4 Data Splitting . . . 19

3.1.5 Data Augmentation . . . 20

3.1.6 Audio Splitting . . . 20

3.1.7 Feature Extraction . . . 20

3.1.8 Convolutional Neural Networks (CNN) . . . 22

3.1.9 Long Short-Term Memory Networks (LSTM) 23
3.1.10 Combined CNN and LSTM (CNN+LSTM) 24
3.1.11 Hyperparameter Optimization using Genetic Algorithm 25
3.1.12 Algorithms Used 27
3.2 Working Mechanism for Identification using Image 29
3.2.1 Dataset 30
3.2.2 Dataset Overview for Image 30
3.2.3 Data Preprocessing 31
3.2.4 VGG16 31
3.3 System Diagrams 33
3.4 Software Development Model 33
4 Epilogue **35**
References **36**

LIST OF FIGURES

1.1 Gantt Chart 6
3.1 Block diagram for the working mechanism of the system 15
3.2 Training Dataset 17
3.3 Validation Dataset 17
3.4 Testing Dataset 18
3.5 Sample of the Data 18
3.6 Feature Extraction Using Spectrogram and MFCC 21
3.7 CNN architecture 22
3.8 LSTM architecture 24
3.9 Block diagram for the working mechanism of the system 29
3.10 Dataset distribution for Top 50 Bird Species among 200 Species 30
3.11 VGG16 architecture 32
3.12 UseCase Diagram 33
3.13 Incremental Model for development of SnapTag 34

ABBREVIATIONS

CNN	Convolutional Neural Network
CUB	Caltech-UCSD Birds
DCNN	Deep Convolutional Neural Network
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
GA	Genetic Algorithm
GPS	Global Positioning System
GRU	Gated Recurrent Network
LSTM	Long Short-Term Memory
MAP	Mean Average Precision
MFCCs	Mel-Frequency Cepstral Coefficients
MLP	Multilayer Perceptron
RGB	Reg Green Blue
ROC	Receiver Operating characteristics
RNN	Recurrent Neural Network
STFT	Short-Time Fourier Transform
UCSD	University of California San Diego
VGG	Visual Geometry Group

CHAPTER 1
INTRODUCTION

1.1 Background

Globally, the avian kingdom is vast, with over 11,000 species, a testament to nature's complexity and evolutionary prowess. This figure, sourced from the International Ornithological Committee as of April 2023, merely scratches the surface of avian diversity, each species a unique entity with its own ecological role and evolutionary story.[1]

Turning our gaze to Nepal, a country of remarkable biodiversity and varied ecosystems, from the lowland Terai to the towering Himalayas, it is home to more than 887 bird species, as reported by the Himalayan Nature organization. This represents more than 8% of the world's known bird species, a significant figure given Nepal's relatively small geographical footprint.[2]

Among these, a number are endangered, their existence threatened by habitat loss, climate change, and human activities. The National Red List of Nepal's Birds, a comprehensive assessment of the country's avian biodiversity, identifies several species at risk. Specifically, Nepal harbors 168 nationally threatened bird species, including 68 Critically Endangered, 38 Endangered, and 62 Vulnerable species, as detailed in a publication by the Journal of Threatened Taxa.[3]

The plight of these endangered species underscores the urgency of conservation efforts. Technologies such as audio recognition and image classification offer innovative tools for identifying and monitoring bird populations. By analyzing the unique sounds and visual characteristics of birds, researchers can enhance our understanding of species distribution, behavior, and threats. Such technologies not only aid in the conservation of endangered species but also contribute to the broader field of biodiversity research.

1.2 Problem Statement

In Nepal, a hotspot of avian biodiversity, accurately identifying and classifying bird species, particularly those that are endangered, is a critical yet complex task. Traditional observation methods are limited by the vast geographical and ecological diversity of the region, making it challenging to monitor and protect these birds effectively. The necessity for precise identification is paramount for conservation efforts aimed at maintaining ecosystem balance. To address this, there is a pressing need for a method that can overcome these constraints by leveraging advanced technologies capable of distinguishing between the myriad of bird calls and songs, as well as visual markers through image classification. Such a method promises to automate the identification process, enhancing accuracy and efficiency in monitoring endangered species.

Implementing audio recognition and image classification technologies, however, raises several challenges, including the effective training of these systems to recognize the specific calls and visual markers of Nepal's endangered birds. This requires collecting and curating extensive datasets, a task complicated by the elusive nature of many species and the complex acoustics of their habitats. Additionally, integrating these technologies into conservation strategies is crucial for not just identifying but also protecting these species from various threats. A multidisciplinary approach, combining expertise in ornithology, conservation biology, machine learning, and environmental science, is essential to develop a robust system. Such a system could provide reliable data on species distribution, population trends, and habitat use, thereby informing targeted conservation actions and policies to preserve Nepal's rich avian biodiversity.

1.3 Objectives

i To develop and implement an integrated technological solution that utilizes advanced audio recognition and image classification techniques for the accurate identification and monitoring of bird species in Nepal, with an emphasis on endangered species.

1.4 Application Scope

1. **Conservation Efforts:**

   This system will enhance conservation by enabling accurate monitoring of bird populations, helping track endangered species and take a step towards habitat protection.

2. **Biodiversity Monitoring:**

   Automated identification will aid biodiversity monitoring by processing large datasets, helping detect species distribution on bird communities.

3. **Ecological Research:**

   Researchers can use the system to study bird migration, and habitat use, providing crucial data for modeling ecosystems and understanding ecological interactions.

4. **Environmental Education and Awareness:**

   Integrated into educational programs, this tool will raise public awareness about biodiversity and conservation, engaging students and citizen scientists in bird identification.

5. **Bird viewing:**

   Bird enthusiasts will benefit from these systems as they will enhance bird watching experiences by providing instant identification of bird species

1.5 Features

1. **Species Identification Using Audio:**

   The app allows users to record bird sounds in real-time using their device's microphone or upload pre-recorded audio files. Advanced noise filtering techniques isolate bird calls from background noise, and sound wave analysis helps in identifying distinct frequency patterns. Machine learning algorithms, trained on a vast database of bird calls, match the recorded sound to identify the bird species accurately.

2. **Species Identification Using Image:**

   Users can capture photos of birds using the app's camera or upload images from their gallery. The app enhances image quality and analyzes features such as color,

size, shape, and patterns. Utilizing computer vision models, the app identifies the bird species by comparing the image with a comprehensive database of bird images.

3. **Mapping Identified Bird Habitat:**

   The app tags the location of identified birds using GPS, providing detailed habitat information typical of each species. Integrating with mapping services, it displays bird sightings on an interactive map, generating heat maps to show species density and distribution. Additionally, it tracks and visualizes bird migration patterns over time, helping users understand seasonal movements.

4. **Provide Description About the Birds:**

   For each identified bird species, the app offers detailed profiles that include scientific and common names, physical descriptions, and conservation status. It also provides audio and visual media for reference, along with information on the bird's behavior, diet, and typical habitats, enriching the user's understanding of the species.

1.6 Feasibility Study

Before implementation of project design, the feasibility analysis of the project must be done to move any further. The feasibility analysis of the project gives an idea on how the project will perform and its impact in the real world scenario. So, it is of utmost importance.

1.6.1 Economic Feasibility

Our system is economically achievable as a result of the development of several tools, libraries, and frameworks. Since all the software required to construct it is free and readily available online, this project is incredibly cost-effective. Only time and effort are needed to create a worthwhile, genuinely passive system. The project doesn't come at a substantial cost. From an economic standpoint, the project appears successful in this sense.

1.6.2 Technical Feasibility

The software needed to implement a project can be downloaded from a wide variety of online resources. Technically speaking, the project is feasible as the necessary software is easily available. We were able to learn the information we required for the project through a variety of online sources, including classes. All the libraries and data are accessible online for free because this project does not require any licensing costs. It is technically possible if one has the necessary information and resources.

1.6.3 Operational Feasibility

The project aims to enhance bird species identification through audio classification, making bird sound recognition more accessible and efficient. This solution is particularly beneficial for ornithologists, bird watchers, and environmental researchers who require accurate and quick identification of bird species based on their calls. The project leverages advanced audio signal processing and deep learning algorithms to classify bird sounds, ensuring high accuracy and reliability. Given the widespread availability of mobile devices and recording equipment, the project is operationally feasible, as it can be easily integrated into existing workflows and tools used by bird enthusiasts and professionals. By providing an efficient method for bird sound classification, the project supports a sizable community interested in avian studies and conservation, ensuring practical applicability and ease of use.

1.6.4 Schedule Feasibility

The workload of the project is divided amongst the project members. The scheduling is done according to an incremental model where different modules are planned to be assigned to the group members.So, the project fulfills the schedule feasibility requirements.

