Automatic Classification of Multiple Acoustic Events Using Artificial Neural Networks

Dineth Egodage

188015M

Thesis submitted in partial fulfillment of the requirements for the
Degree of Master of Science (Research) in Computer Science and Engineering

Department of Computer Science & Engineering

University of Moratuwa

Sri Lanka

January 2021

DECLARATION

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature of the candidate: Date:

The above candidate has carried out research for the MSc thesis under my supervision.

Name of the supervisor:

Signature of the supervisor: Date :

ACKNOWLEDGEMENTS

First of all, I would like to express my gratitude towards the Department of Computer Science and Engineering offering the Master's Degree programme to perform this kind of work.

Secondly, I would like to express my heartiest gratitude towards my research supervisor Dr. Sulochana Sooriyaarachchi for the guidance and immense support provided in both knowledge and resource-wise to make this research a success.

I would also like to thank Dr. Navinda Kottege and Dr. Chandana Gamage for guiding and advising us to stay on the correct path at the initial stages of this research. The Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia, should also be mentioned thankfully for providing this research idea and Dr. Kottege as a resource person for us.

Furthermore, I would like to convey my gratitude towards Dr. Sampath Seneviratne from Animal Communications, Genetics, Evolutionary Biology section of the University of Colombo for sharing his expertise domain knowledge and spend his valuable time to guide me through the research.

Besides, I am grateful to the progress panel members, Dr. Ranga Rodgrigo and Dr. Charith Chithrarajan, for giving their valuable feedback on the work carried out.

Finally, I would like to express my gratitude towards Dr. Charith Chithrarajan as the Research Coordinator and all other staff members for the knowledge given throughout the last two years.

Finally, I would like to express my heartiest gratitude towards my friends and family for the immense support given throughout the research. Special thanks to my wife, who remained supportive until the end.

ABSTRACT

There are numerous scenarios where similar acoustic events occur multiple times. Acoustic monitoring of migratory birds is an ideal example. Birds make a type of call known as flight calls during migration. A flight call can be considered as *an acoustic event* because it is a short-term, intuitively distinct sound. It is challenging to identify multiple occurrences of extremely short-range acoustic events such as flight calls in real-world recordings using classification techniques that require more computational power. It is mainly due to background noise and complex acoustic environments. This research aims at developing a classification model that reduces the effect of background noise, extract ROIs from continuous recordings, extract suitable features of flight calls and detect multiple occurrences of flight calls. An improved algorithm that can extract features has been developed in this research—by combining a well known Maximally Stable Extremal Regions (MSER) technique with state of the art traditional techniques. Namely Spectral and Temporal Features(SATF) and a combination of SATF and Spectrogram-based Image Frequency Statistics(SIFS). We name this novel algorithm as Spectrogram-based Maximally Stable Extremal Regions (SMSER). Three distinct feature sets have formed such that Featureset-1 created using SATF. Featureset-2 is a blend of SATF and SIFS. Featureset-3 is a combination of SATF, SIFS, and SMSER. The kNN, RF, SVM, and DNN classification techniques evaluated a real-world dataset using the extracted feature sets. Research carried out several tests to find out the best performing combination of classification model and feature set. The results showed that the flight calls' detection accuracy increased when the number of features increased, although high computational power requirement is a disadvantage. The performance of SMSER feature set was the best among almost every classification technique above. It should be because the SMSER Feature set has the highest number of features. Classification of the SMSER feature set from the DNN classifier showed the highest accuracy of 87.67%.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Networks
DNN	Deep Neural Network
DFT	Discrete Fourier Transformation
DWTC	Discrete Wavelet Transform Coefficients
FFT	Fast Fourier Transformation
GMM	Gaussian Mixture Models
GPU	Graphics Processing Unit
HMM	Hidden Markov Models
k-NN	k-Nearest Neighbor
MFCC	Mel Frequency Cepstral Coefficients
PC	Personal Computer
PCA	Principal Component Analysis
RF	Random Forest Algorithm
RFE	Recursive Feature Elimination
RMSE	Root Mean Square Energy
SATF	Spectral and Temporal Features
SIFS	Spectrogram-based Image Frequency Statistics
SMSER	Spectrogram-based Maximally Stable Extremal Regions
SVM	Support Vector Machine
ZCR	Zero-Crossing Rate

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

1.1 Background

There are thousands of sounds generated by different types of sources in our environment. Each sound occurring in the environment can be identified as a combination of acoustic events. An *acoustic event* is any short-term, intuitively distinct occurrence of a sound such as a door knock, a horn sound, or a dog bark. There are numerous scenarios where similar acoustic events occur multiple times in a sound recording. For example, urban sound samples may consist of repeated vehicle horns or engine sounds.

Acoustic events which animals produce are known as bio-acoustic events. Animals are responsible for a considerable amount of bio-acoustic events in the natural environment. Biologists use acoustic behaviours of animals to obtain important perceptions such as census in the animal population, determine the state of our environment, change in biodiversity and make decisions on animal conservation. Bio-acoustics monitoring is the investigation process of animals (including humans) to gather information on the social dynamics of sound-producing animals, which will help to trace habitat fragmentation and climate change. Therefore bio-acoustics monitoring is an essential field for the long-term assessment and evaluation of the environment. Also, bio-acoustics monitoring can be applied for Audio diarization and tourism as well.

Bio-acoustics monitoring is a complex process because sounds generated by different animals are diverse by qualities of sounds and sound-producing mechanism. For example, many fish sounds are produced by rotating bones or teeth against each other. Swim bladder act as a resonating cavity at sometimes. Frogs produce species-specific, highly vocal sounds with the help of air moving between the lungs and mouth. Birds use the syrinx to produce sounds. Syrinx is a special

region which is at the lower end of the trachea. Mammals produce vocalizations with the help of the Larynx. On the other hand, Larynx is a modification of the upper end of the trachea. Both birds and mammal groups use the mouth for filtering sound or possess special out-pouching or resonating of the oesophagus (trachea) which works as resonating cavities[2].

Among different animal types, birds are unique and have rich vocalizations. They usually are scattered over a vast spatial domain—birds vocalizations used for biodiversity monitoring as an excellent indicator. Therefore, bio-acoustics monitoring of birds has been a growing research area in the last decades. Bio-acoustics monitoring of birds is the most challenging area of research in bio-acoustics monitoring. One reason is high complex variations in bird sounds. This complexity occurs because of the biological arrangement of birds' sound-producing systems. As mentioned previously, the main sound produces part of a bird's syrinx shown in Fig. 1.1. It is placed at the lower end of the trachea. In most bird species, the syrinx is bipartite. Birds can produce two sounds at the same time and overlap them by this bipartite syrinx. We can see a wide variation of sounds even within the same bird species because of this complex arrangement.

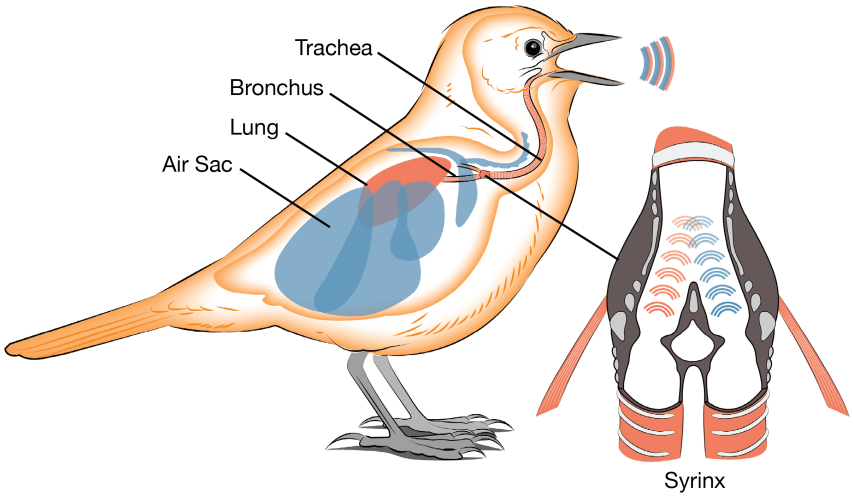

Figure 1.1: Syrinx of a bird and its placement[1]

Birds use several vocalizations in various behavioural contexts for different

purposes, such as maintaining communication with a respective social group, warning about predators, and requesting parental care. Furthermore, birds use vocalizations for territorial defence. They are different calls: alarm calls, begging calls, contact calls, bird songs, and flight calls.

- *Alarm calls* are usually sharp, short and loud. These calls can be heard from a long distance. Often used to alert about a threat or danger they feel threatened.

- *Begging calls* have low sounds. These calls can include chirps, tweets and whines like sounds. Often produced by young chicks to take the attention of their parents.

- *Contact calls* are simple sounds and moderately loud chirps. These calls are used to alert other birds about food sources or to keep in touch with each other in the flock.

- *Bird songs* are familiar to most of us. They can be tuneful and elaborate. Bird songs are distinctive to different species and easy to identify.

- *Flight calls* are often used by birds to manage the flocks together in long sustained flights. These calls are musical.

Many passerine birds and their relative's utter flight calls. Especially during migration, where birds use these species-specific vocalisations in long sustained flights. During migration, multiple bird vocalisations are used in various contexts of behaviours to serve a different purpose. They maintain proximity with the social group, express warnings about potential predators, and request attention for parental care from adults. Birds primarily use these vocalisations during sustained flights, which is a characteristic of migration. The vocalisations used during the migration are known as flight calls. Flight calls have specific characteristics as per [3],

- Flight calls could be tonal or frequency modulated

- Flight calls are generally monosyllabic
- Flight calls can have a duration of 50 - 300 ms
- Flight calls can range from 1 to 11 kHz

Flight calls are different from bird songs and alarm calls. Flight calls are relatively simple vocalisations which exhibit a pattern of rapid frequency sweeps. Therefore flight calls exhibit similarities compared to other complex vocalisations that typically birds generate.

Following are few applications which can be achieved if we can observe multiple occurrences of similar acoustic events,

- Identifying a fingerprint for an event in multiple audio recordings
- Identifying flight calls of migratory birds

In this research, bird flight calls have been considered as a case study for multiple occurrences of similar acoustic events. There is a group of birds that emit flight calls at the same time when a flock of birds are migrating. Improved knowledge on these flight calls could help different applications, such as in mitigating the impact on migratory birds from wind power generation, tracking the movement of species in seasonal migrations, and estimating bird density in-migration from vocalisation counts.

However, the majority of monitoring of these vocalizations are performed by expert humans manually. Human monitoring can have many obstacles: (1) variable observer skills which will affect the detection and classification, (2) varying observer effort and biological phenomena can lead to temporal mismatches, and (3) varying observer coverage and biological distributions can lead to spatial mismatches. Automated techniques are required to overcome these limitations in bio-acoustics monitoring using existing manual methods.

Based on previous studies, most of the tools available to investigate bird migration depend on various information sources to generate characteristics of bird gestures. The sources are,

- *Weather surveillance radar* has been used to supply awareness of the direction, speed of bird movements and density. However there is lack of information regarding the species certainly migrating [4].

- *Crowd-sourced human observations* can only be conducted during the daytime and have a lot of limitations when studying nocturnal migration [5].

- *Tracking devices* have been used to supply information about locale and activity of an individual. However there is a lack of information about the behaviour of the population.

- *Video processing* can be used to track the density during the diurnal migrations but fails in nocturnal migrations.

- *Automatic bio-acoustic monitoring* can be used to track the density, composition, population behaviour, diurnal migration and nocturnal migration as well.

