

# A Domain and Noise Adversarial Bird Tune Classification Pipeline Using Deep Neural Network

By

Aparna Sarker Riya

18301194

Arpita Roy

18101332

Md. Abrar Fahim

18301006

Zarin Tasnim

18101352

Rakibul Islam

17101478

A thesis submitted to the Department of Computer Science and Engineering  
in partial fulfillment of the requirements for the degree of  
B.Sc. in Computer Science

Department of Computer Science and Engineering  
Brac University  
September 2022

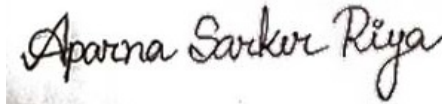
© 2022. Brac University  
All rights reserved.

# Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

**Student's Full Name & Signature:**



---

Aparna Sarker Riya

18301194



---

Arpita Roy

18101332

Abrar Fahim

---

Md. Abrar Fahim

18301006

Rakibul Islam

---

Rakibul Islam

17101478

Zarin Tasnim

---

Zarin Tasnim

18101352

# Approval

The thesis titled “A Domain and Noise Adversarial Bird Tune Classification Pipeline Using Deep Neural Network” submitted by

1. Aparna Sarker Riya (18301194)
2. Arpita Roy (18101332)
3. Md. Abrar Fahim (18301006)
4. Zarin Tasnim (18101352)
5. Rakibul Islam (17101478)

Of Summer, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on September 20, 2022.

## Examining Committee:

Supervisor:  
(Member)

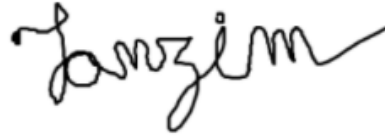


---

Moin Mostakim

Senior Lecturer  
Department of Computer Science and Engineering  
BRAC University

Co-Supervisor:  
(Member)



---

Md Tanzim Reza

Lecturer  
Department of Computer Science and Engineering  
BRAC University

Program Coordinator:  
(Member)

---

Md Golam Rabiul Alam

Assistant Professor  
Department of Computer Science and Engineering  
BRAC University

Head of Department:  
(Chair)

---

Dr. Sadia Hamid Kazi

Chairperson and Associate Professor  
Department of Computer Science and Engineering  
BRAC University

# Abstract

Birds are an important category of animals that ecologists keep track of utilizing autonomous recording units as a key indication of environmental health. Because of the consequences of climate change and the rising number of endangered species, many experts suggested developing an animal species recognition system to help them in specialized research. Researchers can improve their ability to assess the state of biodiversity and its patterns in crucial ecosystems by precise sound detection and categorization, which is supported by machine learning, allowing them to better support global conservation efforts. However, producing analysis outputs with high precision and recall remains a difficulty. Due to a lack of appropriate methods for efficient and accurate extraction of interest signals, the vast bulk of data remains unexplored (e.g., bird calls). Moreover, due to strong source-domain specific features and artificial/natural noises, these acquired raw data create different distributions in datasets. So, to ensure a generalized feature learning, domain adaptation [1] techniques will be implemented in this work to make the networks familiar towards both acquisition sensor noises and background noises without having to do intensive dataset specific augmentations. We used 3 popular and powerful DNN models, including CNN, VGG19 and ResNet50. Out of them, for the bird species classification task VGG19 achieved the best accuracy of 96.02% in testing and 94.01% in training. To the best of our knowledge, this will guide towards convenient and deployable in real life models which will allow future works into the pipeline to ensure better coverage.

**Keywords:** biodiversity, domain adaptation, classification, CNN, VGG19, RESNET50, ReLU

# Acknowledgement

Alhamdulillah.

Firstly, all praise to the Great Allah for whom our thesis has been completed without any major interruption. Secondly, to our supervisor and co-advisor sir for his kind support and advice in our work. He helped us whenever we needed help. And finally to our parents without their throughout support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

# Table of Contents

Declaration	i
Approval	iii
Abstract	v
Acknowledgment	vi
Table of Contents	vii
List of Figures	ix
Nomenclature	x
<b>1 Introduction</b>	<b>1</b>
1.1 Research Motivation . . . . .	1
1.2 Research Objectives . . . . .	1
1.3 The Challenges Regarding Bird Tune Identification . . . . .	2
<b>2 Literature Review</b>	<b>3</b>
<b>3 Proposed Method</b>	<b>6</b>
3.1 Model Details . . . . .	6
3.1.1 Transfer Learning . . . . .	6
3.1.2 Learning Rate . . . . .	6
3.1.3 Convolutional Neural Network (CNN) . . . . .	6
3.1.4 Optimizer . . . . .	7
3.1.5 Activation Functions . . . . .	7
3.1.6 VGG19 . . . . .	8
3.1.7 Resnet-50 . . . . .	9
3.2 Work Plan . . . . .	9
<b>4 Dataset Description</b>	<b>12</b>
4.1 Introduction . . . . .	12
4.2 Dataset Details . . . . .	12
4.3 Data Preprocessing . . . . .	12
<b>5 Result and Analysis</b>	<b>15</b>
5.1 Simple CNN . . . . .	15
5.2 VGG19 . . . . .	19



5.3	ResNet-50 . . . . .	19
<b>6</b>		<b>27</b>
6.1	Conclusion . . . . .	27
6.2	Limitations . . . . .	27
6.3	Future Work . . . . .	27
	<b>Bibliography</b>	<b>30</b>

# List of Figures

3.1	Convolutional Neural Network . . . . .	7
3.2	ReLU activation function [17] . . . . .	8
3.3	VGG19 architecture . . . . .	9
3.4	Resnet50 architecture . . . . .	10
3.5	Flowchart of Implementation . . . . .	11
4.1	Continent wise species distribution . . . . .	14
5.1	Accuracy and loss diagram for Africa(CNN) . . . . .	16
5.2	Accuracy and loss diagram for Asia(CNN) . . . . .	16
5.3	Accuracy and loss diagram for Europe(CNN) . . . . .	17
5.4	Accuracy and loss diagram for North America(CNN) . . . . .	17
5.5	Accuracy and loss diagram for Oceania(CNN) . . . . .	18
5.6	Accuracy and loss diagram for South America(CNN) . . . . .	18
5.7	Accuracy and loss diagram for Africa(VGG19) . . . . .	19
5.8	Accuracy and loss diagram for Asia(VGG19) . . . . .	20
5.9	Accuracy and loss diagram for Europe(VGG19) . . . . .	20
5.10	Accuracy and loss diagram for North America(VGG19) . . . . .	21
5.11	Accuracy and loss diagram for Oceania(VGG19) . . . . .	21
5.12	Accuracy and loss diagram for South America(VGG19) . . . . .	22
5.13	Accuracy and loss diagram for Africa(ResNet50) . . . . .	22
5.14	Accuracy and loss diagram for Asia(ResNet50) . . . . .	23
5.15	Accuracy and loss diagram for Europe(ResNet50) . . . . .	23
5.16	Accuracy and loss diagram for North America(ResNet50) . . . . .	24
5.17	Accuracy and loss diagram for Oceania(ResNet50) . . . . .	24
5.18	Accuracy and loss diagram for South America(ResNet50) . . . . .	25
5.19	Model Accuracy graph . . . . .	25

# Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

*Adagrad* Adaptive Gradient Algorithm

*ANN* Artificial Neural Network

*ARU* Autonomous Recording Units

*CNN* Convolutional Neural Network

*DNN* Deep Neural Network

*DTW* Dynamic Time Warping

*FCN* Fully Convolutional Network

*FLOPS* Floating-Point Operations per Second

*GMM* Gaussian Mixture Model

*HMM* Hidden Markov Model

*LSTM* Long Short-Term Memory

*MLP* Multilayer Perceptrons

*ReLU* Rectified Linear Unit

*ResNet* ResNet : Residual Neural Network

*RMSProp* Root Mean Squared Propagation

*SOM* The Self Organizing Map

*SVM* Support Vector Machine

*VGG* Visual Geometry Group

*WPD* Wavelet Packet Decomposition

# Chapter 1

## Introduction

Birds are one of the most important groups of animals that environmentalists monitor with audio recordings, as birds are an important indicator of biodiversity. Bird songs or bird calls can be split into hierarchical layers of phrases, syllables, and components or musical notes [2]. Syllables are made up of multiple elements, but they are an effective unit for bird identification because they can be detected more correctly from continuous records than from elements. More regional and individual variances can be seen in phrases and songs than in syllables. Birds can act as sentinels, umbrella species, model organisms, and flagship species.

### 1.1 Research Motivation

On this planet, there are 50 billion to 430 billion birds. Since 1500, 129 different bird species have disappeared from the planet. 10–20% of all avian species have been lost due to human activity [3]. Though the topic of our research is not new and already many works have been done, we chose this topic to create the awareness of protecting birds. Due to global warming and climate change many bird species are going extinct. It is very hard to locate endangered species in an acoustic environment manually or by setting a camera. As it can only cover a limited area and it is not feasible to know the accurate species through visualization. To solve this problem, we will try to develop a system which will be able to identify birds based on their audio recordings. Thus, people can easily know about the regional birds and their species. Hence, the endangered bird can be protected and future generations will be able to know about them.

### 1.2 Research Objectives

Our research is basically based on audio. To work with audio sounds it is really hard to capture all of a species bird's voice. By this research we are detecting voice and identifying the desired bird by using some algorithms. Working on too much data is not effective as external noise can make a hassle in voice identification. In an open place environment many sounds can come in a record which can not be hearable in person. For this reason, our plan is to work with our dataset in an effective way. Besides, there are some objectives in it. They are :

1. Generating spectrograms from audio datasets.

2. Studying performance of the proposed models using change in input modality.
3. Proposing a better performing model for bird tune classification.

### **1.3 The Challenges Regarding Bird Tune Identification**

“There are three stages to learn bird tune identification – (a) you cannot tell them apart; (b) you think you can tell them apart; (c) you realize you cannot tell them apart” - Anon

Nearly all bird species produce unique sounds that are used for communication between themselves as well as other species . The syrinx, an organ only found in birds is the main source of bird noises. This organ is complex in structure and function but the diversity of organs in birds is also great. Therefore, the range of different sounds that birds can produce is also very wide, which poses great challenges for the development of automatic sound classification systems. In addition, birds make incomplete, brief, or extended calls in a number of circumstances. For instance, during the breeding season, birds make brief cries because they are preoccupied with brooding or raising their eggs. The challenges of automatic classification of birds' voices from using data from different sources of acoustic recordings are influenced by the following things:

1. High interspecies and intraspecies variability of bird tune
2. Fine grain classification needed for some similar bird tune
3. Acoustic sensor noise patterns injected by acquisition device
4. External noise ( artificial and natural ) Overlapping bird voices

# Chapter 2

## Literature Review

According to [4] published in 2021, a DNN called BirdNET has been developed which is able to recognize 984 different bird species in North America and Europe based on sound. They evaluated the model with three distinct sets of data and discovered that extending domain-specific data is essential for creating models which are resistant to rising levels of background noise and can handle intersecting sound. They additionally discovered that the high temporal resolution of the input spectrograms (short length of the FFT window) increases efficiency of bird sound classification. To train a deep neural network, they used a workflow that started with the collection of a large amount of audio data, went through pre-processing to create pictorial interpretations of audio, augmented these illustrations, and then trained a sophisticated model architecture with about 27 million trainable variables. Independent validation and test splits of the obtained data were used for inference. In summary, for single species records, BirdNET generated an average accuracy of 0.791.