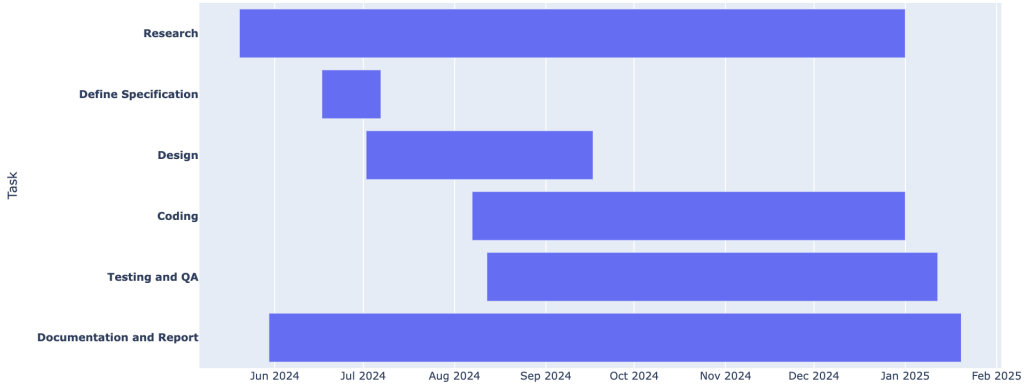

Figure 1.1: Gantt Chart

1.7 System Requirements

1.7.1 Development Requirements

Table 1.1: Development Requirements

Software Requirements	**Hardware Requirements**
Programming Language: Python, Dart, Javascript	Camera: >= 12 Megapixels
Design Tools: Figma, Canva, Draw.io	RAM: >= 8 GB
Libraries: Librosa, PyAudio, Pytorch	CPU: i5 10th (Recommended)
Framework: Flutter, Django RestFramework	GPU: P100
IDEs: VSCode, Android Studio	Storage: >= 50 GB

1.7.2 Deployment Requirements

Table 1.2: Deployment Requirements

Software Requirements	**Hardware Requirements**
Android: >= 10	Camera: >= 12 Megapixels
Read/Write FileSystem	RAM: > 4 GB
Internet Accessibility	Storage: >= 20 GB
Database: Sqlfite	Recording Quality >= 256 Kbps, 48 KHz

CHAPTER 2
LITERATURE REVIEW

Regardless of whether they are documented or not, every project has helped to shape the world as it is today. Other researchers can benefit from documented projects by learning specifics about problems and how to solve them. Additionally, they boostproject efficiency by removing the need to start the project from the beginning and specifying the starting point.

2.1 Related Works

BirdNET is a cutting-edge research platform developed through collaboration between the K. Lisa Yang Center for Conservation Bioacoustics at the Cornell Lab of Ornithology and the Chair of Media Informatics at Chemnitz University of Technology. Its primary aim is to detect and classify bird sounds using machine learning technologies, serving both experts and citizen scientists in their efforts to monitor and protect bird populations.

BirdNET can identify around 3,000 of the world's most common bird species, with plans to expand this number. Features such as a live submissions map and a Twitter bot are included to engage the community and share real-time data. The project is supported by donations and collaborations, offering opportunities for researchers and developers to contribute to its growth. BirdNET serves as an invaluable tool for bird enthusiasts, conservationists, and biologists alike, providing innovative solutions for large-scale acoustic monitoring and contributing to the conservation and understanding of avian biodiversity.

The BirdCLEF 2023 competition on Kaggle is a significant data science challenge that falls under the broader LifeCLEF initiative, aimed at pushing the boundaries of species identification and biodiversity monitoring through technological innovation. This particular competition focuses on the development of machine learning models that can identify bird species based on audio recordings. It presents a complex and realistic challenge due to the diversity of the audio recordings, which are collected from various

environments and feature a wide range of bird species.

2.2 Related Research

The research paper 'Audio Classifier for Automatic Identification of Endangered Bird Species of Nepal' focuses on developing an audio classifier to identify endangered bird species in Nepal using deep learning techniques. The dataset, collected from xeno-canto.org, comprises 2215 audio recordings of 41 bird species, 38 of which are endangered. This dataset was expanded to 6733 recordings through 10-second audio splitting and Gaussian noise augmentation, with 5407 recordings used for training, 639 for validation, and 687 for testing. The methodology involved handling imbalanced class distribution through data augmentation, employing Mel spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs) for feature extraction, and developing a custom Convolutional Neural Network (CNN) model and an EfficientNet model. The hyperparameters of these models were optimized using a genetic algorithm. The Mel spectrograms were created using Short-Time Fourier Transform, converting amplitudes to decibel scale, and applying Mel filter banks to the spectrograms. Similarly, MFCCs were derived by framing the audio signals, applying Discrete Fourier Transform, logarithmic scaling, Mel scaling, and Discrete Cosine Transform. The EfficientNet architecture utilized compound scaling for network depth, width, and resolution. The findings indicated that the proposed approach achieved satisfactory results in classifying the bird species. However, limitations include the relatively small dataset size and the need for further enhancement in model robustness and accuracy. Future enhancements could involve expanding the dataset, exploring additional feature extraction techniques, and incorporating more advanced deep learning models to improve classification performance. This research contributes significantly to the conservation efforts by providing a reliable method for automatic bird species identification, aiding in monitoring and protecting endangered species.[4]

The paper 'Audio Bird Classification with Inception-v4 extended with Time and Time-Frequency Attention Mechanisms' presents an innovative adaptation of the Inception-

v4 deep convolutional network for bioacoustic classification, focusing specifically on bird sound recognition. The datasets employed include various bird sounds, prominently from the BirdClef2017 challenge, consisting of 1500 bird species recordings. The methodology revolves around treating bird sound classification as an image classification problem through transfer learning. The Inception-v4 model, initially pre-trained on ImageNet, was adapted to process time-frequency representations of bird sounds by converting these sounds into RGB images using three log-spectrograms generated via fast Fourier transform at different scales (128, 512, 2048 bins). Data augmentation techniques, common in computer vision, were applied to these spectrograms to enhance the robustness of the model. The findings demonstrate that the model, termed 'Soundception', integrates time and time-frequency attention mechanisms effectively, significantly improving classification accuracy. The results highlight Soundception's outstanding performance, achieving a mean average precision (MAP) of 0.714 in classifying 1500 bird species, 0.616 MAP for background species, and 0.288 MAP for soundscapes with time-codes, making it the top model in the BirdClef2017 challenge across multiple tasks. However, limitations include the incomplete convergence of the model due to computational constraints and the extensive GPU resources required for training, which restricted the full potential realization within the challenge's timeframe. The paper concludes with a discussion on future improvements, such as exploring different scalable optimizations and incorporating stacked GRU layers for better audio-to-image representation learning, underscoring the potential of transfer learning from advanced image classification models to acoustic domains.[5]

The research paper 'Bird Species Identification using Deep Learning' presents a methodology using Deep Convolutional Neural Networks (DCNNs) to classify bird species, leveraging the Caltech-UCSD Birds 200 (CUB-200-2011) dataset, which contains 11,788 annotated images of 200 bird species. The methodology involves converting images to grey scale to reduce computational complexity, followed by the application of DCNNs using TensorFlow to extract hierarchical features from images such as edges, textures, and complex patterns. The neural network architecture includes convolutional layers for feature extraction, pooling layers for dimensionality reduction, activation lay-

ers for non-linearity, and fully connected layers for classification. Key findings show the DCNN model achieved a testing accuracy of 80% and training accuracy of 93%, with a validation accuracy of around 75%, indicating robust performance across different data splits. The study highlights the effectiveness of combining multiple features (head, body, color, beak) over single-feature classification, with generated autographs and score sheets facilitating identification. The system's usability is enhanced through a web interface for image uploads, and future directions propose mobile app development and cloud integration to improve accessibility and scalability. However, the research also identifies limitations such as the dependency on the quality and diversity of the dataset, potential overfitting due to the complexity of the model, and the computational resources required for training deep learning models. Addressing these limitations could further enhance the model's accuracy and applicability in real-world scenarios.[6]

The research paper 'Bird species classification from an Image using VGG-16 Network' utilizes machine learning and deep learning techniques has shown significant advancements in both methodology and accuracy. The primary focus has been on the creation and utilization of diverse datasets, with a notable example being the dataset of 1600 images across 27 bird species, as well as the Caltech-UCSD Birds-200-2011 Dataset. Methodologically, the use of Convolutional Neural Networks (CNNs), particularly the VGG-16 model, has been prevalent due to its effectiveness in feature extraction from images. This approach has been complemented by transfer learning techniques, which have notably improved classification accuracies, as seen in the use of pre-trained networks like AlexNet and VGG-16, achieving accuracies up to 85.4% and 92.13% respectively. The application of machine learning algorithms such as SVM with a linear kernel and KNN has also been explored, with SVM achieving an accuracy of 89% after parameter optimization. Additionally, the integration of computer vision methods to extract Histogram Oriented Gradients and RGB histogram values has further enhanced classification performance. Despite these advancements, the studies have encountered limitations, including the challenge of accurately classifying bird species from various angles and positions, and the need for extensive preprocessing to remove noise

from datasets. Overall, the body of work demonstrates a trend towards higher accuracy in bird species classification through the innovative use of deep learning techniques and sophisticated feature extraction methods, though challenges remain in dealing with the variability of natural images and the computational demands of processing large datasets.[7]