	Density	Composition	Population Behaviour	Diurnal Migration	Nocturnal Migration
Weather surveillance radar monitoring	✓	X	X	✓	✓
Tracking devices	X	X	X	✓	✓
Video processing	✓	X	X	✓	X
Audio processing	✓	✓	✓	✓	✓

Figure 1.2: Comparison of current tools to study migration

As shown in Fig. 1.2 *Automatic bio-acoustic monitoring* is the solution that can be used to evaluate all the aspects of migration. Furthermore, automatic bio-acoustic monitoring has been identified in [6] as a scalable solution to produce species-specific information and population behaviours, which were unfeasible to gain from any of the above procedures.

When it comes to automatic classification of avian migratory vocalizations, several methods have been identified for this task. Most of these methods used to gain details of migrating patterns in human interest areas such as wind farms. Analyses of flight calls from automation, however, has been only partially explored. Lack of available databases and domain knowledge of several domains lead to lack of automation in this area. Up to now, most of the studies have merged automatic and manual proceedings, to extract species-specific migrating data [7] and estimates of species richness[8], and to permit comparative analyses among species[9]. However, a more efficient analysis tool for flight call will enhance the species monitoring across a sizable ecological scale. The automatic classification of flight calls will cover a great use case. It will increase the analysis efficiency, facilitating the conservation and management of migratory birds efficiently due to enhanced knowledge.

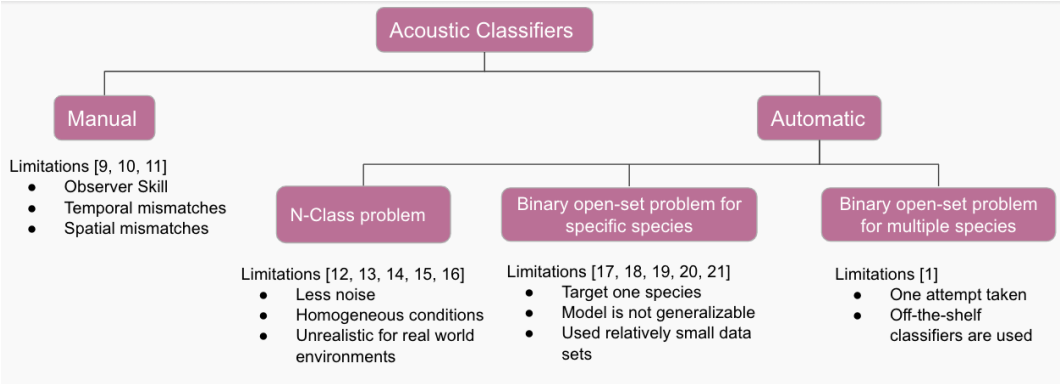

Figure 1.3: Breakdown of previous research done in flight call classification

Fig. 1.3 shows a breakdown of previous research done in the area of flight call classification. Acoustic classifiers found with both manual and automatic approaches. Manual classification [10, 11, 12] is done using expert human knowledge of the domain, and thereby no systems involved in the classification process. The automatic acoustic classifiers can be found addressing three types of problems. A number of systems target to address the N-class problem[13, 14, 15, 16, 17]. Also the binary open-set problem for specific species[18, 19, 20, 21, 22]. Last type is a binary open-set problem for multiple species[2].

Training a model to classify a given clip to N classes assuming that the dataset

contains clips of only N number of species is known as N-class problem where there is much research done in this area. Classification accuracy of 97.6% has been obtained in [14] for N class problem scenario owing to the use of feature learning techniques. The Binary open-set problem for specific species is, to classify whether a vocalization of a specific species is present in a set of sound clips or not. Where this method also has been explored by a lot of research work. Finally, the binary open-set problem for several species is, to classify a set of sound clips on the presence of vocalization of multiple species (in general) or not.

Sara et al. [3] shows that flight calls are extremely short, can be modulated in frequency or tone, and are generally monosyllabic, ranging from 1 to 11 kHz in her evaluation of approaches for classification based on similarities of flight calls of four species. To classify the flight calls in general regardless of the species in the real world would be falling under *Binary open-set problem for several species*. A general classifier that can identify the flight calls' similarities must be implemented to solve the binary open set problem for several species. The closest attempt in previous research related to the above situation is [2]. They have used energy-based detectors, spherical k-means, support vector machines and deep convolutional networks to classify the flight calls after augmentation and reached a classification accuracy of 95% binary accuracy.

This area has been less explored with novel features, feature extraction techniques, Region of interest(RoI) extraction techniques, noise reduction techniques, Etc. This study aims to find novel approaches that can apply to the same scenario with less computational power that could be used in developing countries as well. When addressing the real-world scenarios monitoring a continuous recording and noise reduction will also be a challenge. Noise will play a big role in a continuous recording in the real-world environment. Moreover, to process a continuous recording would be hard without segmenting it to small chunks of Regions of interest.

1.2 Problem

Considering all the aspects mentioned above the problem of this study can be defined as follows,

- In general terms: Implementing a novel approach for automatic acoustic based detection of *multiple occurrences* of *similar acoustic events* in *continuous recordings* with varying background noise conditions using signal processing and classification techniques with less computational power.
- Use case: Implementing a novel approach to detect the existence of a flight call generally regardless of species in continuous recordings with varying background noise conditions and using classification techniques with less computational power.

1.3 Research Objectives

- To identify suitable techniques to reduce the effect of varying background noise conditions
- To identify suitable techniques to extract ROI from a continuous recording
- To identify suitable novel features extraction techniques to detect a flight call
- To design and develop Classification Models to detect flight calls in real world environment
- To evaluate the performance of the developed technique using datasets

1.4 Contributions

The current research work has been accepted by the IESL Journal. Moreover, it is waiting to be published right now. In this research work, a novel feature extraction method has been identified, and it requires less computational power

due to the novel approach. There is a lot more to be explored in this specific area of study.

1.5 Thesis Outline

Chapter one of the thesis elaborates on the introduction of the work and explains 'What is it all about?'. Chapter 2 starts with the literature review of the past research done in this area. The methodology of the project is discussed in Chapter 3. Chapter 4 outlines the experimental evaluation and discussion of results. Chapter 5 concludes the work done. Finally, all the references have been listed.

Chapter 2

LITERATURE REVIEW

'Automatic acoustic detection of animals' has been a popular research area throughout the last few decades. Prior work on 'automatic acoustic detection of animals' has been conducted targeting several areas such as birds, amphibians, and marine mammals. As per the introduction, this research work is about 'Automatic detection of multiple occurrences of similar acoustic events' where the use-case would be to evaluate: 'detect the existence of a flight call generally regardless of species'. Therefore this section summarizes and critically evaluates recent literature in preprocessing, segmentation, feature extraction and classification techniques which are the standard procedures observed in similar work.

2.1 Flight calls

'Flight calls' are uttered by many passerine birds and their relatives. Passerines include more than half of all bird species. They are perching birds known as songbirds. Most of them can sing very well and are small in size. These species-specific vocalizations are known as 'flight calls' emitted primarily during long sustained flights, especially during migration.

More details about flight calls for further conservation research of migratory birds has been evaluated in-depth in [23]. They describe flight calls as species-specific vocalizations. Those vocalizations can be pure to several syllables or frequency modulated that normally lie between 1kHz to 9kHz frequency band. The span of the flight call has a duration between 50 to 300 ms. Further [23] states that flight calls are different from bird songs. Flight calls are also different from other known bird vocalization types like short calls such as alarm calls and 'chip' notes.

2.2 Preprocessing and Segmentation

Preprocessing is an integral step in 'Automatic acoustic detection' related prior work. An audio recording will consist of incomplete and inconsistent noisy data when an audio recording is captured from a real-world environment. The preprocessing operation has helped reduce the noise components of a signal in prior work as well. The research in the 'bird vocalization processing domain' also has used various methodologies for preprocessing. Some research work has used high-quality audio for bird vocalization recognition with high signal to noise ratio. [24] has focused on segmentation and parametric representation of bird songs with high-quality audios. In [25] semi-automatic classification of bird, song vocalization did use high-quality audio without considering preprocessing. As this is not the case in the real world, preprocessing the audio is an integral step for identifying species in a complex acoustic environment.

Bardeli et al. [20] employs a down-sampling mechanism to increase the effectiveness. Detection of Savi's warbler(bird) with noise has reached an overall detection rate of 79.06%, and 92% overall detection rate has reached for Savi's warbler calls without noise. The above scenario shows how noise reduction plays a role in increasing the detection rate. The noise component can be eliminated by subtracting the estimated noise from a low-pass filter, form signal. Therefore it is necessary to focus on preprocessing, to improve recognition rates in automatic bird sound identification.

Ruiz, Mauricio and Castellanos used a combination of STFT decomposition and two-dimensional Wiener filter to a smoothed and denoised spectrum[26]. They have tested this method with 15 bird species. According to their results, there was no significant difference between the proposed method and its baseline.

Another research has used a continuous time-domain algorithm which mitigates the consequence of varying background noise conditions to do segmentation while reducing the effect of background noise [27]. In their proposed method the initial step would be to calculate the decibel scaled energy envelope of a signal. They have used a frame size of 3 ms, an overlap between frames of 50% and the

global energy minimum was used to estimate the noise. Moreover, the syllables were threshold-ed by half of the noise estimate. The noise estimates have been calculated using non-syllable frames. The estimated new threshold has been used to search for new syllables.

Research work has been done to evaluate the effect of the bird's distance from the monitoring station (field microphone) as expressed by different signal-to-noise ratios, to the species classification performance [28]. After the audio acquisition stage, the signal was decomposed to overlapping feature vectors of constant length, using spectral and temporal audio parameterization algorithms. The sequence of feature vectors has been used as input to the bird activity detection block. The audio signal was binarily segmented to intervals with or without bird vocalizations. When the signal-to-noise ratio is low (-6dB), they have achieved a maximum of 76.58% accuracy and, when it is high (20dB), they reached up to 93.02% accuracy.

Automatic Detection and Recognition of Tonal Bird Sounds in Noisy Environments [29] paper presents short-time magnitude spectrum and sinusoidal distance-based method to do preprocessing. To detect the region of interest in the time domain, which consists of bird vocalizations; the spectral shape has been used to identify sinusoidal components in the short-time spectrum. Experiments conducted with both real-world and white noise corrupted sound recordings of birds. The detection method proposed has shown a high accuracy of detection for bird call segments.

In the classification of Killer Whale calls [30] audio signals segmented manually. However, this approach is not valid in fully automatic audio characterization and classification methods.

Segmentation of an audio recording to the level of individual syllables has been conducted in [31], with the help of an iterative time-domain algorithm. Initially, the signal's light energy envelope was constructed; then, minimum energy(global) selected. Variable NdB has been defined as the initial background noise level estimate. Moreover, the initial threshold TdB has been set to 1/2 into the initial noise level. The initial noise level is assumed as the lowest signal envelope energy.

Until the noise level is sufficiently stable; the noise level and threshold updated by the algorithm defined below until convergence.

- Identify the regions which surpass syllable threshold TdB (syllable candidates).
- Update NdB based on the difference between the syllable candidates.
- Adjust the threshold accordingly, ex:- TdB=NdB/2, and Finally return to step 1.

Almost all of the above-mentioned prior research has conducted preprocessing methods and segmentation methods. It shows that conduct of the preprocessing and segmentation steps are vital. Most of the researchers above do not cover specifically on flight calls. Therefore in this research focus will be given to preprocessing and segmentation of audio clips containing flight calls. Finally, it can be evaluated by the techniques which increased the detection rates of flight calls.