For situations where training data is limited and classes are uneven, another work [5] proposes a few-shot learning task where an algorithm must make predictions given only a few instances of each class. Comparing the effectiveness of several Siamese 13 Neural Networks at learning metrics for the set of Cassini's Vireo syllables, the network features were repeated. Tan2013 dataset has been used which has a total 1116 tokens divided across 64 classes and each class has a range of 1 to 73 tokens. Utilizing a Python package (praatio), recordings were segmented and annotated by people, using the corresponding pairs of audio files as inputs. The output of this approach is a set of labeled audio segments with a sample rate of 22,050 Hz and a resolution of 32 bits per segment. The Keras framework was used to test and develop four siamese networks with the following feature extractors: CNN, fully-connected (FCN), LSTM, and bidirectional-LSTM networks for  $k=[1,3,5,7]$ . A zero neural network and siamese  $k$  shot learners' performance were compared (Zero). With seven attempts, the bidirectional LSTM siamese networks' highest accuracy was 91.31%. The spike-based bird species recognition model, which addresses the process of distinguishing bird species based on vocalization, was introduced by Ricky, Bandi, and Sandeep in 2020 [6]. The vocalizations of 14 different bird species are included in the dataset. Prior to preprocessing, a minute-long recording of birds was separated into 12 identical frames of 5 seconds each using the technique of silence removal. Discrete wavelet transforms are used to remove noise from frames. Short Time Fourier Transforms was used to generate spectrograms for the segmentation of frames. Ap-

plying syllable segmentation, syllables were retrieved from this spectrogram based on the energy. Features were extracted and then the procedure of feature standardization was carried out. The classification and identification of bird species using the Spiking Neural Network with the Permutation Pair Frequency Matrix was done after that.

Brooker et al. stated [7] that ensemble approach, used with properly trained individual recognition systems, has the potential to finally open up ARU as a means of automatically recognizing the occurrence of target species and identifying patterns in vocal activity over time in acoustic environments. This approach resulted in an average of 74% of singing events being recognized across all five song types compared with 59 % . [7]. An additional paper proposed an unsupervised dictionary learning approach with a deep convolutional neural network architecture which outperformed the dictionary learning approach for bird sound classification [8] . The accuracy was 96.0%.

Zhao et al. described a new approach for automated field recording analysis in 2017 that included improved automated feature extraction and robust bird species categorization [9]. They used a Gaussian Mixture Model (GMM)-based frame selection with an event-energy-based sifting process that selected representative acoustic events and SVM algorithms for classification. Advanced robustness in real-world scenarios is achieved, as evidenced by a significant improvement in the F-score metric for "unknown" activities from 0.632 to 0.928. Because of the gain, the proposed method is more appropriate for automated field recording analysis.

Erika Vilches, Ivan A. Escobar, Edgar E. Vallejo, Charles E. Taylor explored the application of data mining techniques to the problem of acoustic recognition of bird species. They advised automatic bird species and individual recognition through acoustic records in conjunction with the existing technology of sensor networks. To perform data mining, cross validation techniques need to be applied. 70% of the total samples will be on the training part of the database and the rest will be on the testing part. Several data mining algorithms were explored during the creation of this project, and some of them were taken into consideration and used to the data obtained from the audio samples. The probabilistic classifier Naive-Bayes, the decision tree-based ID3 and J4.8, and vector quantization were chosen as the algorithms. For classification, to minimize the problem's dimensionality and to get rid of duplicate data sets, decision tree-based methods were used. The highest accurate algorithm was J4.8 (without Naïve-Bayes) obtaining 98.39% accuracy [10].

Again, one article introduces a novel method to effectively detect transient and in-harmonic bird sounds [11]. The detection algorithm consists of feature extraction by wavelet decomposition and detection using a supervised (MLP) or unsupervised (SOM) classifier. The SOM, a clustering algorithm, was used to examine the distinguishability of bird species and the sound data was classified using the MLP. After preprocessing the audio files are divided into small segments which are checked manually during post processing. Employing the wavelet packet decomposition, all of the sounds were broken down into wavelet coefficients (WPD). These wavelet coefficients were used to determine the features which are maximum energy, position, spread, and width. Then the feature vectors were put together. During the training phase, the MLP and SOM networks were provided with the training data's feature vectors. The recognition results were then analyzed after both networks had been tested on independent testing data. All stages of the recognition process were au-

omatic, with the exception of the manual sound-checking step. The results were reassuring: the SOM network recognized 78% and the MLP network 96% of the test sounds properly.

In 1998, Joseph and Daniel compared [12] the performance of two techniques for automated recognition of bird song units from continuous recordings. On a huge database of specific bird's song, the advantages and limits of dynamic temporal warping (DTW) and hidden Markov models (HMMs) are analyzed. Under difficult circumstances, such as noisy recordings or the presence of perplexing short-duration calls, the DTW-based technique requires careful template selection. As HMMs are trained, HMMs can achieve similar or even superior performance based solely on segmentation and classification of vocalizations. HMM provided better performance with many more training examples than DTW. The proposed system gives accuracy about 80% for five templates per class and to 92.5% for the larger set of templates. In 1997, Alex L. McIlraith and Howard C. Card presented backpropagation learning in two-layer perceptrons as well as several well-known methods from multivariate statistics for bird species recognition [13]. There are two different datasets. Each contained records made up of ten variables taken from either the 512- or 2048-sample window, as well as the song length. Higher order derivative information or backpropagation without momentum are used as the learning models in this research. An effective network includes ten inputs, twelve hidden nodes, and six outputs, as per cross validation experimentation. A learning rate of 0.2 is used. To enhance learning, target values of 0.0 and 1.0 were modified to 0.2 and 0.8. The distribution of songs between the training and test data sets was done randomly. In each run, 25% of the data was used for testing. For each type of dataset, 10 training and ten test sets are created. For both datasets, the ten test and ten training sets were generated in the same sequence to understand differences clearly. To determine the bird's identity, the species class with the highest count was deemed the winner. After the recognition system was trained on a dataset of recordings and on the computing needs of the algorithms, they demonstrated generalization accuracy. When backpropagation was used to learn spectrum data, it seemed to succeed between 82 and 93% of the time.

From the above discussion, we can infer that deep neural networks have the potential to be a feasible approach for bird tune classification but there are aberrant challenges that need to be subdued. Although there have been attempts made such as validation of generalizing performance to address the issues of overfitting, it also decreases a system's ability to adapt to unknown data. Along with this, the scarcity of appropriate methods for efficient and accurate extraction of interest signals, the vast bulk of data remains unexplored. So, it is important to validate generalization performance using a dataset that is different from the training dataset. To summarize, all of these shortcomings leave the door open to work in these aspects to look for models that will yield better accuracy and be well suited for classification of bird tunes.



# Chapter 3

## Proposed Method

### 3.1 Model Details

#### 3.1.1 Transfer Learning

Transfer learning is a technique that allows the reutilization of a pre-trained model's components in a fresh machine learning model. Fundamental knowledge can be shared between the two models if they were created to carry out comparable tasks. It starts with a model that has been trained on a bigger dataset. Second, a custom classifier with multiple layers of trainable variables is added to the model once the parameters in the lower convolutional layers are frozen. Then, classifier layers are trained on available training data and more layers are unfreezed through fine-tune hyperparameters. The advantages of transfer learning include resource savings and increasing efficiency while developing new models. Additionally, it can aid in model training when there are only unlabeled datasets available and instead of training in real-world situations, models can be trained in simulations [14].

#### 3.1.2 Learning Rate

Learning rate specifies how rapidly a neural network updates the information it has learnt. A desired learning rate is one that is both high enough to allow for quick training and low enough to ensure that the network converges to something useful. Due to the smaller weight changes made in each update, smaller learning rates require more training epochs (and therefore more training time), whereas larger learning rates provide quick changes and require fewer training epochs.

#### 3.1.3 Convolutional Neural Network (CNN)

Neural systems are numerical models that store information through the utilization of brain inspired learning components. Comparative to the brain, the neural systems are composed of a few neurons which are associated through various associations. In a variety of applications, neural networks have been used to simulate unknown relationships between various parameters using a huge number of samples [15]. There are mainly three layers of a neural network which is input, hidden and output layers. Data is passed to the input layer and distributed to the hidden layers accordingly. All calculations and conclusions are made here before they are sent to the output layer. Pictures were at first prepared in pixels employing a multi-layer

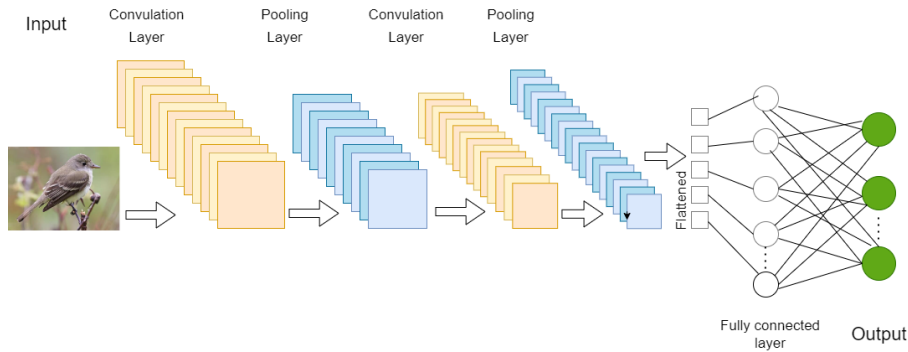


Figure 3.1: Convolutional Neural Network

design known as CNN. Traditional neural networks are not appropriate for image processing since the pictures must be entered as reduced-resolution bits [16]. For signal processing CNN can also be used. CNNs have their nodes, or neurons, organized more just like the frontal flap of the brain, which is mindful for visual stimuli processing in living creatures. This keeps the layers organized so that the whole picture may be handled at once instead of independently. The neurons in the network's layers are organized in three measurements, two of which are input dimensions, and the third of which is the activation volume.

### 3.1.4 Optimizer

An optimizer is a process or technique that modifies neural network properties like weights and learning rates. As a result, it helps in minimizing total loss and improves accuracy. As a deep learning model normally contains millions of parameters, choosing the best weights for the model is a difficult task. That's why, It emphasizes the necessity to select an optimization algorithm that is appropriate for any application. There are different types of optimizers, among them RMSprop and Adam optimizer have been used in this paper. RMSprop is a popular optimizer because it performs well with larger datasets. Additionally, RMS prop improves the AdaGrad optimizer by lowering the monotonically decreasing learning rate. The learning rate for RMSprop is 0.001 by default. Besides, The Adagrad and RMS prop algorithms' traits are both inherited by Adam optimizers. Therefore, compared to other optimization techniques, the method is easier to create, runs faster, uses less memory, and requires less tuning.