This research paper presents significant advancements in the field of 'Automatic bird species identification through the integration of audio signal processing and neural networks'. The study, conducted by Chandu B, Akash Munikoti, Karthik S Murthy, Ganesh Murthy V, and Chaitra Nagaraj from the BNM Institute of Technology, outlines a robust methodology for identifying bird species from audio recordings, leveraging a combination of meticulously curated datasets and sophisticated machine learning techniques. The dataset, a critical component of the study, was manually compiled from both local recordings and online resources such as xeno-canto.com, featuring audio clips from species like cuckoo, sparrow, crow, and laughing dove, as well as ambient noise and human voices to simulate real-world conditions. Pre-processing techniques including pre-emphasis, framing, silence removal, and reconstruction were applied to the audio clips to enhance the relevant frequency components and eliminate unnecessary noise, ensuring the purity of the dataset. Spectrograms of these processed clips were generated and used as input for training a convolutional neural network (CNN), specifically AlexNet, chosen for its high accuracy in image classification tasks. Through transfer learning, AlexNet was adapted to recognize bird species from the spectrograms, achieving a classification accuracy of 97% in controlled environments. However, recognizing the variability of real-world conditions, the researchers retrained the model with datasets containing ambient noise, achieving a real-time classification accuracy of 91%. Despite these promising results, the study acknowledges limitations such as the relatively small size of the dataset and the need for further tuning of performance parameters to improve robustness. The potential for future applications is vast, including the development of mobile applications and hardware implementations for ecological monitoring, highlighting both the scientific and practical significance of automated bird species identification systems.[8]

The paper 'An Ensemble of Convolutional Neural Networks for Audio Classification' delves into a comprehensive study on CNN classification using different architectures, data augmentation techniques, and audio signal representations, aimed at enhancing audio classification tasks across various datasets. The study employs three datasets: BIRDZ, CAT, and ESC-50, each offering unique challenges in audio classification. The methodology involves training five convolutional neural networks (CNNs) with four audio representations combined with six different data augmentation methods, resulting in thirty-five subtypes of ensembles. The audio representations include techniques such as the Discrete Gabor Transform (DGT), Waveform Similarity OverLap Add (WSOLA), and Phase Vocoder. The data augmentation methods encompass procedures like short spectrogram augmentation, random time shift, and frequency masking. The CNN architectures are pre-trained models fine-tuned with these augmented datasets to boost classification accuracy. The findings reveal that the ensemble method outperforms standalone networks, achieving 97% accuracy on the BIRDZ dataset, 90.51% on the CAT dataset, and 88.65% on the ESC-50 dataset. The study also highlights that the best-performing CNNs are VGG16 and VGG19, with DGT as the most effective signal representation. However, the study acknowledges limitations, such as the computational cost of training ensembles and the variability in performance across different augmentation techniques. Notably, no single augmentation protocol consistently outperforms others across all datasets. The results underscore the potential of combining different CNNs and augmentation policies to enhance performance, although this approach demands significant computational resources. Additionally, the study provides the MATLAB source code for reproducibility, contributing to the research community's efforts in advancing audio classification technologies.[9]

The literature review focused on the 'analysis of bird call datasets sourced from Xeno-Canto', comprising 72,172 samples from 264 bird species in 16-bit wav format with a 16 kHz sampling rate. The methodology involved preprocessing the audio data to filter out low-frequency noise and normalize signal amplitude, followed by generating Mel-spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs) as inputs for

deep learning models. The Mel-spectrograms were produced using discrete Fourier transform (DFT), and the MFCCs were derived by applying discrete cosine transform (DCT) to the Mel-spectrogram. The study employed various metrics to evaluate the performance of these methods, including ROC analysis to visualize model effectiveness. Findings indicated that the proposed models showed significant promise in identifying bird species from their calls, with improvements in classification accuracy compared to previous approaches. However, limitations were noted, including potential biases in the dataset due to uneven sample distribution across species and the challenge of background noise affecting signal quality. Future work suggested enhancing noise reduction techniques and exploring more sophisticated neural network architectures to further improve model robustness and accuracy.[10]

The study conducted an in-depth analysis of 'bird species recognition through acoustic monitoring', utilizing a robust dataset of bird sound samples, meticulously annotated and validated for accuracy. The dataset, referred to as SD, comprises multispecies bird sound recordings, each labeled with species name and sample ID, along with corresponding metadata, providing a comprehensive foundation for model training and evaluation. Methodologically, the research employed a spectrogram-based feature extraction approach, leveraging Short-Time Fourier Transform (STFT) to capture the intricate temporal and spectral characteristics of bird sounds. This was followed by the application of a Multilayer Perceptron (MLP) classifier to distinguish between different bird species. The findings reveal that the proposed model achieved high recognition accuracy, with some species being identified with perfect precision, recall, and accuracy (100%), though the performance varied across species, with a few showing lower recognition rates (86.9%) and precision/recall values ranging between 50-75%. The results demonstrated an overall classification accuracy of 96%, with cross-validation accuracy standing at 81.4%, highlighting the model's robustness yet indicating room for improvement in generalizability across diverse datasets. Despite the promising results, the study acknowledges several limitations, including the variability in recognition accuracy among different species and the potential influence of environmental noise on model performance. Future work is suggested to explore feature and model

fusion techniques, integrate the model with cloud-based systems for real-time recognition, and expand the dataset to include a broader range of bird species to enhance the model's applicability and accuracy in practical scenarios.[11]

CHAPTER 3
METHODOLOGY

This chapters describes bird identification using audio and image separately.

3.1 Working Mechanism for Identification using Audio

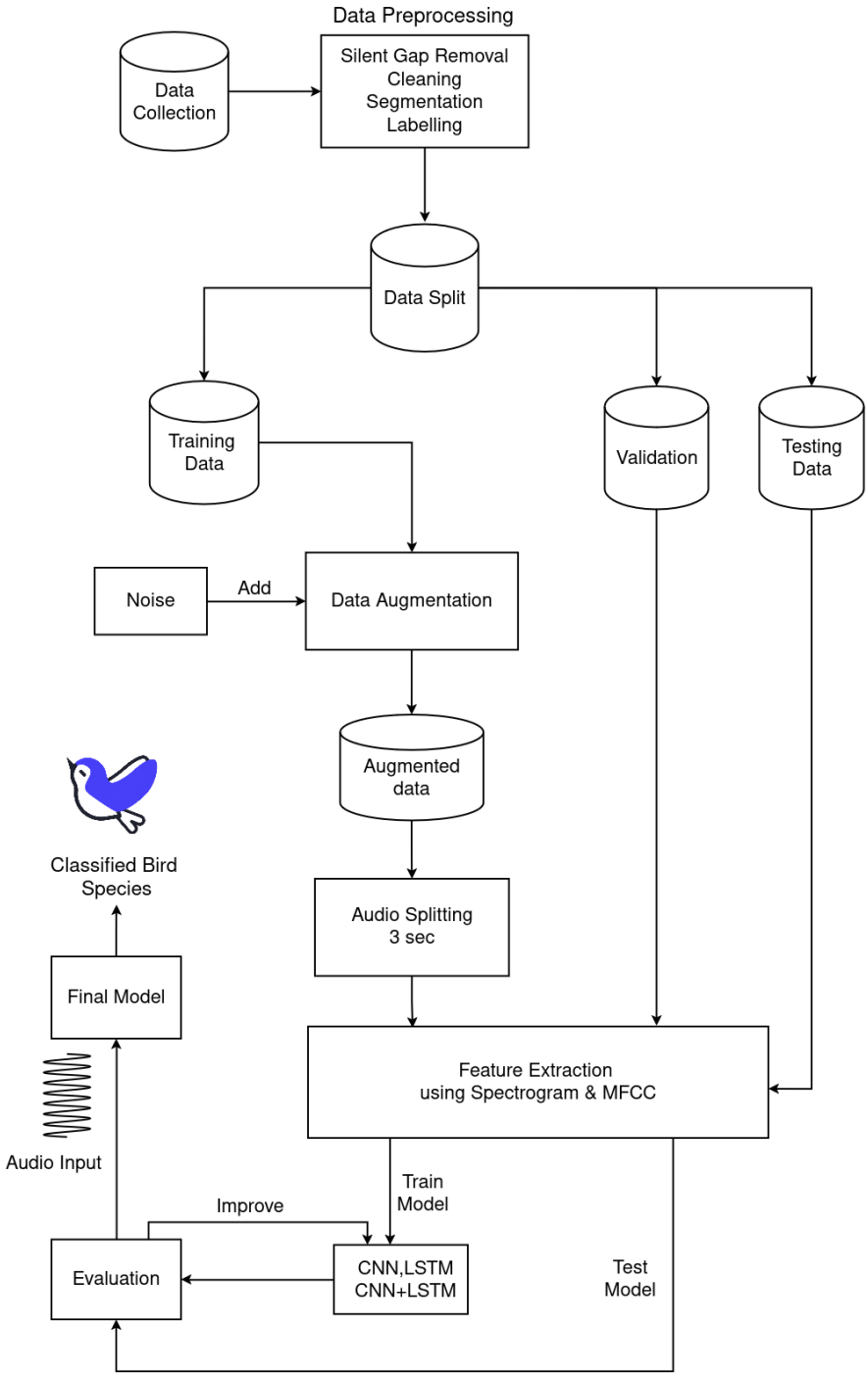

Figure 3.1: Block diagram for the working mechanism of the system

The methodology for the project ”Bird Classification from Audio” as outlined in the provided paper involves a comprehensive approach that spans from data collection to the implementation of a novel deep learning model. This methodology is designed to accurately classify bird species based on their calls, leveraging advanced signal processing and machine learning techniques. The following sections detail the methodology:

3.1.1 Dataset

For this project, we are using the CharaNet dataset, which is specifically curated to include audio files of Nepal's endangered bird species. This dataset was collected from *xeno-canto.org*, a platform where bird sounds from around the world are shared by contributors who travel extensively to capture these sounds. From this source, we gathered 2215 audio recordings representing 41 bird species, 38 of which are listed as endangered in Nepal.

