UNIVERSITY OF CAPE TOWN

FACULTY OF SCIENCE

DEPARTMENT OF STATISTICAL SCIENCES

MASTERS HALF DISSERTATION

Bioacoustic classification of Hainan Gibbon call types using Deep Learning

Author:
Nonhlanhla L. Luphade

Supervisor:
A/Prof Ian Durbach
Co-Supervisor:
Dr Emmanuel Dufourq
Co-Supervisor:
Stefan Britz

Centre for Statistics in Ecology, the Environment, and Conservation

September 20, 2023

Abstract

In Bawangling National Nature Reserve (BNNR), Hainan, China, there exists a critically endangered primate known as the Hainan gibbon *Nomascus hainanus*. Many species, including the Hainan gibbon, are at high risk of extinction due to many factors such as unsustainable hunting, climate change, and deforestation. The Hainan gibbons live in social groups and the ability to discriminate between the group is useful for tracking migration patterns, population management, and identification of new groups. Currently, there has not been any study which attempts to distinguish between the groups. More recently, researchers have begun using deep learning to answer ecological questions, in a similar way that deep learning has successfully been used in computer vision and audio classification tasks. This study is the first attempt at investigating how deep learning can be used to distinguish between the Hainan gibbon social groups using only the acoustic data recorded in BNNR. Two convolutional neural networks (CNNs) were developed, the first was a binary classification model to detect gibbon calls from non-gibbon calls, and the second was a group classifier to distinguish between the social groups in BNNR. The audio data was converted into mel-scale spectrograms, resulting in images used as input to train the CNNs. Two steps were taken to train reliable models. Firstly, data augmentation techniques were explored to increase the amount of data as a means to train reliable models, and secondly, hyperparameter tuning was conducted. The binary classifier obtained a testing accuracy of 86%. The findings reveal that the model is able to distinguish between gibbon calls and non-gibbon calls. The social group model was not able to distinguish between the social groups as the model predicted the majority of the calls as one group. The result of this study demonstrates the usefulness of deep learning in addressing ecological questions that would be otherwise very challenging for a human to achieve.

Acknowledgements

I would like to express my gratitude to my supervisors, Associate Professor Ian Durbach, Dr Emmanuel Dufourq, and Stefan Britz, for their continued encouragement, support, and guidance throughout this journey. Being given an opportunity to do this project has exposed me to many aspects of data science which I was not aware of and for that, I am truly grateful.

I would like to acknowledge Heidi Ma and Samuel Turvey from Institute of Zoology, Zoological Society of London who provided us with the recordings and information about the current state of the Hainan gibbons. Also, I would like to acknowledge the field team in Hainan for collecting the recordings. Furthermore, I am grateful to the centre for Statistics in Ecological, Environmental and Conservation (SEEC) for the exposure and education I received with regard to the application of statistics to ecological tasks.

I would also like to thank my family and friends for their love and support throughout the process. A special thanks to my mother for being my pillar strength, my father for encouraging me, Tavonga Mandava for believing in me, Thapelo Nyathi for cheering me up with food, Gareth Ndlovu for motivating me, Sandisile Moyo for being my greatest cheerleader, and Joregina Mthembu for consistently checking up on me. Above all else, I would like to thank the Almighty Father, who makes all things work out for our own good.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Ecological research provides a means to understand the relationships between organisms and their environment. This research enables ecologists to assist in providing effective monitoring and conservation strategies for most species. Conservation of species is of high importance globally as shown by its inclusion in the United Nations Millennium Development Goals, the sustainable development goals, and also the African Union's Agenda 2063. Many species are at high risk of extinction due to many factors such as unsustainable hunting, over fishing, deforestation and global climate change [Almond et al., 2020]. Thus, there is need for the development of better conservation strategies.
In Bawangling National Nature Reserve (BNNR), Hainan, China, there exists a critically endangered ape species known as the Hainan gibbon (*N. Hainanus*). The Hainan gibbon is listed as one of the world's 25 most endangered species since 2002 [Zhang et al., 2010]. Additionally, according to the International Union for Conservation of Nature Red List, the Hainan gibbon is classified as class one (highest protected species in China [Koh et al., 2021]).
In the 1950s, the Hainan gibbons were widespread across Hainan with an estimated population size of 2000. However, between the 1950s and the 1980s, the species' population experienced a drastic decline due to hunting for traditional Chinese medicine and deforestation which led to the decline in the condition of the Hainan forest. The deforestation was linked to industrialisation which resulted in loss of habitat for the Hainan gibbons to cater for the human population [Liu et al., 1984]. When Zhezhe Liu and his colleagues began field work on the species in the early 1980s, the global population was estimated to be between 30 and 40 with less than seven individuals in BNNR. From 2003 onwards there was no Hainan gibbon population found outside BNNR [Fellowes et al., 2008]. As of 2022, there is a single population of 35 Hainan gibbons in BNNR consisting of only five social groups, these groups are labelled as A, B, C, D, and E [Guo et al., 2020b; Li et al., 2022]. Group A, B, and C are characterized as normal social units as they consist of two or three adult males, two reproducing female adults and offspring while group D consists of one adult male, two adult females and offspring, and group E consist of one adult male, one adult female, and a

young infant [Li et al., 2022].

The Hainan gibbons are known as tropical dwellers, living exclusively in trees [Zhang et al., 2010]. Gibbons are colloquially known as the "Adele" of the animal world due to their distinct vocal behaviour which is species-specific, meaning each gibbon species has a distinct vocal behaviour. Hainan gibbons often sing at dawn to maintain territorial boundaries, advertise to potential mates, and to enhance family bonds [Deng et al., 2014; Dufourq et al., 2021]. The gibbon songs consist of short individual syllables or notes of ca. 0.2-2.75s assembled together into longer phrases consisting of one to six notes [Dufourq et al., 2021]. These notes organised together become songs. The songs produced by the adult males consist of mainly one to three short notes and one to five long notes while the solo adult male song consist of only long notes. Additionally, there is the chorus which is sang by both males and females which consists of long notes sang by the males at the beginning and short "notes" sang by the females at the end. Finally, there is a duet which is sang by a male and females. Most of the songs are male dominant and there are no female solos. [Deng et al., 2014].

Given the perilous state of this rare species, many researchers are working together to come up with effective methods of monitoring. Over the decades, local and international field researchers have gathered data on Hainan gibbons. This data provides a number of key factors on Hainan gibbons such as their biology, behaviour, and ecology [Turvey et al., 2015].

In 2014, a team of researchers in charge of conservation of Hainan gibbons held a workshop to discuss the future of Hainan gibbons. They decided that it would be best to divide the threats into three major groups, namely, Gibbon Group Formation Subgroup, Habitat Availability and Connectivity Subgroup, and Catastrophic Decline Subgroup [Turvey et al., 2015]. The gibbon group formation subgroup was created to deal with the problem of having low rates of group formation caused by lack of suitable habitat to disperse to and the gibbon not being able to find suitable mates from the new groups formed. The habitat availability and connectivity subgroup was created to deal with the low survivorship and habitat connectivity problem caused by the lack of understanding of survivorship of dispersing individuals and lack of suitable quality habitat for new group formation. The catastrophic decline subgroup was created to look into potential future catastrophic declines of the Hainan gibbon population [Turvey et al., 2015]. To address the problems faced by the first two group actions such as improving monitoring of individuals in all social groups, conducting playback experiments in unoccupied forest which entails playing a recording of the target species in the area of interest with hopes of obtaining a response from undetected individuals [Bryant et al., 2016], and employment of new acoustic technologies to support monitoring efforts were proposed [Turvey et al., 2015].

Traditionally, the gibbons have been monitored using acoustic monitoring whereby a team of people would survey the forest and listen for gibbon songs. However, this was dangerous, labour intensive, time consuming and unsustainable. In 2016, the group in charge of employing new acoustic technologies to support monitoring efforts decided to use novel methods such as passive acoustic monitoring [Dufourq et al., 2021]. This was achieved by placing acoustic recorders at specific ge-

ographic locations to record the gibbon calls. A subset of 32 hours of recordings were manually annotated by inspecting spectrograms used to visualise sound as images and by listening to the audio recordings. An automated acoustic classifier that detects gibbon calls using deep learning methods such as convolutional neural networks was developed using these annotated recordings, and used to label the remainder of the recordings [Dufourq et al., 2021]. The developed classifier serves as road map for other researchers to develop their own classifiers or extend the already existing classifier which all contributes towards the improvement of monitoring and conservation of calling species. The primary objective of this dissertation is to develop a classifier to distinguish between the different Hainan gibbon social groups that exist in BNNR. To achieve this objective, we also develop a binary classifier to detect presence (from any social group) and absence of Hainan gibbon calls.

1.2 Problem Statement

Currently, the existing classifier (the binary classifier develop in [Dufourq et al., 2021]) is capable of identifying gibbon calls and non-gibbon calls, and the manual labelling process contains additional information such as from which social group the audio recordings came from. The same data can be used to further extend the already existing classifier to develop a social group classifier. The aim of the developing a classifier for discriminating between groups is to be able to know whether or not different groups can be identified by their calls. Currently this is not known, but as calls are at least partially made to advertise territory, it might be expected that these will be group-specific. Being able to identify which group makes a call would allow researchers to track group movements such as migration to a new territory or to another group's territory. This is important for managing a small population. Additionally, being able to identify which group makes a call would allow researchers to identify any new groups that form.

1.3 Research Purpose and Objectives

The goal of this dissertation is develop a classifier which is able to detect from which social group each call comes from. This will be achieved through the following objectives:

1. Develop a classifier which is able to identify gibbon calls.
2. Develop a classifier which is able to identify from which social group each call comes from.

1.4 Outline of the thesis

This dissertation consists of six chapters. Chapter 2 presents the literature behind bioacoustic processes. Chapter 3 provides the theory behind convolutional neural networks. Chapter 4 provides a description of the data, preparation of the data and the pre-processing methods. Additionally, chapter 4 details the implementation of the models, training, and testing of the final models.

Chapter 5 presents and provides a discussion of the obtained results, with chapter 6 providing the overarching conclusions of the study and suggestions for future work.

Chapter 2

Sound and Ecology

The aim of this research project is to develop classifiers capable of identifying gibbon calls. Consequently, in this chapter, the use of sound in ecology through bioacoustics is introduced. Bioacoustics is a multidisciplinary branch of the sciences, bridging biological and acoustic sciences. Bioacoustics plays an important role in ecology as a tool to study and monitor animal diversity, abundance, behavior, dynamics, distribution, and their relationship with the ecosystem and the environment. Furthermore, the chapter describes how sound is analyzed by using audio feature extraction and spectrograms for sound visualization.

2.1 Animal Communication

Animal communication is the transmission of information or signals from one species to another species, which can be between the same species or different species [Browning et al., 2017]. The signals transmitted carry emotional, physiological, and individual information about animal species [Garcia and Favaro, 2017]. The nature of communication is determined by factors such as behavioral context, social systems, the environment, the nature of the animal species, and the ability to produce, receive and process a given message [Garcia and Favaro, 2017]. Animals can communicate in different ways: acoustic, visual, chemical, electrical, and tactile. The method used in a particular situation is selected by the animal based on the type of information, the environment the signal has to go through, the species abilities, and the biological significance. The study of animal communication is an inter-disciplinary science that covers a wide range of biological sciences. This chapter will review literature related to animal acoustic communications (bioacoustics).

Most animals including birds, crustaceans, arachnids, mammals, amphibians, reptiles, insects, and some fish produce species-specific sound signals daily which are used for many purposes such as communication, echolocation, sexual display, and territorial defense [Vehrencamp and Bradbury, 1998]. Sound emitting animals produce a wide range of distinct sounds which can be categorized by researchers to infer animal distribution, physiological state, abundance, and behavioral characteristics. This makes bioacoustics the perfect tool to monitor biodiversity. Bioacoustics is an

interdisciplinary field which combines biological and acoustic sciences by using sound technologies such as microphones and hydrophones to record, store and analyze large amounts of animal acoustic data [Penar et al., 2020; Mcloughlin et al., 2019]. Historically the interest in the bioacoustics field has been mainly focused on attempts to try [Garcia and Favaro, 2017]:

1. Classify, describe, and examine animal vocalizations used in behavioral contexts for a given species.

2. Understand sensory ecology which is related to how organisms receive, process, and respond to information from their environment. This led to experiments such as playback, a technique of rebroadcasting natural or synthetic signals to an environment with animals and observing their response.

3. Understand the relationship between the physiological and anatomical structures of animals and their vocal features.

All these attempts were meant to understand the full vocal abilities of animal vocalizations. However, these methods did not pick up until the end of the 20^{th} century as a result of the development of better hardware and software technologies that enable the collection and analysis of large acoustic datasets in various field settings.

2.1.1 Types of Acoustics

This section introduces typical terminology used in the bioacoustic domain such as active or passive acoustic monitoring and songs or calls. Many kinds of animal vocalization can be described as made up of syllables, a single utterance that may be one component in a longer vocalization, and phrases defined as series of syllables.

Active versus Passive acoustics

Passive acoustic monitoring (PAM) refers to using acoustic sensors to record sound in the environment which is then used for ecological inference [Kvsn et al., 2020; Brown and Riede, 2017]. PAM provides cost-effective, non-invasive, weather resistant, taxonomically board, and long-term tools for biodiversity monitoring [Sugai et al., 2019; Gibb et al., 2019]. Active acoustic monitoring (AAM) refers to when a sound is transmitted to a specific target and the response is detected by a return echo [Enari et al., 2017]. PAM tools are only restricted to sound-emitting species while AAM tools function regardless of whether the animal species emit sound or not [Mann et al., 2008].

Calls versus Songs

Animal communication for each species is different and often varies in terms of rhythm, pattern and pitch. Calls are described as short and simple phrases, for instance; mating calls, feeding calls, distress calls, and excitement calls [Kvsn et al., 2020]. Songs are described as longer duration

sounds made up of multiple phrases. Hainan gibbon songs usually contain more than five syllables while calls contain two at most [Fellowes et al., 2008].

2.1.2 Applications of bioacoustics in ecology

Bioacoustics tools have been able to solve multiple challenges such as resource constraints, observer bias, and limited data by providing non-invasive methods of monitoring. In recent years, bioacoustic monitoring has become an increasingly important and widely used tool for understanding the relationship between animals, their sounds, and ecology. In this section, the current and emerging uses of bioacoustic monitoring in ecology and their major limitations will be discussed.

Biodiversity research

Human activity has caused many rare critically endangered species to live in hiding, and these animals will probably became extinct before being fully documented [Penar et al., 2020; Wilson, 2017]. In such cases, automated recorders are the only viable solution to document their short stay on earth. Bioacoustic methods have provided scientists with an alternative non-invasive way to properly document the biodiversity of an area [Wilson, 2017]. Studies have shown that the usefulness of bioacoustics techniques in providing information about most sound-emitting species especially those that are rare and difficult to observe [Wrege et al., 2017].

Bioacoustic monitoring of environmental pollution

For studies concerning habitat conservation the relationship between natural and anthropogenic sounds, using an unattended recorder enables one to monitor entire ecosystems over time and space [Browning et al., 2017]. The acoustic data can be used to estimate the impact of climate change to the ecosystem which may help protect species which are most vulnerable to climate change in a timely and effective manner [Penar et al., 2020]. Furthermore, the same acoustic data can be used to assess the impact of human activities on the environment such as blast fishing [Browning et al., 2017; Penar et al., 2020].

Migrations

Bioacoustics can be used to monitor migration patterns using PAM as a tool. PAM is an effective tool for monitoring migration patterns because target species can be monitored over a long period of time and space, additionally, the tool is non-interactive and is not affected by weather conditions [Gibb et al., 2019; Bateman et al., 2021]. Generally, the calling characteristics of the target species have to be known beforehand as this information is used to calibrate the recording tools. Furthermore, a target species survey design is done to decide the most effective way to deploy the acoustic sensors [Aulich et al., 2019; Gibb et al., 2019]. Bioacoustics have been used to understand the migration patterns of terrestrial species, marine species, mammals, and majority

of these species migrate due to climate conditions, habitat disturbance, human activities, foraging, and breeding [Penar et al., 2020; Aulich et al., 2019].

Bioacoustic research in extinct systems

PAM can also be used to document biodiversity worldwide by collecting time series of audio recordings which can be used for future generations to gain knowledge of how the planet's acoustic communities have evolved over time [Sugai and Llusia, 2019]. Currently, the audio recordings of extinct species such as the Ivory-billed woodpecker (*C. Principalis*) can be found in biological collections which provides evidence of the species. By harnessing the advantages PAM we are able to collect recordings over time and then use these recordings as a way of preserving evidence of the current biodiversity, and possibly using these recordings as a benchmark for future research [Pyke and Ehrlich, 2010]. Additionally, animal vocalization recordings have been used in museums and sound libraries, for example, the Macaulay Library which contains mostly repositories of bird vocalizations, and the FonoSound library at the Spanish National Museum of Natural Sciences which contains more than a thousand repositories of frog vocalizations. These sound repositories have supported multiple bioacoustic research areas [Koehler et al., 2017; Goicoechea et al., 2010; Guerra et al., 2018] which in turn assists in the development of efficient monitoring and conservation systems. Furthermore, historic collections of animal sound recordings can be used to estimate the global change in biodiversity [Newbold et al., 2016].