### 3.1.5 Activation Functions

The activation functions of artificial neural networks (ANNs) are crucial in defining the output of deep learning models. By controlling neurons an activation function chooses whether a neuron should be active or not. Each neuron in the network has a function which decides that neuron should be active or not based on the information it receives from the model. Activation function determines the Deep Learning models accuracy and efficiency. The activation function has a value that can be anything from 0 to 1 or -1 to 1.

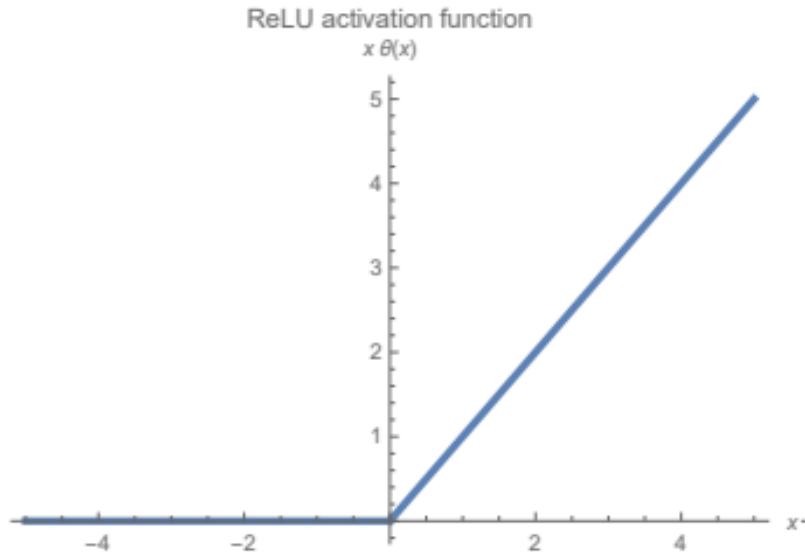


Figure 3.2: ReLU activation function [17]

There are different kinds of activation functions. For example, sigmoid used in the output layer of binary classification. Another activation function TanH which activates neurons depending on the output, Here if the output goes below zero that time the neurons are detached from the network. Another popular non-linear activation function is ReLU. Multiple layers of deep neural networks can be constructed using ReLU. ReLU is used in the hidden layer to deal with vanishing gradient problems. Another activation function, softmax is used in the final output layer.

In this work we have used ReLU and softmax activation functions. The ReLU equation gives the maximum possible output value between 0 and the input esteem utilized. Because of the negative input output shows 0. On the other hand if the input is positive then the output is equal to the input. As ReLU is layer compatible, it can be used as convolutional layers. Pooling layers [18] pool the outputs of convolutional layers to retain higher-level representations. Then the signal is usually sent to a fully connected layer for classification [19].

### 3.1.6 VGG19

There are different variants of VGG models like VGG11, VGG16 and others. Among them VGG19 is the most common model. It contains 19.6 billion FLOPs, which makes VGG19 more precise than other VGG models. As VGG19 is an advanced version of CNN it works with pre-trained layers. It has been trained on millions of diverse images. It is a deep CNN used to classify images. The model consists of 19 layers, of which 16 are convolutional layers, 3 are fully connected layers, 5 are MaxPool layers, and 1 is a SoftMax layer. The model uses small  $3 \times 3$  filters in all convolutional layers which helps to reduce parameters. In VGG19, when the filters are increased, convolution layers also increase. It has 6 main structures and each structure is composed of multiple connected convolutional layers. Model Filters size is  $3 \times 3$  and the input

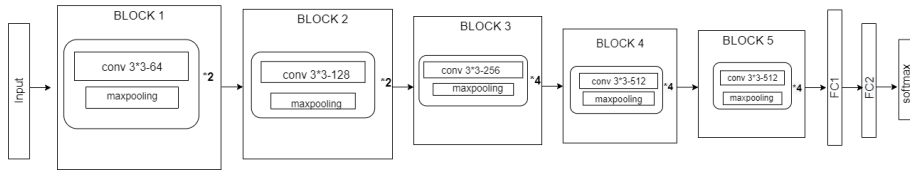


Figure 3.3: VGG19 architecture

size is  $224 \times 224 \times 3$ . This model is used as a pre-processing model. Using an alternating layer structure, which is better than a single layer, can improve the ability to extricate picture highlights and utilize max pooling for downsampling. The activation function ReLU selects the maximum value of the image region as the pooled esteem. To improve the anti-distortion ability of the image network, downsampling is used while preserving the main features of the sample [20].

### 3.1.7 Resnet-50

The convolutional neural network ResNet-50 consists of 50 deep layers. This architecture can be used for image classification, object localisation, object detection. It consists of 48 Convolution layers, 1 MaxPool and 1 Average Pool layer. It has high Floating point operations. As in this thesis we are dealing with a large number of datasets, we needed this type of model. This model enabled us to train very deep neural networks with over 150 layers. The ResNets were used to recognise images but they can be used for tasks other than computer vision to improve performance. ResNet50 architecture contains five stages. Each stage has a convolution and Identity block. Both blocks have 3 convolution layers each. Utilizing skip association resnet-50 includes the yield from an prior layer to a afterward layer, which makes a difference to relieve the vanishing gradient issue. Skip connection helps to skip over some layers of the model and because of this output changes [21] [22].

## 3.2 Work Plan

From Figure (3.5), We have a clear picture of the steps we need to perform in order to classify bird calls. Raw audio files are considered input data here. Spectrograms are generated from these audio files. The knowledge of transfer learning is used here. The whole dataset is divided between six continents and randomly splitted into train sets (75%) and test sets (25%). We will train our dataset based on 3 existing models i.e CNN, VGG19 and ResNet50. Based on training and testing accuracy, we have tried to identify the accurate model for our dataset.