To augment the dataset, we employed a 10-second split method, which expanded the dataset to 6733 audio recordings. Additionally, for bird species with fewer than 30 recordings, we added Gaussian noise to further increase the number of samples. This resulted in a robust dataset comprising 5407 audio recordings for training, 639 for validation, and 687 for testing. This comprehensive dataset ensures that our model can learn effectively and generalize well to new, unseen data.

3.1.2 Dataset Overview for Audio

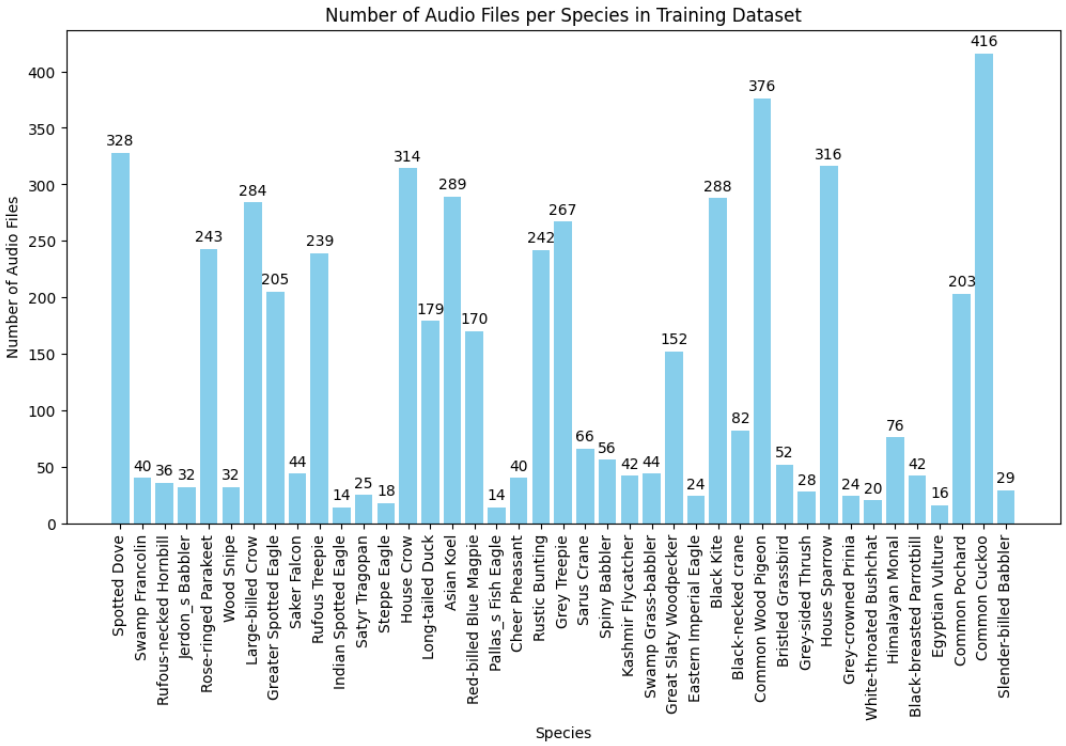

Figure 3.2: Training Dataset

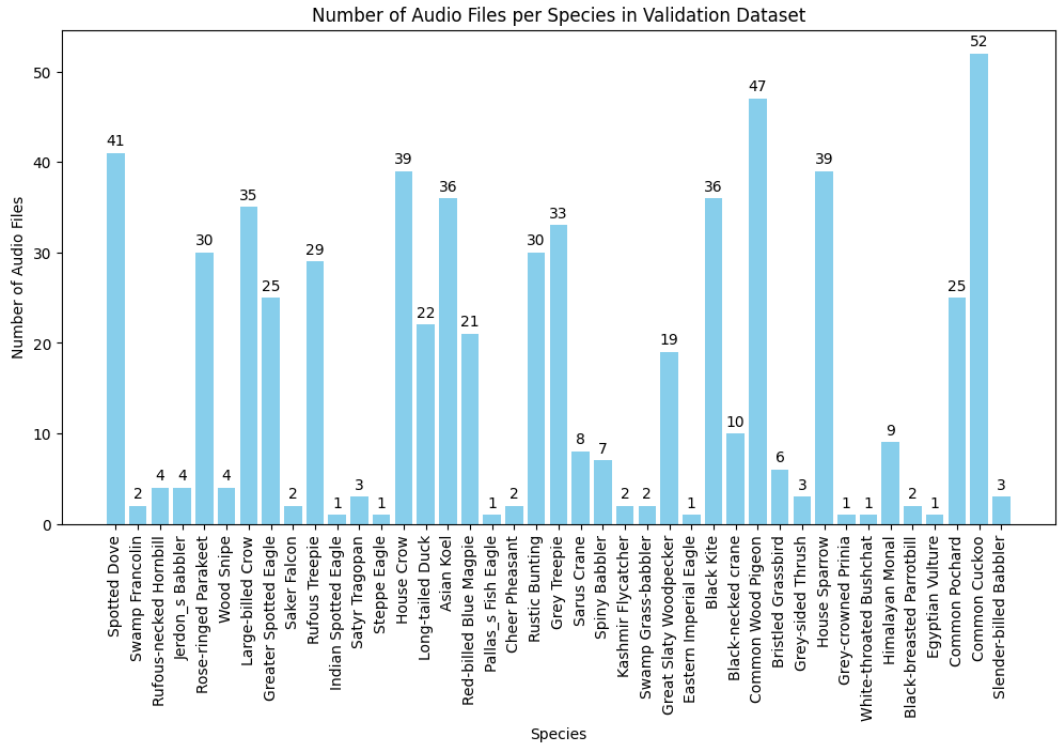

Figure 3.3: Validation Dataset

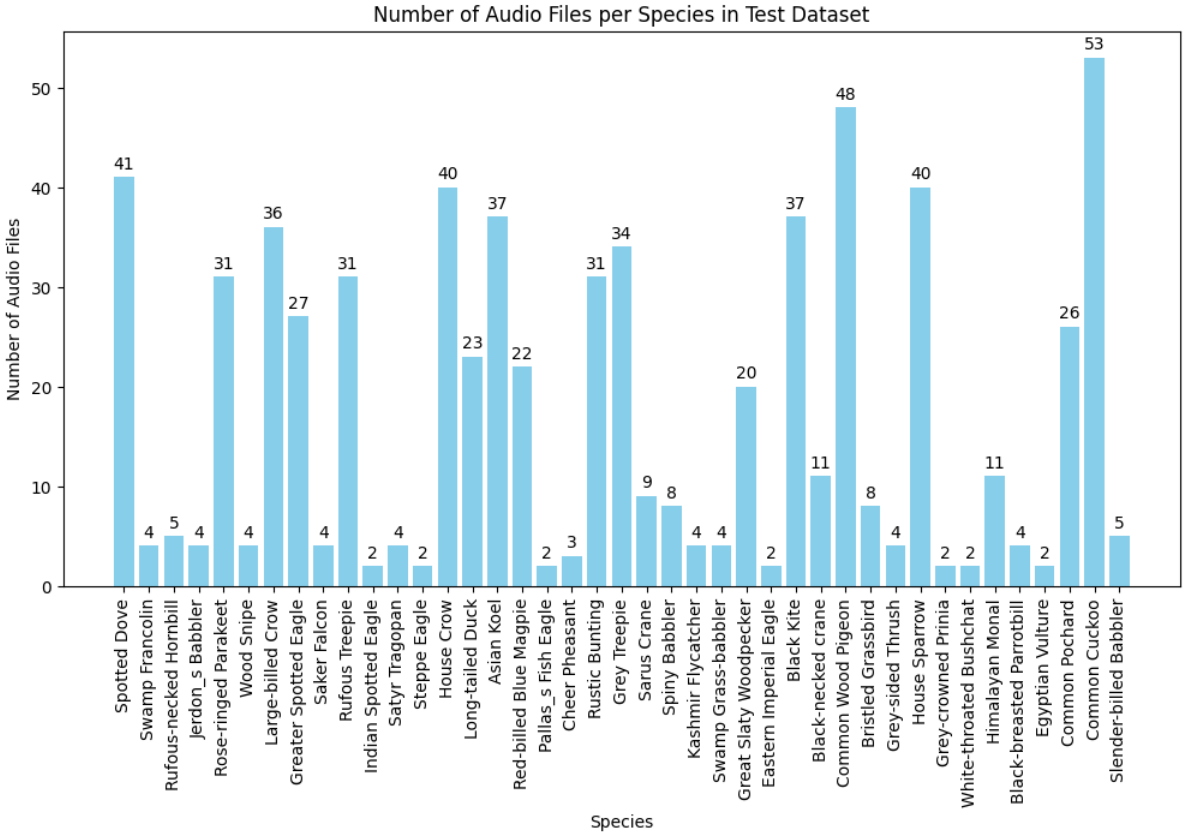

Figure 3.4: Testing Dataset

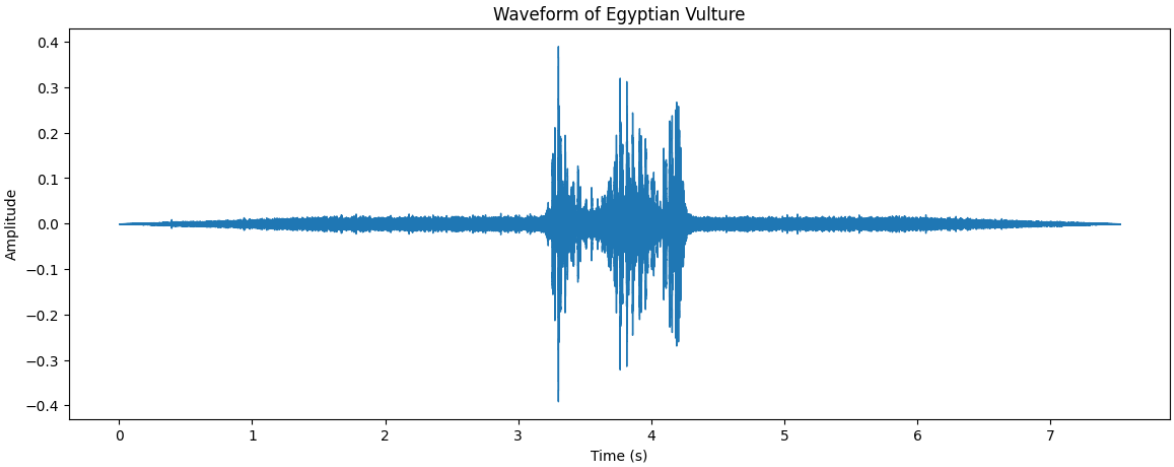

Figure 3.5: Sample of the Data

3.1.3 Data Preprocessing

The collected audio data undergoes thorough preprocessing to ensure its suitability for model training. This process includes:

- **Silence Gap Removal:** After observing some of the data samples, it was found that the audio has some silence gaps in either of the ends as seen in **??**. These could be removed employing the silence removal algorithm. The processing starts from both ends and moves toward the center. In that processing, the local mean of the window segment of the audio wave is calculated and compared with the audio's global mean. If the local mean is smaller than the global mean, the window segment is considered to contain insignificant data, thus the segment is clipped from the original audio[12]. Algorithm 3.1.12.1 clarifies this.
- **Segmentation:** Dividing the continuous audio recordings into smaller, more manageable segments, ensuring consistency in input length.
- **Cleaning:** Removing any background noise and irrelevant sounds that could interfere with the training process.
- **Labeling:** Assigning the correct bird species labels to each audio segment, which is crucial for supervised learning.

3.1.4 Data Splitting

After preprocessing, the data is split into three sets:

- **Training Data:** Used to train the machine learning models.
- **Testing Data:** Used to evaluate the model's performance during training.
- **Validation Data:** Used to validate the final model's performance and prevent overfitting.

This structured approach ensures that the model can learn effectively from the training data while being evaluated on unseen data to measure its generalization capabilities.

3.1.5 Data Augmentation

To further enhance the robustness of the model, data augmentation techniques are applied to the training data. This includes adding various types of noise to the audio segments to create more diverse training samples and prevent the model from overfitting to specific patterns in the data.

$$\text{Augmented Data} = \text{Original Data} + \text{Noise} \tag{3.1}$$

By introducing these variations, the model becomes more resilient to different audio conditions and better at generalizing to new recordings.

3.1.6 Audio Splitting