2.3 Feature Extraction

After all, preprocessing a parametric representation of the segments is needed. Feature extraction and selection is a crucial step when considering automatic bioacoustics monitoring as a classification problem. Many combinations of features have been used in previous research. Some of these features are standard features used in signal processing and music information retrieval. Some have introduced novelty features needed for recognition. In general, feature types can be divided into spectral, temporal, low-level signal parameters and novel. Different representations have advantages and disadvantages based on the sound type they are trying to represent.

2.3.1 Mel-Frequency Cepstral Coefficients (MFCC)

Even though bird species' sound production mechanism, well known with the availability of bipartisan syrinx; specific features based on the production mechanism of bird sound have not been found. MFCCs are implemented based on

the mechanism of sound production of human vocalization. Since MFCCs are used widely in automatic speech and speaker recognition[32]. Sounds and speech generated by humans are filtered by the vocal tract's shape, including the tongue, teeth and other organs. This shape determines what sound comes out. If there is a mechanism to determine the vocal tract's shape accurately, it should give us an accurate representation of the produced phoneme.

This paragraph explains how MFCC works. The first step of the MFCC calculation signal will divide into small frames. If the time-domain signal represented as $S(n)$ let framed signal represented as $S_i(n)$ where i is the frame number. Then discrete Fourier Transform of signal $(S_i(k))$ is taken. Using the values obtained by the discrete Fourier Transform the Periodogram estimate of the power spectrum $(P_i(k))$ calculated.

$$S_i(k) = \sum_{m=1}^{N} s_i(n)h(n)e^{j2\pi kn/N}$$

$$P_i(k) = \frac{1}{N}|S_i(k)|^2 \tag{2.1}$$

Periodogram estimate of the power spectrum has motivated the human cochlea (an organ in the ear), which vibrates at different spots depending on the incoming sounds' frequency. Depending on the location, in the cochlea that vibrates various nerves; fire informing the brain that certain-frequencies are present. Periodogram estimate performs a similar job. The next step is Mel-spaced filterbank calculation. These consist of triangular filters. Steps of creating Mel-spaced filter banks is as follows:

1. Using equation 2.2, convert the upper and lower frequencies to Mels. If M filter bank is needed; M+2 equally spaced points in the Mel scale should generate in between lower and upper frequencies.

$$M_i(f) = 1125ln(1 + f/700) \tag{2.2}$$

2. Then those equal spaced Mel scale values are converted back to Hertz using

to equation 2.3. As the frequency resolution may not have all the M+2 values, those values are rounded off to nearest frequency bins. To convert the frequencies to FFT bin numbers sample rate needed to be known.

$$M^{-1}(m) = 700(exp(m/1125) - 1) \tag{2.3}$$

3. As the final step, triangular filter-banks are calculated using equation 2.4. For example, the first filterbank will start at the first point, reach its peak at the second point, and then return to zero at the 3rd point. A graphical representation of Mel-filter bank is shown in Fig 2.1

$$H_m(k) = \begin{cases} 0 & ; \quad k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)} & ; \quad f(m-1) \leq k \leq f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)} & ; \quad f(m) \leq k \leq f(m+1) \\ 0 & ; \quad k > f(m+1) \end{cases} \tag{2.4}$$

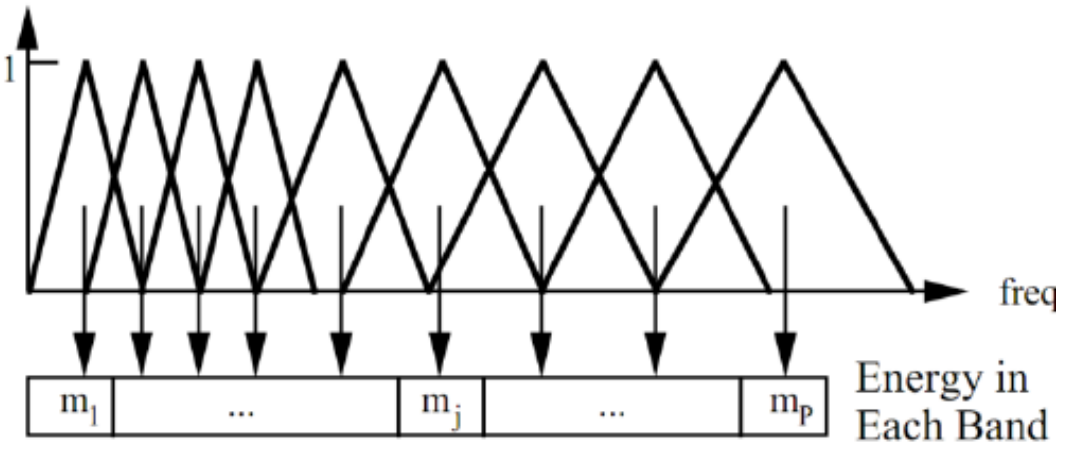

Figure 2.1: Mel-Filterbank

After calculating the Mel-filter bank to calculate filterbank energies, each filterbank multiplied with the power spectrum; coefficients are added. The resulting M values will represent how much energy presented in the filter bank. Then log of each M coefficient will take. Then Discrete Cosine Transform (DCT) of the M log filterbank energies to give M cepstral coefficients. In Human voice recog-

nition, only first 12-13 coefficients have been used as higher coefficients degrade Automatic Speech Recognition (ASR) performance [33].

Though the birds sound production mechanism is complex than human sound production, the MFCC technique has tried out in previous research. Maina and Ciira Wa[34] have used 19-dimensional Mel-Frequency Cepstral Coefficients (MFCC) to identify 19 bird species with an accuracy of 50%. 645 recordings of 19 bird species available in the African continent have split between training and testing dataset. GMM has used as a classification algorithm. MFFC coefficients based approach was used to classify the four bird species in 2016[35] with an overall accuracy of 64%. Before conducting feature extraction, they have windowed the signal using the hamming function to make the speech signal nearly stationary to the audio signal. Twenty triangular filters have used for extracting MFCCs; they have used the first 13 coefficients as features. 70% of the total collected samples have used as training data while 30% have used for testing. When we critically evaluate the above two kinds of research [34] [35] we can conclude using higher MFCC coefficients above 13 degrading results as it has happened in ASR.

2.3.2 Standard Temporal and Spectral Features

Apart from MFCCs, other temporal and spectral features also have been used in bioacoustic monitoring applications. A comparison between MFCCs with Spectral ridge features were conducted with 24 bird species in 2015[36]. The first step is, with the use of short-time Fourier transform(STFT), all the recordings have converted to spectrograms. The second step is to conduct a ridge detection method to the spectrogram and detect points of interest. A new representation region has been explored based on detected points of interest. The new representation region consists of various bird vocalizations and a local descriptor. The local descriptor includes information calculated for each sub-region: frequency bin entropy, temporal entropy, and histogram of counts of four ridge directions. Above parametric representation has been used as a six-dimensional feature vec-

tor. Details of the features as follows:

1. Temporal entropy (H_t): The ridge magnitudes are summed frame-wise over all frames in the region and the N values normalized to unit sum. H_t calculated as:
$$H_t = \sum_{i=1}^{N} \pi \log_2 \pi \tag{2.5}$$

2. Frequency bin entropy (H_f): Similar to the calculation of H_t except that the ridge magnitudes are summed bin-wise over all bins in the region. H_t and H_f describe the spatial distribution of spectral ridges in a region.

3. Histogram of counts of four ridge directions: A bird call has been divided into a grid of non-overlapping square blocks of size 11 by 11, termed as regions. Four-dimensional vector has derived from the counts of region cells belonging to ridges having direction 0, $\pi/4$, $\pi/2$ and $3\pi/4$.

To speed up the retrieval process; indexing has been carried out, and retrieved results have been ranked according to similarity scores. The experimental results show that the proposed feature set can achieve 0.71 in terms of retrieval success rate, which outperforms spectral ridge features alone (55%) and Mel frequency cepstral coefficients (36%).

A mixture of 13 MFCCs with temporal features and spectral features have been proposed in[37] which tested 33 bird species with a maximum accuracy of 83.88%. In here a feature vector of length twenty-seven has been used. The feature vector consists of four-pitch features(standard deviation, maximum, minimum and average pitch values), four energy features (standard deviation, maximum, minimum and average energy values), four duration features, 13 MFCC coefficients and two tempo features.

A model with MFCCs, three spectral features (spectral roll-off, spectral centroid and spectral flux) and one temporal feature (zero-crossing rate) has been tested with seven bird species with a maximum accuracy of 84.59% [31]. In here, syllables have been divided into overlapping frames because less information has

obtained at the tapered ends when using a window function. They have suggested using a sliding window with an overlap as a way to fix that.

S. Fagerlund[38] has compared three feature sets for the same dataset. One feature set: consists of six Spectral features (Signal bandwidth, Spectral centroid, Spectral roll-off frequency, Spectral flatness, Spectral flux, Frequency range) and three Temporal features (Short time energy, Zero crossing rate and Modulation spectrum). Second feature set: consists of twelve MFCC coefficients, delta coefficients and delta-delta coefficients. The third feature set is a mixture of the first and second feature set. Best overall recognition result has been obtained from mixed feature sets. In 2014, Evangelista [39] has used another combination of other features with MFCC. The feature set was 64-dimensional and also includes the twelve initial Mel-frequency Cepstral coefficients (MFCC), and the means and variances of timbral features calculated in the pulses. The timbral features: including the centroid of the frequency spectrum, the spectral flux, that estimates the variability of magnitudes in this spectrum, and the roll-off. They have used 1,619 bird recordings of 75 species in the Southern Atlantic Coast of South America. They have tried out a novelty segmentation method for automatic segmentation and have obtained an overall accuracy of 59.75% of species detection.

A proposed model in[40] to identify 88 insect species using only 18 MFCCs as features have reached maximum 99.08% accuracy. Insect sounds are less computationally complex sounds even with human vocalizations.

2.3.3 Novel Features

A novel set of features by detecting a temporal pattern, introduced in 2010 in[41], can be used to detect 'Eurasian bittern' calls and 'Savi's Webber' calls. However, they are specialized features that are not generalized for bird sound detection or any other purpose. They are focused only on inherent characteristics of *Savis Webber* and *Eurasian bittern* calls. First, *Savis Webber* detecting algorithm focuses on detecting simple events. This algorithm can detect exactly simple events

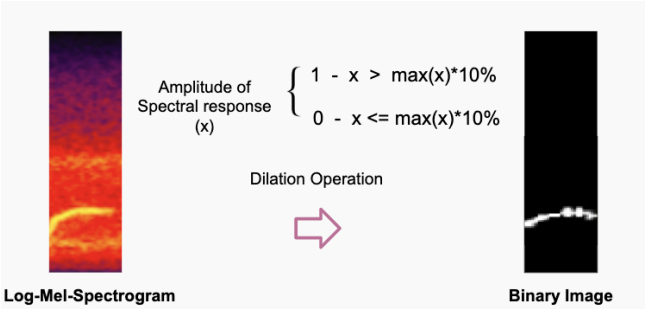

Figure 2.2: Conversion of Log-Mel-Spectrogram into a Binary Image.

in the spectrum, even with the presence of broad noise. Describing this in short, windowed power spectrum ($S(\omega, t)$) of a down-sampled input signal, was taken and then energy weighted, novelty measure also was taken. A squared difference of $S(\omega, t)$ has been chosen for this novelty estimate to enhance considerable energy variations over tiny ones. The measuring of sound energy has been used to achieve peak selection. Using this feature is more powerful in situations where light noise creates sharpened peaks in the novel estimate. Second *Savis Webber* algorithm focuses on the fact that *Savis Webber* calls repeat the element at a regular repeating rate of 50Hz. These two algorithm authors have been able to gain 79.06% detection rate of *Savis Webber* calls in a noisy environment.