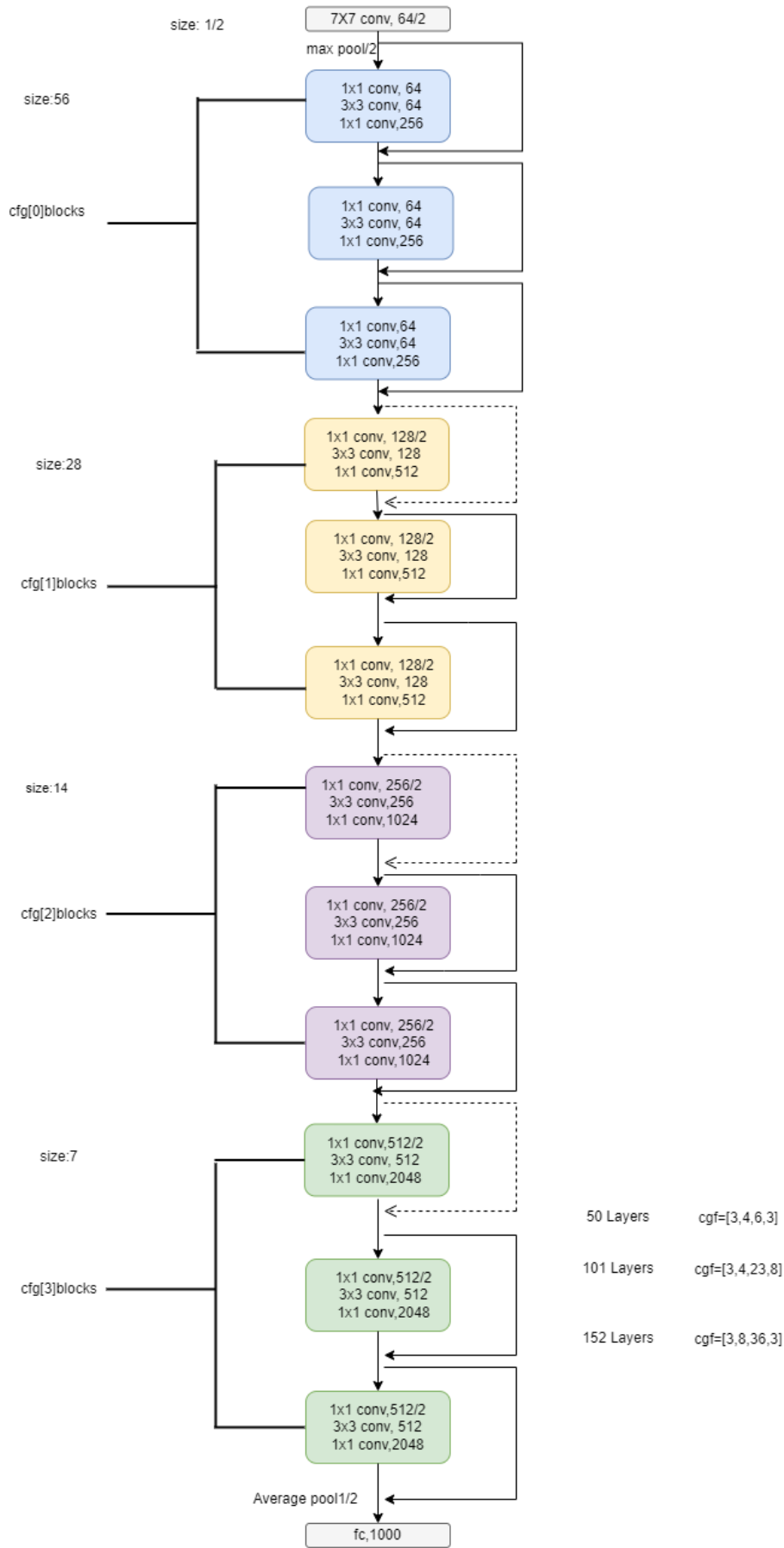


Figure 3.4: Resnet50 architecture

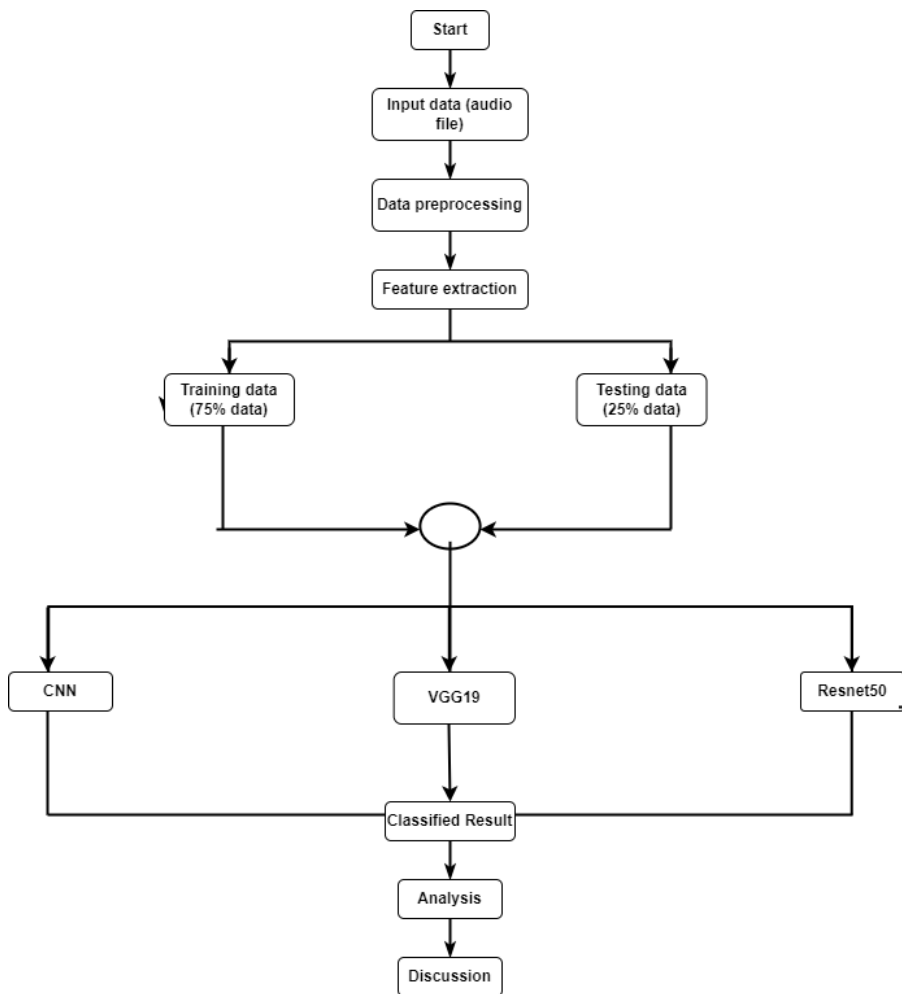


Figure 3.5: Flowchart of Implementation

# Chapter 4

## Dataset Description

### 4.1 Introduction

An online database called Xeno-canto (XC) gives users access to audio recordings of birds from all around the world. An expanding group of amateur and professional birdwatchers from all over the world, exchange the recordings. The goal of Xeno-canto is to reflect all bird sounds, which includes all taxa, down to the level of subspecies, their whole repertory, all of the geographic variations, and at all developmental phases. In 2005, Xeno-canto's website went live. It is managed by the Netherlands-based Xeno-canto Foundation for Nature Sounds, with assistance from the Naturalis Biodiversity Center.

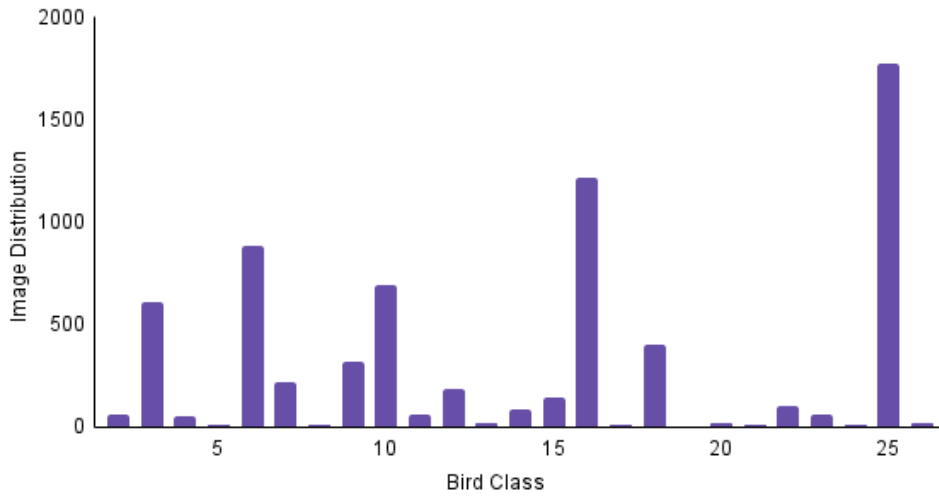
### 4.2 Dataset Details

The xeno canto dataset that we are using has 23,784 recordings from 259 different species. There were few recordings for 37 species; we eliminate them from the dataset. So, our final dataset contains 222 species and 23,646 recordings. The sample frequency of audio files varies from 16.00 kHz to 44.1 kHz.

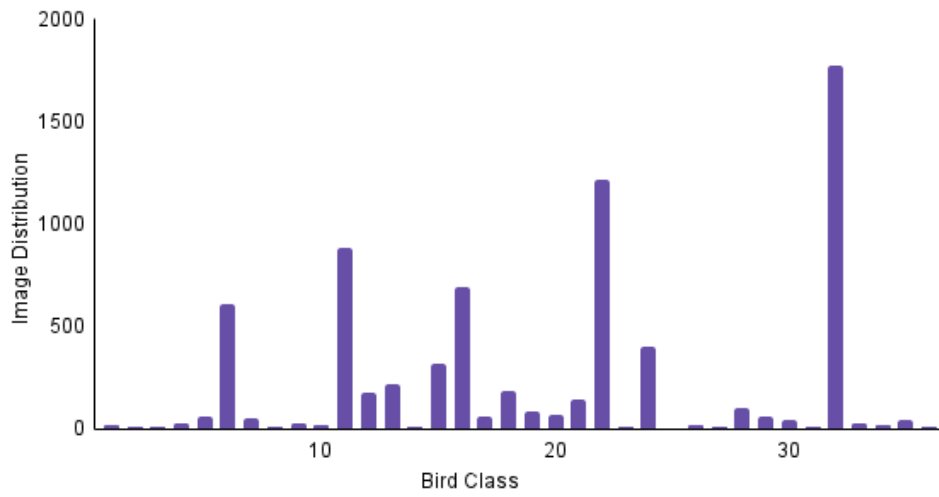
### 4.3 Data Preprocessing

The next step was processing the input audio file. All the audio files were converted into spectrogram visuals, with the frequency (vertical y-axis, Hz) changing over time (horizontal x-axis, sec). The data were then divided into 6 folders based on their continent i.e Africa, Asia, Europe, North America, Oceania and South America. The following figure(4.1) represents the uneven distribution of bird classes for each region. Due to the uneven nature, during supervised training any class with a greater sample count has more significance, and the classifier's output is biased towards the correlated class. Some data or species were repeated because they are found in more than one region.

Class Distribution for Africa



Class Distribution for Asia



Class Distribution for Europe





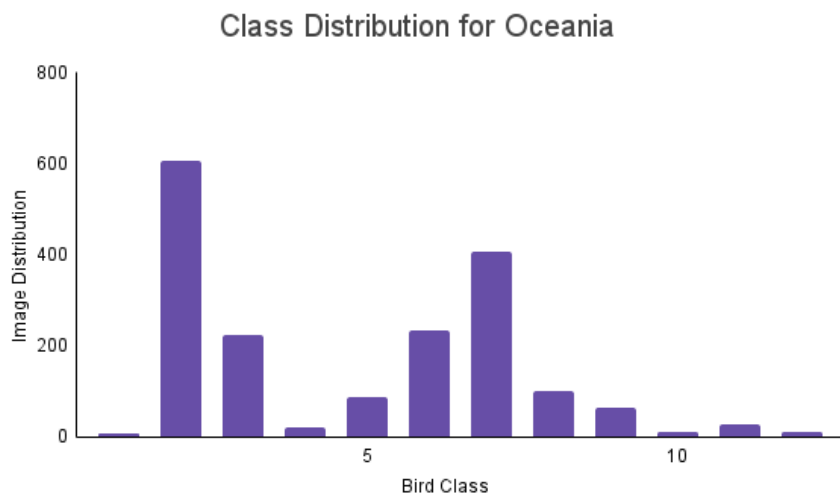
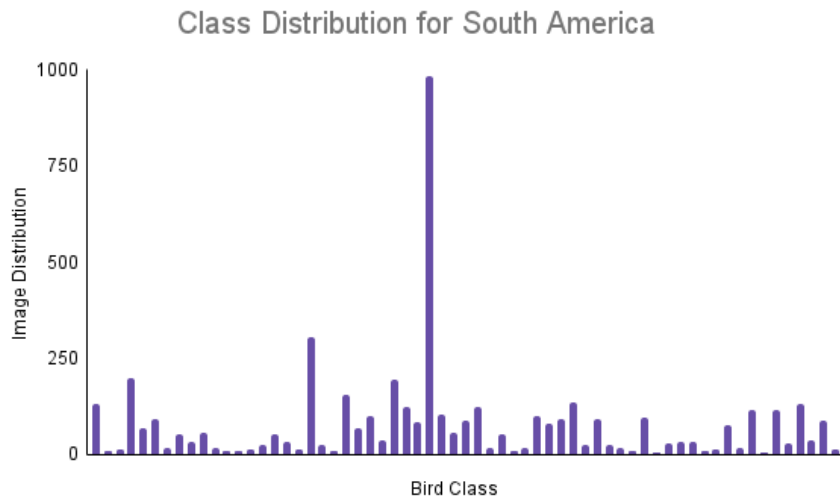
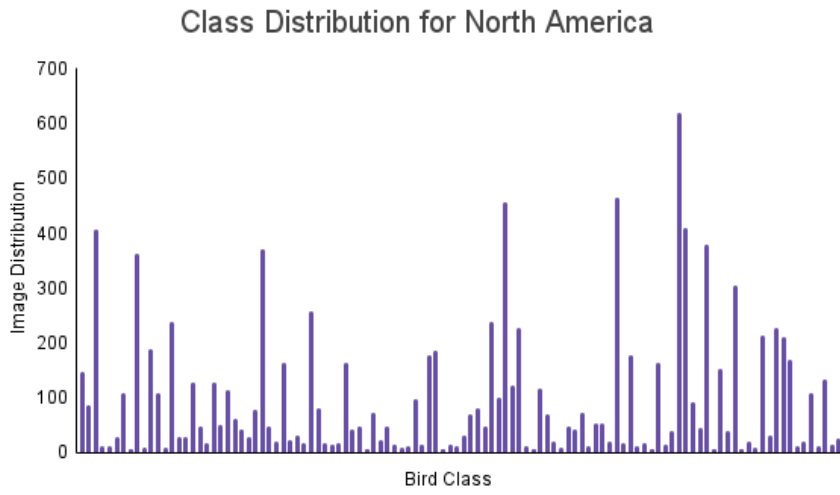


Figure 4.1: Continent wise species distribution

# Chapter 5

## Result and Analysis

### 5.1 Simple CNN

When it comes to identifying sound classification, deep learning is now significantly contributing. For deep learning, there is no need for manual or expert based feature selection. Instead, raw data can be used to create an efficient optimal representation that incorporates that data. The convolution neural network is distinct from the basic neural network in that it has 140 or more layers. This feature of CNN makes a difference to bargain with huge datasets and improves accuracy across multiple disciplines. We implemented The CNN model which was composed of a Sequential class. This class was stacked with three Conv2D network layers with a continuously rising number of filters such as 32, 64, and 128. These Conv2D layers contribute to the feature extraction process from input images. We kept the padding the same among all 3 layers and utilized the ReLU function in all Conv2D networks. There were 3 MaxPool2D layers with a pool size of 2, 2 between the Conv2D layers and for all the layers, we kept the stride size 2. After the Conv2D and MaxPool2D layer, we used a Flatten function to reduce the varying dimensional tensors to one dimensional. Additionally, a Dense function was applied with a softmax activation function and number of classes. The learning rate for the model was 0.001 and the loss function was categorical-crossentropy. Lastly, RMSprop optimizer was utilized so that we can increase our learning rate. The batch size is set to 4 and the image target size is set to 128x128. The model was then trained for 10 epochs and we got desired results.