The augmented audio data is then split into smaller 3-second clips. This standardization ensures that the input size remains consistent, which is essential for the feature extraction and modeling stages.

3.1.7 Feature Extraction

Feature extraction is a critical step where the audio clips are transformed into a format that can be fed into machine learning models. We use two primary techniques for feature extraction: Spectrograms and Mel-Frequency Cepstral Coefficients (MFCC).

3.1.7.1 Spectrogram

A spectrogram is a visual representation of the spectrum of frequencies in a sound signal as they vary with time. It is generated by applying the Short-Time Fourier Transform (STFT) to the audio signal. This transformation provides insight into how the frequency

content of the signal changes over time.

$$X(n) = \sum_{m=0}^{N-1} x(m) \cdot w(n-m) \tag{3.2}$$

where $x(m)$ is the audio signal and $w(n)$ is the window function.

3.1.7.2 Mel-Frequency Cepstral Coefficients (MFCC)

MFCCs are coefficients that collectively describe the short-term power spectrum of a sound signal. The process of obtaining MFCCs involves several steps:

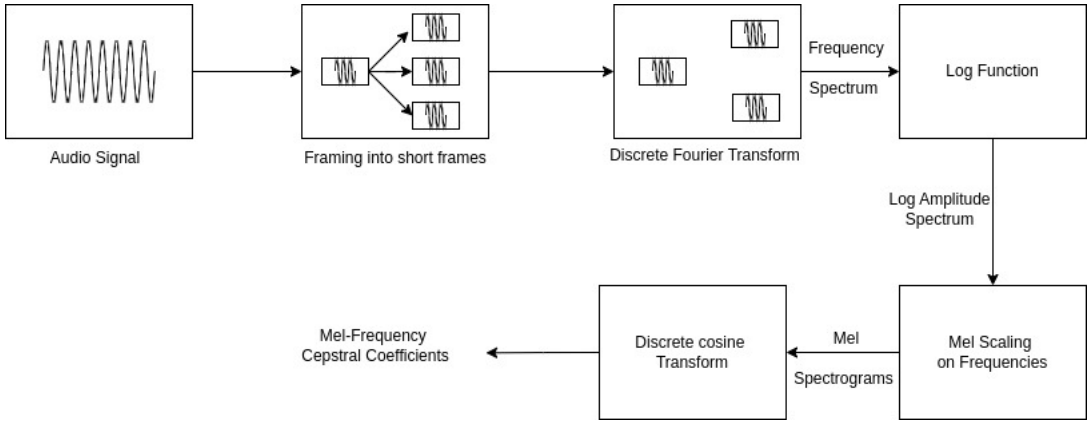

Figure 3.6: Feature Extraction Using Spectrogram and MFCC

1. **Framing:** Divide the audio signal into short frames.
2. **Discrete Fourier Transform (DFT):** Convert each frame to the frequency domain.

$$X(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-j\frac{2\pi}{N}kn} \tag{3.3}$$

3. **Log Function:** Apply a logarithm to the amplitude spectrum.

$$S_{\log}(k) = \log(|X(k)|) \tag{3.4}$$

4. **Mel-Scaling:** Map the frequencies to the Mel scale, which better represents how

humans perceive sound.

$$f_{\text{mel}} = 2595 \cdot \log_{10}(1 + \frac{f}{700}) \tag{3.5}$$

5. **Discrete Cosine Transform (DCT):** Convert the Mel spectrum to the cepstral domain, yielding the MFCC features.

$$C(n) = \sum_{k=0}^{K-1} S_{\text{mel}}(k) \cdot \cos\left(\frac{\pi n(k+0.5)}{K}\right) \tag{3.6}$$

3.1.8 Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) are highly effective for extracting spatial features from spectrograms, making them well-suited for audio classification tasks. CNNs use convolutional layers to detect patterns and features in the input data by applying convolutional filters across the input spectrogram.

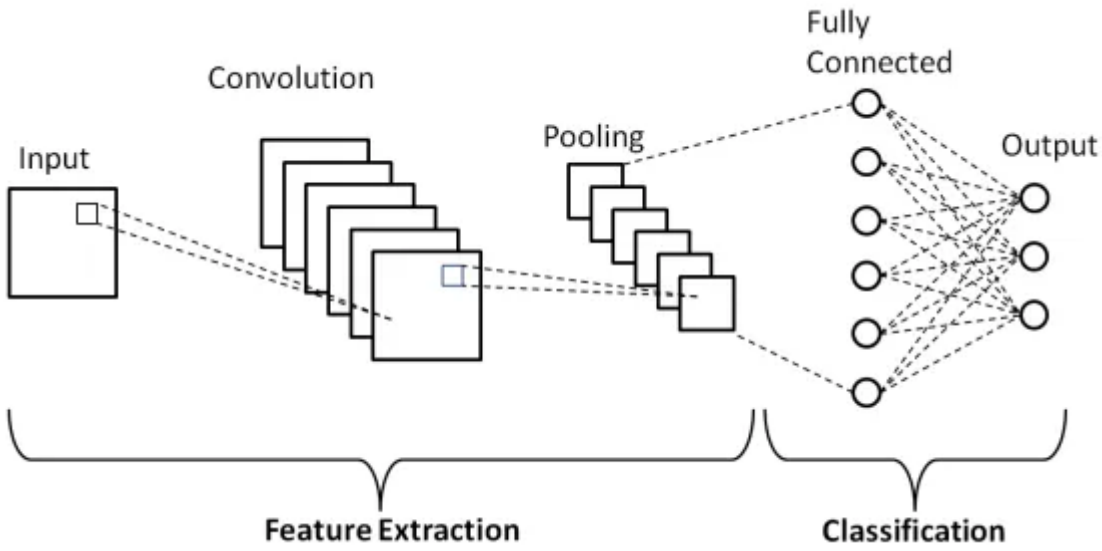

Figure 3.7: CNN architecture

The basic operations involved in a CNN include:

- **Convolution:** This operation involves applying a set of learnable filters (kernels)

across the input spectrogram to produce feature maps.

$$(f * x)(t) = \sum_{\tau=-\infty}^{\infty} x(\tau) \cdot f(t - \tau) \tag{3.7}$$

- **Activation Function:** The rectified linear unit (ReLU) activation function is applied to introduce non-linearity into the model.

$$f(x) = \max(0, x) \tag{3.8}$$

- **Pooling:** Pooling layers reduce the spatial dimensions of the feature maps, typically using max pooling to retain the most significant features.
- **Fully Connected Layers:** After several convolutional and pooling layers, the feature maps are flattened and passed through fully connected layers to produce the final classification output.