To classify bird flight calls another researcher has used a novel feature extraction algorithm in [15]. After constructing a spectrogram using audio data; a series of FTTs in the time domain applied to all of the data segments in a window. Based on observations, this feature set has been implemented based on the following reasons according to the author. Flight call attributes,

- The lower frequencies consist of the highest amplitude noise.
- In the time-frequency domain calls are most obviously distinguishable.
- Contains the highest amplitude in the signal.

The Log-Mel-Spectrogram of each call has been extracted as the first step. Then the bottom-most frequency responses have been filtered out. If the amplitude of the spectral responses is greater than or equal to 10% of the spectrogram's overall maximum spectral amplitude, the amplitude will be replaced by 1; and all

the other responses replaced by 0. Further, a dilation operation has performed to enhance the continuity in the call. Fig. 2.2 shows the binary image after the above-explained steps conducted. Then the clip has been subjected to extract the features in it. Highest frequency, the lowest frequency, median frequency and mean frequency calculated for the total, initial 3/7, second 3/7 and third 3/7 of the binary image as to features. 16 features extracted from an audio clip from the SIFS method. The author shows that the mixed set of novel features with MFCC performs well and performs individually for classification. The author has evaluated four feature sets with four different classifiers.

When summarizing past literature on the feature extraction part of birds call identification, we can say that automatic bird identification using only MFCCs for bird identification has shown low accuracy rates where the combination of MFCCs with other spectral and temporal features has shown improved results. That may be because MFCCs based on the mechanism of human sound generation and bird vocal tracts have a significant difference with that of humans. Introduction of novelty features has good accuracy, but the problem is that introduced novel features are specialized for certain classes of birds. Many feature combinations can be found when MFCCs are mixed with other temporal, spectral and novel features. So feature weighting and feature selection can be used to select the best set of features and use them to improve performance.

2.4 Classification

The recent research has identified automatic bioacoustics monitoring as a classification problem, and emerging machine learning techniques are extensively applied. Numerous classification algorithms are available for application to classification problems.

2.4.1 Classification of bird vocalization as Time-series

In the early stages, Dynamic Time Warping (DTW) algorithm was used for classification. DTW is a low complex algorithm because what it does is a simple

template matching. It can detect a similarity between two temporal sequences regardless of speed differences and phase differences. DTW returns the optimal alignment between the two-time series. It has used classifying instances where there is less regional variation in vocalizations as an example in automatic classification of *Killer Whale* vocalization[42]. For the same reason of having local variations, DTW has not been used much in automatic detection of bird sounds.

Hidden Markov Model (HMM) which is known for temporal pattern recognition has been used instead of DTW algorithm[43] for frequency modulated bioacoustic signals. In probability theory, the Markov Model is a stochastic model used to model randomly changing systems. Markov Model made on top of two assumptions. In the Limited Horizon assumption, an order-k Markov process assumes conditional independence of state z_t from only the states that are $k+1$ time steps before it. A stationary process assumption assumes that the conditional distribution over the next state, given the current state, does not change over time. The Hidden Markov Model is a particular case of the Markov Model when we cannot observe the states themselves but only the result of some states' probabilistic function.

A recent research [44] has done trying to identify CAVI calls using a template-based technique. The particular technique suits even the sound clip recorded in a noisy environment with a limited training dataset. The proposed algorithm uses prominent time-frequency regions of training data and dynamic time-warping (DTW) to obtain templates. The accuracy rates have been compared with existing hidden Markov (HMM) and DTW models with various training data sets on several test conditions. The DTW model has performed well even with a lack of data for the training set where HMMs are doing better with more data in the training set. However, both models suffer from severe background noise availability. Their proposed model has gained 84% accuracy in classifying only CAVI calls and outperformed both DTW and HMMs models with a high margin even with severe background noise levels on most of the training and testing conditions.

2.4.2 Segment-wise Classification of Bird Vocalizations

Support Vector Machine(SVM) is a crucial machine learning algorithm performed exceptionally well in general. SVM is a supervised machine learning algorithm, which is used for both regression and classification problems. In a two-class classification problem with n number of features, first data plotted on n-dimensional space and optimum hyperplane selected with the help of Support Vectors. For multi-class classification, the same method used to find hyper-plane with one vs rest approach where iteratively every class is considered one class and a combination of all the other classes, as the other class. In a case where data is not linearly separable, a kernel function can be used. Kernel function maps the data points to a higher dimension making them linearly separable. SVM has been used frequently in bioacoustic monitoring tasks for classification. SVM has been used to classify seven bird species with four SVM kernels separately[31]. They have used radial base, linear, cubic and quadratic kernels. Maximum accuracy of 84.59% has been gained from the Quadratic kernel. SVM has been used to classify 88 insects with high accuracy (99.08%)[40]. Several methods used to build SVM classifiers when classifying more than two classes other than the previously mentioned method. To classify seven bird species[38] as a multi-class classifier a decision tree with SVM classifier has been used. The decision tree contains a lot of nodes; at every node it has an SVM binary classifier. In this method, each layer in the decision tree accounts to a class. Therefore each layer rejects one class when going to the next layer. There will be only one class remaining in the last layer, which will be the winning class. This approach has shown a maximum accuracy of 90% in classifying between the seven bird species.

Apart from SVM, many other classification models have also been used to classify vocalizations of different species. Different variations of the following Machine Learning algorithms have used in different scenarios:

- **k Nearest Neighbor:** used for classification of a multidimensional feature vector where each row is accountable for a class label. Store the class labels and feature vectors of training data done in the training phase of kNN. The

k is a constant which can be defined by the user in the classification phase. When a query is executed; at a point, an unlabeled vector of length k is classified by assigning the labels of all the nearest training points. The most frequent label will become the classification result of that query point.

- **Random Forest:** classifier constructed by multiple decision trees. In the classification process, it returns the mode of the classes and in the regression process; it returns the mean of prediction. RF is an ensemble method of classification. For building a decision tree, the tree bagging technique has been used. Because of that, a large forest composed of shallow trees will generate as a result. For large training data-sets, individual trees are less likely to over-fit.

- **Neural Networks:** consist of simple processing elements or units called highly interconnected neurons. The layers have neurons. These neurons are in layers, input, hidden and output layers which converts an input vector into output. Multiple variations of a neural network such as self-organizing maps, multilayer perceptron neural networks and feedforward networks have been used.

[45] have compared the classification accuracy of three machine learning algorithms. The first algorithm is a decision tree, the Second algorithm is linear discriminant analysis, and the Third algorithm is support vector machines(SVM). Nine frogs and three bird species were the target species group of the research. Acoustic classification of frog's and the bird's calls have been done. This research is addressing the N-Class problem. Altogether 10,061 isolated calls have been used to train and test the system. The average accuracy was: 89.20% for the decision tree, 94.95% for support vector machine and 71.45% for linear discriminant analysis. In retrieval of wild animal vocalizations [46] k-Nearest Neighbor Classification (k=11) and Multilayer perceptron (MLP) feed-forward neural network has been used.

Generally, the past literature points are: (1) Signal quality could improve through preprocessing. (2) Effective segmentation can improve recognition rates

of bird species. (3) Many feature combinations are available; Feature selection and Feature reduction can improve accuracy. (4) Existing Machine learning algorithms are performing well with bioacoustics data.

Furthermore, the Literature review shows that recognition rates of bird vocalizations identification are yet to be improved. Moreover, continuous monitoring can increase more informative predictions using bioacoustics data. kNN, SVM, Random Forest and Neural Networks have performed exceptionally well with bioacoustics data of many biological species. Therefore, this research focuses on developing an effective preprocessing method and Novel Features that will enhance the visibility of flight calls and improve the accuracy of automatic flight call detection.

Chapter 3

METHODOLOGY

3.1 Overview of Methodology

The methodology of the system consists of 4 main processing sections until classification:

1. Pre-processing
2. Feature Extraction
3. Classification

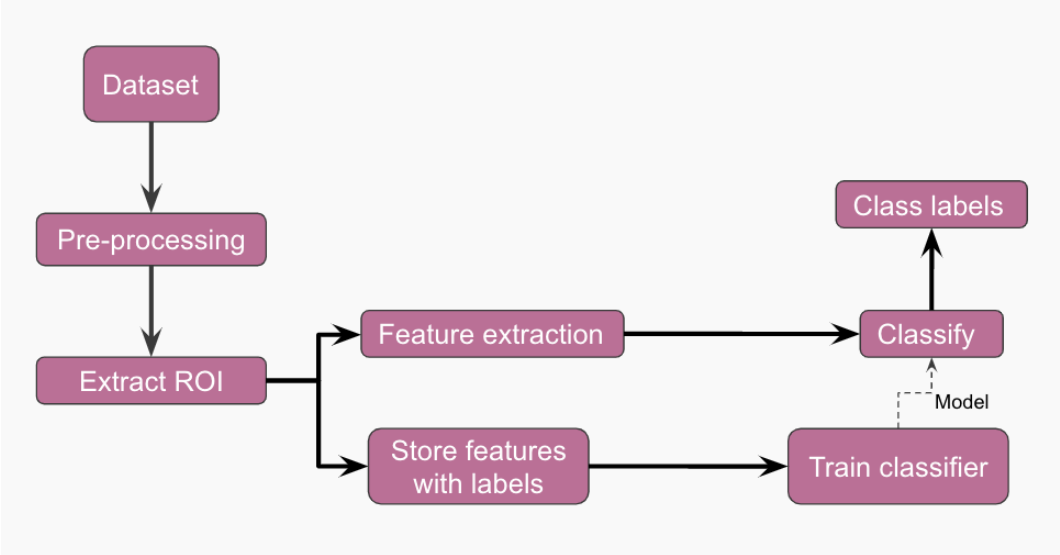

Figure 3.1: Overall system design

Fig. 3.1 summarizes the overall design of the system with main keywords at each step. Pre-processing step is conducted to Noise reduction. Then identify regions of interest. Feature extraction and classification are done.

3.2 Dataset

3.2.1 CLO-43SD

This is a dataset that has been created by Andrew et al. Dataset consists of 5428 flight calls. Forty-three different North American wood-warbler species have been used to create the Dataset. All the sound clips contain a flight call of any of the 43 different species. Each call is labelled for the existence of a species' flight call; therefore, this Dataset can be used to evaluate classification models that address the N-Class problem.

3.2.2 CLO-SWTH and CLO-WTSP

The two of these datasets have been created by Justin et al. These datasets can be used for Binary open-set scenarios for specific species classification. They consist of 16,703 flight calls of White-Throated Sparrow (WTSP) and 179,111 flight calls of Swainson's Thrush (SWTH). Every clip in this dataset will either contain a flight call of the respective species or not.

3.2.3 BirdVox-full-night

This dataset can be used for Binary open-set scenarios for multiple species classification. It contains six far-field audio clips recorded during a full night, which contains 35,000 flight calls of 25 species. The dataset has been annotated individually by experts with the time of birds called existence. All the audio clips in the dataset consist of either a flight call of any species or not.

Since the only dataset that can be used to address the Binary open set problem for multiple species was BirdVox-fullnight, that dataset has been selected to implement the general classifier to recognize flight calls in an audio clip in this research.