Studying species and population dynamics

Many species have specific calls which may vary in age, gender, and size, thus, bioacoustic monitoring is a useful tool in monitoring species populations [Kvsn et al., 2020; Penar et al., 2020]. Recording animal vocalizations allows researchers to establish baselines, detect variation over time, and monitor species presence in different areas driving effective conservation strategies, including land preservation [Browning et al., 2017].

Species identification and classification

Every species has unique biological characteristics that are specific to it, which means each species can be clustered based on similar characteristics [Kvsn et al., 2020; Browning et al., 2017]. Bioacoustics have been used to assist in the classification of species based on their unique vocal characteristics. Species identification can be used to classify species into genus, species, and individually [Kvsn et al., 2020]. Furthermore, studies conducted show the usefulness of bioacoustics in discovering identification errors in the teleost taxa [Raick et al., 2020]. Some species have been identified as related based on similar external characteristics, however, bioacoustic studies prove that they are different because of different vocal behaviours. Additionally, bioacoustics monitoring has been used successfully to overcome the challenges faced when monitoring cryptic species [Williams et al., 2018; Raick et al., 2020; Teixeira et al., 2019]. Cryptic species refers to species which are difficult to observe or are rare, such as birds or bats which are visually cryptic bird, and insects which are small

and difficult to find [Penar et al., 2020; Browning et al., 2017]. Bioacoustics tools have been quite useful in detecting cryptic species such as the sounds of Cory's shearwaters (*Calonectris diomedea*) which are detectable at night and tend to breed on inaccessible locations and the Piranhas (*genus Pygocentrus*) which are hard to distinguish based on their external features, however, it is possible to distinguish them based on their vocal behavior which is different [Raick et al., 2020; Williams et al., 2018; Penar et al., 2020].

Animal Emotion and Welfare

Vocalizations are a reflection of the internal state of the caller, thus, bioacoustics can be used to infer animal emotion [Friel et al., 2019]. For major farm livestock, bioacoustics have been used to understand the emotional context of the different calls, for example, using pig coughing sounds for early detection of respiratory diseases [Penar et al., 2020], mapping vocalization of goats to particular situations [Briefer et al., 2015], mapping chicken distress calls to environmental stressors [Herborn et al., 2020], and using cattle vocalization in response to calf separation to assess [Green et al., 2018]. Currently, many farmers are adapting bioacoustics methodologies as a way of monitoring cattle as they provide a cost-effective, and non-invasive alternative to traditional measures of welfare [Green et al., 2018]. These methods have improved farm management systems by providing inferences about how the herds are coping in different situations which assists in the development of efficient feeding and reproductive systems, and improvement of the general cattle welfare [Green et al., 2018].

Limitations of Bioacoustics

Although bioacoustics have been able to address many issues in ecology and conservation, there remain several challenges and limitations to tackle such as:

- Passive acoustic monitoring is only possible for calling animals, a relatively small subset of all species. In many species only one sex calls (either exclusively or predominantly) [Browning et al., 2017].
- The development of a fully automated system which does not require manual interventions. This is necessary to reduce both analysis time cost and manual analysis biases [Kvsn et al., 2020; Browning et al., 2017].
- Creation of user-friendly software which can be understood by ecologists and conservation researchers to develop project-specific tools best suited for their own data [Browning et al., 2017].
- The development of classification models which are able to recognize multiple species like the BirdNET developed by Kahl et al. [2021] which uses sound to identify 984 North American and European bird species would be beneficial [Kvsn et al., 2020]. Currently, the majority of the techniques used have only one species present in a segment of the recording.

- A development of effective pre-processing methods as most bioacoustics data is affected by anthropogenic noises [Penar et al., 2020]. Additionally, bioacoustics methods are affected by weather variability which affects the recording quality and makes it difficult to detect certain frequencies [Goerlitz, 2018].

2.2 Bioacoustic Analysis Process

The utilization of bioacoustics requires a process flow for the interpretation and understanding of sound as indicated in figure 2.1, whereby data collection is succeeded by acoustic signal extraction, signal pre-processing, feature extraction, and bioacoustics analysis which is split into deep learning methods and other methods such as traditional statistical methods and machine learning methods.

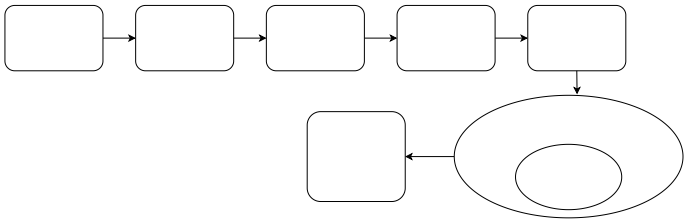


Figure 2.1: Common bioacoustics process flow.

2.2.1 Data Collection

The data collection process involves the recording of animal sounds and obtaining the recorded audio data in digital format.

Sound production and transmission

Sound is defined as the propagation of waves of pressure through a medium, which may be gas, liquid or solid. The waves are produced by vibrations from a sound emitting object (such as the larynx of an animal). The vibrations are alternating compressions and rarefactions of the medium creating waves of alternating high and low pressure that propagate outward from the emitter [Vehrencamp and Bradbury, 1998].
A sound wave consists of several key properties which are shown in figure 2.2 and defined as:

- **Wavelength** is length of a complete cycle.
- **Frequency** is number of cycles per unit time. It is measure in hertz (Hz, cycles per second). The frequency of a wave is inversely proportional to its wavelength.
- **Amplitude** is proportional to the amount of energy contained within a sound wave, and it is usually measured in decibel units (dB).

The amplitude of the sound wave progressively reduces as the sound's energy dissipates into the environment as they propagate outward from the source, usually referred to as attenuation [Brown-

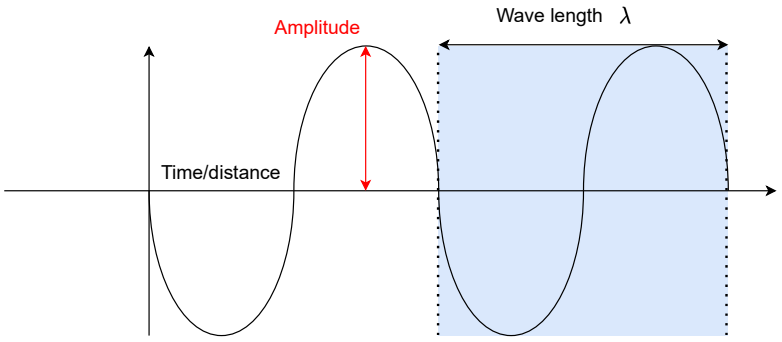


Figure 2.2: The characteristics of a sinusoidal sound wave.

ing et al., 2017]. Sound waves with lower frequencies experience less attenuation than those with higher frequencies which means that they can travel longer distances and still be recognized. This means that if two animals call at the same amplitude but at different frequencies, the animal with the lower frequency will be recognized at a greater distance than the the one at a higher frequency. On the other hand, animals calling at high amplitudes or loudly can be detected at a greater distance that those calling at lower amplitudes. All these cases have serious implications for the detection of signals produced by sound emitting animals.

The likelihood of sound to be detected by a sensor can also be affected by the properties of the medium. Sound waves travel faster through water than air because water has higher density. Factors such as pressure, water depth, temperature, and density affect the distance travelled by a sound wave, thus, the environmental factors also have implications for monitoring animals through acoustics, since they affect sound detection.

Sound reception

The sound is produced by the source (an animal calling) travels through the medium, and the vibrations of the sound wave are converted into a corresponding electric signal by a transducer [Kvsn et al., 2020]. A transducer is a device that converts one energy form to another. Microphones contain transducers which are responsible for converting sound into electrical signals [Kvsn et al., 2020]. In addition to the transducers, microphones contain other parameters which affect the quality of the recording such as efficiency, self-noise, polar pattern and frequency. There are variety of transducers that exist and each is sensitive to particular frequencies. For example, humans frequencies within $20 - 20000Hz$ are known as the audible range [Browning et al., 2017]. Sounds above this frequency range such as bat echolocation calls (ultrasonic) are usually undetectable to humans and require ultrasonic detectors such as the piezo-electric transducers to be recorded. Similarly, sounds lower than the human audible range, such as elephant rumbles (infrasonic) require a transducer that is able to detect those frequencies. For optimal detection, one has to consider frequency sensitivity and select the most suitable microphone [Browning et al., 2017].

Sound recording and visualization

As mentioned earlier, microphones contain transducers which convert the sound vibrations into an electrical signal which is recorded during sound recording. There exist two types of sound recorders: analog and digital. However, analog sound recorders have been replaced by digital recorders which are now universally utilized in bioacoustic research [Brown and Riede, 2017; Kvsn et al., 2020]. Digital recorders are practically advantageous over analog recorders because they allow for longer recording hours, one can program recording schedules, the sound recordings are easily accessible (can be downloaded to a computer for analysis), and they allow for the process automated (from recording to analysis) [Kvsn et al., 2020]. During digital recording, there are different factors to consider such as sampling rate (typically measured in thousands of samples per second, kHz) and the bit depth (the number of possible amplitudes that can be measured, $2^{16} = 65536$ bit system). The sampling rate is defined as the rate at which the amplitude of the electrical signal is sampled at [Dyer and Harms, 1993; Sueur et al., 2018]. The bit-depth refers to the process of assigning numerical value based on a bit-scale to each sample according to its amplitude, this process is known as quantisation and can be considered as a rounding off process. For example, values rounded off to a higher number of significant figures are considered as accurate, hence, higher quantisation would lead to values closer to reality [Thompson et al., 2017; Sueur et al., 2018]. The sampling rate and bit-depth affect the analysis later, the bit-depth affects the amplitude resolution and the sampling rate affects the the frequency resolution [Brown and Riede, 2017]. In order to retain sufficient frequency information of a sound, a signal must be sampled in accordance with the Nyquist theorem which states that the sampling rate has to be twice as high as the highest frequency in the continuous signal [Rabiner and Gold, 1975; Sanchez-Gendriz, 2021]. If the sample rate is below the Nyquist limit, a phenomena known as aliasing occurs which results in incorrect amplitudes and lower frequencies in the reconstructed signal [Tan and Jiang, 2018].
The audio files from the digital recorders are usually stored in uncompressed wave (*.wav*) format which is preferred as it stores the full information unlike the lossy compressed (*.mp3*) format which reduces the amount of information [Sueur et al., 2018]. Once the audio files are stored they can be visualized using different tools such as oscillograms and spectrograms. Oscillograms are graphical representation of an audio wave form with time in the horizontal axis and amplitude on the vertical axis. Spectrograms are visual representation of audio signals with time on the horizontal axis, frequency on the vertical axis, and amplitude represented by the intensity of the colour [Kvsn et al., 2020].

2.2.2 Preprocessing: Signal denoising

After the audio data is collected, the next process is signal denoising. Signal denoising is an important preprocessing step which improves the performance of the subsequent bioacoustic analysis steps such as feature extraction [Xie et al., 2020]. Many signal denoising methods have been developed over the years, and they are often grouped by: time domain, frequency domain, and

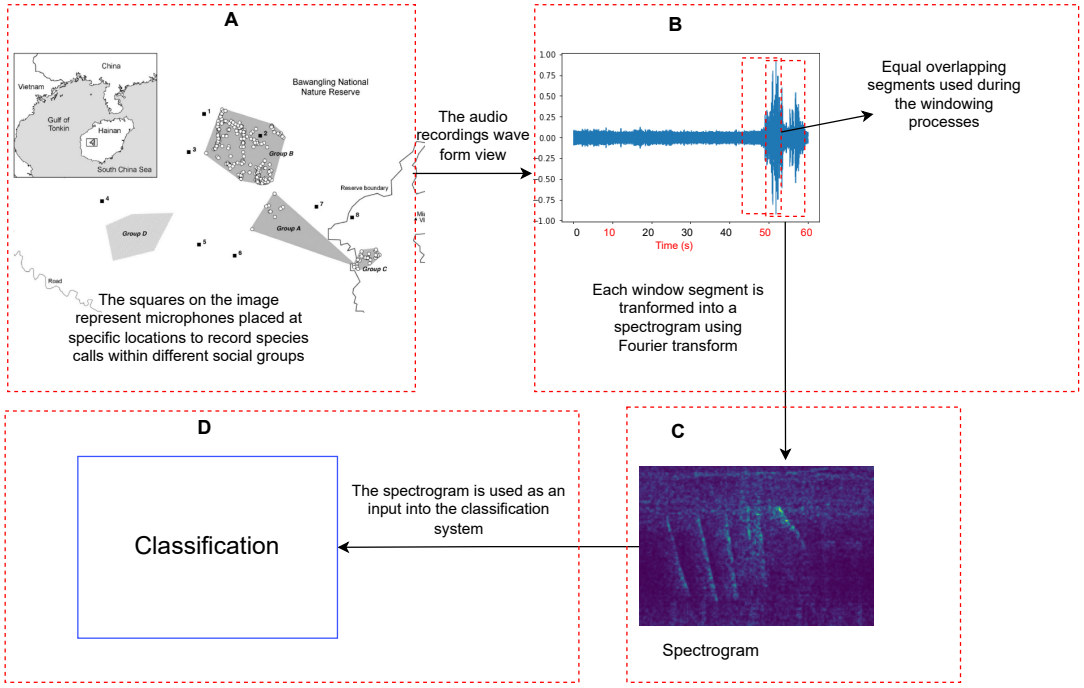


Figure 2.3: Summary of the development of bioacoustic classifier. A refers to the data collection phase, the microphones record audio data from the different locations within the nature reserve. B refers to the signal extraction and signal processing process, the audio recordings are analysed using signal processes such as Fourier transforms. C shows an example spectrogram obtained from the acoustic recording. D refers to the classification system, the spectrograms are fed into the classification system as an input.

time-frequency domain. Some of the common methods used include: Kalman filter methods, Wiener-Kolmogorov filter methods, signal subspace methods, and statistical methods [Kvsn et al., 2020].

Noise in the bioacoustic context can be defined as unwanted sounds that distort or conceal the sounds which are being recorded. For example, a recorder placed close to a busy highway will record low-frequency noise that can obscure low-frequency calls from animals vocalizing nearby which makes it difficult to obtain the calls. Additionally, all calling species often have a specific calling frequency range which makes it possible to remove any information outside the target frequency range [Brown and Riede, 2017; Xie et al., 2020]. For example, Hainan gibbons calls have a maximum frequency range of 2000 Hz, thus, any frequency information above 2000 Hz can be removed using audio filters [Dufourq et al., 2021]. Audio filters remove some frequencies from a signal while letting other frequencies pass unaltered. In general, audio filters can be classified into three categories: lowpass filters, highpass filters, and bandpass filters [Xie et al., 2020]. These filtering methods are often used as preprocessing steps in bioacoustic signal denoising and only work if the recording has species with the same frequency range [Xie et al., 2020].

The low pass filter also known as high-cut filters are used to reduce the amplitude of an audio signal below a specified frequency while allowing lower frequencies to pass through [Xie et al.,

2020]. For example, a low pass filter can be used to remove high frequency noise in a recording. The low pass filters prevent the aliasing phenomenon mentioned in section 2.2.1 by ensuring that the frequency components of the input signal are below the Nyquist limit [Tan and Jiang, 2018]. The high pass filter also known as low-cut filters are used to reduce the amplitude of an audio signal below a specified frequency while allowing higher frequencies to pass through. For example, a high pass filter can be used to remove low frequency wind noise in a recording. The band pass filter is a combination of a high pass filter and a low pass filter, only allowing signals between two specified frequencies to pass through [Xie et al., 2020].

2.2.3 Feature extraction: Signal processing and frequency analysis

After the audio data is collected and pre-processed, the next step is to extract meaningful information from the sound signal. This process is known as audio feature extraction [Kvsn et al., 2020; Browning et al., 2017]. Audio features can be extracted in different domains: time domain, time-frequency domain, and frequency domain. There are several methods used for audio feature extraction and method selection is based on the researcher.

Time Domain

Sound is recorded in the time domain as shown in figure 2.3 which can be described as waveform representation of the raw audio. Examples of audio features extracted from the time domain include:

- **Zero crossing rate**: Zero-crossing rate is the number of times a waveform crosses the horizontal time axis. The feature has been used for detection of voices in noisy environments and also for the recognition of pitched sounds. The zero-crossing rate feature is typically combined with other features when used for analysis [Mcloughlin et al., 2019].
- **Amplitude Envelop**: Amplitude envelope feature provides a rough idea of loudness, however, it is sensitive to outliers. It is mostly used in music genre classification [Mcloughlin et al., 2019].