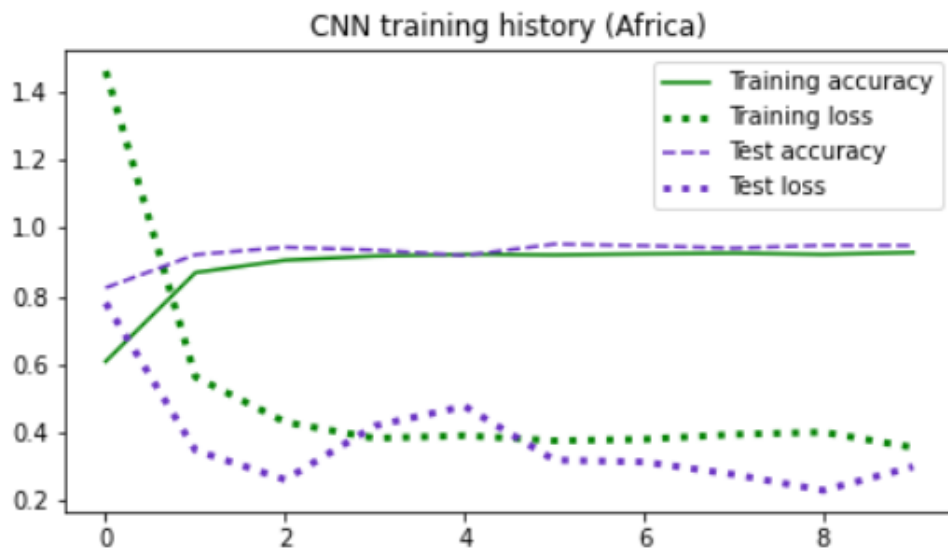


Figure 5.1: Accuracy and loss diagram for Africa(CNN)

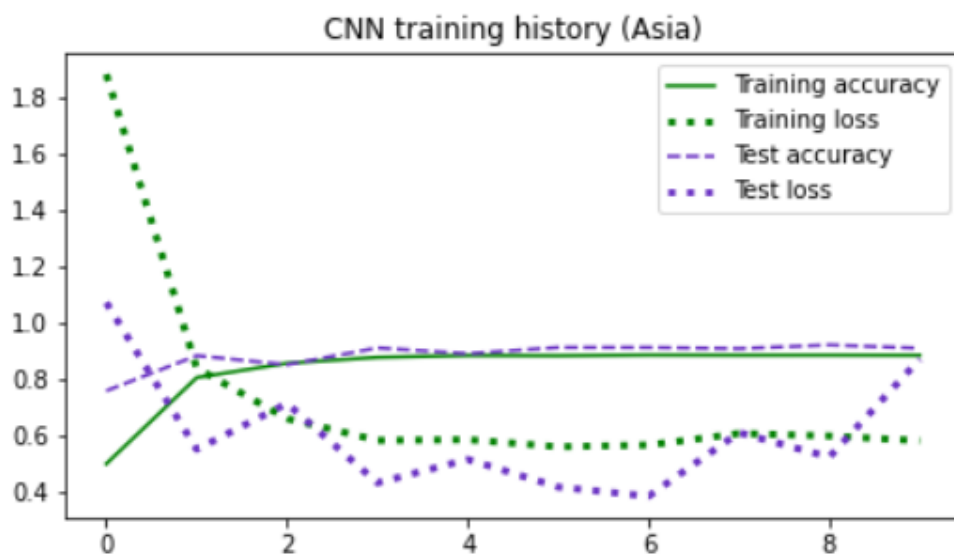


Figure 5.2: Accuracy and loss diagram for Asia(CNN)

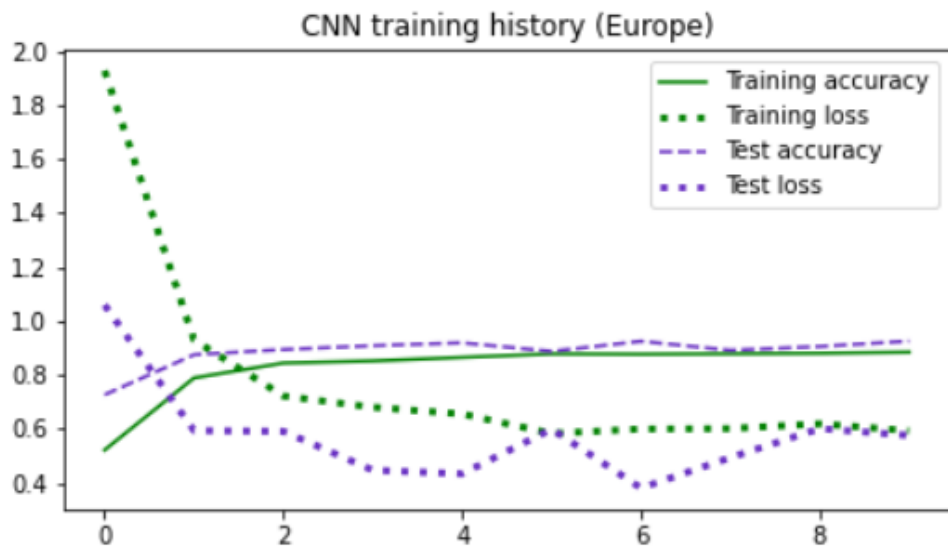


Figure 5.3: Accuracy and loss diagram for Europe(CNN)

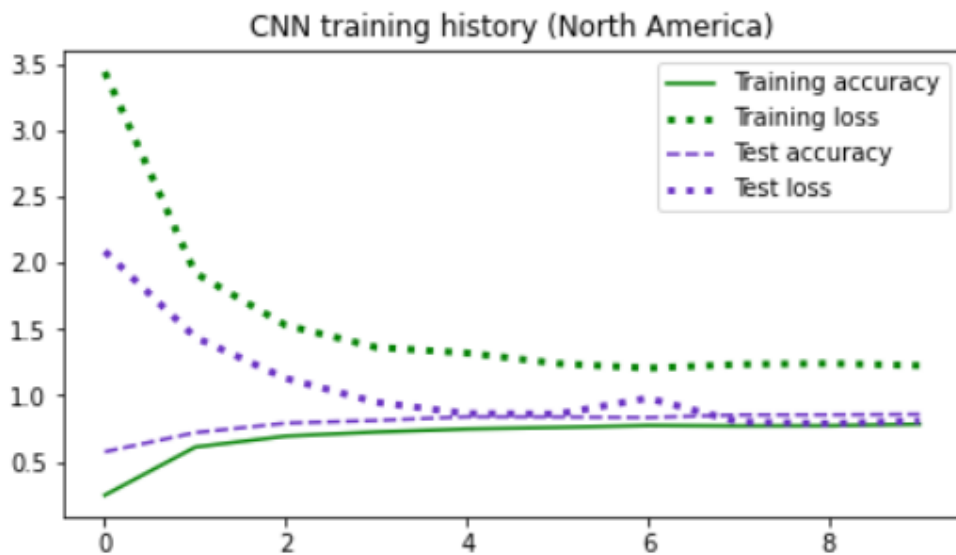


Figure 5.4: Accuracy and loss diagram for North America(CNN)

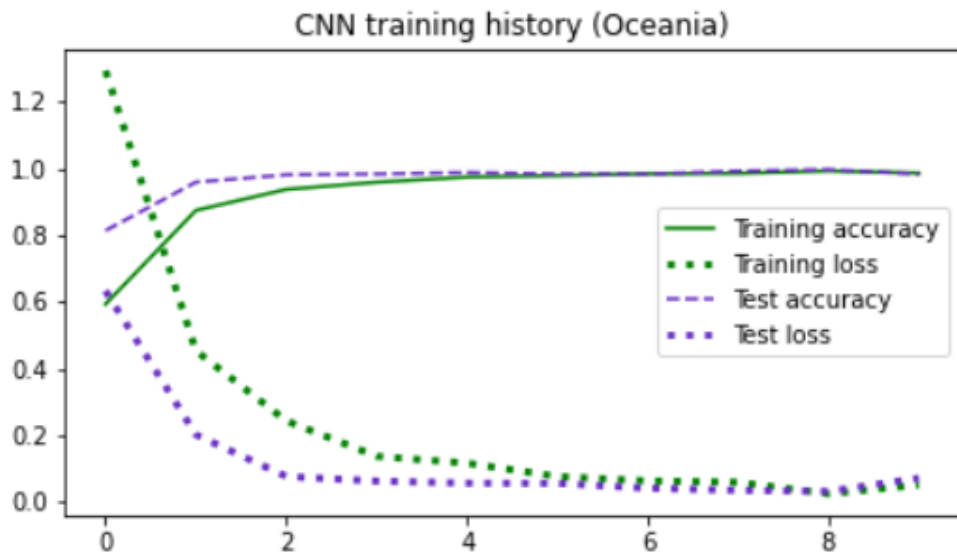


Figure 5.5: Accuracy and loss diagram for Oceania(CNN)

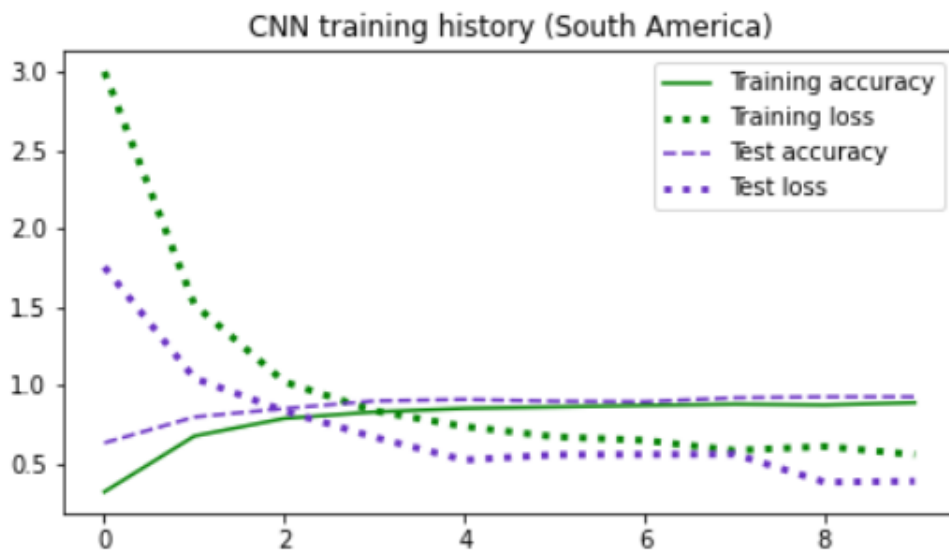


Figure 5.6: Accuracy and loss diagram for South America(CNN)