Details about the data set:

- Data set name : *BirdVox-70k*
- Captured by : ROBIN autonomous recording units

- Captured on : Night of September 23rd, 2015
- Location : Ithaca, NY, USA during the fall 2015
- Species : 25 different species
- Captured data : Avian flight calls
- Sample Rate : 22,000 Hz
- # chanels : Single channel (mono)
- Length of a clip : 0.5 Seconds
- Total # clips : 70,804 clips
- Flight calls : 35,402 clips
- Non-Flight calls : 35,402 clips

The dataset contains 70,804 audio clips. Total of 35,402 flight calls of 25 different species are available. And the rest of the audio clips contain sounds of the real world environment.

Acknowledgement to authors of the dataset : Vincent Lostanlen, Justin Salamon, Andrew Farnsworth, Steve Kelling, and Juan Pablo Bello.

3.3 Pre-processing and Noise reduction

As mentioned in previous sections, real-world recordings consist of different types of background noises.

1. Biogenic noise
2. Anthropogenic noise
3. Noise due to wind and rain

Biogenic noise is the noise made by other animals. When considering bird monitoring, all sounds made by other animals are considered as noises. In real-world recordings, the noise of insects and amphibians can be identified.

Anthropogenic noise is noise resulting from human activities. This type of noise may not be present in forests. However, if recording devices are located near any human-made construction, there is a possibility of anthropogenic noise in the recordings.

Noise due to wind and rain is the other primary type of noise in real-world audio recordings. This noise can be seen in almost all the recordings.

Noise reduction is the first primary process done. Purpose of this step is to reduce the existing noise in ROIs and get the high signal to noise ratio. Proper noise reduction method is needed to do features accurately. Therefore, this step mainly focused on improving the accuracy of the classification results.

Pre-processing and noise reduction is essential to increase the signal to noise ratio, and increase the feature set's accuracy. Firstly 1 kHz Butterworth High Pass Filter to filter out the background noise below 1 kHz. Since a flight call is between 1 and 11 kHz, there is no harm—secondly Spectral Subtraction, where an average noise spectrum is subtracted from the signal to remove recurring noise. The average noise spectrum was estimated from a period, where the signal is absent.

3.3.1 High-pass Butterworth Filter

Since a large portion of wind noise is below 1 kHz has been identified. Therefore, a high-pass *butterworth filter* has been used to cancel out all of the sounds below 1 kHz. This step does not affect the original bird flight calls because most birds' fundamental frequency lies above 1kHz [23]. After applying this filter, a considerable amount of background noise was reduced. At the same time, this filter cuts down the noise added by the microphone as well.

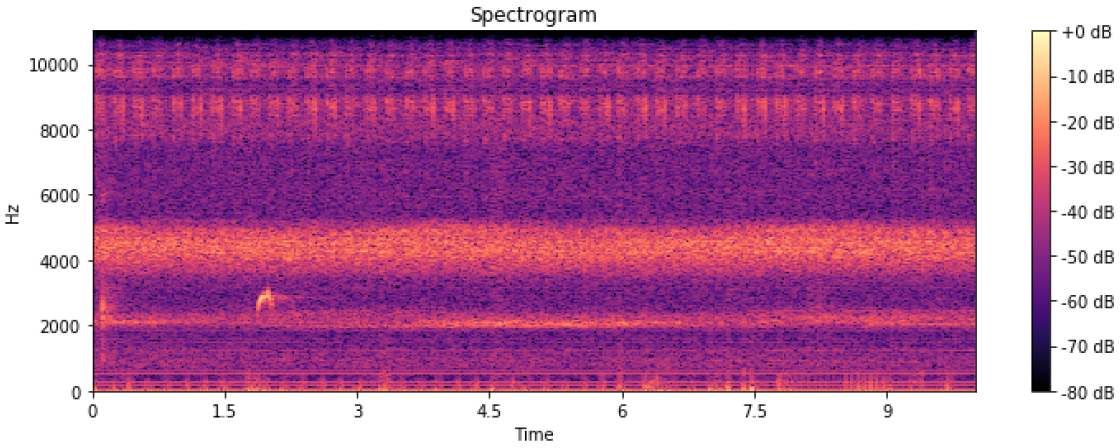

Figure 3.2: Spectrogram of Flight call with Noise

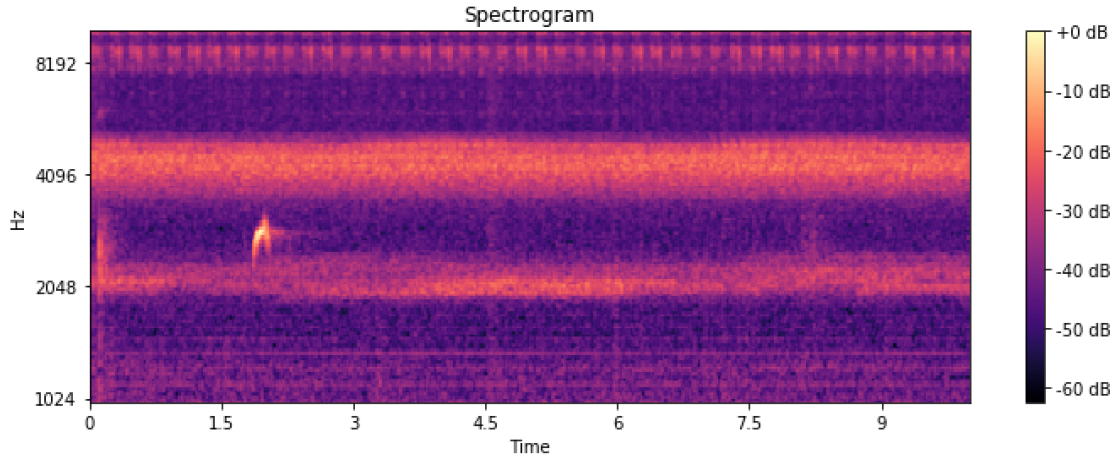

Figure 3.3: Spectrogram of Flight call after applying High-pass Butterworth Filter and Spectral Subtraction

3.3.2 Spectral Subtraction

Spectral subtraction is the second step in the noise reduction phase. The technique called *Spectral Subtraction* was used to make noise reduction of the signal. This method is used in speech processing applications to do noise removal.

The Spectral Subtraction's central concept is to remove the recurring or average magnitude of the noise spectrum from the noisy spectrum. The computing of the noisy spectrum is done by Fast Fourier Transformation(FFT). Python was used to implement the noise removal algorithm. Average magnitude noise of the clip was calculated using the initial 12 ms period of the clip. Here, the assumption

that no bird call syllable is present in the first 12 ms period has been taken. The first step would be to store the noisy RoI data frames as Hanning time-widowed half-overlapped data buffers. After that, each form has been converted to the corresponding spectrum using FFT. Soon after subtracting the estimated noise average from the noisy frames, the spectrum is reconstructed again by Inverse Fast Fourier Transformation (IFFT).

Fig. 3.2 shows the spectrogram of a flight call before applying noise reduction and spectral subtraction, and Fig. 3.3 shows the spectrogram of the same flight call after applying high-pass Butterworth filter and spectral subtraction. The clarity of the flight call has been increased in the second spectrogram than the first spectrogram. It means the power of the background noise is reduced considerably. However, bird flight call syllables remain the same. Therefore, this method's effect on our interested segment, which is bird flight call syllables is negligible.

3.4 ROI Extraction

The intention of this process was to extract regions of interest from the audio stream of 500 ms. In this approach, the energy of the signal concerning the time domain has been calculated.

Here are several methodologies to figure out acoustic events.

1. Zero Crossing Rate
2. Instantaneous Energy Threshold
3. Short Time Energy Threshold

3.4.1 Zero Crossing Rate(ZCR)

Zero crossing rate measures the number of zero axis crossings by the signal in a certain period. A zero-crossing is defined within a specific context like discrete-time signals, for successive samples when there are different algebraic signs. ZCR is an uncomplicated type measure of the frequency content of the signals.

Here is the definition of the zero-crossing rate.

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]| w(n-m) \tag{3.1}$$

where

$$sgn[x(n)] = \begin{cases} -1 & ; \quad x(n) < 0 \\ 1 & ; \quad x(n) \geq 0 \end{cases}$$

and

$$w(n) = \begin{cases} \frac{1}{2N} & ; \quad 0 \leq n \leq N-1 \\ 0 & ; \quad otherwise \end{cases}$$

Zero crossing rate is high for silent regions and low for non-silent regions (when there is an acoustics event). Since low frequencies imply low zero-crossing rate and high frequencies imply high zero-crossing rate from that, it could be figured out that a correlation between ZCR and energy distribution with frequency. As a generalized conclusion, a zero-crossing rate is high when there are no acoustics events, while if the zero-crossing rate is low, there is an acoustics event.

Zero crossing rate was robust when there was a weak signal/noise ratio known as SNR. However, it cannot guarantee a weak signal/noise ratio found in a real-world record. Therefore Zero Crossing Rate did not suit the requirement for this research.

3.4.2 Instantaneous Energy Threshold

The energy was calculated using the square product of amplitude. It was a less complicated method which can be used for acoustic event detection. Here is the definition of instantaneous energy.

$$E_s = \sum_{n=-\infty}^{\infty} |x(n)|^2 \tag{3.2}$$

E_s = *Instantaneous energy*

$x(n)$ = *Amplitude at time instance n*

This fixed level threshold depends on bird species because the energy of bird flight calls varies with bird species, so that defined fixed threshold cannot be used to extract regions of interest for any bird flight call.

3.4.3 Short Time Energy Threshold

The signal was divided into frames and measured the average energy concerning each frame. The individual amplitude of an acoustic signal deviates significantly amidst time. The amplitude of silent signal segments (when there are no acoustics events) are generally much lower than the amplitude of voiced segments. Short time energy will provide a smooth representation than amplitude variations. Here is the definition of short-time energy.

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \tag{3.3}$$

E_n = *Energy of the frame*

$x(m)$ = *Amplitude at time instance m*

$w(n-m)$ = *Window function*

Short Time Energy Threshold which sets the threshold dynamically, by observing the signal for a short time. This technique was good enough for the particular scenario where the energy of a bird flight call varies with bird species.

3.5 Feature Extraction and Selection

To identify the existence of a flight call of a bird species using the obtained segments; it is necessary to have a parametric representation of the signal/segment. This section describes how features were extracted and selected from the phrases

extracted from the previous section. The details of features that have been used are mentioned in section 3.5.1.

3.5.1 Feature set 1

Features can be classified as temporal, spectral and novel features. Temporal features are capable of capturing temporal characteristics in a signal. Spectral features can capture frequency base characteristics in a signal with the Fourier transform of the signal. Different combinations of mixtures of temporal, spectral and novel features with MFCC have performed well in previous bioacoustics monitoring.