Frequency Domain

Representations of audio signals in the time domain or frequency domain provide important information about temporal and spectral characteristics which are important for acoustic analysis, however, the frequency domain representation seems to be the most popular way to explore audio signals in most ecological studies [Sanchez-Gendriz, 2021; Sueur et al., 2018]. One way to switch between the time domain and frequency domain and vice versa is by the use of Fourier transforms. A Fourier transform is a reversible mathematical technique that decomposes signals into sinusoidal functions with definite frequencies [Hopp et al., 2012]. Signals can be characterized as either continuous or discrete and periodic or aperiodic, and the choice of Fourier technique used depends on a combination of these features. In particular, for discrete signals the discrete Fourier

transform (DFT) is used [Hopp et al., 2012]. A DFT is specific form of the Fourier transform which transforms a signal from the time domain to the frequency domain by using the amplitude and frequency of the signal to build the frequency spectrum of the original time signal [Smith, 2008; Sueur et al., 2018; Sanchez-Gendriz, 2021]. The direct implementation of the DFT is computationally inefficient, thus, the DFT is applied through an efficient algorithm known as the fast Fourier transform [Cochran et al., 1967]. The fast Fourier transform reduces the computational time and rounding off errors. The DFT through the fast Fourier transform allows one to obtain frequency components of a signal [Mcloughlin et al., 2019; Sanchez-Gendriz, 2021]. Examples of audio features extracted from the frequency domain include:

1. **Mel Frequency Cepstrum Coffiecients (MFCCs)**: MFCCs are a popular feature implemented in many speech and birdsong studies [e.g. Chao et al., 2019]. The MFCCs feature extraction technique involves transforming a signal from the time-domain to frequency domain and mapping the transformed signal to the mel scale, a logarithmic scale in which bin widths widen at higher frequencies, motivated by non-linearities in the way humans perceive sound. The MFCCs offer several advantages: they are suitable for many deep learning algorithms, they are easy to use, computational efficient, fast, have good accuracy, and can be used for different call types. On the other hand they are prone to interference from background noise [Kvsn et al., 2020; Mcloughlin et al., 2019].

2. **Fundamental Frequency**: The fundamental frequency is known as the lowest frequency that can be obtained in a signal. It has been used in many applications, for example, to assess if pitch can be characterized by body size, weight, age, and sex. The feature is easy to conceptually understand compared to other features, however, it is computationally expensive [Taylor and Reby, 2010; Mcloughlin et al., 2019].

Time-Frequency Domain

The time-frequency domain contains features with both the time and the frequency components of the audio signal. So far we have seen that DFT provides frequency resolution only, however, to obtain a time-frequency domain we require both the temporal and frequency resolution. For such cases the short time Fourier transform (STFT) should be used. The STFT allows one to compute the frequency variation of a signal over time by sliding a window along a signal and applying the DFT on each window frame [Smith, 2008]. The audio signal is divided into overlapping segments of equal length then applying the fast Fourier transform on each windowed segment to create the STFTs, and a spectrogram is then created by plotting the STFTs [Sueur et al., 2018]. The horizontal axis of a spectrogram represents time. It provides information such as the duration of the sound, how many calls occur in a given time frame, and the sound rate. For example, for animals that call rapidly the calls spectra would appear closer to each other. The vertical axis of a spectrogram represents frequency. Frequency can be interpreted as pitch, hence, high pitched sounds would have a high frequency compared to low pitched sound. In a

spectrogram the amplitude is represented by the contrast between the background and the sound spectra. The more the difference, the louder the sound [Kvsn et al., 2020]. Additionally, the resolution of the spectrogram depends on the choice of the STFT window size, increasing the window duration improves the frequency resolution and decreases the time resolution and vice versa. One way of improving the visualization of the time resolution without compromising on the frequency resolution is by using overlapping segments [Proakis, 2001; Sanchez-Gendriz, 2021].
A Mel spectrogram Stevens et al. [1937] is a spectrogram whose frequencies are converted to mel scale. Many studies have shown that humans perceive frequencies non-linearly. Humans can more easily differentiate between audio signals with low frequencies than audio signals with higher frequencies. For example, we can easily tell the difference between 500Hz and 1000Hz than 10000Hz and 10500hz, although, the difference between them is the same. Hence, the introduction of the mel scale. The mel scale is a logarithmic scale which is based on the principle that equal distances in pitch sounded equally distant to the listener [Aly and Alotaibi, 2022; Leitner and Thornton, 2019]. A mathematical operation which is described below is performed on the frequencies to convert them to the mel scale, where M represents Mels and f represents Hertz:

$$M = 2595 \cdot \log\left(1 + \frac{f}{500}\right) \tag{2.2.1}$$

The mel spectrogram is commonly used as input for deep learning algorithms, this is due to the pitch shifts that corresponds to linear shifts on a logarithmic scale which aligns with the ability to reliably detect linearly shifted features that exists in deep learning methods [Xie et al., 2019; Stowell, 2022]. Additionally, using the mel scale instead of frequency on the y-axis reduces the input size which makes mel spectrograms more computationally practical compared to regular spectrograms.

2.2.4 Analysis tools for acoustic data

In addition to the signal detection, signal preprocessing and feature extraction, there are many other bioacoustic computational analysis tools such as machine learning which have been developed to solve real-world ecological applications such as the ones mentioned in section 2.1.2. As discussed in section 2.1.1, call or song types vary from species to species, thus, can be used for species identification. Traditional machine learning methods such as those listed in table 2.1 have been popular in species identification applications, however, within the machine learning space deep learning methods have been gaining popularity in the bioacoustic and ecology world. Deep learning methods such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have the ability to make predictions in complex problem settings, such as speech recognition or visual object recognition or visual object recognition. In addition, they have the capability to model temporal relation which makes them more favourable over the traditional machine learning methods [Li et al., 2017a; Xie et al., 2020]. Additionally, deep learning methods can automatically detect and extract features, unlike traditional machine learning methods where feature extraction and method selection is often manual and separate, thus, relies on experts in the field. The method separation

and the expert reliance also makes traditional machine learning methods inappropriate for online or real time analysis [Mao et al., 2021]. Furthermore, these manually extracted features fail to extracted the complex features and are usually subject specific which results in low generalization [Mao et al., 2021]. The deep learning methods make use of multi-layer structures composed of multiple linear and non-linear transformations to extract features of the input data by proceeding from low-level to high-level abstraction enabling the extraction of complex features. Deep learning methods are advantageous because of their efficient feature extraction abilities, scalability, and real time processing capabilities, also, from a work load perspective, the automatic feature extraction ability reduces the work load for analyst which entail reduces costs [Stowell, 2022]. This section reviews previous work done on the use of deep learning methods in bioacoustics which will guide the research methodology.

Technique	Species	Bioacoustic Application	Reference
CART	Chickens	Animal welfare - classification of distress calls.	[Mcloughlin et al., 2019]
Discriminant Function Analysis	Zebra	Sex-based classification of zebra finch distance calls.	[Sahu et al., 2022]
	Wolves	Identification of individual wolves and subspecies using wolf howls.	[Larsen et al., 2022]
Support Vector Machine (SVM)	Birds	Identifying patterns of human activities and bird activities over time.	[Li et al., 2019]
	Termites	Termite detection.	[Nanda et al., 2018]
Aritficial Neural Networks (ANN)	Frogs	Identification and classification of frogs using bioacoustic data.	[Chao et al., 2019]
	Cats	Classification of cat breed using bioacoustic features.	[Raccagni and Ntalampiras, 2021]
Hidden Markov Model	Elephants	Classification of elephant vocalizations to determine context.	[Kvsn et al., 2020]
	Cetaceans	Multispecies discrimination of whales using Hidden Markov Models.	[Trawicki, 2021]
k-NN	Birds	Classification of bird and anuran species.	[Akbal et al., 2022]

Table 2.1: Summary of different techniques used in bioacoustic analysis for different animal groups.

Bioacoustic classification using deep learning

In bioacoustics, deep learning has been used for mostly classification tasks, which means classifying an item as being part of a particular class based on certain characteristics [Stowell, 2022]. For example, suppose there is a recording containing calls known to be produced by certain species, would it be possible to correctly assign those calls to the correct species? In this dissertation, an attempt will be made to develop a bioacoustic classifier which can correctly assign calls to the right social group using deep learning methods. Many researchers have developed automated classifiers for different species such as birds [Gupta et al., 2021; Salamon and Bello, 2017], bats [Zualkernan et al., 2020], frogs [Xie et al., 2022], primates [Dufourq et al., 2021], cattle [Jung et al., 2021], fish [Guyot et al., 2021], whales [Jiang et al., 2019; Padovese et al., 2021], however, limited number of studies focused on social group classification.

Prior work on bioacoustic classification using deep learning has typically adopted a standard recipe which follows: signal detection, signal preprocessing, feature extraction, and classification [Knight et al., 2020; Xie et al., 2016]. The process typically begins with a full length audio file which contains a mixture of animal vocalizations and noise, the full audio file which goes through some preprocessing using signal processing methods described in the previous sections. The signal preprocessing process steps involves transformation of the audio file into a visual representation such

as a spectrogram, segmentation and filtering unwanted background noise. Segmentation involves dividing the audio file into fixed sizes of one to ten seconds because it is easier and computationally efficient to develop a classifier system using segments than the full raw audio file [Stowell, 2022]. The next step, feature extraction, involves extracting acoustic features from each segment. The vast majority of studies surveyed used the mel spectrogram mentioned in section 2.2.3 as input data. The use of the spectrogram is preferred in deep learning systems because the spectrogram has a similar format to a digital image, thus, taking advantage of all the benefits and developments taking place in image classification deep learning systems. There have been debates on the benefits of using the standard spectrogram which uses linear frequency axis or one that uses logarithmic frequency axis such as the mel spectrogram. Most researchers argue that the mel spectrogram contains various useful acoustic components [Xie et al., 2019], however, literature presents no consensus in terms of which is best between standard spectrogram and the mel spectrogram. The decision seems to be driven by the researchers' aim and the type of deep learning method they intend to use. Furthermore, so far there is no consistently best acoustic feature across all tasks and species. Prior the deep learning era, MFCCs were the most used feature and have been used in some bioacoustic deep learning tasks [Colonna et al., 2016; Jung et al., 2021], however, there are typically outperformed by mel spectrograms [Elliott et al., 2021]. Alternatively, the raw waveform can used as an input which removes the spectrogram transformation step, however, deep learning with raw waveforms tends to require larger amounts of data for training compared to spectrograms [Stowell, 2022].

The final step is classification. This involves developing a classifier which takes an input which can be either mel spectrograms or MFCCs and assigns each mel spectrogram or MFCCs to a particular class depending on the problem. The most commonly used deep learning algorithms for bioacoustic classification tasks are CNNs which are described in detail in section 3.2 and CRNNs which is a CNN with a recurrent layer which were designed to capture both the spatial and temporal characteristics [Himawan et al., 2018; Tzirakis et al., 2020; Gupta et al., 2021]. The use of CNNs in bioacoustics was inspired by the first application of CNNs to audio recordings for automatic speech recognition [Deng et al., 2013] which led to the development bioacoustic classification systems for many species such as birds and whales [Salamon and Bello, 2017; Stowell, 2018]. The main two reasons why CNNs were proposed for bioacoustic classification were the location invariant property in CNNs which means that a CNN is capable of classifying a pattern regardless of the location of the spectrographic image, hence, the algorithm is unaffected by changes in time and frequency of the signal within the recording, and the other reason is that the CNNs operates directly on the spectrogram which removes the manual feature extraction steps. However, this means that the quality of spectrogram is important, thus, the choice of spectrogram parameters is important [Knight et al., 2020].

Based on literature reviewed, the classification process is made up of three main steps; preprocessing step, architecture development, and model performance [Stowell, 2022]. The preprocessing step contains data splitting and data augmentation. Data splitting is the processing of splitting the data

into three data sets, namely, training, validation, and testing sets. The training set is used to train the CNN model, then the validation set is used to evaluate the performance of the models and is used for model selection purposes, and the testing set usually referred to as the unseen data which is used to estimate how well the classifier generalizes [Stowell et al., 2019]. Data augmentation is a technique used to increase small training data by creating synthetic data. For audio, this involves noise blending [Dufourq et al., 2021], time shifting, time or frequency axis warping [Lasseck, 2018] and sound mixing . Data augmentation helps with diversifying the data set, increasing the amount of data, and also ensuring there is equal representation within the data set which improves model performance and generalization [Lasseck, 2018; Stowell, 2022].

Architecture development refers to the selection of CNNs architectures, some researchers prefer developing their own CNN architectures, however, currently, there is a strong move towards using already existing CNN architectures such as AlexNet [Krizhevsky et al., 2012]. These off-shelf architectures are already pretrained on standard datasets and are very convenient because they reduce the number of computations [Gupta et al., 2021; Knight et al., 2020]. Lastly, the standard metrics used for model evaluation are accuracy, precision, recall, F-score, and area under the curve [Stowell, 2022]. The next chapter goes into detail about how CNNs work and how a CNN takes an image input and gives an output of what class the image belongs to, which will assist in the development of our own social group bioacoustic classifier.

Chapter 3

Neural Networks

This chapter consists of two sections. The first section provides literature about feed-forward neural networks including the different parts that build up a neural network such as activation functions, loss functions, backpropagation algorithms, and regularization techniques. The second section provides background material on convolutional neural networks which are used for image classification.

3.1 Basic Artificial Neural Networks

Historically, the interest in neural networks was driven by the desire to understand the principle by which the human brain works with the hope that by mimicking the brain's structure, one might capture some of its capability, and create machines capable of performing complex tasks for which sequentially operating computers are not well suited for [Müller et al., 1995]. This was inspired by two researchers namely Warren McCullough and Walter Pitts who discovered the first mathematical model of a biological neuron which suggested that the human brain could be thought of as a computing device [McCulloch and Pitts, 1943].

The human brain can be defined as a biological neural network that consists of networks of neurons whose function is to receive and send signals. Each neuron has dendrites that receive input signals and based on these input signals, they produce output signals to other neurons through axons. Artificial neural networks (ANN) are computational tools structured similarly to the organization of neurons in the brain [Rosenblatt, 1961]. The biological neural network and ANN share a similar functionality which is processing and transmitting information. Additionally, they have similar capabilities such as the ability to learn through training.

In an ANN, dendrites are represented as inputs received at each node, and axons are depicted by arrows. The neurons are depicted as nodes through which data and computation flow. The artificial neuron receives input signals from raw data or previous layers of the neural network, performs some calculations, and sends output signals to other neurons in the neural network through a synapse [Aggarwal et al., 2018].

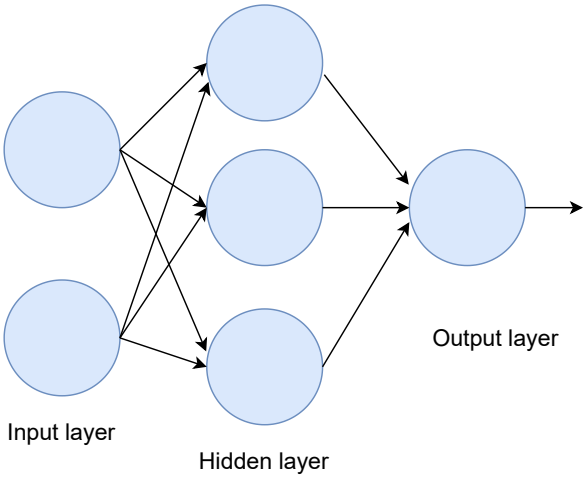


Figure 3.1: An example of a multilayer perceptron with one hidden layer, two input variables, and one output node.

3.1.1 Feed Forward Neural Networks

A feed-forward neural network also known as a multilayer perceptron (MLP) is the simplest form of the neural network developed by Frank Rosenblatt [Rosenblatt, 1958]. The MLP is an artificial neural network that consists of multiple connected perceptrons. MLPs are organized into layers, generally, one input layer, zero to many hidden layers, and one output layer. A neural network with no hidden layer is known as a perceptron and one with at least two hidden layers is called a deep neural network (DNN). An example of a feedforward neural network with one hidden layer is illustrated in figure 3.1. The feed-forward neural networks have a property that information is always fed forward from one layer to the next, there are no loops in the network. Compared to recurrent neural networks which have feedback loops, feed-forward neural networks have a simple learning algorithm. The learning process of ANNs is illustrated in figure 3.2.

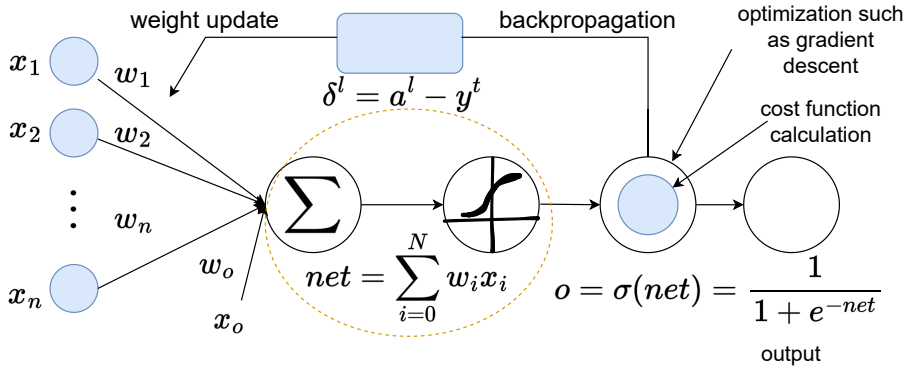


Figure 3.2: A diagram illustrating the ANN learning algorithm.