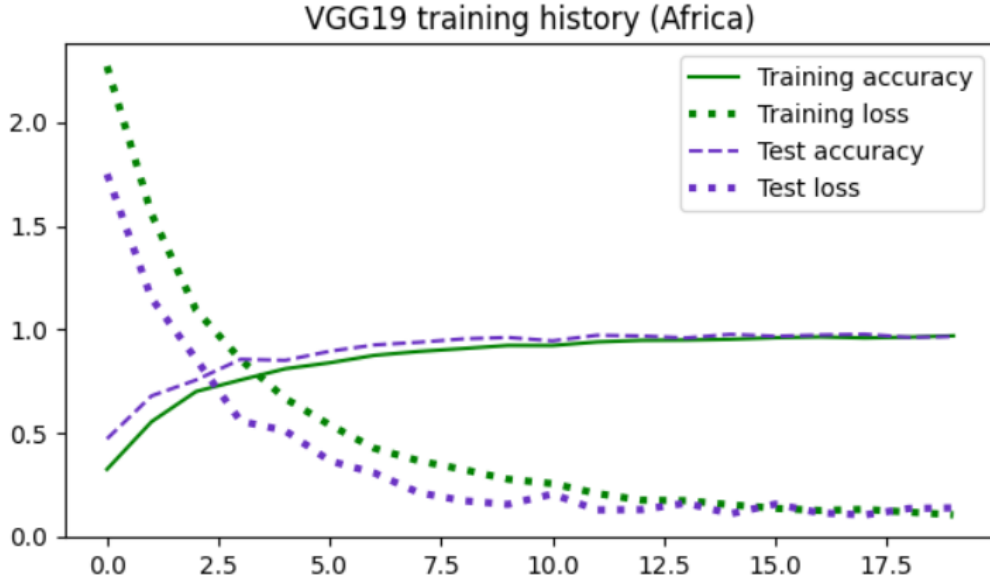


Figure 5.7: Accuracy and loss diagram for Africa(VGG19)

## 5.2 VGG19

For better depth of this paper, we have used another model VGG19 which is a variant of the VGG model. There is a VGG16 CNN architecture as well. Even though VGG19 requires more RAM than VGG16, it performs much better. We used VGG19 to make predictions more precise. In our work, we created a dense layer with unit size of the number of classes in the particular continent and the softmax activation function. We have set 0.00001 as the learning rate of the Adam optimizer and used categorical-crossentropy as the loss parameter. For this model, the batch size was different from region to region and the image size was set to 100 by 100. After that, the model was trained for 20 epochs. Figure 5.7 to 5.12 shows the training graphs of VGG.

## 5.3 ResNet-50

ResNet-50, a 50-layer convolutional neural network, has also been deployed. Since our data-set is fairly larger than typical datasets, we selected this model. We required a model that would enable us to train more than 150 layers of deep neural networks. In this work, we created a dense layer similar to CNN and VGG19 models. We again used the Adam optimizer technique with the same learning rate as VGG19 and the loss parameter was categorical-cross-entropy. The image resolution was set to 100 by 100 and The batch size varied from one region to the other. At last, we trained the model for 10 epochs. The given figures from 5.13 to 5.18 illustrates the training's.

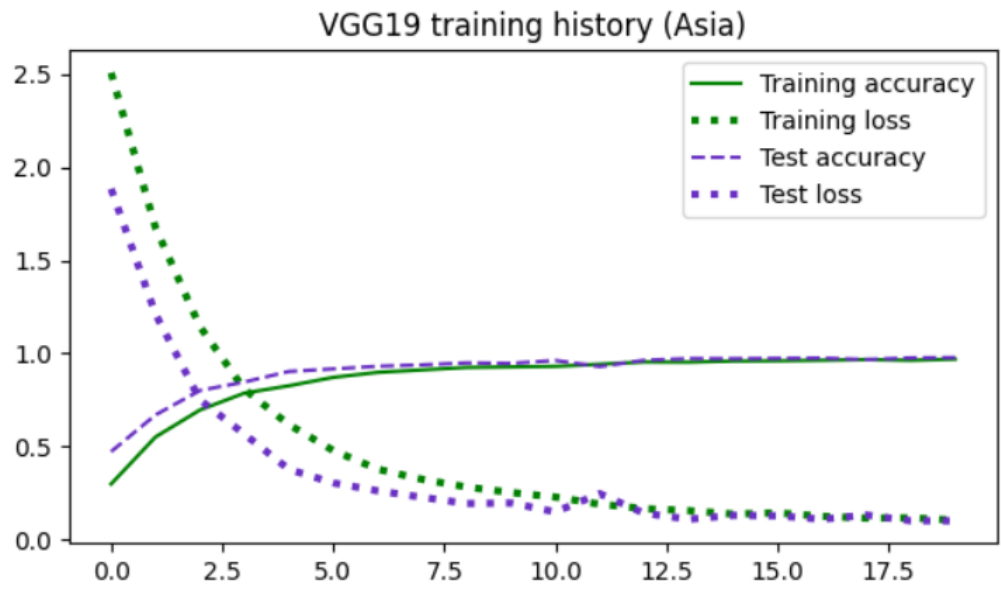


Figure 5.8: Accuracy and loss diagram for Asia(VGG19)

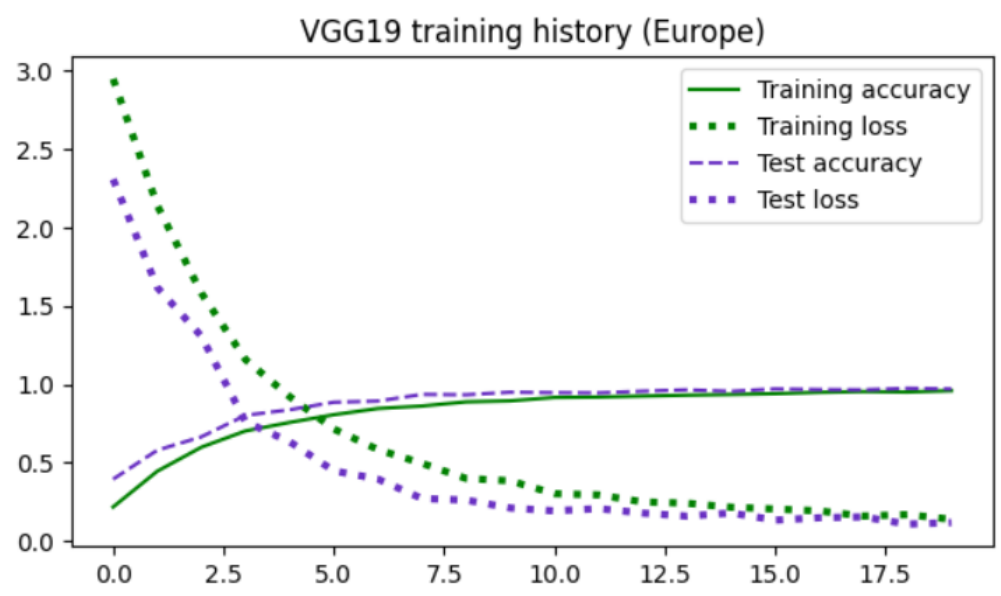


Figure 5.9: Accuracy and loss diagram for Europe(VGG19)

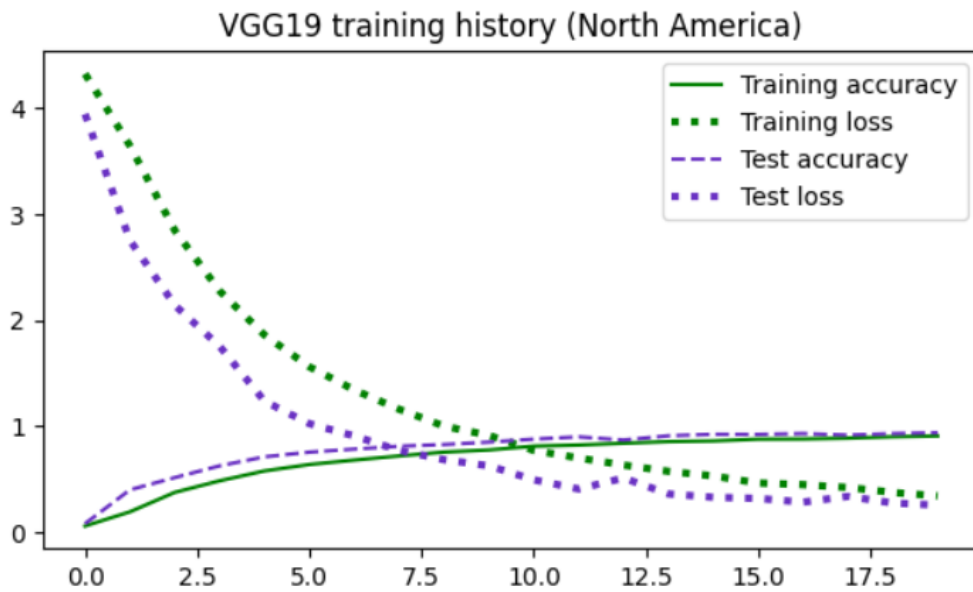


Figure 5.10: Accuracy and loss diagram for North America(VGG19)

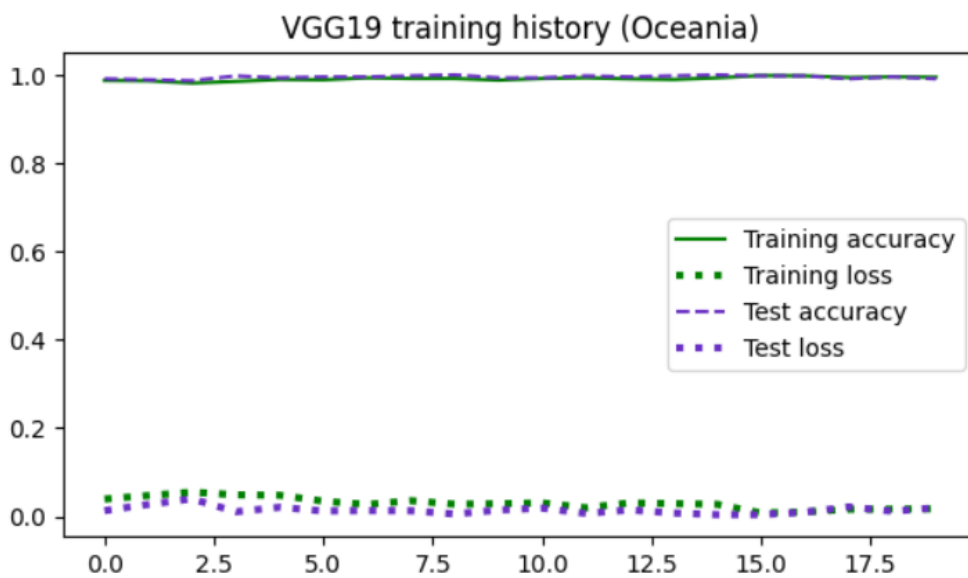


Figure 5.11: Accuracy and loss diagram for Oceania(VGG19)