Though some research have introduced novelty features specific to a particular scenario for general purpose. Firstly tried to classify with standard features used in speech recognition and music information retrieval. The 20 features used for classification are as follows:

Zero Crossing Rate(ZCR)

As discussed previously, considering a particular clip's small duration, the rate of sign-changes of the signal; it can be considered as an inexpensive temporal feature of a signal.

$$ZCR = \sum_{n=0}^{M-1} |sgn(x(n)) - sgn(x(n+1))| \tag{3.4}$$

Energy

The sum of squares of the signal values within a frame, and then normalized by the respective frame length.

$$E_s = \sum_{n=-\infty}^{\infty} |x(n)|^2 \tag{3.5}$$

Spectral Centroid

Spectral Centroid represents the center of gravity of the spectrum. In detail it gives out the frequency where the energy of the signal is concentrated.

$$SC = \frac{\sum_{n=0}^{M} n|X(n)|^2}{\sum_{n=0}^{M} |X(n)|^2} \tag{3.6}$$

Spectral Bandwidth

The Spectral Bandwidth is known as the second central moment of the spectrum. It represents how the energy spreads throughout the frequency spectrum.

$$\sigma^2 = \int (x - \mu)^2 . p(x) \delta x \tag{3.7}$$

Spectral Contrast

Let's assume that each frame of a spectrogram is partitioned into sub-bands. Then in every sub-band, energy variation is assessed by correlating individual mean energy within the top quantile known as peak energy to the bottom quintile known as valley energy. Higher contrast values usually match to clear narrow-band signals, while low contrast values match up broad-band noise.

Spectral Flatness

Spectral flatness is known as the tonality coefficient, which is used to quantify how much noise-like a sound is present. A spectral flatness which is closer to 1.0 is defined as a high spectral flatness, which indicates the spectrum is similar to white noise. Spectral flatness is often converted to decibel scale.

Spectral Rolloff

Spectral roll-off measures the spectral skewness of the spectral shape of the signal. It gives out the frequency below which some percentile of the cumulative spectral

power resides. 90% percentile was used for the calculations.

$$SRF = max\left(SRF| \sum_{n=0}^{SRF} |X(n)|^2 < Th \sum_{n=0}^{M} |X(n)|^2 \right) \tag{3.8}$$

MFCCs

MFC, which is known as Mel-frequency cepstrum, represents the short-term energy spectrum of a given sound. The MFC has derived upon the basics of the human vocal tract. Mel-frequency cepstral coefficients(MFCCs) try to determine the human vocal tract's shape through several coefficients. MFCCs are widely used in speech and speaker recognition. Though bird vocal tracts are more complex than humans, MFCCs can still be mixed with other features. In this research, to calculate 13 MFCCs as features for a single frame used 27 filter banks.

Altogether, 20 features have been used as mentioned above. The exact way of extracting 40 features from a segment is described in section 3.5.4.

Feature set 1 - Spectral and Temporal Features (SATF) were created by combining all the features mentioned below:

- Spectral and Temporal Features (SATF)
  - Zero crossing rate
  - Energy (RMS)
  - Spectral Centroid
  - Spectral Bandwidth
  - Spectral Contrast
  - Spectral Flatness
  - Spectral Rolloff
  - 13 MFCCs

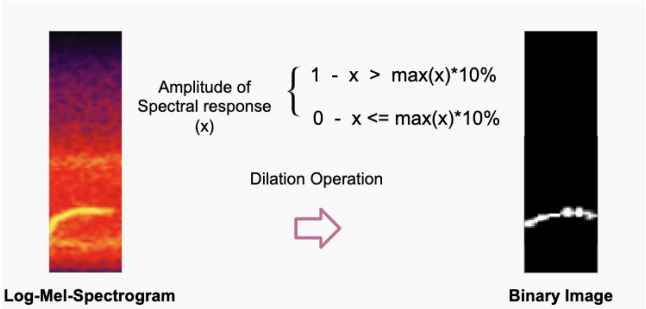

Figure 3.4: Conversion of Log-Mel-Spectrogram into a Binary Image.

3.5.2 Feature set 2

The 16 features are extracted by converting the spectrogram into a binary image using the Spectrogram-based Image Frequency Statistics (SIFS) technique. Selin et al. [15] have used SIFS features to evaluate the classification of bird flight calls. This feature set has been implemented based on the following reasons, according to the author.

1. Flight calls can be characterized by the highest amplitude of the signal
2. In time-frequency domain it is most obviously distinguishable
3. The highest amplitude noise is present in low frequencies

The proposed algorithm has been used in this research after customizing accordingly. The Log-Mel-Spectrogram of each call is extracted as the first step. Then the bottom-most frequency responses have been filtered out. All the spectral response amplitudes greater than 10% of the spectrogram's overall maximum spectral amplitude have been set to a value of one; all others were assigned to zero value. Further, a dilation operation has been performed to enhance the continuity in the call. Fig. 3.4 shows the binary image after conducting the above-explained steps. Then the clip has been subjected to feature extraction. The highest, lowest, median, and mean frequencies have been calculated for the initial 3/7, second 3/7, third 3/7, and the whole of the binary image as to features. Sixteen features in total can be extracted from an audio clip using SIFS.

Altogether 56 features have been used in Feature set 2. It was done by Combining 16 SIFS features in this section, and all the features in Feature set 1. The

decision to create a combined feature set was taken as an inspiration from the work done by Selin et al. [15]. They have shown that a combination of spectral features with novel features will increase bird flight calls' classification accuracy.

Feature set 2 - Spectral and Temporal Features (SATF) and Spectrogram-based Image Frequency Statistics (SIFS) were created by combining all the features mentioned below:

- Spectral and Temporal Features (SATF)
  - Zero crossing rate
  - Energy (RMS)
  - Spectral Centroid
  - Spectral Bandwidth
  - Spectral Contrast
  - Spectral Flatness
  - Spectral Rolloff
  - 13 MFCCs
- Spectrogram-based Image Frequency Statistics (SIFS)
  - The lowest frequency of the whole binary image
  - The highest frequency of the whole binary image
  - The mean frequency of the whole binary image
  - The median frequency of the whole binary image
  - The lowest frequency of the first 3/7 binary image
  - The highest frequency of the first 3/7 binary image
  - The mean frequency of the first 3/7 binary image
  - The median frequency of the first 3/7 binary image
  - The lowest frequency of the second 3/7 binary image
  - The highest frequency of the second 3/7 binary image

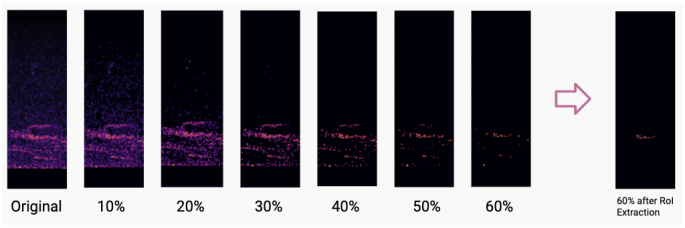

Figure 3.5: Static margin from 10% to 60% and ROI extracted from 60% margin

- The mean frequency of the second 3/7 binary image
- The median frequency of the second 3/7 binary image
- The lowest frequency of the third 3/7 binary image
- The highest frequency of the third 3/7 binary image
- The mean frequency of the third 3/7 binary image
- The median frequency of the third 3/7 binary image

3.5.3 Feature set 3

Spectrogram-based Maximally Stable Extremal Regions (SMSER), the novel feature extraction technique proposed by this research, has 25 features. Maximally Stable Extremal Regions (MSER) is a method used for image processing and has been inspired to create this feature set. In the computer vision domain, Maximally Stable Extremal Regions (MSER) is frequently used to detect a region in an image. MSER can be used to find objects in a snap. Similarly, this method has been used to detect flight call regions in a binary image.

As discussed in the SIFS method in Section 3.5.2 above, the Log-Mel-Spectrogram of a sound clip was converted to the binary image after the dilation operation. Discovering that 10% static margin which was used in the above step was not applicable for all the scenarios in this research. The Fig. 3.5 shows that 10% static margin gives more distraction to detect the flight call in the clip. After further investigation, at 60% margin the flight call was clearly visible and all the distracted noise has been removed.

This scenario led to the implementation of the new algorithm (SMSER) where the percentage was dynamically chosen. The Fig. 3.6 shows the simplified version

```
#getFeatures function finads the percentage that exactly has one contour and
extract features from that percentage

def getFeatures (File):
    lowerBound = 100
#lowerBound holds last known lower bound percentage where zero contours found
   upperBound = 0
#upperBound holds last known upper bound percentage where >=1 contours found
   difference = 10
#difference holds the value that will increased or reduced in every itteration
according to the direction
   direction  = 1
#there are two directions {100% ===> 0%} = 1 and {100% <=== 0%} = 0
   no_of_contours = 0
#use to store no. of contours returned by MSER which has area of a flight call
   percentage = lowerBound
#percentage will be started with having the value of the lower bound
   while(no_of_contours != 1):
#use to find percentage using no_of_contours

      no_of_contours = getNoOfContours(File, percentage)
#using MSER algorithm to find contours which match the area of flight call
      if (no_of_contours == 1):
          break
        else:
            if (direction):
                if (no_of_contours == 0):
                    percentage = percentage - difference
                    lowerBound = percentage
#while the lower bound is always with zero contours this will excute
            else:
                    upperBound = percentage
                    direction  = 0
                    difference = difference / 10
                    percentage = percentage + difference
#if percentage find more than zero contours the direction will be changed and
difference will be reduced 10 times to find closer point of interest
         else:
                if (no_of_contours != 0):
                    percentage = percentage + difference
                    upperBound = percentage
#while the upper bound is always with more than zero contours this will excute
            else:
                    lowerBound = percentage
                    direction  = 1
                    difference = difference / 10
                    percentage = percentage - difference
#if percentage find zero contours the direction will be changed and difference
will be reduced 10 times to find closer point of interest
```

Figure 3.6: Pseudo code of SMSER algorithm

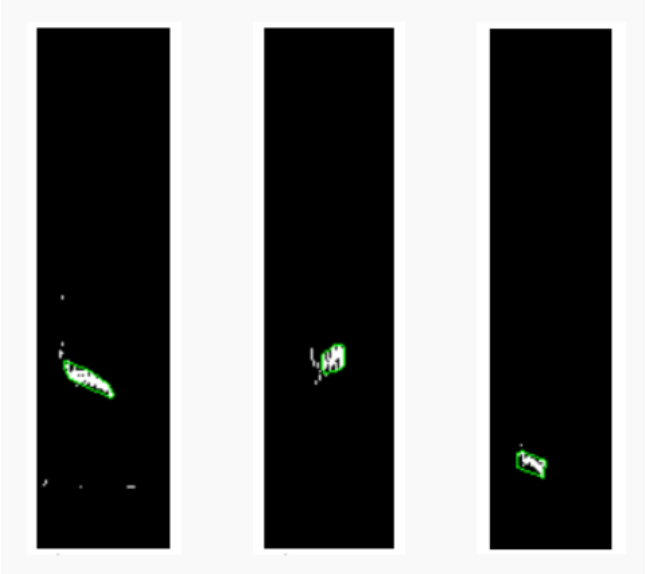

Figure 3.7: Samples of detected contours using the SMSER algorithm.

of that algorithm which produces the feature extraction. The ultimate goal of the algorithm is to determine the best percentage to capture the flight call which

has the highest amplitude of the signal. The binary image consisted of 44 pixels along the frequency axis and 257 pixels along the time axis. The getNumberOf-Contours() function returns the count of all the contours that have a min area of 103 and max area of 1848. These min and max areas were calculated using the assumptions stated in [3]. The while loop of the algorithm will break after 100 iterations, if it can't find a percentage which gives only 1 contour having the given conditions.