- In the context of supervised learning which has data in the form x and y, and the goal is to create a mapping $f : X \mapsto Y$, the forward pass can be summarized in the following steps:

1. In the input layer each node in the neural network receives an input x from an external source.
2. All nodes from the input layer are fully connected to all the nodes in the next layer, each connection has an associated weight w based on its relative importance against other inputs. All the weighted inputs are summed together.
3. The net input S of each weight and input, and a bias term b which is a constant value included in the neural network whose main purpose is to translate the activation function to the direction of the sign of the constant value, thereby introducing flexibility in the model is passed through an activation function (discussed in section 3.1.2) f which transforms the sum of weights into an output $\hat{y}$.
4. The difference between the true value y and network output $\hat{y}$ using the loss function (discussed in section 3.1.3) are calculated.

- The backward pass (discussed in section 3.1.4) can be summarized in the following steps:

1. The gradient of the loss function is calculated and the weights are updated using back-propagation.
2. The neural network is optimized to improve the model predictions.

Mathematically, a feedforward neural network is defined as follows:

$$S_{(l)}^{j} = \sum_{i=0}^{p^{(l-1)}} w_{ij}^{(l)} x_i^{(l-1)} + b_j^{(l)} \tag{3.1.1}$$

$$z_j^{(l)} = f(S_{(l)}^{j}) \tag{3.1.2}$$

where:

l denotes the layer index. $l = 0$ - input layer , . . . , L - output layer.

$p^{(l-1)}$ denotes the number of nodes in the $(l-1)^{th}$ layer.

$w_{ij}^{(l)}$ denotes the weight from the i^{th} node in the $(l-1)^{th}$ layer to the j^th node in the (l) layer.

$x_i^{(l-1)}$ denotes the input value from the i^{th} node in the $(l-1)^{th}$ layer.

$b_j^{(l)}$ denotes the bias term for the j^{th} node in the l^{th} layer.

$S_j^{(l)}$ denotes the sum of the weighted input, that is input to the j^{th} neuron in the l^{th} layer

$f(.)$ denotes the activation function applied element-wise

$z_k^{(l)}$ denotes the weighted input in the $(l+1)^{th}$ layer or output of the j^{th} node in the l^{th} layer

3.1.2 Activation Functions

Activation functions play an important role in neural network design. They are what enables the communication between the different layers in an ANN. The activation function defines an output given an input. Activation functions are critical in ANNs because most of them add non-linearity into the neural network which makes the network more dynamic and capable to learn complex mappings [Sharma and Sharma, 2017]. Activation functions must be differentiable so

that techniques like backpropagation and optimization can be implemented. The most popular activation functions are sigmoid functions, Hyperbolic Tangent (Tanh) functions, and Rectifier functions (ReLU) [Fukushima, 1975]. These are illustrated in figure 3.3.

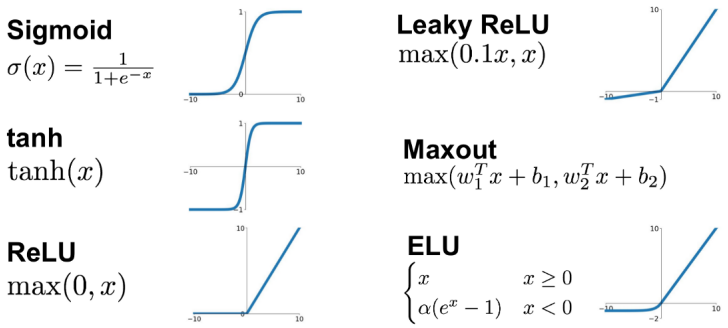


Figure 3.3: Common activation functions.

1. **Sigmoid Function**: The sigmoid function is a popular activation function commonly used in solving binary classification problems. The sigmoid function transforms the output signal into values ranging between 0 and 1 [Sharma and Sharma, 2017; Goodfellow et al., 2017]. Mathematically it is defined as:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.1.3}$$

2. **Hyperbolic Tangent Function**: The hyperbolic tangent (Tanh) activation function is similar to the sigmoid activation function. The only difference is that the tanh activation function transforms the output signal into values ranging between -1 and 1 which means the values are not restricted to vary in one direction. The tanh function is continuous and differentiable [Sharma and Sharma, 2017; Goodfellow et al., 2017]. Mathematically it is defined as:

$$Tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{3.1.4}$$

3. **Rectifier Linear Function**: The ReLU function is a piece-wise linear activation function that is widely used in deep learning [Hahnloser et al., 2000; Goodfellow et al., 2017]. The ReLU function does not activate all neurons at the same time which makes it more efficient compared to other functions. ReLU is linear for positive values and zero for negative values, thus, negative values are not mapped properly and weights of values with a gradient of zero will not update during backpropagation [Sharma and Sharma, 2017; Goodfellow et al., 2017]. There are other improved variants of ReLU namely, Leaky ReLU and Exponential Linear Unit (ELU) [Xu et al., 2015; Clevert et al., 2015]. Leaky ReLU and ELU resolve the problem of the gradient being zero for negative inputs by increasing the range of the ReLU function [Sharma and Sharma, 2017]. The ReLU functions can be expressed mathematically as:

$$r(x) = \max(0, x) \tag{3.1.5}$$

$$r(x) = \max(ax, x) \tag{3.1.6}$$

$$r(x) = \max(a(e^x - 1), x) \tag{3.1.7}$$

4. **Softmax Function** The softmax function is an activation function commonly used for multi-class classification problems. In multiclass problems, the output needs to be a probability distribution containing the probability of each class [Sharma and Sharma, 2017; Goodfellow et al., 2017]. The softmax function is a combination of multiple sigmoid functions, hence, the output signals of each net range between 0 and 1. The softmax function for a K-class classification problem can be expressed mathematically as:

$$softmax(x_k) = \frac{e^{x_k}}{\sum_{i=1}^{K} e^{x_i}}, k = 1, 2, 3 \dots K \tag{3.1.8}$$

3.1.3 Loss Functions

Loss functions also known as objective or cost functions are used to measure the performance of the network [Aggarwal et al., 2018]. They provide a mathematical way to measure how well the model is mapping a set of inputs x to their respective targets y by calculating the difference between the model's output $\hat{y}$ and the target y. This concept is used in:

- Optimization (described in section 3.1.4) to help find weights that minimize the loss and to provide accurate predictions.
- Backpropagation (described in section 3.1.4) to update the neural network's weights.
- Selecting model architecture and hyperparameter tuning (described in section 3.1.5) to find the best model architecture and user-defined parameters that minimize the loss.
- Final model assessment to find how well the final model selected is performing.

Choosing a correct loss function for a particular application is tightly coupled with the choice of the output unit. For example, when working with regression applications with numeric outputs, the output layer will make use of the linear activation function coupled with the Mean Squared Error (MSE) loss function. For classification applications, the sigmoid activation function is used for binary classification problems and the softmax activation function is used for classification problems with more than two classes. The loss function used for both cases is the categorical cross-entropy.

The MSE is the mean square difference between the predicted and model. The MSE formula can be expressed as:

$$E(\hat{y}, y) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.1.9}$$

where i is index of sample, $\hat{y}$ is the predicted value, y is the target value, and n is the number of samples.

Cross-entropy is a method used to measure the difference between two probability distributions. In classification problems, the interest is mapping input variables to a class label. This can be

model as predicting the probability of an example belonging to each class. Thus, cross-entropy can be used to find the difference between the probability distribution of the model's predictions and the probability distributions in the training data set [Mehlig, 2019; Goodfellow et al., 2016].
The categorical cross-entropy loss function for a classification problem with K number of classes and N observations can be expressed mathematically as:

$$E(\hat{y}, y) = -\frac{1}{n}\sum_{n=0}^{N}\sum_{k=1}^{K}(y_k \ln \hat{y}_k + (1 - \hat{y_k}) \ln (1 - \hat{y_k})) \tag{3.1.10}$$

3.1.4 Optimization

Backpropagation

The general flow in a neural network is randomly initializing the weights, a forward process of computing the network outputs using the inputs and the activation functions, calculation of the errors using the loss functions, and the backward iterative processes of updating the weights to minimize loss. Backpropagation is an algorithm used to train neural networks using gradient descent and chain rule [LeCun et al., 1988]. It was introduced in the 1960s and became famous in 1986 after the publication of a paper called "Learning representations by back-propagating errors" [Rumelhart et al., 1986]. The backpropagation algorithm recursively calculates the partial derivative or the gradient of loss function E with respect to the neural network's weights w. The goal of this method is to minimize the loss function by adjusting the model parameters. The level of adjustments is determined by the gradients of the loss function with respect to the model parameters [Goodfellow et al., 2016].
Let Σ_j^l represent the error in the j^th neuron in the l^th layer, E represent the loss function, w_{ij}^l represent the weights and g represent the activation function.
The derivative of the loss function E with respect to the weights can be defined as follows:

$$\frac{\partial E}{\partial w_{ji}^{(l)}} = \frac{\partial E}{\partial a_j^{(l)}} \times \frac{\partial a_j^{(l)}}{\partial w_{ij}^{(l)}} = \delta_j^{(l)} \times x_i^{(l-1)} \tag{3.1.11}$$

The error in the output layer is defined as:

$$\begin{aligned} \delta_j^{(L)} &= \frac{\partial E}{\partial z_j^{(L)}} = \frac{\partial E}{\partial a_j^{(L)}} \times \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \\ &= \frac{\partial E}{\partial a_j^{(L)}} \times f'(z_j^{(L)}) \end{aligned} \tag{3.1.12}$$

And the error in the other inner layers can be defined recursively as:

$$\begin{aligned} \delta_i^{(l-1)} &= \frac{\partial E}{\partial z_i^{(l-1)}} = \frac{\partial E}{\partial a_i^{(l-1)}} \times \frac{\partial a_i^{(l-1)}}{\partial z_j^{(l-1)}} \\ &= f'(z_i^{(l-1)}) \sum_{j=0}^{p^{(l)}} \frac{\partial E}{\partial z_j^{(l)}} \times \frac{\partial z_j^{(l)}}{\partial a_j^{(l)}} \\ &= f'(z_i^{(l-1)}) \sum_{j=0}^{p^{(l)}} \delta_j^{(l)} \times w_{ij}^{(l)} \end{aligned} \tag{3.1.13}$$

Optimization Functions

When training neural networks we began by randomly initializing the weights which means we begin training with a bad performing model. The goal is to obtain a good model with the highest accuracy or lowest loss at the end of training by adjusting the weights of the model using backpropagation to minimize loss [Aggarwal et al., 2018]. This process is called optimization. The most common optimization functions used in machine learning are gradient descent, stochastic gradient descent, Adaptive Moment Estimation (Adam), and RMSprop [Robbins and Monro, 1951; Kingma and Ba, 2014; Tieleman and Hinton, 2012].

1. **First-order optimization algorithms**: gradient descent and stochastic gradient descent are both iterative first-order optimization algorithms because they both depend on the first-order derivative of the loss function. In both algorithms, the goal is to minimize the loss E by computing the gradient of the loss function with respect to the weight w and taking small steps η in the opposite direction of the gradient until the gradient converges to a local minimum. In simple terms, the loss is transferred from one layer to another through back-propagation and the weights are modified depending on the losses. The stochastic gradient descent is an improved variant of the gradient descent which updates the model's parameters more frequently than the gradient descent which updates the model parameters after calculating the gradient on the whole data set. Thus, stochastic gradient descent converges faster than gradient descent. The gradient descent updating rule can be expressed mathematically as:

$$w_{updated} = w_{old} - \eta \frac{\partial E}{\partial W}, \text{ where } \eta \text{ is the learning rate and } E \text{ is the loss function.} \quad (3.1.14)$$

2. **Second-order optimization algorithms**: All types of gradient descent algorithms have some challenges such as selecting the best learning rate value, all parameters having a constant learning rate, and the possibility of the local minima being trapped. Thus, second-order optimization algorithms such as RMSprop and Adam were developed. Both methods make use of momentum which reduces the high variance of parameters and increases convergence by reducing fluctuations in the opposite direction. However, Adam is the best of the two because it is a combination of RMSprop and momentum. The Adam algorithm computes adaptive learning rates for each parameter. In addition to storing an exponentially decaying average of past squared gradients updates like RMSprop, Adam also uses the first m_t and second v_t moments of past gradients for the current update step. Bias correction is applied to m_t and v_t using decay rate parameters β_1 and β_2. The Adam updating rule can be expressed mathematically as:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \frac{\partial E}{\partial W} \quad (3.1.15)$$

$$v_t = \beta_1 m_{t-1} + (1 - \beta_1) \frac{\partial E^2}{\partial W} \quad (3.1.16)$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \tag{3.1.17}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{3.1.18}$$

$$w_{updated} = w_{old} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \epsilon}} \tag{3.1.19}$$

3.1.5 Training Protocols

Many decisions have to be made for one to obtain a good neural network model and these decisions are made by the designer of the model. Firstly, one has to decide how to randomly split the data set into a training, validation, and test sets. The most commonly used split is the 70%:20%:10% split and the decision usually depends on the amount of data available. Another option is using a re-sampling technique called cross-validation which splits the data into a fixed number of folds, runs the analysis of each fold, and averages the overall estimated error. This technique is usually used in small data sets.

Hyperparameters are parameters that are defined by the designer of the model. Examples of hyperparameters related to the network structure which are defined before training are; the number of layers of the network, dropout rate, network weight initialization, and activation functions. Examples of hyperparameters related to the training algorithm are; learning rate, momentum, number of epochs, and batch size. The selection of hyperparameters is a critical step in all machine learning algorithms because hyperparameters affect model performance. Hyperparameter tuning techniques such as grid search, random search, and Bayesian optimization are commonly used to select the best hyperparameter configuration. Searching for the best configuration involves training different models and evaluating the performance of each model using the validation set. The model with the configuration that results in the best validation accuracy is selected. The model is evaluated on the test data to assess how well it performs, and if there is a need for regularization which is a technique used when a model overfits. Finally, the final model is presented and used on unseen data.

3.1.6 Regularization

Generalization is the ability of a trained model to apply what it has learned to unseen data. When a model has a low training error and a very high validation error it is overfitting meaning that the model is memorizing the training data and failing to generalize [Chollet, 2017]. Regularization is a technique that introduces slight additions to the learning algorithm such that the model generalizes better, thus, improving the model's ability to adapt to new data. Regularization is an important technique in deep learning because the networks tend to be complex and prone to overfitting [Goodfellow et al., 2016]. Regularization techniques reduce the complexity of deep neural networks which reduces overfitting by changing the structure of the network or changing the network parameters [Chollet, 2017]. Regularization techniques will be explained in detail in the subsequent chapters.

Weight Regularization

Complex models are prone to overfitting. Weight regularization reduces overfitting by penalizing the complexity in a network which entails forcing the values of the weight matrices to decrease [Chollet, 2017]. To achieve this, the loss function E is updated to E_{reg} which includes an additional term known as the regularization parameter. This regularization parameter is controlled by a hyper-parameter λ. The main reason for this is there already exists a process that minimizes the loss function during training. Hence, by adding the regularization parameter in the same process we can also minimize the weights. The Lasso (L1) and Ridge (L2) are the most commonly used weight regularization techniques [Goodfellow et al., 2016]. In the L2 regularization, the square of the value of the weights is penalized which forces the weights to decrease towards zero. In the L1 regularization, the absolute value of the weights is penalized. Unlike L2, weights can be reduced to zero in L1. Thus, L1 is preferred when trying to compress the model compared to L2.

1. L1 regularization:

$$E_{reg} = E + \lambda \sum w_i \tag{3.1.20}$$

2. L2 regularization:

$$E_{reg} = E + \lambda \sum w_i^2 \tag{3.1.21}$$

Early Stopping

Training neural networks is challenging because one has to train the network long enough such that it is capable of learning the mapping from inputs to outputs but not training it too long such that it overfits. Early stopping is a regularization technique that provides guidance on the number of epochs the model has to go through before it starts to overfit [Caruana et al., 2000]. During training the model stores weights after every epoch and updates every time the validation error decreases. Once the model performance stops improving, the model training processes are triggered to stop and the model weights with the lowest validation error will be selected as the final weights [Goodfellow et al., 2016; Ying, 2019].

Dropout

The original dropout method was introduced by Hinton in 2012 [Hinton et al., 2012]. Dropout is an effective method of preventing overfitting in neural networks by promoting neuron independence and accumulation of independent learning. What happens in dropout is that neurons with a probability p are omitted during training [Srivastava et al., 2014]. The choice of which neurons are omitted is random which reduces neuron interdependent learning by teaching each hidden neuron to work with a randomly selected sample of other neurons [Labach et al., 2019]. This ensures that the model learns multiple independent representations of the same data.