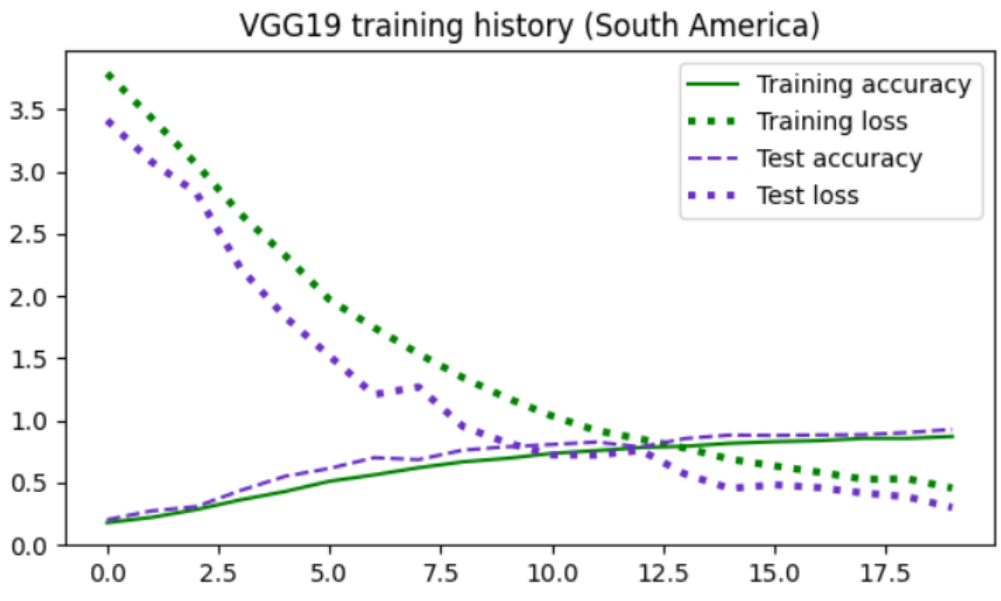


Figure 5.12: Accuracy and loss diagram for South America(VGG19)

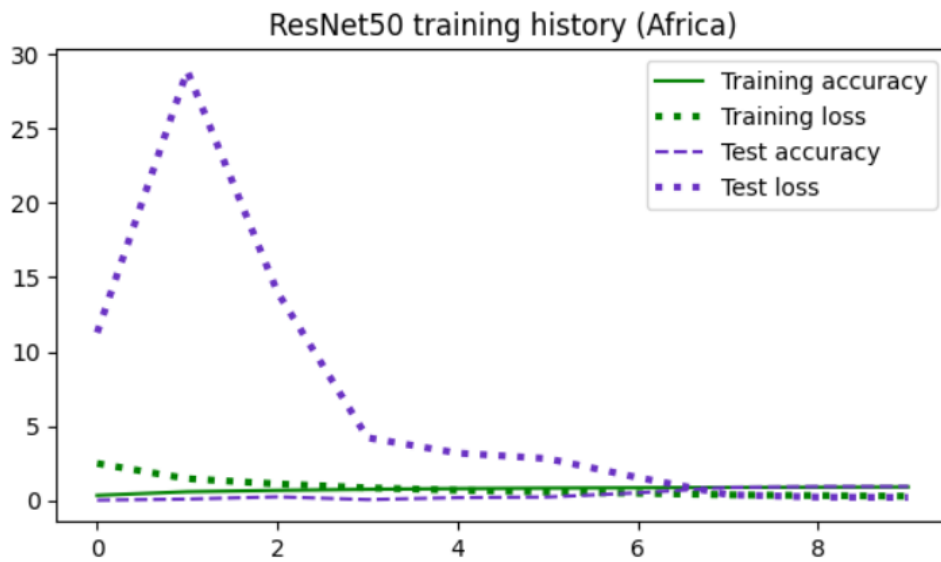


Figure 5.13: Accuracy and loss diagram for Africa(ResNet50)

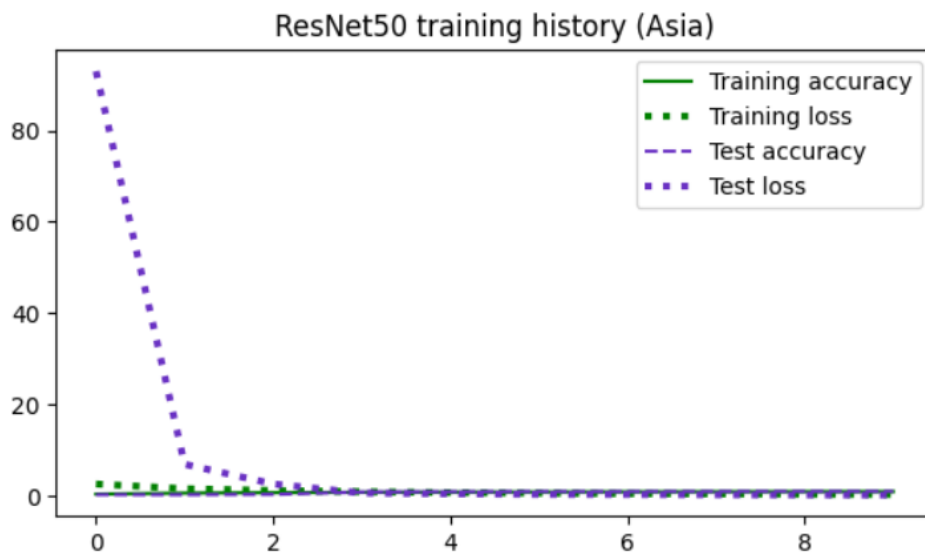


Figure 5.14: Accuracy and loss diagram for Asia(ResNet50)

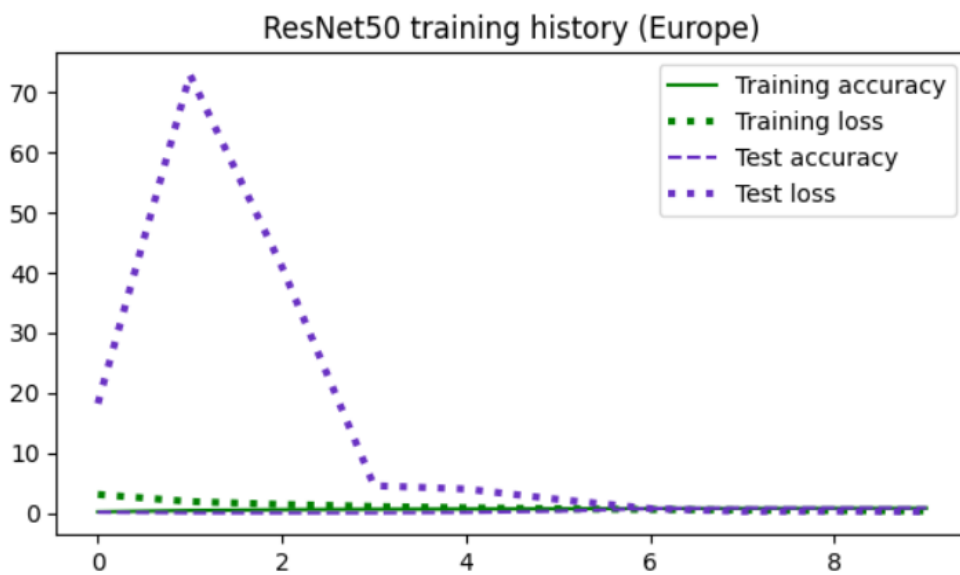


Figure 5.15: Accuracy and loss diagram for Europe(ResNet50)

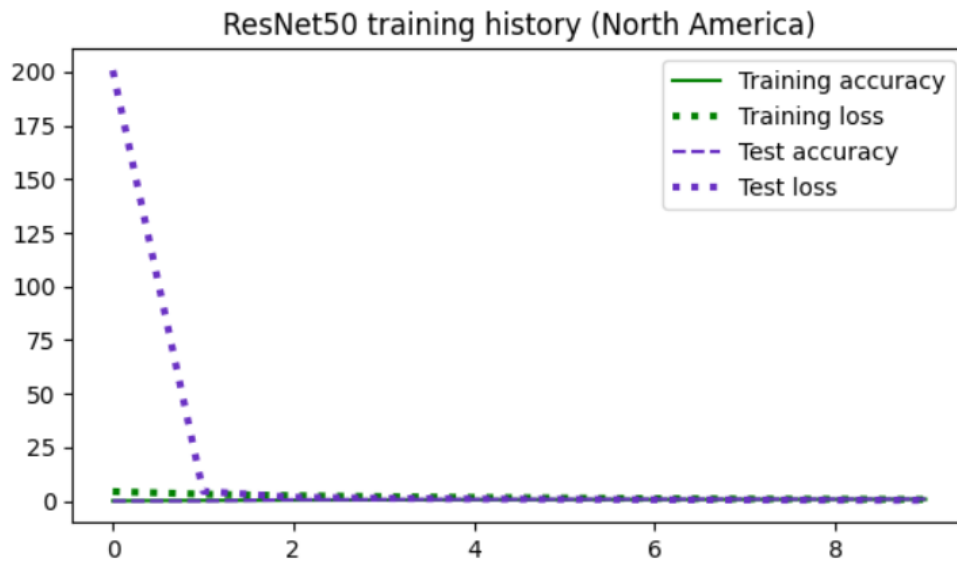


Figure 5.16: Accuracy and loss diagram for North America(ResNet50)

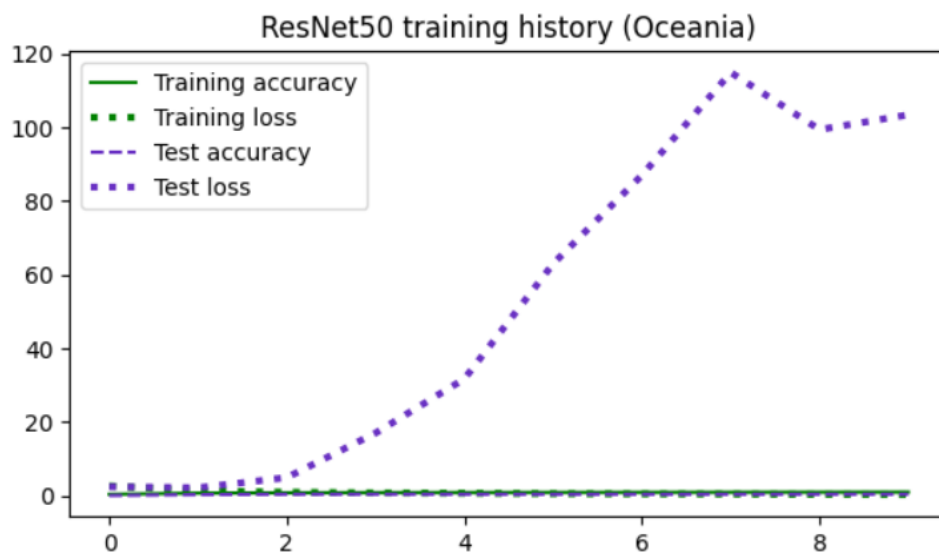


Figure 5.17: Accuracy and loss diagram for Oceania(ResNet50)