The Fig. 3.7 shows a few sample flight calls which have been captured by the algorithm. After this process the image inside the contour will be isolated by another operation. After that the following features will be calculated,

- Moments(2): center of mass of the object
- Contour area(1)
- Contour perimeter(1)
- Contour bounding rectangle attributes(4): top left coordinates, height and width
- Aspect ratio of the bounding rectangle(1)
- Extent of the contour area compared to bounding rectangle area(1)
- Leftmost, rightmost, topmost and bottom-most coordinates of the contour(8)
- Radius and center coordinates of the minimum enclosing circle of the contour(3)
- Starting coordinate and ending coordinate of the contour fit-line(4)

Combining the 56 features that were discussed in Feature set 2 with these 25 features all together there are 81 features in Feature set 3.

Feature set 3 - Spectral and Temporal Features (SATF), Spectrogram-based Image Frequency Statistics (SIFS) and Spectrogram-based Maximally Stable Extremal Regions(SMSER) were created by combining all the features mentioned below:

- Spectral and Temporal Features (SATF)
  - Zero crossing rate
  - Energy (RMS)
  - Spectral Centroid
  - Spectral Bandwidth
  - Spectral Contrast
  - Spectral Flatness
  - Spectral Rolloff
  - 13 MFCCs
- Spectrogram-based Image Frequency Statistics (SIFS)
  - The lowest frequency of the whole binary image
  - The highest frequency of the whole binary image
  - The mean frequency of the whole binary image
  - The median frequency of the whole binary image
  - The lowest frequency of the first 3/7 binary image
  - The highest frequency of the first 3/7 binary image
  - The mean frequency of the first 3/7 binary image
  - The median frequency of the first 3/7 binary image
  - The lowest frequency of the second 3/7 binary image
  - The highest frequency of the second 3/7 binary image
  - The mean frequency of the second 3/7 binary image

  - The median frequency of the second 3/7 binary image
  - The lowest frequency of the third 3/7 binary image
  - The highest frequency of the third 3/7 binary image
  - The mean frequency of the third 3/7 binary image
  - The median frequency of the third 3/7 binary image

- Spectrogram-based Maximally Stable Extremal Regions(SMSER)
  - Moments(2): center of mass of the object
  - Contour area(1)
  - Contour perimeter(1)
  - Contour bounding rectangle attributes(4): top left coordinates, height and width
  - Aspect ratio of the bounding rectangle(1)
  - Extent of the contour area compared to bounding rectangle area(1)
  - Leftmost, rightmost, topmost and bottom-most coordinates of the contour(8)
  - Radius and center coordinates of the minimum enclosing circle of the contour(3)
  - Starting coordinate and ending coordinate of the contour fit-line(4)

3.5.4 Frame Base Extraction of Features

All the 20 features mentioned in section 3.5.1 are frame based features where features are calculated for a certain frame. So in order to extract features from a segment, two approaches can be used

1. Take the segment as a one single frame and do the feature extraction.
2. Divide segments into small frames and do the feature extraction on each small frame.

An audio signal is a signal that constantly changes, therefore it can not be considered that on short time scales that audio signal does not vary much. But in the 1st approach the segment would sometimes take a length of more than 41 frames. So it is hard to get an accurate spectral or temporal measurement with a large segment compared to a small frame. So the 2nd approach of framing the segment into small frames before feature extraction was chosen.

To obtain a reliable spectral or temporal estimate a frame size of 12 ms was used. So that maximum of 41 frames could be generated by a 500 ms long sound clip. Where sample frequency of 44100 Hz. Then for each feature a feature vector of dynamic length can be generated depending on the length of the segment. This causes the following problems as identified:

1. Having a feature vector of dynamic length is an issue for training a classifier because it cannot map features of every segment into a common model.

2. When each and every 20 features have long length feature vectors the overall feature set will have a very high dimensionality. This may lead to low accuracy rates because of the curse of dimensionality

To overcome the above problems, modeled the distribution of each 20 features over the segment using mean and standard deviation (Fig 3.8). So, the final feature vector used for classification consists of 40 features.

3.5.5 Feature Reduction and Selection

First the feature set 1 was used (40 features) as it is for classification models for training and identifying bird flight calls. Then tried the following three methods for feature reduction and selection in order to improve accuracy.

Recursive Feature Elimination (RFE)

RFE is a technique that recursively removes features and returns a set of features with highest ranking. Recursive feature elimination algorithm depends on the model that needs to conduct the feature reduction. Recursive feature elimination

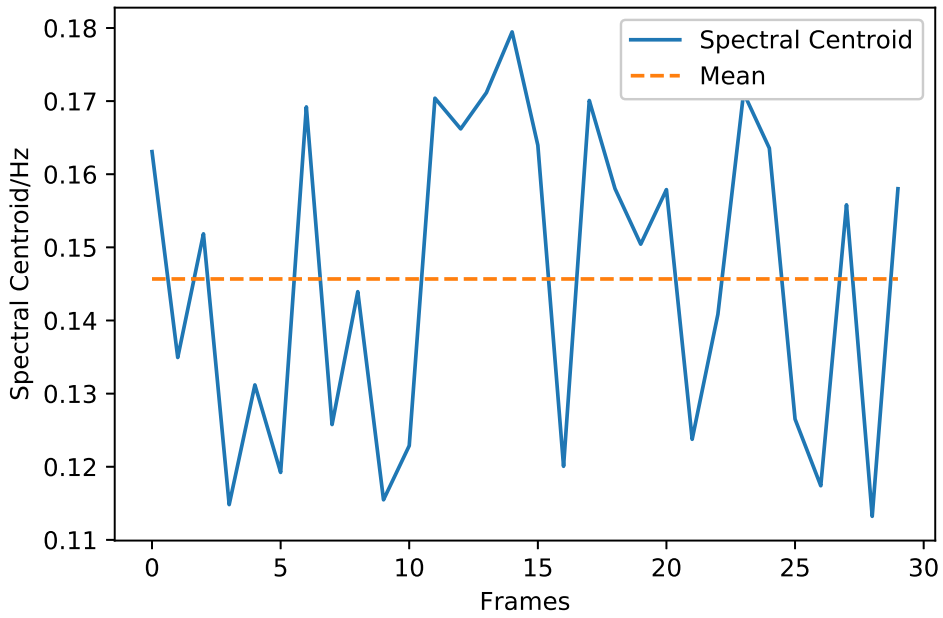

Figure 3.8: Spectral Centroid distribution over a 30 frames segment

was conducted on both SVM and RF models used. Starting from a small number of desired features and improving it iteratively and finding out the best combination of features that gives out the best cross validation score and best accuracy rate on test results. Results and evaluation of the results of RFE can be found in the sub section 4.3.3.

Random Forest Algorithm (RF)

Random Forest, as the name implies, consists of a large number of decision trees. These trees are operated together(ensemble). Each individual decision tree points out a class. The class with the most number of votes becomes the model's prediction. After running the algorithm the bi product of feature importance score as a percentage could be found. Evaluation of that score helped to find out the best features. Results and evaluation of the results of RF can be found in the sub section 4.3.3.

Principal Component Analysis (PCA)

PCA that can be used for dimensionality reduction is a linear transformation algorithm. It emphasizes variation of the data by finding mutually orthogonal principal components which will maximize variations of the data. Starting from a small number of Principal components and improving it iteratively. The best combination of principal components that gives out the best cross validation score and best accuracy rate on test results can be figured out by this.

Another use of PCA is, high dimensional data can be visualized and can get an idea about distribution of data points. Therefore it can summarize and try to find out trends in data for information retrieval. Fig. 4.6 and Fig 4.7 show a visualization on the training data set obtained from all the segments of flight calls. Results and evaluation of the results of PCA can be found in the sub section 4.3.3.

3.6 Classification Models

3.6.1 Recognition of Flight Calls

When considering automatic flight call identification as a classification problem, the last step is to apply classification for identification of flight calls. When it comes to this step parametric representation of segments with labels are ready. Therefore it is preferred to use a supervised learning approach. Since SVM, RF, DNN and kNN have shown good performance in bioacoustics based classification all of them were tried out. Accuracy on the testing data set was used to select the most efficient model. Results on model selection is described in sub sections 4.3.4 and 4.3.5.

Support Vector Machine (SVM)

SVM is a supervised learning algorithm which is based on the concept of decision planes which will determine decision boundaries. There are some benefits of SVM that caused for selecting it for the classification process.

- SVM performs effectively in high dimensional spaces as well.
- Memory efficient because this employs a subset of training spots in each decision function (termed as support vectors).
- Though training time of SVM is high in classification it is less complex.
- In a linearly inseparable situation, kernels can be used to map data points to higher dimensions, improving the model.

Polynomial kernel with different degrees, linear kernel, sigmoid kernel and rbf kernel with their different parameters were tested out. Accuracy of the Test data set against the polynomial kernel degree parameter classified from the Feature set 1 has been shown in Fig. 4.8. Table 4.1 shows the results obtained for each of the kernels for the Feature set 1. Results of SVM classification are described in subsection 4.3.4. The best kernel was chosen for the final comparison.

Random Forest (RF)

Random Forest is an ensemble machine learning algorithm on multiple decision trees which is considered as a good classifier in complex classification tasks. Following reasons were considered when choosing RF:

- RF being an ensemble model reduces the problem of overfitting.
- Good classifier for large datasets.

RF algorithm has a configurable variable called number of trees. The Fig. 4.9 shows how the number of trees will affect the classification results of the RF algorithm using the Feature set 1. The optimum number of trees was 180 according to the results. That amount was finalised for all other evaluations thereafter. The table 4.3 shows how the RF classifier performed with each of the Feature sets. Results of RF classification are described in subsection 4.3.4.

k-Nearest Neighbour (kNN)

k-Nearest Neighbour is an algorithm that can be used for classification as well as regression. kNN relies upon identified input data to learn a function. Then that function generates a suitable output when a new unlabeled data is provided. kNN assumes similar things exist nearby. Following reasons considered when choosing kNN:

- Easy and simplicity in terms of implementation
- No need to build a model, make additional assumptions, or tune several parameters.
- Algorithms remain versatile.

There is one drawback when we use kNN. Unlike other methods, the algorithm becomes significantly more delayed as each amount of predictors variables grow. The kNN algorithm has a configurable variable called value of k. The Fig. 4.10 shows how the value of k will affect the classification results of the kNN algorithm using the Feature set 1. The optimum value for k was 11 according to the results. That amount was finalised for all other evaluations thereafter. The table 4.4 shows how the kNN classifier performed with each of the Feature sets. Results of kNN classification are described in subsection 4.3.4.

Deep Neural Networks (DNN)

DNN is also a special type of artificial neural network where it consists of multiple layers of neurons in between the input layer and the output layer that can be used for classification. DNN is capable of finding the appropriate mathematical function to map input to output regardless if it has a linear or non-linear relationship. DNN will predict a probability for an output. Following reasons were considered when choosing DNN:

- Can predict linear or non-linear relationship
- DNN has performed well in many domains.

There is one drawback when we use DNN. Unlike other methods the algorithm gets significantly over-fitted with less number of training iterations when we have a small training dataset. DNN algorithm has a lot of configurable variables such as,

- Learning rate
- Momentum
- Number of hidden units
- Number of layers
- Number of epoch
- Learning rate decay
- optimization algorithm

It is a known fact training a bigger network always doesn't hurt. So that number of hidden units and number of layers were set to a high amount. Fig. 3.9 shows the decision making algorithm which has been used when tuning the classifier. Adam optimization algorithm was selected because it showed better accuracy rates in less number of epochs. The Fig. 4.12 shows how the train accuracy and test accuracy vary to the number of epochs when classification of DNN algorithm for Feature set 1. The optimum number of epochs was 12 according to the results. That amount was finalised for all other evaluations thereafter. The table 4.5 shows how DNN classifier performed with each of the Feature sets. Results of DNN classification are described in subsection 4.3.5.