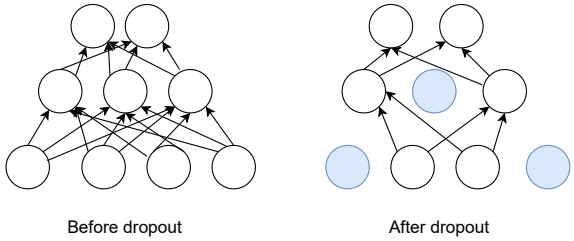


Figure 3.4: An illustration of how dropout works. The first image is a feedforward network before dropout and the second image is after dropout is applies. The blue nodes in the second image have been omitted and will not be used in training

3.2 Computer vision

The concept of computer vision was established in 1960s with the aim to enable systems and computers to understand and retrieve information from visual inputs such as images or videos in a similar manner as the human brain [Huang, 1996]. In 1980, Yann LeCun inspired by the invention of the first basic image recognition neural network by Dr. Kunihiko Fukushima created the first convolutional neural network (CNN) which was known as LeNet; a CNN which was capable of deciphering handwritten text and was widely utilized in the postal and banking industry for deciphering data such as zip codes and digits [Khan et al., 2020; LeCun et al., 1989]. Although LeCun's work set the standard for today's computer vision and image classification applications, the breakthrough in computer vision was inspired by a CNN called AlexNet which surpassed the best image recognition algorithms by a large margin in an ImageNet computer vision contest [Krizhevsky et al., 2012]. The availability of large data sets with labelled pictures such as ImageNet and computer resources have enabled researchers to build more complex complex CNNs which can perform different computer vision tasks which were impossible in the past. In the recent years, CNNs have played a pivotal role in many computer vision applications [Rawat and Wang, 2017].

3.2.1 Convolutional neural networks

A CNN is a specialized type of neural network widely used in deep learning for computer vision tasks. CNNs are used to process grid-like topology such as images. An image is made up of pixels arranged in a grid-like fashion. The pixels have values ranging from $[0, 255]$. Images can be represented as grayscale images or color images. A color image has three channels (red, blue, and green channels stacked on top of each other) which can be considered as 3 matrices stacked on top of each other. A grayscale image has one channel. The traditional MLPs described in the previous sections can be used for image processing, however, they are not ideal because to process a color image, a traditional MLP would require a large number of parameters which is computationally expensive. Unlike the traditional MLPs, CNNs have mitigated these limitations by utilizing a linear mathematical operation known as a discrete convolution in place of the general matrix multiplication. This operation leverages three important ideas that have been useful in

computer vision research: sparse interaction, parameter sharing, and shift-invariance [Goodfellow et al., 2016]. The CNN architecture consists of three layers: a convolutional layer, a pooling layer, and a fully connected layer described in the following sections.

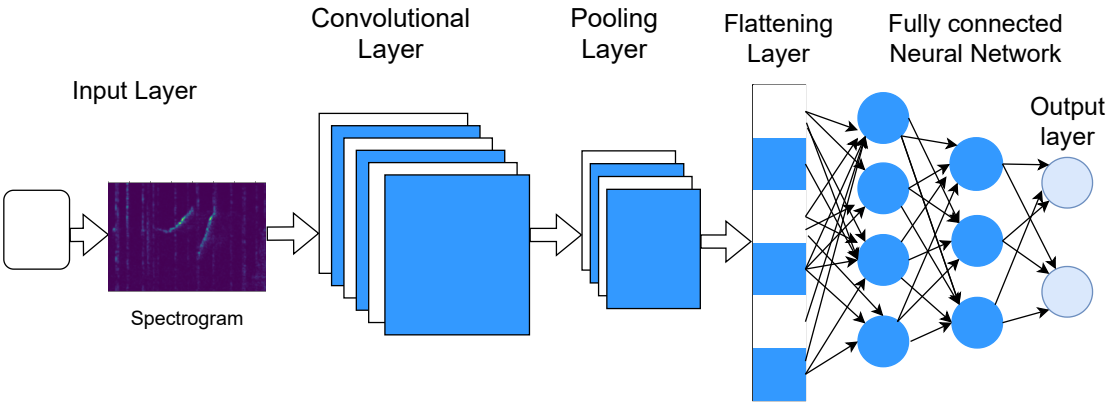


Figure 3.5: An example of a CNN architecture

The convolutional layer

The convolutional layer is the "heart" of a CNN as it carries most of the network's computational load. The word "convolutional" was inspired by the process of convolving a filter or kernel on the input image to produce a feature map. The convolution step is used to extract useful visual features of the image. The mathematical representation of the convolution operation is as follows:

$$F = I * K \tag{3.2.1}$$

where F represents the output feature map, I represents the input image, K represents the filter, and $*$ represents the convolution operation which is a dot product between two matrices.
The filter is a matrix of weights. The values in the filter matrix are updated each time the network performs backpropagation. However, the dimensions of the filter matrix are user-defined hyperparameters. The filter weights have to be optimized using an optimizer. A filter is applied to the image to extract certain visual features of the image. For example, in edge detection, a filter is applied to obtain the edges of the input image [Goodfellow et al., 2016]. Filters are usually smaller than the input image which allows for the extraction of small useful local features of the image rather than using the full image, thus, storing fewer parameters and increasing the efficiency of the model. This is known as sparse interaction [Goodfellow et al., 2016]. During training, the filter slides over the input image to produce a feature map. The same filter is applied to every region of the input image thereby reducing the complexity of the network by lowering the number of parameters (parameter sharing) [Rawat and Wang, 2017]. The parameter sharing feature makes it possible for CNNs to have the shift-invariance property, for example, an image of a bag remains a bag even if the location is adjusted.

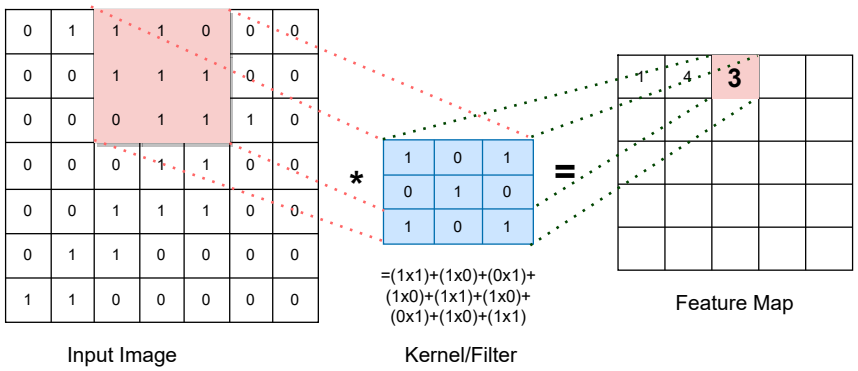


Figure 3.6: Visualization of how the convolution operation works.

The pooling layer

The output from the convolution layer known as feature maps is passed to the pooling layer. The purpose of the pooling layer is to reduce the dimensionality of the feature maps while preserving the important features, hence, decreasing the number of parameters and the amount of computation. Furthermore, pooling decreases the training time and controls overfitting [Rawat and Wang, 2017]. The pooling operation involves replacing the output of a feature map at given positions with a summary statistic. The pooling layer requires three hyper-parameters, namely, width, height, and stride [Goodfellow et al., 2016]. There are several pooling functions such as max pooling, average pooling, and stochastic pooling. Max pooling takes the maximum number in each window, thus, the feature map is decreased while keeping the most important information.

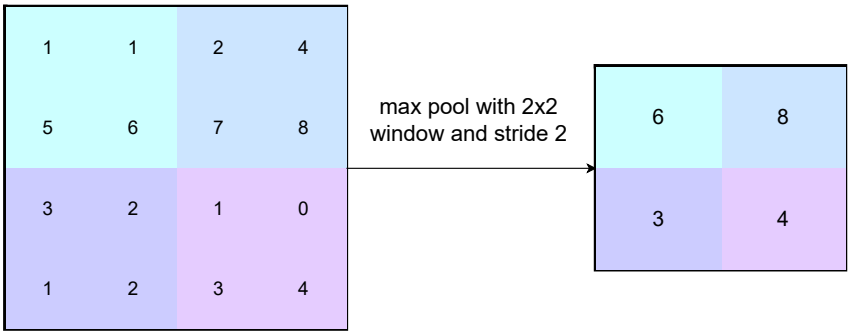


Figure 3.7: Example of max-pooling

The fully connected layer

After the alternating sequence of convolutional and pooling layers extracts all the image features, a traditional MLP is used to classify the images. However, to connect the final output of the convolutional layer to the MLP, the output feature map or pooled feature map is flattened. Flattening refers to turning a multidimensional input into a one-dimensional input. For example, if the final output of the convolutional layer is a $(28, 28, 3)$ tensor. Flattening this output will result in a long one-dimensional $(2352, 1)$ array. The 2352 comes from $28 \times 28 \times 3$. The flattened feature map is

used as the input to the fully connected network which works the same as a normal feed-forward ANN for classification problems.

3.2.2 Transfer learning

Transfer learning is a deep learning technique where a trained model for task A is re-purposed on a different but related task B to improve performance when modelling task B [Pan and Yang, 2009; Chollet, 2017]. In transfer learning, instead of training from scratch, the pre-trained model serves as a starting point for the new task. Transfer learning is useful in cases where there is a lack of training data. Furthermore, transfer learning saves training time and reduces overfitting [Weiss et al., 2016; Shin et al., 2016].
Approaches to transfer learning:

1. **Training a model to reuse it**: For example, we want to solve task A but there is insufficient training data for a DNN. One solution is to find a related task B with abundant data, train the DNN on task B then use the knowledge gained as a starting point for solving task A. This means task B which was previously trained would provide better initial weights for task A.

2. **Using a pre-trained model**: This approach involves leveraging some popular pre-trained models such as VGG-16, VGG-19, Inception V3 and ResNet-50 [Simonyan and Zisserman, 2014; He et al., 2016; Szegedy et al., 2015, 2016].

3. **Feature Extraction**: This approach involves freezing certain layers of the pre-trained models or fine tuning layers in the process to use them as feature extractors. Freezing layers refers to obtaining pre-trained weights and biases of a network, then deciding on whether to freeze a layer or not. Fine-tuning refers to replacing the final layer of a pre-trained network with an appropriate final layer of the new task

3.3 Deep Learning and Ecology

Deep learning techniques have been successfully applied in numerous applications such as speech recognition, image recognition, natural language processing, and finance [Kamath et al., 2019; Bashar et al., 2019; Chai and Li, 2019; Huang et al., 2020]. In ecology, deep learning has provided an effective and efficient way for ecologists to process large and complex datasets [Norouzzadeh et al., 2018]. Currently, deep learning is utilized for automated species identification, environmental monitoring, and wildlife behavioral studies [Mcloughlin et al., 2019]. Amongst the several deep learning techniques, CNNs are the most widely used [Borowiec et al., 2022]. The next sections provide examples of the uses of deep learning in ecological disciplines.

Identification and classification

Deep learning methods such as CNNs and RNNs have been popular in most identification and classification problems [Guo et al., 2020a]. CNNs have been used for the development of plant species classification models [Wäldchen and Mäder, 2018], plant disease classification [Chen et al., 2020], and phenotyping [Mochida et al., 2019]. Furthermore, other deep learning tools such as transfer learning have contributed towards the development of efficient models capable of classifying large amounts of species images [Chen et al., 2020; Guo et al., 2020a]. Tools such as camera traps allow for remote monitoring of animals [Browning et al., 2017]. CNNs are applied to camera trap data to aid in the automated classification of images of different species, which reduces the need for human involvement and saves time [Tabak et al., 2019; Guo et al., 2020a]. Deep learning has also been applied to acoustic classification of animal sounds such as bird songs [Salamon and Bello, 2017], mosquito sounds [Kiskin et al., 2021], and marine mammals vocalizations [Bermant et al., 2019], consequently assisting in creating automated species identification systems, behavioral monitoring systems, and biodiversity monitoring systems [Norouzzadeh et al., 2018].

Behavioural studies

Deep learning has also been used to automatically track animal behaviour. Camera trap images and acoustic data have been used to classify animal species' behavioural patterns [Norouzzadeh et al., 2018]. CNNs have been used to study behavioural and social interactions of bees [Wild et al., 2018]. CNNs coupled with Global Position System (GPS) systems have been used to understand migration and foraging patterns [Browning et al., 2018; Christin et al., 2019]. RNNs have also been used to analyse videos to model worm behaviours [Li et al., 2017b; Christin et al., 2019]. Furthermore, CNNs have been used to understand courtship rituals in animal species [Hulse et al., 2022; Janisch et al., 2021].

Population monitoring

Deep learning techniques can also be extended to estimate and monitor populations variations. Deep learning has been successfully utilized to monitor endangered species [Browning et al., 2017]. Furthermore, deep learning tools such as CNNs have also been widely applied in disease and welfare monitoring in wild plants, crops, and animal population [Christin et al., 2019].

Ecosystem management and conservation

Understanding the changes in the whole ecosystem is an important task for ecologists in monitoring, management and conservation of the ecosystem [Christin et al., 2019]. Deep learning techniques have been used to analyse acoustic data to understand the interactions in food web models and monitoring climate change indicators such as bats and birds which are known to be sensitive to habit and climate change [Brown and Riede, 2017; Christin et al., 2019]. Furthermore, deep learning has been used for landscape analysis, for instance, aquatic monitoring of coral reefs and freshwater

habitats [Browning et al., 2017]. Beyond species monitoring, deep learning has been used to track the impact of human activities on biodiversity. Kroodsma et al. [2018] used CNNs to track the effect of industrial fishing vessels on fish taxa. Di Minin et al. [2018] made use of deep learning techniques for the detection of illegal wildlife products such as rhino horns, pangolin scales, and elephant ivory on e-commerce platforms, the dark web, and social media. The usefulness of deep learning in ecology has been seen in the recent years and coupled with automated sensors, drones, and robots, we may be able to create fully automated management systems which will allow for allow for continuous ecosystem management without requiring much human intervention.

Chapter 4

Methodolody

The purpose of this dissertation is to use deep learning for bioacoustic classification of Hainan gibbon calls types. To achieve this, two classifiers will be developed:

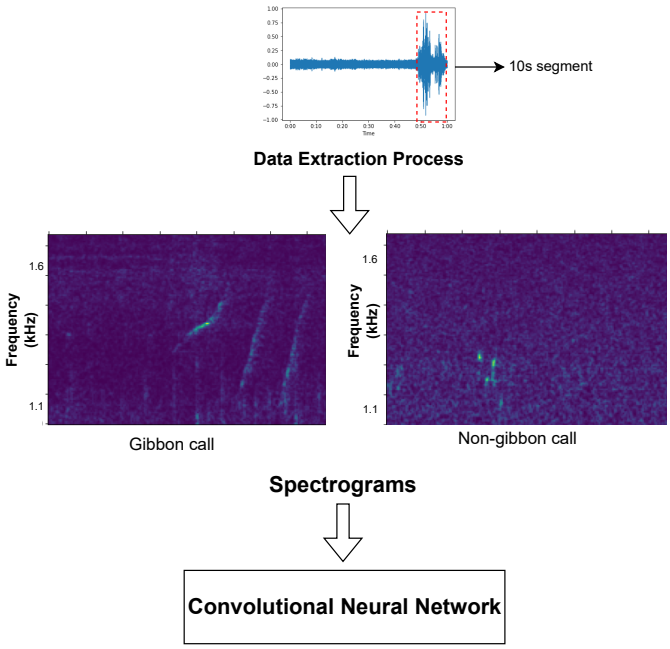
- **Classifier 1**: a classifier capable of distinguishing between a gibbon call and a non-gibbon call as illustrated in Figure 4.1a.
- **Classifier 2**: a classifier which can distinguish from which social group each gibbon call comes from as illustrated in Figure 4.1b.

This chapter provides the methodology used to develop the classifiers arranged into three sections. The first section presents the description of the data. The description provides a summary breakdown of how many *.wav* files were used from each social group. The second section provides a description of how the data extraction, feature extraction, and pre-processes techniques as described in Chapter 2 were implemented on the Hanain gibbon acoustic files. The third section provides the modelling processes such as model architectures of the CNNs used to develop the classifier, and how the algorithms were trained and evaluated.

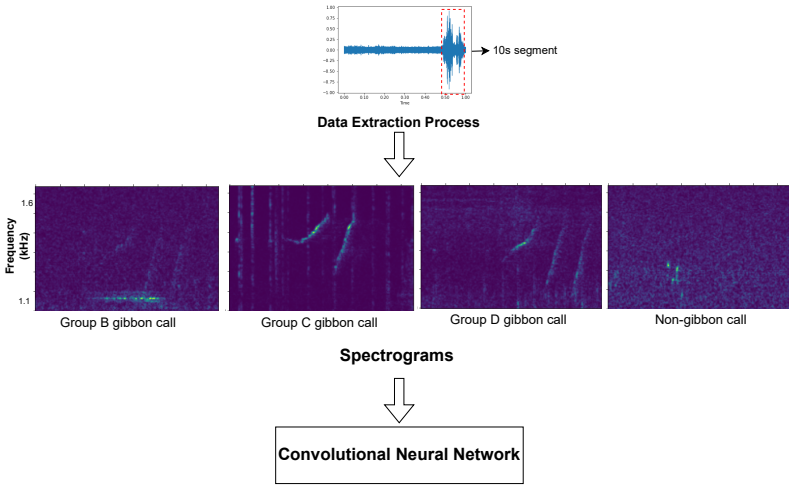
4.1 Data Description

The dataset consists of audio files obtained from the recorders situated within the known home ranges of the three Hainan gibbon social groups existing currently, namely groups B, C, and D (see Figure 4.2). The recorders were used to monitor the Hainan gibbons and were set to record the peak Hainan gibbon calling period (06:00-07:00) and were programmed to record for eight hours after [Bryant et al., 2016]. The majority of the recordings were made with a sampling rate of $9600Hz$ and a bit depth of 16 [Dufourq et al., 2021].