3.1.9 Long Short-Term Memory Networks (LSTM)

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) capable of capturing temporal dependencies in sequential data. This makes them well-suited for processing audio signals where the order of data points matters. LSTMs are designed to overcome the limitations of traditional RNNs by addressing the vanishing gradient problem through the use of gates that regulate the flow of information.

The key components of an LSTM cell include:

- **Forget Gate:** Determines which information from the previous cell state should be discarded.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{3.9}$$

- **Input Gate:** Decides which new information should be added to the cell state.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3.10}$$

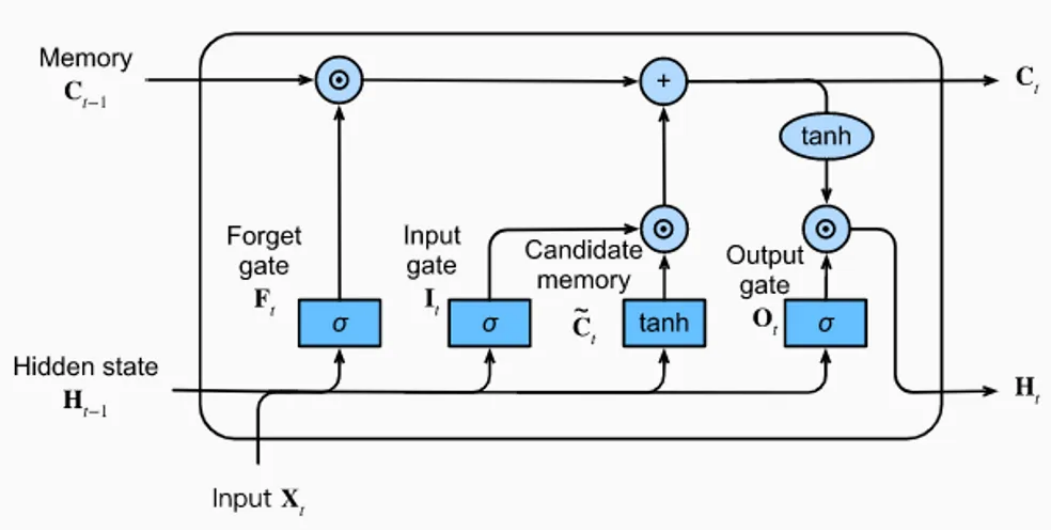

Figure 3.8: LSTM architecture

- **Candidate Cell State:** Creates new candidate values that could be added to the cell state.

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{3.11}$$

- **Cell State:** The cell state is updated based on the input gate and the forget gate.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{3.12}$$

- **Output Gate:** Determines the output of the LSTM cell.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{3.13}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{3.14}$$

3.1.10 Combined CNN and LSTM (CNN+LSTM)

A hybrid model that leverages the strengths of both CNN and LSTM architectures is developed. In this combined model, the CNN processes spectrograms to extract spatial features, which are then fed into LSTM layers to capture temporal patterns. The integration of these two models aims to utilize both spatial and temporal information,

thereby improving the overall classification accuracy. Fully connected layers at the end perform the final classification based on the features extracted by both CNN and LSTM components.

3.1.11 Hyperparameter Optimization using Genetic Algorithm

Genetic algorithms (GAs) are a powerful method for optimizing hyperparameters in machine learning models. Genetic Algorithm have proven to significantly improve the performance metrics of the CNN model instead of using hand tuned approach for hyperparameters. This section outlines the steps involved in using GAs for hyperparameter optimization[13].

- **Encoding the Hyperparameters**
  - Hyperparameters are represented as a chromosome, where each hyperparameter is a gene in the chromosome.
  - For example, in a neural network, a chromosome might include genes for the learning rate, number of layers, number of neurons per layer, and activation functions.
- **Initial Population**
  - An initial population of chromosomes is generated randomly, with each chromosome representing a different set of hyperparameters.
- **Fitness Function**
  - A fitness function is defined to evaluate the performance of each set of hyperparameters.
  - This typically involves training the model with the given hyperparameters and measuring its performance on a validation set.
- **Selection**
  - Selection involves choosing the best-performing chromosomes to serve as parents for the next generation.
  - Various selection methods can be employed, such as tournament selection, roulette wheel selection, or rank-based selection.

- **Crossover (Recombination)**
  - Crossover combines pairs of parent chromosomes to produce offspring for the next generation.
  - This is done by swapping segments of parent chromosomes to create new chromosomes, thereby combining features of both parents.
- **Mutation**
  - Mutation introduces random changes to some of the genes in the offspring chromosomes.
  - This helps maintain genetic diversity in the population and allows the algorithm to explore a broader search space.
- **Replacement**
  - The current population is partially or entirely replaced with the new generation of chromosomes, ensuring that better solutions are carried forward while allowing for exploration of new possibilities.
- **Termination**
  - The process of selection, crossover, mutation, and replacement is repeated until a termination criterion is met.
  - This could be a set number of generations, convergence of fitness scores, or achieving a satisfactory performance level.
- **Best Solution**
  - The best chromosome at the end of the process represents the optimal or near-optimal set of hyperparameters for the model.

3.1.12 Algorithms Used

3.1.12.1 Silent Gaps Removal

Algorithm 1 Clipping of silent gaps from both ends

wav $\leftarrow$ sampled audio signal
$\Delta \leftarrow$ appropriate window length
/* In our code, $\Delta = 500$ for 16KHz sampling rate */
INPUT: wav, Δ
PROCESS:
wavAvg $\leftarrow$ Average($|$wav$|$)
$N \leftarrow$ Length(wav)
/* Removing the silent gap from the start */
for idx $= 0, \Delta, 2\Delta, \ldots, N - \Delta$ **do**
    win $\leftarrow$ wav[idx : idx $+ \Delta$]
    winAvg $\leftarrow$ Average($|$win$|$)
    if winAvg $>$ wavAvg **then**
        wav $\leftarrow$ wav[idx :]
        break
    end if
end for
/* Removing the silent gap from the end */
for idx $= N - \Delta, N - 2\Delta, \ldots, 0$ **do**
    win $\leftarrow$ wav[idx : idx $+ \Delta$]
    winAvg $\leftarrow$ Average($|$win$|$)
    if winAvg $>$ wavAvg **then**
        wav $\leftarrow$ wav[: idx]
        break
    end if
end for
OUTPUT: processed_wav $\leftarrow$ wav

3.1.12.2 Genetic Algorithm

Algorithm 2 Genetic Algorithm for Hyperparameter Optimization

Initialize the population with random hyperparameters.
for generation = 1 to N **do**
    Evaluate the fitness of each individual in the population.
    Select individuals to be parents based on their fitness scores.
    Generate offspring through crossover.
    Apply mutation to the offspring.
    Replace the old population with the new generation.
end for
Return the best solution found.

3.1.12.3 Fitness Function

The fitness function evaluates the performance of hyperparameters by training and validating the model:

Algorithm 3 Fitness Function

1: Train the model with given hyperparameters.
2: Evaluate the model's performance on a validation set.
3: **return** Performance metric (e.g., accuracy, F1 score).

3.1.12.4 Mel Spectrogram

Algorithm 4 Mel Spectrogram Extraction

1: **Input:** Audio signal $x(t)$.
2: **Output:** Mel spectrogram S_{mel}.
3: Apply STFT to generate spectrogram $S(f, t)$.
4: Convert amplitudes of $S(f, t)$ to dB scale, obtaining $S_{\text{dB}}(f, t)$.
5: **Convert frequencies to Mel scale**.
6: Choose number of mel bands N_{mel}.
7: **Construct mel filter banks**.
8: Convert f_{min} and f_{max} of $S_{\text{dB}}(f, t)$ to Mel scale.
9: Divide Mel scale range into N_{mel} intervals.
10: Convert center frequencies of Mel bands back to Hertz.
11: Round center frequencies to nearest bins.
12: Design triangular band pass filters for each Mel band.
13: Apply mel filter banks to $S_{\text{dB}}(f, t)$ to obtain S_{mel}.

3.1.12.5 Mel-Frequency Cepstral Coefficients Extraction

Algorithm 5 MFCC Extraction

1: **Input:** Audio signal $x(t)$
2: **Output:** MFCC coefficients.
3: Frame the signal into short frames.
4: Apply a window function to the frames.
5: Apply DFT to generate the frequency spectrum of each frame.
6: Apply logarithm to the spectrum to get log amplitude spectrum.
7: Perform Mel scaling using filter banks to get Mel spectrogram.
8: Apply DCT to the Mel spectrogram to get MFCCs.