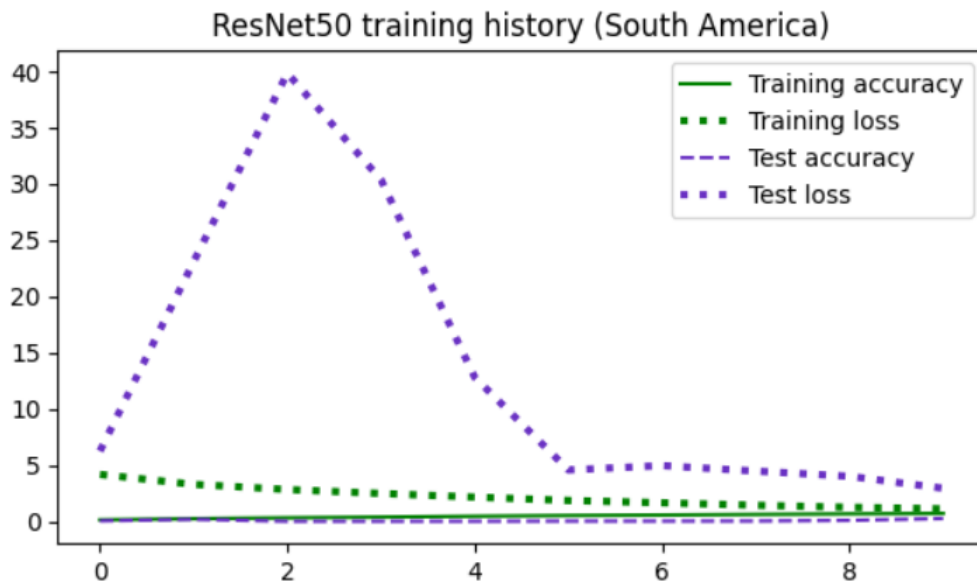


Figure 5.18: Accuracy and loss diagram for South America(ResNet50)

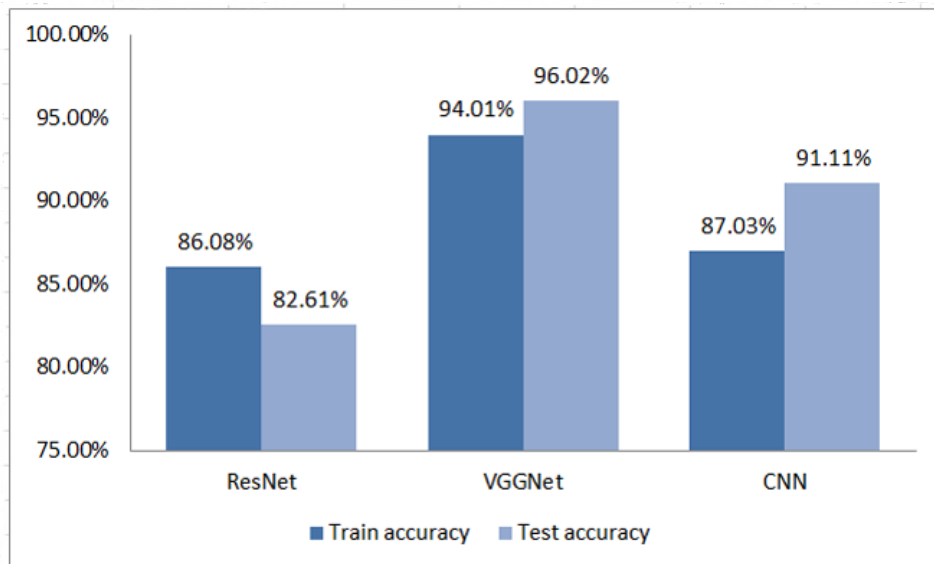


Figure 5.19: Model Accuracy graph

## Experimental Analysis

As we can see from the graphical representation, the test accuracy provided by VGGNet(96.02%), which is comparatively the highest point in this graph. The train accuracy in this instance is 94.01%. Next, CNN earns the second-highest score with test accuracy of 91.11%. With an accuracy of 87.03%, the train accuracy is quite low for this model. Last but not the least, Resnet50 offers test accuracy that is lower than train accuracy. The train accuracy is 86.08%, and accuracy in the test is 82.61%. The datasets were examined as zone-based structures. First of all, we have used a simple CNN model which is way lighter than VGG19 or ResNet50. Though CNN is lighter, it provides almost the same accuracy as the heavier model (VGG19). In comparison, VGG is producing a nice outcome even though it is heavy. In contrast, VGG is producing positive results despite its large weight. Resnet is a fairly complex model and it's hard to train this model. If we want to achieve decent accuracy then we need a higher amount of data-sets. Additionally, By increasing the number of epochs the performance of this model could be better. So, we discovered that the VGG model was superior to all the 3 models. The light weighted CNN is more affordable and simpler to deploy, however it does not provide the best accuracy in comparison. Here, we've demonstrated a respectable level of accuracy using less deploy-able data-sets in figure 5.19.

# Chapter 6

## 6.1 Conclusion

In ecology, birds are an important element as they play a vital role in our ecosystem. As for the food chain and biodiversity ecosystem, birds are not in good numbers. We are losing so many species of birds, the data is not easy to collect. In our research, we are working to correctly detect and classify tunes of birds from all over the world in a noise agnostic and generalized way. This work will also pave the way to do ornithology research using readily available sound acquisition devices as we are training our models invariant to acquisition devices and also noises. So less sensitive microphones can also contribute to the existing knowledge base and even crowd sourcing of datasets will be possible through this research, which was not possible previously due to dependence on sophisticated recording devices currently used. This work can also contribute to the field of ‘machine listening’ research in general and these methodologies and techniques can also be implemented for similar tasks such as animal call classification, acoustic scene classification etc [23][24][25][26][27].

## 6.2 Limitations

Bird call identification has always been a time consuming operation which has to deal with a lot of noisy data. Though there is a lot of dataset available but most of them are not downloadable. It is also seen that audio dataset which are downloadable are very noisy or poor in quality. We encountered a lot of difficulties throughout this time. First off, being a heavy model, the implementation of VGG and Resnet models are time and resource consuming. So, we could not implement the models by changing the ratio of train, test data. Secondly, the limitations with devices are also a responsible caution to work with. As all the training was done on Spectrogram, there were hardware limitations, which made it take much longer than expected.

## 6.3 Future Work

Planning and preparing for the future has long been thought to be a human influence, but at least one bird species has been discovered to do so as well. The discovery also offers the intriguing notion that, like us, birds experience future anxiety. We can make an application in the future to reserve the audios of several uncommon bird species. As the number of species declines, it will be easier to introduce birds to future generations. People can utilize the data to change their studies in other

areas. We are unable to apply our study since many species of birds have been lost. As a result, others may not suffer the same problem in the future.

Future research will expand on this information for not only acoustic settings but also occurrences. By using the same guidelines for recording and annotating, other groups are encouraged to contribute to the dataset.

# Bibliography

- [1] Y. Ganin, E. Ustinova, H. Ajakan, *et al.*, “Domain-adversarial training of neural networks,” *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.
- [2] C. K. Catchpole and P. J. Slater, *Bird song: biological themes and variations*. Cambridge university press, 2003.
- [3] R. A. Fuller, J. Tratalos, and K. J. Gaston, “How many birds are there in a city of half a million people?” *Diversity and Distributions*, vol. 15, no. 2, pp. 328–337, 2009.
- [4] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, “Birdnet: A deep learning solution for avian diversity monitoring,” *Ecological Informatics*, vol. 61, p. 101 236, 2021.
- [5] S. Renteria, E. E. Vallejo, and C. E. Taylor, “Birdsong phrase verification and classification using siamese neural networks,” *bioRxiv*, 2021.
- [6] R. Mohanty, B. K. Mallik, and S. S. Solanki, “Recognition of bird species based on spike model using bird dataset,” *Data in brief*, vol. 29, p. 105 301, 2020.
- [7] S. A. Brooker, P. A. Stephens, M. J. Whittingham, and S. G. Willis, “Automated detection and classification of birdsong: An ensemble approach,” *Ecological Indicators*, vol. 117, p. 106 609, 2020.
- [8] J. Salamon, J. P. Bello, A. Farnsworth, and S. Kelling, “Fusing shallow and deep learning for bioacoustic bird species classification,” in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2017, pp. 141–145.
- [9] Z. Zhao, S.-h. Zhang, Z.-y. Xu, *et al.*, “Automated bird acoustic event detection and robust species classification,” *Ecological Informatics*, vol. 39, pp. 99–108, 2017.
- [10] E. Vilches, I. A. Escobar, E. E. Vallejo, and C. E. Taylor, “Data mining applied to acoustic bird species recognition,” in *18th International Conference on Pattern Recognition (ICPR’06)*, IEEE, vol. 3, 2006, pp. 400–403.
- [11] A. Selin, J. Turunen, and J. T. Tantt, “Wavelets in recognition of bird sounds,” *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1–9, 2006.
- [12] J. A. Kogan and D. Margoliash, “Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden markov models: A comparative study,” *The Journal of the Acoustical Society of America*, vol. 103, no. 4, pp. 2185–2196, 1998.



- [13] A. L. McIlraith and H. C. Card, “Birdsong recognition using backpropagation and multivariate statistics,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2740–2748, 1997.
- [14] M. Hussain, J. J. Bird, and D. R. Faria, “A study on cnn transfer learning for image classification,” in *UK Workshop on computational Intelligence*, Springer, 2018, pp. 191–202.
- [15] D. Yu, W. Xiong, J. Droppo, *et al.*, “Deep convolutional neural networks with layer-wise context expansion and attention.,” in *Interspeech*, 2016, pp. 17–21.
- [16] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.
- [17] M. Straat, “On-line learning in neural networks with relu activations,” Ph.D. dissertation, 2018.
- [18] J. Brownlee, “A gentle introduction to pooling layers for convolutional neural networks,” *Machine learning mastery*, vol. 22, 2019.
- [19] R. Tharsanee, R. Soundariya, A. S. Kumar, M. Karthiga, and S. Sountharajan, “Deep convolutional neural network-based image classification for covid-19 diagnosis,” in *Data Science for COVID-19*, Elsevier, 2021, pp. 117–145.
- [20] J. Xiao, J. Wang, S. Cao, and B. Li, “Application of a novel and improved vgg-19 network in the detection of workers wearing masks,” in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1518, 2020, p. 012 041.
- [21] S. Mascarenhas and M. Agarwal, “A comparison between vgg16, vgg19 and resnet50 architecture frameworks for image classification,” in *2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON)*, IEEE, vol. 1, 2021, pp. 96–99.
- [22] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [23] J. Xie, M. Towsey, J. Zhang, and P. Roe, “Frog call classification: A survey,” *Artificial Intelligence Review*, vol. 49, no. 3, pp. 375–391, 2018.
- [24] E. C. Garland, M. Castellote, and C. L. Berchok, “Beluga whale (*delphinapterus leucas*) vocalizations and call classification from the eastern beaufort sea population,” *The Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. 3054–3067, 2015.
- [25] V. Demartsev, N. Gordon, A. Barocas, *et al.*, “The “law of brevity” in animal communication: Sex-specific signaling optimization is determined by call amplitude rather than duration,” *Evolution letters*, vol. 3, no. 6, pp. 623–634, 2019.
- [26] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, “Acoustic scene classification: Classifying environments from the sounds they produce,” *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.
- [27] A. Mesaros, T. Heittola, and T. Virtanen, “Tut database for acoustic scene classification and sound event detection,” in *2016 24th European Signal Processing Conference (EUSIPCO)*, IEEE, 2016, pp. 1128–1132.