The audio files used for this research were originally annotated for a presence and absence model and were sufficient for that model, however, to ensure that there is class balance for the current model, additional files were annotated. In total, 54 audio files were used for the analysis (see Table 4.1). According to Table 4.1 we can notice that there were more C and D audio files; this was



(a) Binary gibbon classifier.



(b) Social gibbon classifier.

Figure 4.1: Summary of the methodology.

because most C and D audio files contained fewer gibbon calls compared to B audio files thus more audio files were required for group C and D to ensure that there is class balance. Additionally, only audio files obtained from recorders that could be unambiguously and with high confidence associated with a single group's territory were used to ensure that any calls in that audio file belonged to the same group. All audio files from recorders placed in between groups were excluded to avoid interference in calls. Furthermore, the audio files were accompanied by annotation files that were manually annotated by subject experts. The annotated files contained information about the start and end times, and the number of notes of each observed gibbon phrase in that audio file.

Social group	Number of audio files
B	13
C	20
D	21

Table 4.1: The number of audio files in each social group.

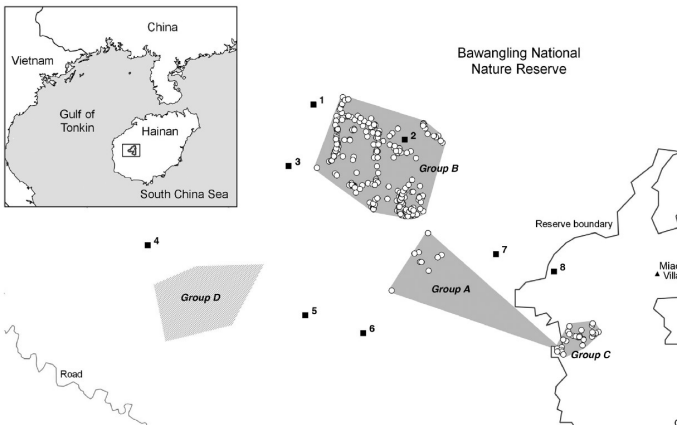


Figure 4.2: Location of eight Song Meter recorders (labelled 1–8) which are PAM devices used to detect gibbons in 2016 in BNNR in relation to the estimated distributions of the four Hainan social groups (A–D) at that time period [Bryant et al., 2017]. In this research only groups (B–D) was considered due to changes in the social group distributions in 2020, however, currently there exist five social groups [Li et al., 2022].

4.2 Bioacoustic Classification System

This section provides implementation details of the bioacoustic classification system illustrated in Figure 4.3. The input to the system are audio files containing Hainan Gibbon calls, calls refer to all the different syllables that exist in the audio file, and are used as the unit of the analysis (refer to figure 4.6). The calls go through signal pre-processing and feature extraction processes reviewed in chapter 2. The outputs of the system are the calls classified according to their social groups.

4.2.1 Audio Signal Extraction

The process begins with obtaining the timestamps of gibbon calls and non-gibbon calls from the annotation file provided for each audio file. The annotation file provides start and end times of the gibbon calls, the duration of the call and the type of call within that time period. The start and end times were used to segment the audio files into gibbon calls and non-gibbon calls as illustrated in Figure 2.2.4. The start and end times were converted by multiplying the times by the sampling rate.

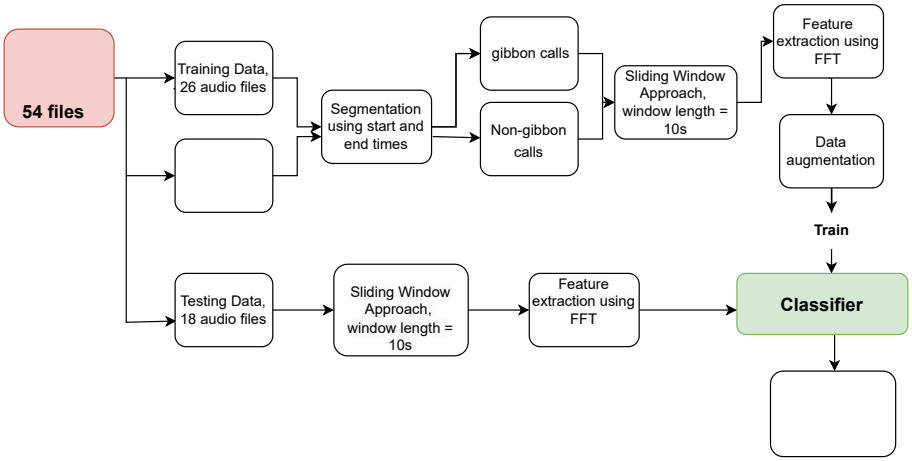


Figure 4.3: Bioacoustic classification system

To extract all the gibbon calls, the sliding window approach with a ten second window length defined as α was used. The ten second window length was selected because it fits the longest phrase (eight seconds) within a single segment [Dufourq et al., 2021]. The sliding approach tries to extract the gibbon call by shifting a small amount α. For example, suppose we have a gibbon call segment that has a start time of 15:02 and an end time of 15:03. The sliding window with size α will slide to the right by one second until it gets to the end time as illustrated in Figure 4.4. The same approach is applied to the non-gibbon segments.

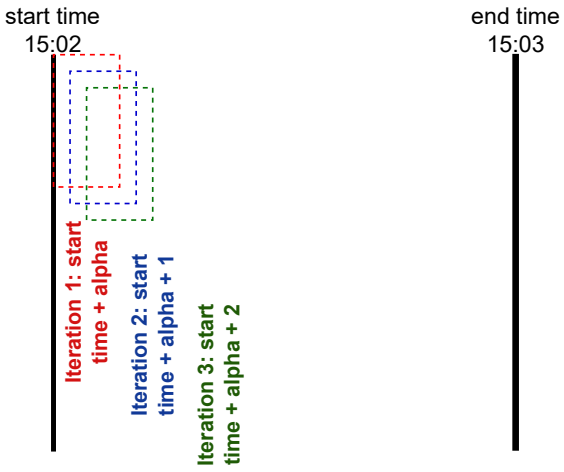


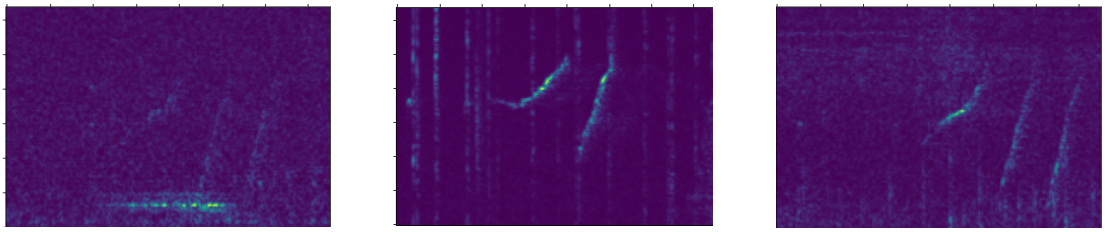
Figure 4.4: Sliding window approach.

Feature extraction

Mirroring Dufourq et al. [2021], each ten second segment was downsampled to $4800Hz$, which is higher than twice the maximum frequency of Hainan gibbon calls ($2000Hz$) and thus above the Nyquist limit described in chapter 2. The fast Fourier transform was applied to each windowed segment to create the mel-scale spectrogram using a window size of $1024/9600s$, a hop size of $256/9600s$, 128 mel frequency bins with centres uniformly spaced between 1 and $2kHz$, and

$(4800/256) * 10s$ number of hops. Therefore, a spectrogram of size 128×188 was obtained and used as the input to the 2-D CNN described in section 4.3.

The final dataset consisted of 4355 gibbon call images and a similar size of non-gibbon call images to ensure that we have an almost balanced dataset. The gibbon calls were stored based on which social group the audio file recording was from and the distribution is illustrated in Table 4.3.



(a) Group B. (b) Group C. (c) Group D.

Figure 4.5: Examples of randomly selected spectrograms for the different social groups.

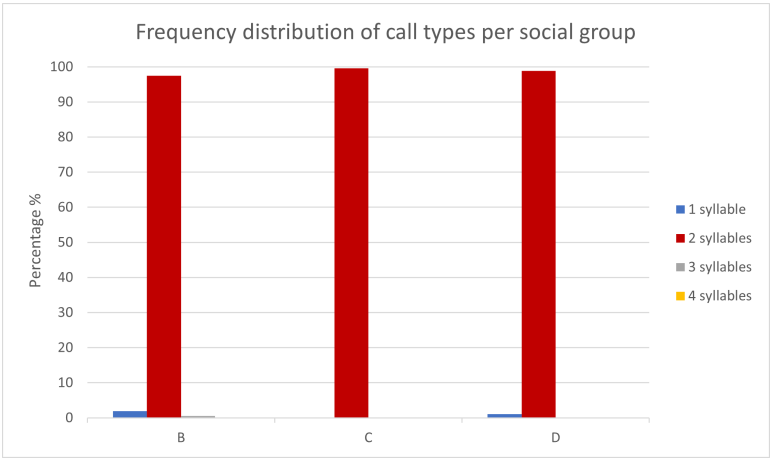


Figure 4.6: The distribution of call types per social group.

4.3 Image Classification

4.3.1 Pre-processing Methods

Data augmentation is used to boost the sample size by adding artificial samples through manipulation of existing samples [Salamon and Bello, 2017; Dufourq et al., 2021]. This method helps improve classifier performance for tasks with small training data sets such as our current problem. Data augmentation is used to create ten new copies of each ten second segment in both the presence and absence classes. The steps followed for the data augmentation process are as follows:

1. For each gibbon call segment $x^{(call)}$, ten noise samples are selected randomly ${x_i}^{(noise)}$, $i = 1,\ldots,10$.

2. A shifted segment ${x_i}^{(shift)}$ is created by randomly shifting the start time of each noise

segment forward by $0 < t_i < 9$ seconds, with the noise segment wrapping back on itself so that it remains ten seconds long.

3. The gibbon call segment is blended with the shifted segment to created augmented call presences ${x_i}^{(aug)} = \alpha x^{(call)} + (1 - \alpha){x_i}^{(shift)}$, where α is the mixing parameter.

4. The same process is followed for the noise segments – each noise segment is blended with a created blend of noise backgrounds.

A total of 30160 non-gibbon call segments and 33280 gibbon call segments were obtained after augmentation. The dataset was split using the number of gibbon calls before data augmentation into training and validation at a ratio of 77%:23% (refer to Table 4.2 and 4.3). Of the 54 audio files, 26 were used for training, 10 for validation, and 18 were for testing (six audio files per social group). Table 4.2 and 4.3 show how the training and validation data was constructed and augmented and not the testing data. This was because some of the test data was annotated (mostly group B and D) while there was lack of annotated data for majority of group C files, thus, the tables only show training and validation sets. Additionally, augmentation is not applied to the test set.

	Before Augmentation			**After Augmentation**		
	Train	Validation	Total	Train	Validation	Total
Gibbon call	2565	763	3328	25650	7630	33280
Non-gibbon call	2262	754	3016	22620	7540	30160

Table 4.2: Binary model data split before and after data augmentation.

	Before Augmentation			**After Augmentation**		
Social Group	Train	Validation	Total	Train	Validation	Total
B	995	278	1273	9950	2780	12730
C	749	237	986	7490	2370	9860
D	821	248	1069	8210	2480	10690
Non-gibbon call	2262	754	3016	22620	7540	30160

Table 4.3: Social group model data split before and after data augmentation.

4.3.2 Neural Network Methods

The purpose of this study is to develop classifiers capable of classifying Hainan gibbon calls using deep learning methods. Section 2.2.4, presented literature on the successful applications of CNNs in bioacoustic classification problems. In this section, we will discuss how CNNs were used to develop the binary classifier and the social group classifier.

Model architecture

Following Dufourq et al. [2021], a 2-D CNN architecture was considered. As mentioned in chapter 3, CNNs specialized neural networks used to process images, thus, the selection of CNNs was based on the input of the analysis which were images of frequencies.

The spectrograms images obtained from the ten seconds preprocessed amplitude segment were used as the input. The 2-D CNN contains two convolutional layers each followed by a max pooling layer which reduces the required amount of computations and weights. Each convolutional layer used 16×16 convolutional kernels and contained one dense layer (see Figure 4.7). The final layer uses a softmax activation function for both the binary and the multi-class classification task.

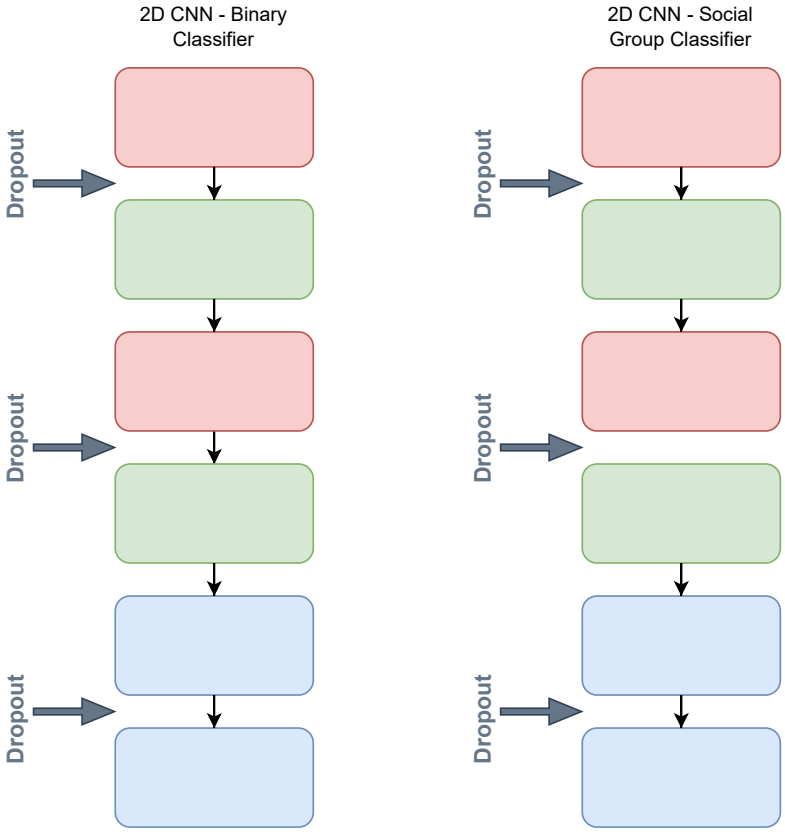


Figure 4.7: CNN architectures for the binary and the social group case.

Hyperparameter Tuning

Hyperparameter tuning discussed in section 3.1.5 was performed to obtain the optimal model parameters. The hyperparameters considered were the batch size and the dropout rate as shown in Table 4.4. Only a few hyperparameters were used due to computational constraints. The first model was trained for 30 epochs, however, no improvement in the performance was observed after the seventh epoch, thus, early stopping was applied and all the other models were trained for eight epochs using the Adam optimizer with a learning rate of 0.001.

Hyperparameters	Binary Classifier	Group classifier
Batch Size	8, 16	8, 16, 32, 64
Dropout Rate	0, 0.2, 0.4	0.1, 0.2, 0.4

Table 4.4: The different hyperparameters used for hyperparameter tuning.

Model Selection and Testing

The best model was determined based on the F1 score, which measures the quality of the classifier's predictions. These performance measures were obtained using the validation set. Finally, the best model was used to evaluate the overall performance of the classifiers by applying it on the test data. For the social group model, the best model was used to predict full audio files. The full eight hour audio files in the test files were broken into ten second segments using the sliding window approach and each segment was converted into a spectrogram using the FFT method described in chapter 2. The best model was applied on the spectrograms which predicted from which social group each audio file originated. The model should be able to predict all the Hainan gibbon calls of each audio file as one group, for example, Hainan gibbon calls from an audio file from a group D should all be predicted as D since we know that this audio file is from recorders in group D's territory.

4.4 Software Packages

The Python package librosa was used to process and extract the spectrograms from the audio files. Keras APIs were used for the deep learning procedures. All the modelling computations were performed on Google Colab using Python 3. All the code and anaylsis scripts are stored in a repository on GitHub (link found in the appendix A.1).

Chapter 5

Results and Discussion

The purpose of this research is to develop a social group classifier. To achieve this we developed a binary classifier that is able to detect gibbon calls and non-gibbon calls. The best binary classifier achieved an accuracy of 96% on the validation set, and 88% on the test set. Consequently, we developed a multi-class classifier which is able to detect from which social group each gibbon calls originates, the best social group classifier achieved an accuracy of 76% on the validation set and 63% on the test set.

This chapter presents and discusses the results and observations of the development process of the binary and social group classifier. Section 5.1 presents the outcomes of the binary classifier task, specifically, the model selection process, and the performance of the model on the test data. Section 5.2 presents the outcomes of the social group model, model selection from the hyperparameter process, final model performance on test data, and the model's ability to predict gibbon calls from a full length audio.