3.2 Working Mechanism for Identification using Image

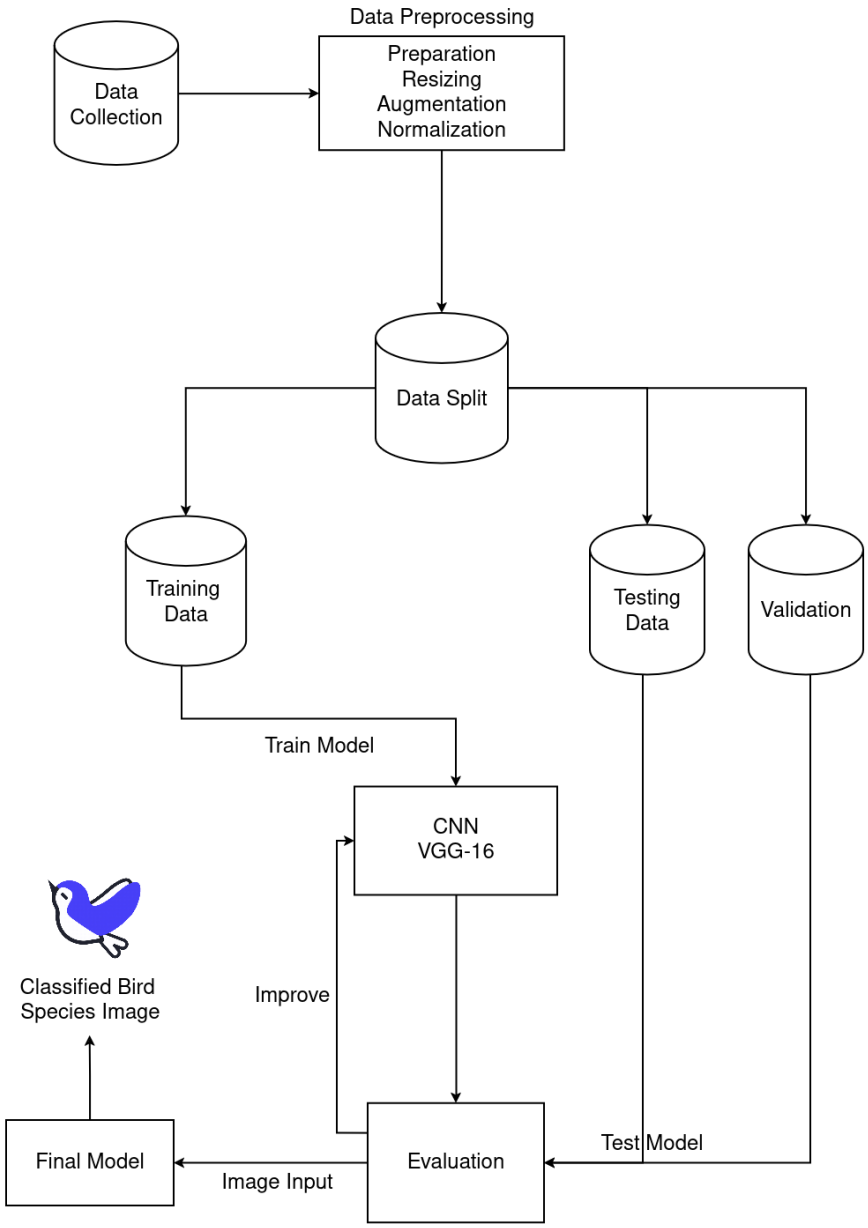

Figure 3.9: Block diagram for the working mechanism of the system

The working mechanism for the bird identification using image is to be done by VGG16. VGG16 have been proven to classify bird speices with accuracy of 92.13% using Caltech-UCSD Birds-200-2011 (CUB-200-2011) dataset which had 1600 images across 27 bird species[7]. Our project aims to achieve similar result for 200 bird species.

3.2.1 Dataset

For the identification of bird species using image, we are using CUB-200-2011 dataset, which is an extended version of the CUB-200 dataset. It consists of 11,788 number of images across 200 species categories[14].

3.2.2 Dataset Overview for Image

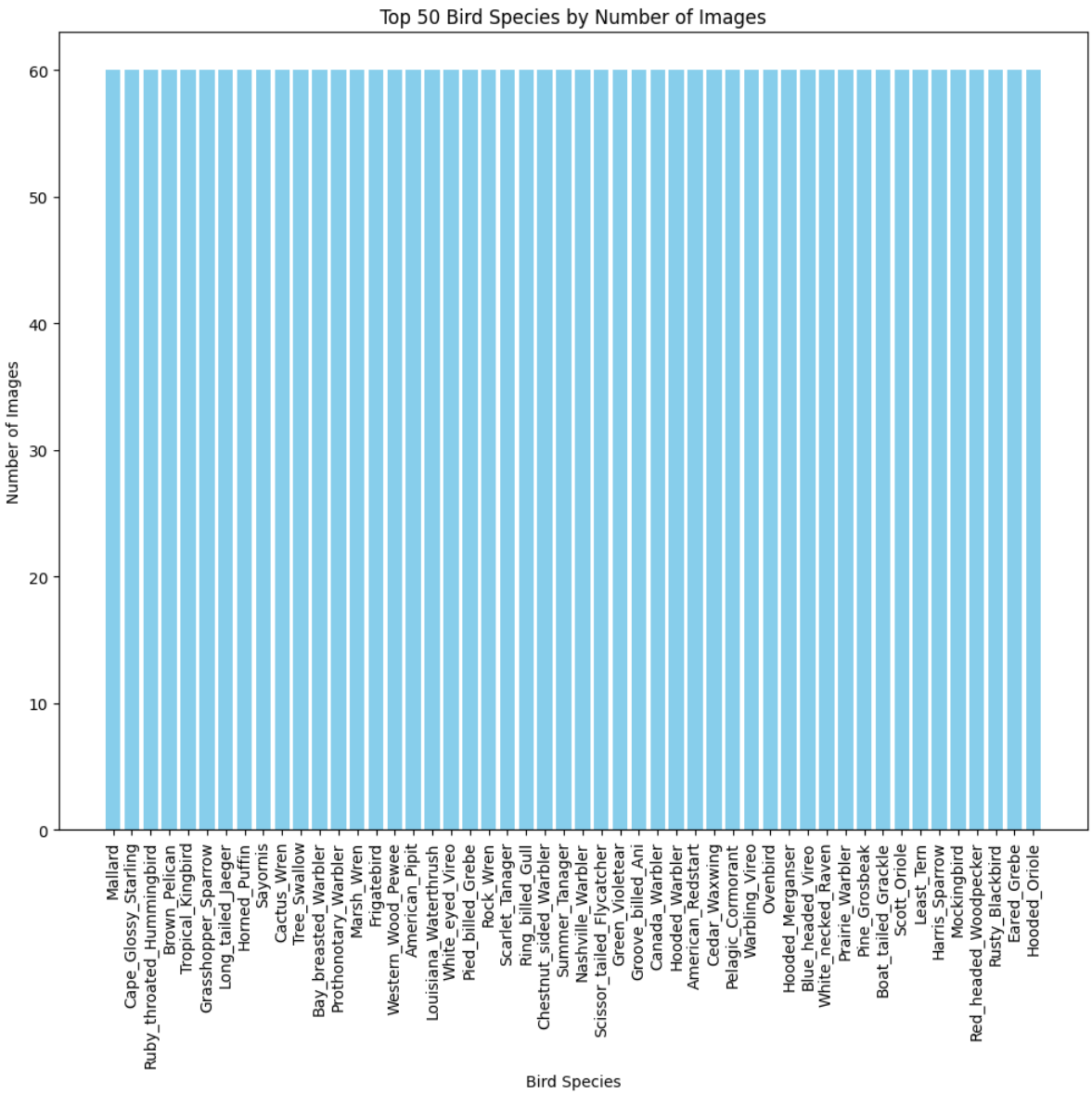

Figure 3.10: Dataset distribution for Top 50 Bird Species among 200 Species

After observing the distribution of the images across the species, most of them have 60 images per species while few of them ranges from 45 to 60.

3.2.3 Data Preprocessing

For the image classification of bird species, the data preprocessing stage is crucial for enhancing model performance and ensuring the input data is in the correct format. Initially, all images are resized to 224x224 pixels with 3 channels (RGB) to match the input size requirement of the model architecture. This uniformity is essential for the model to process the images efficiently. Following resizing, images undergo normalization, where pixel values are scaled to a range of 0 to 1. This step is vital as it helps in speeding up the convergence by reducing the variability in input data, making the optimization landscape smoother. Additionally, to augment the dataset and introduce more variability, data augmentation techniques such as flipping, rescaling, and shearing are applied. These techniques not only help in generating a more robust dataset but also prevent overfitting by simulating various perspectives and conditions. Finally, the dataset is split into training, validation, and test sets, ensuring that the model can learn effectively, validate its learning without bias, and finally, be tested on unseen data to evaluate its generalization capability. This comprehensive preprocessing pipeline is instrumental in preparing the dataset for effective training and evaluation of the model for bird species classification.