5.1 Image Classification: Binary Classifier

The best binary classifier had a batch size of 16 and dropout rate of 0.4. Table 5.1 gives a summary of the top four models with the validation accuracies ranging from 94% to 97%. The best model achieved an accuracy of 96.2%. This model was considered the best because it had the highest F1-score for both the gibbon and non-gibbon call, which means that the model has both high precision and recall. The best model is both precise and robust, i.e., it has a low number of false positives and false negatives. The confusion matrix in Figure 5.1 illustrates the best model's ability to detect gibbon and non-gibbon calls in the test data. The test set used for the binary model made use of only the annotated files.

5.2 Image Classification: Social Group Classifier

Table 5.2 presents the results obtained from the hyperparameter tuning process. In this study, batch size $(8, 16, 32, 64)$ and dropout rate $(0.1, 0.2, 0.4)$ were the only hyperparameters considered.

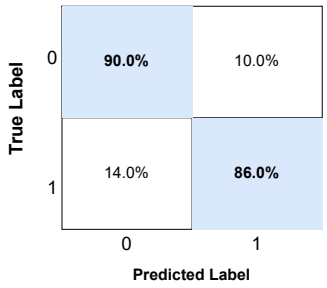


Figure 5.1: Confusion matrix for classifications made by best Hainan gibbon classification model (Model 4) on test data.

	Hyperparameters		**Model Performance Measures**		
	Batch Size	Dropout Rate	Validation accuracy	F1 Score Gibbon Call	F1 Score Non-gibbon Call
Model 1	8	0.2	97.0%	77.0%	80.0%
Model 2	16	0.2	94.4%	81.0%	84.0%
Model 3	8	0.4	96.7%	82.0%	87.0%
Model 4	16	0.4	96.2%	87.0%	90.0%

Table 5.1: Summary of the top four performing hyperparameter tuning configurations for the binary model.

We chose to explore only these hyperparamters because the network architectures were explored in the previous Hainan gibbon paper [Dufourq et al., 2021]. Therefore, we decided to focus on these hyperparameters, which were not explored in that paper.

Based on the results in Table 5.2, when the dropout rate is kept constant, increasing or decreasing the batch size yields an increase in the validation accuracy. Additionally, when the batch size is kept constant, decreasing the dropout rate results in an increase in the validation accuracy. To ensure that the results obtained were not based on weight initialisation, four runs were done for each hyperparameter and the results were the same for each run. The validation accuracy for all the top six models ranged from 72%–76%, with the best model (Model 5) obtaining a validation accuracy of 76.3%. The best model obtained the highest F1 score for all the social groups. This is important because the best model should be able to correctly predict all the social groups. Table 5.2 illustrates that the best model tends to do well at predicting groups C, D, and non-gibbon call. However, the best model struggles with classifying group B as it has the lowest F1 score. Furthermore, Figure 5.2 shows that the best model tends to correctly classify groups C, D, and the non-gibbon group compared to group B, which is usually misclassified as either D or non-gibbon. The misclassification may have been caused by the call types distributions for the three social groups being similar (see Figure 4.6).

The main aim of this study is to predict from which social group each call originates, thus, the best model was applied to 18 full eight-hour audio files from the different social groups. The eight-hour audio files were broken into ten second segments and each segment was predicted as either B, C, D or non-gibbon. Table 5.3 shows how the model predicted each segment. The majority of

the segments were predicted as non-gibbon because the Hainan gibbon call peak activity occurs shortly after dawn and drops rapidly throughout the morning, thus, we expect fewer gibbon calls compared to non-gibbon calls. The model output results for the test set in Table 5.3 contradict the results shown in the confusion matrix shown in Figure 5.2, since the majority of the gibbon call segments were predicted as C for audio files from social groups B and D. The model performs well on the validation data and poorly on the real test data which is somewhat strange since both the validation and test sets are unseen files (not used for training). The contradiction may have been caused by the call type distribution being similar between the three social groups. There is also a possibility that the model is incapable of distinguishing between the social group, therefore it predicts all the gibbon calls as one social group. Furthermore, as mentioned in chapter 4.3.2 there was a lack of annotated data for most group C audio files and some group B and D audio files which means there was no way of telling how many calls were in these audio files. However, what is known is that any call that comes from an audio file from a certain social group territory can only come from that social group, for example, it is known that all audio files from group C's territory can only come from there. Thus, these audio files can be used to make some inference about the model's ability, even though they are not annotated. It is important to note this is unlike most datasets, it is unique to the survey of this gibbon population. There is no way to assess true calls missed by the model (false negatives), and there is limited ability to infer if what is being predicted is in fact actual calls (false positives). All that is known is that there are some calls in the audio recordings and they should be classified as group C.

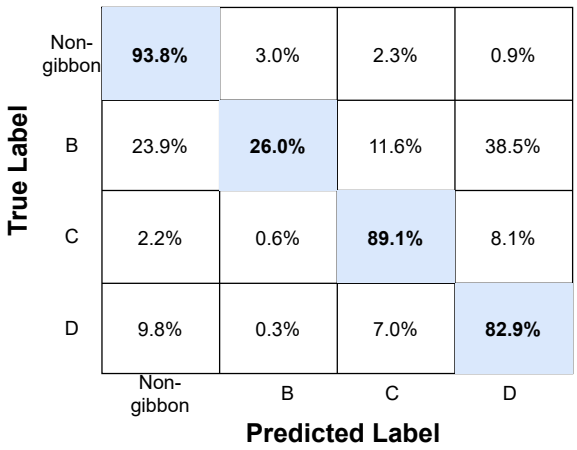


Figure 5.2: Confusion matrix for classifications made by best Hainan gibbon classification model (Model 4) on validation data.

	Hyperparameters		Model Performance Measure				
	Batch Size	Dropout Rate	Validation accuracy	F1 Score B	F1 Score C	F1 Score D	F1 Score Non-gibbon Call
Model 2	8	0.2	75.7%	45.0%	70.6%	60.4%	90.8%
Model 3	16	0.4	74.6%	45.8%	62.0%	60.1%	90.7%
Model 4	8	0.4	73.0%	54.1%	38.8%	61.3%	90.3%
Model 5	16	0.1	76.3%	54.2%	67.8%	68.9%	89.6%
Model 7	64	0.2	76.5%	39.8%	75.5%	61.1%	91.6%
Model 8	64	0.4	72.7%	41.8%	45.4%	61.5%	91.8%

Table 5.2: Summary of the top six performing hyperparameter tuning configurations for the multi-class model.

		Model Output			
Audio Files	**Social Group**	**Non-gibbon**	**B**	**C**	**D**
HGSM3A_0+1_20150625_050700	D	28653	0	138	0
HGSM3A_0+1_20160314_055200	D	28750	0	38	3
HGSM3A_0+1_20160316_055100	D	27239	0	1	0
HGSM3A_0+1_20160317_055000	D	28791	0	0	0
HGSM3D_0+1_20160404_053500	D	28726	0	45	20
HGSM3D_0+1_20160429_051600	D	28637	0	147	7
HGSM3B_0+1_20150803_052000	B	28351	0	440	0
HGSM3B_0+1_20150919_053100	B	28246	0	545	0
HGSM3B_0+1_20150920_053200	B	28720	0	71	0
HGSM3B_0+1_20150923_053200	B	28385	0	406	0
HGSM3B_0+1_20160315_055200	B	28356	43	363	27
HGSM3B_0+1_20160317_055000	B	28783	0	3	5
HGSM3C_0+1_20150808_052200	C	27906	0	884	1
HGSM3C_0+1_20150907_052900	C	24914	0	3877	0
HGSM3C_0+1_20150908_052900	C	25048	0	3743	0
HGSM3C_0+1_20150909_052900	C	23961	0	4830	0
HGSM3C_0+1_20150910_053000	C	24708	0	4082	1
HGSM3C_0+1_20150911_053000	C	25355	0	3425	11

Table 5.3: Summary of full audio file predictions.

Chapter 6

Conclusion

6.1 Summary and Conclusion

The aim of this dissertation was to develop a classifier which is able to detect Hainan gibbon calls from audio files and classify them according to their social groups. The trained classifier was trained to differentiate between gibbon calls from different social groups. Calls are made to advertise territory and the hypothesis was that the calls would be specific to each group.

The research question was addressed by constructing two acoustic classification models – one to detect the presence and absence of gibbon calls and another that identifies which social group each call originates from. Practically, developing the acoustic classifiers involved converting raw audio recordings from Bawangling National Nature Reserve into spectrograms, augmenting the data and splitting the data into training, validation, and test sets, deciding on the appropriate CNN architectures, hyperparameter tuning; and model selection based on the performance on the validation set.

Independent CNN models with different architectures were used for both bioacoustic classification tasks. The best model for the binary case achieved a validation accuracy of 96.2%, F1 score of 87% for the gibbon calls, and a test accuracy of 86%. The results obtained for the binary task attest to the model's ability to distinguish between the presence and absence of gibbon calls. The best model for the multi-class case achieved a validation accuracy of 76.3%, and F1 scores of 54.2%, 67.8%, and 68.9% for groups B, C, and D respectively. Furthermore, the best model was used to assess whether it was capable of predicting gibbon calls from full length audio files and to distinguish between the social groups. The model predicted almost all of the gibbon calls as group C.

To conclude, there was no prior knowledge pertaining to differences in the calls made by the different gibbon groups. The best social group model was not able to discriminate between the social groups based on the performance on the test data, however, it was able to find differences in the validation data. This means that there might be some differences between the social groups, but our model was incapable of detecting them in the real test data, perhaps more data and modelling would result in a better test accuracy.

Classifiers capable of identifying which group each call originates from are valuable in ecology as they assist researchers in tracking group movements such as groups moving into new territories or moving into another group's territory. Additionally, these classifiers allow researchers to identify new groups. Overall, these classifiers are valuable in managing small populations like the Hainan gibbon and can be extended to other endangered species.

6.2 Directions for Future Work

Although the social group classifier developed in our study was incapable of discriminating between social groups based on their calls, we believe that the performance can be improved in the future by:

- Ensuring that there is a balanced call type distribution between social groups.
- Increasing the amount of data. CNNs tend to require huge amounts of data to generalize better, thus, increasing the amount of data may improve the model performance.
- Incorporating other data augmentation techniques, which would increase the size of the data set and help the model generalize better Stowell [2022].
- Incorporating other noise removal methods, which may assist in removing unwanted sounds, thereby improving the quality of the images [Xie et al., 2020].
- Revising the placement of recorders in BNNR to determine the best location for particular groups. Although it was thought that each recorder fell into a single group's territory, it is possible that the audio files used in our study contained calls from multiple groups.
- Enhancing computational power which would allow the use of a larger grid search to find the best hyperparamters.
- Stacking spectrograms into a multichannel input which would be processed the same way as RGB images, thereby providing the CNN with more information [Xie et al., 2022].

Appendix A

Appendices

A.1 Research Data and Code

A subset of acoustic recordings, including training and testing labels, has been stored in the Zenodo link: https://zenodo.org/record/3991714#.YQbtDWgzbIU

All the code used can be found in the GitHub repository link: https://github.com/shelovescode000/Automated-classification-of-Hainan-Gibbon-call-types-using-deep-learning

Bibliography

Aggarwal, C. C. et al. (2018). Neural networks and deep learning. *Springer*, 10:978–3.

Akbal, E., Dogan, S., and Tuncer, T. (2022). An automated multispecies bioacoustics sound classification method based on a nonlinear pattern: Twine-pat. *Ecological Informatics*, 68:101529.

Almond, R. E., Grooten, M., and Peterson, T. (2020). *Living Planet Report 2020-Bending the curve of biodiversity loss*. World Wildlife Fund.

Aly, M. and Alotaibi, N. S. (2022). A novel deep learning model to detect covid-19 based on wavelet features extracted from mel-scale spectrogram of patients' cough and breathing sounds. *Informatics in Medicine Unlocked*, 32:101049.

Aulich, M. G., McCauley, R. D., Saunders, B. J., and Parsons, M. J. (2019). Fin whale (balaenoptera physalus) migration in Australian waters using passive acoustic monitoring. *Scientific Reports*, 9(1):1–12.

Bashar, A. et al. (2019). Survey on evolving deep learning neural network architectures. *Journal of Artificial Intelligence*, 1(02):73–82.

Bateman, H. L., Riddle, S. B., and Cubley, E. S. (2021). Using bioacoustics to examine vocal phenology of neotropical migratory birds on a wild and scenic river in Arizona. *Birds*, 2(3):261–274.

Bermant, P. C., Bronstein, M. M., Wood, R. J., Gero, S., and Gruber, D. F. (2019). Deep machine learning techniques for the detection and classification of sperm whale bioacoustics. *Scientific reports*, 9(1):1–10.

Borowiec, M. L., Dikow, R. B., Frandsen, P. B., McKeeken, A., Valentini, G., and White, A. E. (2022). Deep learning as a tool for ecology and evolution. *Methods in Ecology and Evolution*, 13(8):1640–1660.

Briefer, E. F., Tettamanti, F., and McElligott, A. G. (2015). Emotions in goats: mapping physiological, behavioural and vocal profiles. *Animal Behaviour*, 99:131–143.

Brown, C. and Riede, T. (2017). *Comparative bioacoustics: An overview*. Bentham Science Publishers.

Browning, E., Bolton, M., Owen, E., Shoji, A., Guilford, T., and Freeman, R. (2018). Predicting animal behaviour using deep learning: GPS data alone accurately predict diving in seabirds. *Methods in Ecology and Evolution*, 9(3):681–692.

Browning, E., Gibb, R., Glover-Kapfer, P., and Jones, K. E. (2017). Passive acoustic monitoring in ecology and conservation.

Bryant, J. V., Brulé, A., Wong, M. H., Hong, X., Zhou, Z., Han, W., Jeffree, T. E., and Turvey, S. T. (2016). Detection of a new Hainan gibbon *(N. Hainanus)* group using acoustic call playback. *International Journal of Primatology*, 37(4):534–547.

Bryant, J. V., Zeng, X., Hong, X., Chatterjee, H. J., and Turvey, S. T. (2017). Spatiotemporal requirements of the Hainan gibbon: Does home range constrain recovery of the world's rarest ape? *American journal of primatology*, 79(3):e22617.

Caruana, R., Lawrence, S., and Giles, C. (2000). Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. *Advances in neural information processing systems*, 13.

Chai, J. and Li, A. (2019). Deep learning in natural language processing: A state-of-the-art survey. In *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, pages 1–6.

Chao, K.-W., Chao, Y.-C., Su, C.-K., Hu, N.-Z., and Chiu, W.-H. (2019). Using machine learning method to identify for frog classification. In *2019 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE)*, pages 168–171.

Chen, J., Chen, J., Zhang, D., Sun, Y., and Nanehkaran, Y. A. (2020). Using deep transfer learning for image-based plant disease identification. *Computers and Electronics in Agriculture*, 173:105393.

Chollet, F. (2017). *Deep learning with Python*. Simon and Schuster.

Christin, S., Hervet, É., and Lecomte, N. (2019). Applications for deep learning in ecology. *Methods in Ecology and Evolution*, 10(10):1632–1644.

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*.

Cochran, W. T., Cooley, J. W., Favin, D. L., Helms, H. D., Kaenel, R. A., Lang, W. W., Maling, G. C., Nelson, D. E., Rader, C. M., and Welch, P. D. (1967). What is the fast fourier transform? *Proceedings of the IEEE*, 55(10):1664–1674.

Colonna, J., Peet, T., Ferreira, C. A., Jorge, A. M., Gomes, E. F., and Gama, J. (2016). Automatic classification of anuran sounds using convolutional neural networks. In *Proceedings of the ninth international c* conference on computer science & software engineering*, pages 73–78.

Deng, H., Zhou, J., and Yang, Y. (2014). Sound spectrum characteristics of songs of Hainan gibbon *(N. Hainanus)*. *International Journal of Primatology*, 35(2):547–556.

Deng, L., Hinton, G., and Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8599–8603. IEEE.

Di Minin, E., Fink, C., Tenkanen, H., and Hiippala, T. (2018). Machine learning for tracking illegal wildlife trade on social media. *Nature Ecology & Evolution*, 2(3):406–407.

Dufourq, E., Durbach, I., Hansford, J. P., Hoepfner, A., Ma, H., Bryant, J. V., Stender, C. S., Li, W., Liu, Z., Chen, Q., et al. (2021). Automated detection of Hainan gibbon calls for passive acoustic monitoring. *Remote Sensing in Ecology and Conservation*.

Dyer, S. A. and Harms, B. K. (1993). Digital signal processing. *Advances in Computers*, 37:59–117.

Elliott, D., Otero, C. E., Wyatt, S., and Martino, E. (2021). Tiny transformers for environmental sound classification at the edge. *arXiv preprint arXiv:2103.12157*.

Enari, H., Enari, H., Okuda, K., Yoshita, M., Kuno, T., and Okuda, K. (2017). Feasibility assessment of active and passive acoustic monitoring of sika deer populations. *Ecological Indicators*, 79:155–162.

Fellowes, J. R., Chan, B., Zhou, J., Chen, S., Yang, S., and Ng, S. (2008). Current status of the Hainan gibbon *(N. Hainanus)*: progress of population monitoring and other priority actions. *Asian Primates Journal*, 1(1):2–11.

Friel, M., Kunc, H. P., Griffin, K., Asher, L., and Collins, L. M. (2019). Positive and negative contexts predict duration of pig vocalisations. *Scientific reports*, 9(1):1–7.

Fukushima, K. (1975). Cognitron: A self-organizing multilayered neural network. *Biological Cybernetics*, 20(3):121–136.

Garcia, M. and Favaro, L. (2017). Animal vocal communication: function, structures, and production mechanisms.

Gibb, R., Browning, E., Glover-Kapfer, P., and Jones, K. E. (2019). Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution*, 10(2):169–185.

Goerlitz, H. R. (2018). Weather conditions determine attenuation and speed of sound: Environmental limitations for monitoring and analyzing bat echolocation. *Ecology and Evolution*, 8(10):5090–5100.

Goicoechea, N., De La Riva, I., and Padial, J. M. (2010). Recovering phylogenetic signal from frog mating calls. *Zoologica Scripta*, 39(2):141–154.

Goodfellow, I., Bengio, Y., and Courville, A. (2017). Deep learning (adaptive computation and machine learning series). *Cambridge Massachusetts*, pages 321–359.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

Green, A., Johnston, I., and Clark, C. (2018). Invited review: The evolution of cattle bioacoustics and application for advanced dairy systems. *animal*, 12(6):1250–1259.

Guerra, V., Llusia, D., Gambale, P. G., Morais, A. R. d., Marquez, R., and Bastos, R. P. (2018). The advertisement calls of brazilian anurans: Historical review, current knowledge and future directions. *PLoS One*, 13(1):e0191691.

Guo, Q., Jin, S., Li, M., Yang, Q., Xu, K., Ju, Y., Zhang, J., Xuan, J., Liu, J., Su, Y., et al. (2020a). Application of deep learning in ecological resource research: Theories, methods, and challenges. *Science China Earth Sciences*, 63(10):1457–1474.

Guo, Y., Chang, J., Han, L., Liu, T., Li, G., Garber, P. A., Xiao, N., and Zhou, J. (2020b). The genetic status of the critically endangered Hainan gibbon *(N. Hainanus)*: a species moving toward extinction. *Frontiers in genetics*, 11.

Gupta, G., Kshirsagar, M., Zhong, M., Gholami, S., and Ferres, J. L. (2021). Comparing recurrent convolutional neural networks for large scale bird species classification. *Scientific reports*, 11(1):1–12.

Guyot, P., Alix, F., Guerin, T., Lambeaux, E., and Rotureau, A. (2021). Fish migration monitoring from audio detection with cnns. In *Audio Mostly 2021*, pages 244–247.

Hahnloser, R. H., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., and Seung, H. S. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *nature*, 405(6789):947–951.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Herborn, K. A., McElligott, A. G., Mitchell, M. A., Sandilands, V., Bradshaw, B., and Asher, L. (2020). Spectral entropy of early-life distress calls as an iceberg indicator of chicken welfare. *Journal of the Royal Society Interface*, 17(167):20200086.

Himawan, I., Towsey, M., Law, B., and Roe, P. (2018). Deep learning techniques for koala activity detection. In *INTERSPEECH*, pages 2107–2111.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hopp, S. L., Owren, M. J., and Evans, C. S. (2012). *Animal acoustic communication: sound analysis and research methods*. Springer Science & Business Media.

Huang, J., Chai, J., and Cho, S. (2020). Deep learning in finance and banking: A literature review and classification. *Frontiers of Business Research in China*, 14(1):1–24.

Huang, T. (1996). Computer vision: Evolution and promise. *Detectors and Experimental Techniques*.

Hulse, S. V., Renoult, J. P., and Mendelson, T. C. (2022). Using deep neural networks to model similarity between visual patterns: Application to fish sexual signals. *Ecological Informatics*, 67:101486.

Janisch, J., Mitoyen, C., Perinot, E., Spezie, G., Fusani, L., and Quigley, C. (2021). Video recording and analysis of avian movements and behavior: Insights from courtship case studies. *Integrative and comparative biology*, 61(4):1378–1393.

Jiang, J.-j., Bu, L.-r., Duan, F.-j., Wang, X.-q., Liu, W., Sun, Z.-b., and Li, C.-y. (2019). Whistle detection and classification for whales based on convolutional neural networks. *Applied Acoustics*, 150:169–178.

Jung, D.-H., Kim, N. Y., Moon, S. H., Jhin, C., Kim, H.-J., Yang, J.-S., Kim, H. S., Lee, T. S., Lee, J. Y., and Park, S. H. (2021). Deep learning-based cattle vocal classification model and real-time livestock monitoring system with noise filtering. *Animals*, 11(2):357.

Kahl, S., Wood, C. M., Eibl, M., and Klinck, H. (2021). Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236.

Kamath, U., Liu, J., and Whitaker, J. (2019). *Deep learning for NLP and speech recognition*, volume 84. Springer.

Khan, A., Sohail, A., Zahoora, U., and Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8):5455–5516.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kiskin, I., Sinka, M., Cobb, A. D., Rafique, W., Wang, L., Zilli, D., Gutteridge, B., Dam, R., Marinos, T., Li, Y., et al. (2021). Humbugdb: a large-scale acoustic mosquito dataset. *arXiv preprint arXiv:2110.07607*.

Knight, E. C., Poo Hernandez, S., Bayne, E. M., Bulitko, V., and Tucker, B. V. (2020). Preprocessing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks. *Bioacoustics*, 29(3):337–355.

Koehler, J., Jansen, M., Rodriguez, A., Kok, P. J., Toledo, L. F., Emmrich, M., Glaw, F., Haddad, C. F., Roedel, M.-O., and Vences, M. (2017). The use of bioacoustics in anuran taxonomy: theory, terminology, methods and recommendations for best practice. *Zootaxa*, 4251(1):1–124.

Koh, L. P., Li, Y., and Lee, J. S. H. (2021). The value of china's ban on wildlife trade and consumption. *Nature Sustainability*, 4(1):2–4.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Kroodsma, D. A., Mayorga, J., Hochberg, T., Miller, N. A., Boerder, K., Ferretti, F., Wilson, A., Bergman, B., White, T. D., Block, B. A., et al. (2018). Tracking the global footprint of fisheries. *Science*, 359(6378):904–908.

Kvsn, R. R., Montgomery, J., Garg, S., and Charleston, M. (2020). Bioacoustics data analysis–a taxonomy, survey and open challenges. *IEEE Access*, 8:57684–57708.

Labach, A., Salehinejad, H., and Valaee, S. (2019). Survey of dropout methods for deep neural networks. *arXiv preprint arXiv:1904.13310*.

Larsen, H. L., Pertoldi, C., Madsen, N., Randi, E., Stronen, A. V., Root-Gutteridge, H., and Pagh, S. (2022). Bioacoustic detection of wolves: Identifying subspecies and individuals by howls. *Animals*, 12(5):631.

Lasseck, M. (2018). Audio-based bird species identification with deep convolutional neural networks. *CLEF (working notes)*, 2125.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

LeCun, Y., Touresky, D., Hinton, G., and Sejnowski, T. (1988). A theoretical framework for backpropagation. In *Proceedings of the 1988 connectionist models summer school*, volume 1, pages 21–28.

Leitner, B. Z. J. and Thornton, S. (2019). Audio recognition using mel spectrograms and convolution neural networks.

Li, J., Dai, W., Metze, F., Qu, S., and Das, S. (2017a). A comparison of deep learning methods for environmental sound detection. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 126–130. IEEE.

Li, K., Javer, A., Keaveny, E. E., and Brown, A. E. (2017b). Recurrent neural networks with interpretable cells predict and classify worm behaviour. *bioRxiv*.

Li, P., Garber, P. A., Bi, Y., Jin, K., Qi, X., and Zhou, J. (2022). Diverse grouping and mating strategies in the critically endangered Hainan gibbon (*Nomascus hainanus*). *Primates*, 63(3):237–243.

Li, R., Garg, S., and Brown, A. (2019). Identifying patterns of human and bird activities using bioacoustic data. *Forests*, 10(10):917.

Liu, Z., Yu, S., and Yuan, X. (1984). Resources of the Hainan black gibbon and its present situation. *Chinese Wildlife*, 6:1–4.

Mann, D. A., Hawkins, A. D., and Jech, J. M. (2008). Active and passive acoustics to locate and study fish. *Fish bioacoustics*, pages 279–309.

Mao, A., Huang, E., Gan, H., Parkes, R. S., Xu, W., and Liu, K. (2021). Cross-modality interaction network for equine activity recognition using imbalanced multi-modal data. *Sensors*, 21(17):5818.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

Mcloughlin, M. P., Stewart, R., and McElligott, A. G. (2019). Automated bioacoustics: methods in ecology and conservation and their potential for animal welfare monitoring. *Journal of the Royal Society Interface*, 16(155):20190225.

Mehlig, B. (2019). Artificial neural networks. *arXiv e-prints*, pages arXiv–1901.

Mochida, K., Koda, S., Inoue, K., Hirayama, T., Tanaka, S., Nishii, R., and Melgani, F. (2019). Computer vision-based phenotyping for improvement of plant productivity: a machine learning perspective. *GigaScience*, 8(1):giy153.

Müller, B., Reinhardt, J., and Strickland, M. T. (1995). *Neural networks: an introduction*. Springer Science & Business Media.

Nanda, M. A., Seminar, K. B., Nandika, D., and Maddu, A. (2018). A comparison study of kernel functions in the support vector machine and its application for termite detection. *Information*, 9(1):5.

Newbold, T., Hudson, L. N., Arnell, A. P., Contu, S., De Palma, A., Ferrier, S., Hill, S. L., Hoskins, A. J., Lysenko, I., Phillips, H. R., et al. (2016). Has land use pushed terrestrial biodiversity beyond the planetary boundary? a global assessment. *Science*, 353(6296):288–291.

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., and Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25):E5716–E5725.

Padovese, B., Frazao, F., Kirsebom, O. S., and Matwin, S. (2021). Data augmentation for the classification of north atlantic right whales upcalls a. *The Journal of the Acoustical Society of America*, 149(4):2520–2530.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Penar, W., Magiera, A., and Klocek, C. (2020). Applications of bioacoustics in animal ecology. *Ecological Complexity*, 43:100847.

Proakis, J. G. (2001). *Digital signal processing: principles algorithms and applications*. Pearson Education India.

Pyke, G. H. and Ehrlich, P. R. (2010). Biological collections and ecological/environmental research: a review, some observations and a look to the future. *Biological reviews*, 85(2):247–266.

Rabiner, L. R. and Gold, B. (1975). Theory and application of digital signal processing. *Englewood Cliffs: Prentice-Hall*.

Raccagni, W. and Ntalampiras, S. (2021). Acoustic classification of cat breed based on time and frequency domain features. In *2021 30th Conference of Open Innovations Association FRUCT*, pages 184–189.

Raick, X., Huby, A., Kurchevski, G., Godinho, A. L., and Parmentier, É. (2020). Use of bioacoustics in species identification: Piranhas from genus pygocentrus (teleostei: Serrasalmidae) as a case study. *PLoS One*, 15(10):e0241316.

Rawat, W. and Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Rosenblatt, F. (1961). Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo NY.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Sahu, P. K., Campbell, K. A., Oprea, A., Phillmore, L. S., and Sturdy, C. B. (2022). Comparing methodologies for classification of zebra finch distance calls. *The Journal of the Acoustical Society of America*, 151(5):3305–3314.

Salamon, J. and Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal processing letters*, 24(3):279–283.

Sanchez-Gendriz, I. (2021). Signal processing basics applied to ecoacoustics. *Ecological Informatics*, 66:101445.

Sharma, S. and Sharma, S. (2017). Activation functions in neural networks. *Towards Data Science*, 6(12):310–316.

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., and Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Smith, J. O. (2008). *Mathematics of the discrete Fourier transform (DFT): with audio applications*. Julius Smith.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, 8(3):185–190.

Stowell, D. (2018). Computational bioacoustic scene analysis. In *Computational analysis of sound scenes and events*, pages 303–333. Springer.

Stowell, D. (2022). Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10:e13152.

Stowell, D., Wood, M. D., Pamuła, H., Stylianou, Y., and Glotin, H. (2019). Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. *Methods in Ecology and Evolution*, 10(3):368–380.

Sueur, J. et al. (2018). *Sound analysis and synthesis with R*. Springer.

Sugai, L. S. M. and Llusia, D. (2019). Bioacoustic time capsules: Using acoustic monitoring to document biodiversity. *Ecological Indicators*, 99:149–152.

Sugai, L. S. M., Silva, T. S. F., Ribeiro Jr, J. W., and Llusia, D. (2019). Terrestrial passive acoustic monitoring: review and perspectives. *BioScience*, 69(1):15–25.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Tabak, M. A., Norouzzadeh, M. S., Wolfson, D. W., Sweeney, S. J., VerCauteren, K. C., Snow, N. P., Halseth, J. M., Di Salvo, P. A., Lewis, J. S., White, M. D., et al. (2019). Machine learning

to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, 10(4):585–590.

Tan, L. and Jiang, J. (2018). *Digital signal processing: fundamentals and applications*. Academic Press.

Taylor, A. M. and Reby, D. (2010). The contribution of source–filter theory to mammal vocal communication research. *Journal of Zoology*, 280(3):221–236.

Teixeira, D., Maron, M., and van Rensburg, B. J. (2019). Bioacoustic monitoring of animal vocal behavior for conservation. *Conservation Science and Practice*, 1(8):e72.

Thompson, A. R., Moran, J. M., and Swenson, G. W. (2017). *Digital Signal Processing*, pages 309–390. Springer International Publishing, Cham.

Tieleman, T. and Hinton, G. (2012). Lecture 6.5-rmsprop, coursera: Neural networks for machine learning. *University of Toronto, Technical Report*, 6.

Trawicki, M. B. (2021). Multispecies discrimination of whales (cetaceans) using hidden markov models (hmms). *Ecological Informatics*, 61:101223.

Turvey, S., Traylor-Holzer, K., Wong, M., Bryant, J., Zeng, X., Hong, X., and Long, Y. (2015). International conservation planning workshop for the Hainan gibbon: final report. *London/Apple Valley MN: Zoological Society of London/IUCN SSC Conservation Breeding Specialist Group*.

Tzirakis, P., Shiarella, A., Ewers, R., and Schuller, B. W. (2020). Computer audition for continuous rainforest occupancy monitoring: the case of bornean gibbons' call detection. In *Interspeech*.

Vehrencamp, S. L. and Bradbury, J. (1998). *Principles of animal communication*. Sinauer Associates Sunderland.

Wäldchen, J. and Mäder, P. (2018). Plant species identification using computer vision techniques: A systematic literature review. *Archives of Computational Methods in Engineering*, 25(2):507–543.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, 3(1):1–40.

Wild, B., Sixt, L., and Landgraf, T. (2018). Automatic localization and decoding of honeybee markers using deep convolutional neural networks. *arXiv preprint arXiv:1802.04557*.

Williams, E. M., O'Donnell, C. F., and Armstrong, D. P. (2018). Cost-benefit analysis of acoustic recorders as a solution to sampling challenges experienced monitoring cryptic species. *Ecology and Evolution*, 8(13):6839–6848.

Wilson, E. O. (2017). Biodiversity research requires more boots on the ground. *Nature Ecology & Evolution*, 1(11):1590–1591.

Wrege, P. H., Rowland, E. D., Keen, S., and Shiu, Y. (2017). Acoustic monitoring for conservation in tropical forests: examples from forest elephants. *Methods in Ecology and Evolution*, 8(10):1292–1301.

Xie, J., Colonna, J. G., and Zhang, J. (2020). Bioacoustic signal denoising: a review. *Artificial Intelligence Review*, pages 1–23.

Xie, J., Hu, K., Zhu, M., Yu, J., and Zhu, Q. (2019). Investigation of different cnn-based models for improved bird sound classification. *IEEE Access*, 7:175353–175361.

Xie, J., Towsey, M., Zhang, J., and Roe, P. (2016). Acoustic classification of australian frogs based on enhanced features and machine learning algorithms. *Applied Acoustics*, 113:193–201.

Xie, J., Zhu, M., Hu, K., Zhang, J., Hines, H., and Guo, Y. (2022). Frog calling activity detection using lightweight cnn with multi-view spectrogram: a case study on kroombit tinker frog. *Machine Learning with Applications*, 7:100202.

Xu, B., Wang, N., Chen, T., and Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*.

Ying, X. (2019). An overview of overfitting and its solutions. In *Journal of Physics: Conference Series*, volume 1168, page 022022. IOP Publishing.

Zhang, M., Fellowes, J. R., Jiang, X., Wang, W., Chan, B. P., Ren, G., and Zhu, J. (2010). Degradation of tropical forest in Hainan, China, 1991–2008: Conservation implications for Hainan gibbon *(N. Hainanus)*. *Biological Conservation*, 143(6):1397–1404.

Zualkernan, I., Judas, J., Mahbub, T., Bhagwagar, A., and Chand, P. (2020). A tiny cnn architecture for identifying bat species from echolocation calls. In *2020 IEEE/ITU International Conference on Artificial Intelligence for Good (AI4G)*, pages 81–86. IEEE.