3.2.4 VGG16

For bird species image classification, the VGG-16 model serves as a powerful convolutional neural network (CNN) framework, originally developed by the Visual Geometry Group (VGG) at the University of Oxford. This model is distinguished by its depth, comprising 16 layers, which include 13 convolutional layers followed by 3 fully connected layers at the end. The architecture of VGG-16 is celebrated for its straightforward yet highly effective design, making it a robust tool for various computer vision tasks, notably in the realm of image classification and object detection. In the context of bird species classification, VGG-16's architecture is particularly beneficial. Its multiple convolutional layers, arranged in a deep stack followed by max-pooling layers, allow the model to capture and learn from the complex, hierarchical visual features specific to different bird species. This capability enables the model to distinguish between

species with high accuracy, leveraging the depth of the network to process and recognize subtle differences in plumage, size, and shape among birds. Despite the emergence of newer models, VGG-16's combination of depth, simplicity, and performance ensures its continued relevance and effectiveness in classifying bird species from images, making it a preferred choice for researchers and practitioners in the field of ornithology and computer vision alike.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Figure 3.11: VGG16 architecture

3.3 System Diagrams

The usecase diagram for Bird species identification from Audio and Image is given below:

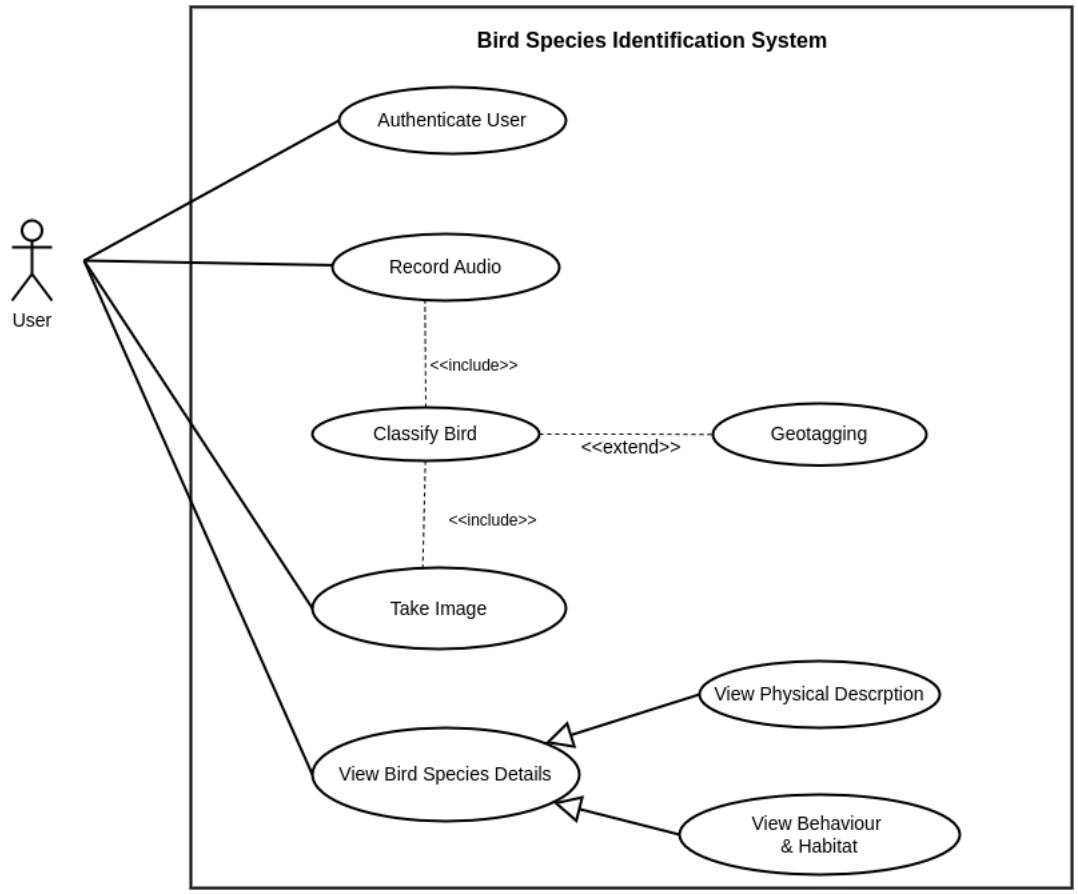

Figure 3.12: UseCase Diagram

3.4 Software Development Model

This project is developed using an incremental methodology since it offers a functioning prototype at an early stage of development. The requirements and scope of the project can be altered as necessary by studying the prototype. The rationale for the preference for this software development strategy is the flexibility offered by adopting the incremental technique. In this paradigm, the project goes through several releases or iterations prior to its official release.

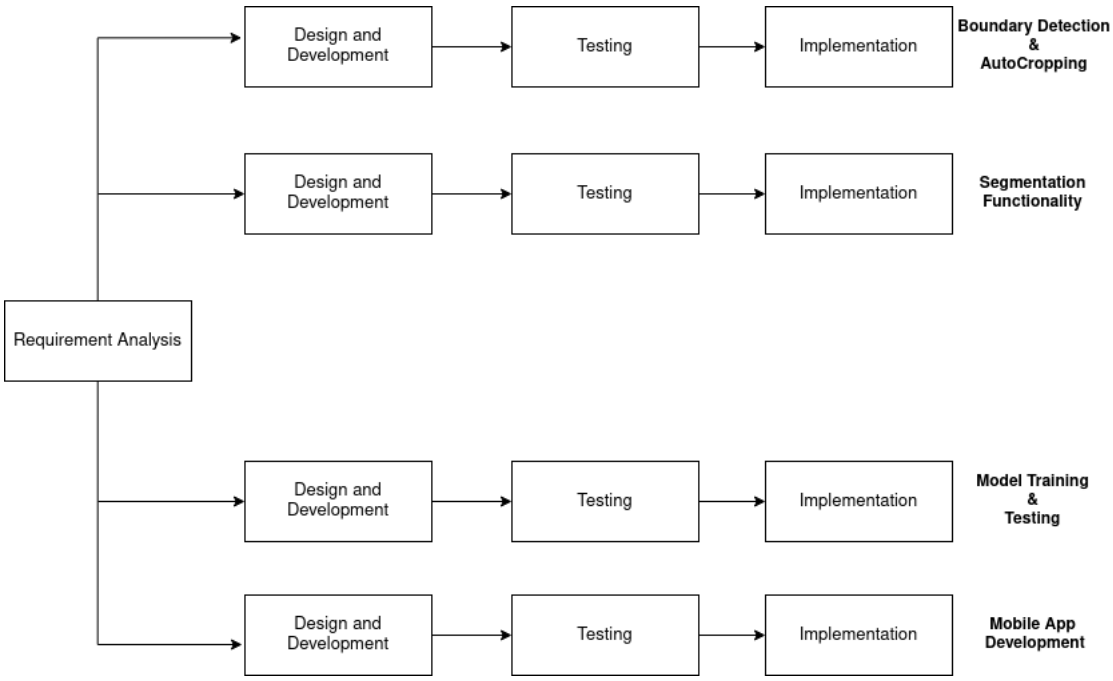

Figure 3.13: Incremental Model for development of SnapTag

CHAPTER 4
EPILOGUE

REFERENCES

[1] I. O. Union, "Ioc world bird list," https://www.worldbirdnames.org/new/updates/, 2024, accessed: 2024-06-04.

[2] H. Nature, "National red list of nepal's birds," https://www.himalayannature.org/works/projects/national-red-list-of-nepals-birds/, 2024, accessed: 2024-06-03.

[3] C. Inskipp, H. S. Baral, T. Inskipp, A. P. Khatiwada, M. P. Khatiwada, L. P. Poudyal, and R. Amin, "Nepal's national red list of birds," *Journal of Threatened Taxa*, vol. 9, no. 1, pp. 9700–9722, 2017.

[4] R. Gautam, B. Khatiwada, B. P. Subedi, N. Duwal, and K. C. Dahal, "Audio classifier for automatic identification of endangered bird species of nepal," 2023.

[5] A. Sevilla and H. Glotin, "Audio bird classification with inception-v4 extended with time and time-frequency attention mechanisms." *CLEF (Working Notes)*, vol. 1866, pp. 1–8, 2017.

[6] P. Gavali, P. A. Mhetre, N. C. Patil, N. S. Bamane, and H. D. Buva, "Bird species identification using deep learning," *International Journal of Engineering Research and Technology*, vol. 8, no. 4, April 2019.

[7] S. Islam, S. I. A. Khan, M. M. Abedin, K. M. Habibullah, and A. K. Das, "Bird species classification from an image using vgg-16 network," in *Proceedings of the 7th international conference on computer and communications management*, 2019, pp. 38–42.

[8] B. Chandu, A. Munikoti, K. S. Murthy, G. Murthy, and C. Nagaraj, "Automated bird species identification using audio signal processing and neural networks," in *2020 International Conference on Artificial Intelligence and Signal Processing (AISP)*. IEEE, 2020, pp. 1–5.

[9] 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.

[10] H. Wang, Y. Xu, Y. Yu, Y. Lin, and J. Ran, "An efficient model for a vast number of bird species identification based on acoustic features," *Animals*, vol. 12, no. 18, p. 2434, 2022.

[11] R. Pahuja and A. Kumar, "Sound-spectrogram based automatic bird species recognition using mlp classifier," *Applied Acoustics*, vol. 180, p. 108077, 2021.

[12] M. Dhakal, A. Chhetri, A. K. Gupta, P. Lamichhane, S. Pandey, and S. Shakya, "Automatic speech recognition for the nepali language using cnn, bidirectional lstm and resnet," pp. 515–521, 2022.

[13] A. Reiling, W. Mitchell, S. Westberg, E. Balster, and T. Taha, "Cnn optimization with a genetic algorithm," in *2019 IEEE National Aerospace and Electronics Conference (NAECON)*, 2019, pp. 340–344.

[14] Tech. Rep.