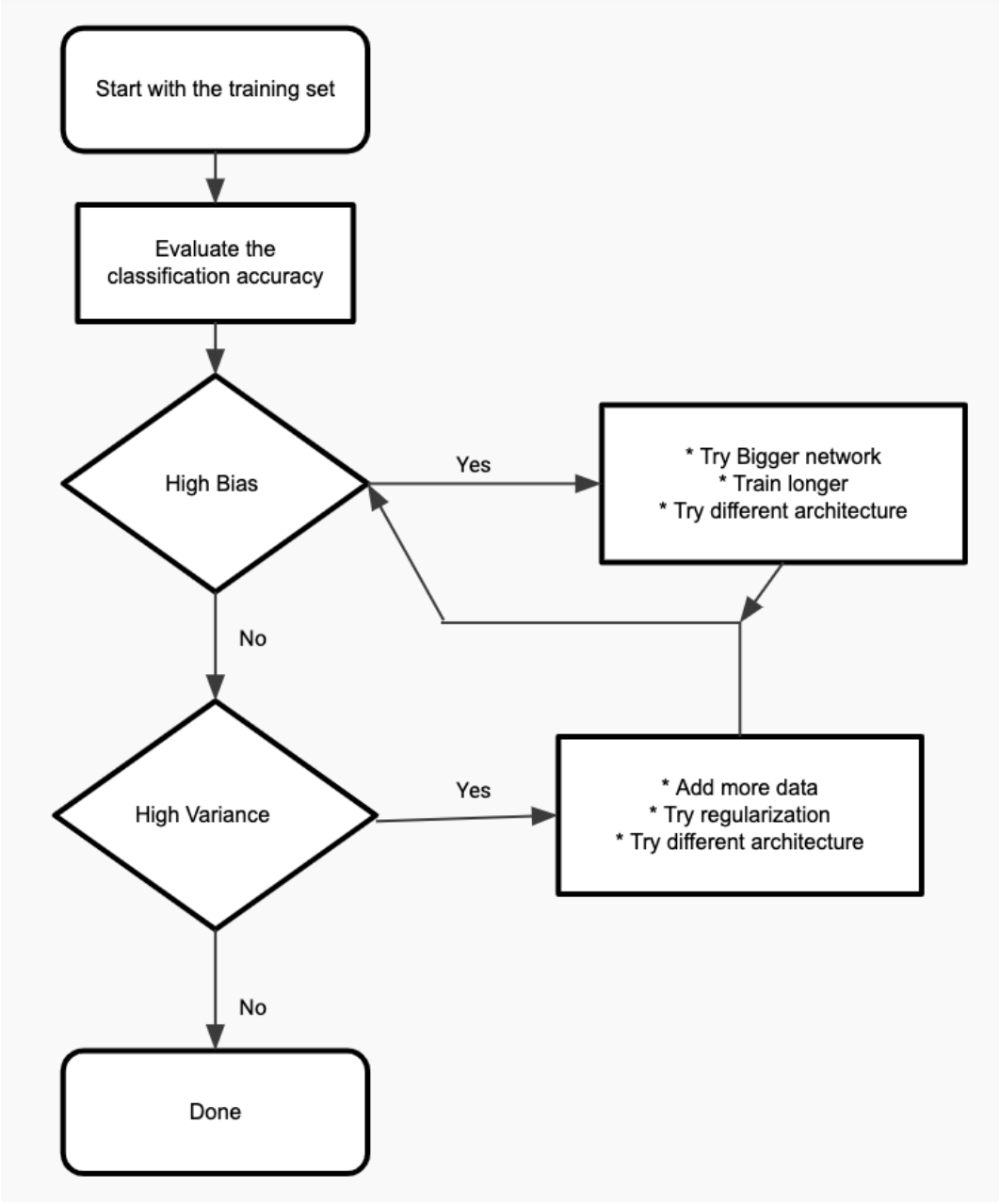

Figure 3.9: Basic recipe for Neural Networks

Chapter 4

EXPERIMENTAL EVALUATION AND DISCUSSION OF RESULTS

This chapter focuses on describing the dataset used for building classification models, experiments conducted while implementing the final system and results of different approaches and evaluation of those results.

4.1 Data Set

After evaluating 4 different data sets containing flight calls namely CLO-43SD, CLO-WTSP, CLO-SWTH and BirdVox-full-night. From 4 of these data sets BirdVox-full-night was selected for our research since it contains flight calls of 25 different species at real world environments, further BirdVox-full-night is an ideal dataset to evaluate multiple occurrences of similar acoustic events. BirdVox-full-night has provided with six far-field recordings, of full night, containing 35,402 flight calls in the dataset, from 25 species of passerines. The dataset is individually annotated against frequency and time by experts, along with an official evaluation methodology. All together there are 70,804 total clips including 35,402 non-flight calls.

Details about the data set:

- Data set name : *BirdVox-70k*
- Captured by : ROBIN autonomous recording units
- Captured on : Night of September 23rd, 2015
- Location : Ithaca, NY, USA during the fall 2015
- Species : 25 different species
- Captured data : Avian flight calls

Details about the data set:

- Sample Rate : 22,000 Hz
- # chanels : Single channel (mono)
- Length of a clip : 0.5 Seconds
- Total # clips : 70,804 clips
- Flight calls : 35,402 clips
- Non-Flight calls : 35,402 clips

Acknowledgement to authors of the dataset : Vincent Lostanlen, Justin Salamon, Andrew Farnsworth, Steve Kelling, and Juan Pablo Bello.

4.2 Experiments

After evaluating the clips collected from the dataset, low signal to noise ratio was observed in most of the clips. Therefore the initial step of the process was to increase the signal to noise ratio by reducing background noise. 1kHz Butterworth High Pass Filter and Spectral Subtraction techniques was used to reduce the background noise of a clip. This data set was preprocessed before the classification. Since the dataset contains 0.5 second(500ms) audio clips of flight calls and non-flight calls, also a flight call duration generally lies between 50ms - 300ms, therefore removing the unvoiced regions was important for a better performance. Region of Interest was calculated to extract best values for all the features from a clip. RoI extraction techniques were tried out to determine the sound event in all the 0.5 second audio clips. Instantaneous energy threshold and Short time energy threshold techniques were tried out for the extraction. Flight calls have a frequency range between 1kHz to 11 kHz generally, due to the dataset containing clips of non-flight calls as well there were other sound events which occur outside of the 1-11 kHz boundary. Therefore the instantaneous energy threshold was not enough for the RoI extraction. Short time energy threshold worked fine for the requirement.

The next was to extract features before that let me explain what are the tried out methods and feature sets.

- **Classification methods**
  - Support Vector Machine (SVM)
  - Random Forest (RF)
  - K-Nearest Neighbor (kNN)
  - Deep Neural Network (DNN)
- **Feature extraction methods**
  - Spectral and Temporal Features (SATF)
    - Zero crossing rate
    - Energy (RMS)
    - Spectral Centroid
    - Spectral Bandwidth
    - Spectral Contrast
    - Spectral Flatness
    - Spectral Rolloff
    - 13 MFCCs
  - Spectrogram-based Image Frequency Statistics (SIFS)
    - The lowest frequency of the whole binary image
    - The highest frequency of the whole binary image
    - The mean frequency of the whole binary image
    - The median frequency of the whole binary image
    - The lowest frequency of the first 3/7 binary image
    - The highest frequency of the first 3/7 binary image
    - The mean frequency of the first 3/7 binary image
    - The median frequency of the first 3/7 binary image

        * The lowest frequency of the second 3/7 binary image
        * The highest frequency of the second 3/7 binary image
        * The mean frequency of the second 3/7 binary image
        * The median frequency of the second 3/7 binary image
        * The lowest frequency of the third 3/7 binary image
        * The highest frequency of the third 3/7 binary image
        * The mean frequency of the third 3/7 binary image
        * The median frequency of the third 3/7 binary image
    - Spectrogram-based Maximally Stable Extremal Regions(SMSER)
        * Moments(2): center of mass of the object
        * Contour area(1)
        * Contour perimeter(1)
        * Contour bounding rectangle attributes(4): top left coordinates, height and width
        * Aspect ratio of the bounding rectangle(1)
        * Extent of the contour area compared to bounding rectangle area(1)
        * Leftmost, rightmost, topmost and bottom-most coordinates of the contour(8)
        * Radius and center coordinates of the minimum enclosing circle of the contour(3)
        * Starting coordinate and ending coordinate of the contour fit-line(4)

- **Feature sets used**
    - Feature set 1:
        * SATF : 40 features
    - Feature set 2:
        * SATF and SIFS : 56 features (40 + 16)
    - Feature set 3:

∗ SATF, SIFS and SMSER : 81 features (40 + 16 + 25)

The above feature sets were used in different classification models based on the compatibility. Before the classification of Feature set 1, Principal Component Analysis(PCA) and Recursive Feature Elimination(RFE) were done to get a better understanding of all the features in it. 2 component and 3 component Principal Component Analysis was conducted. Recursive Feature Elimination was tested with Random Forest Classifier and SVM with Linear Kernel Classifier.

4.2.1 Supervised Learning Methods

Supervised learning methods were evaluated only with the All the Feature sets. Three popular supervised learning methods were tried out.

Support Vector Machines(SVM)

Support Vector Machines was used with 4 different kernels,

- Linear kernel
- Polynomial kernel
- Gaussian RBF kernel
- Sigmoid kernel

All kernels other than the Polynomial kernel didn't have other parameters. Polynomial kernel was tested multiple times by changing the value of the degree parameter.

Random Forest(RF)

Random Forest algorithm consists of trees, therefore the parameter of no. of trees should be selected. Different no. of trees were tested to get the best outcome. After running the algorithm the bi product of feature importance score as a

percentage could be found. Evaluation of that score helped to find out the best features.

k-Nearest Neighbor(kNN)

kNN is a lazy learning algorithm which is non-parametric. Multiple tests were carried out to see the performance of the parameter K on this algorithm. Different number of feature combinations were tested to find the best accuracy rate. RFE results were helpful in this step to find the best features out of all the 40 features in Feature set 1.

4.2.2 Deep Learning Method

Deep learning methods were evaluated with all the three Feature sets.

Deep Learning with Feature Set 1

In this scenario the deep neural network was optimised for better test accuracy and to try how deep learning will perform on the data set after extracting the spectral and temporal features.

Deep Learning with Feature Set 2

The deep neural network was used to classify the data set after the extraction process of SATF and SIFS features. Initially SATF features were extracted. After that, running the SIFS algorithm to convert the spectrogram into a binary image. Then SIFS features were extracted. By combining both features together(40 + 16) Feature set 2 including 56 features has been created.

Deep Learning with Feature Set 3

The deep neural network was used to classify the data set by using the Feature set 3. Similar to the above step, initially extracting SATF features then running SIFS algorithm to extract SIFS features. Finally running SMSER algorithm to

extract the SMSER features. By combining all the features together (40 + 16 + 25) Features 3 including 81 features has been created.

4.3 Results and Evaluation

This section gives a detailed view of results followed by subsections having evaluation of these results.

4.3.1 Evaluation of Noise reduction techniques

The purpose of using noise reduction techniques was to improve the signal to noise ratio. By doing that the classification system could focus on the signal and the accuracy of the features will also be increased. The 1 kHz butterworth high pass filter was obvious since all the noises lie below the 1kHz region. After the high pass filter spectral subtraction was tried out to eliminate the average noise from the signal.

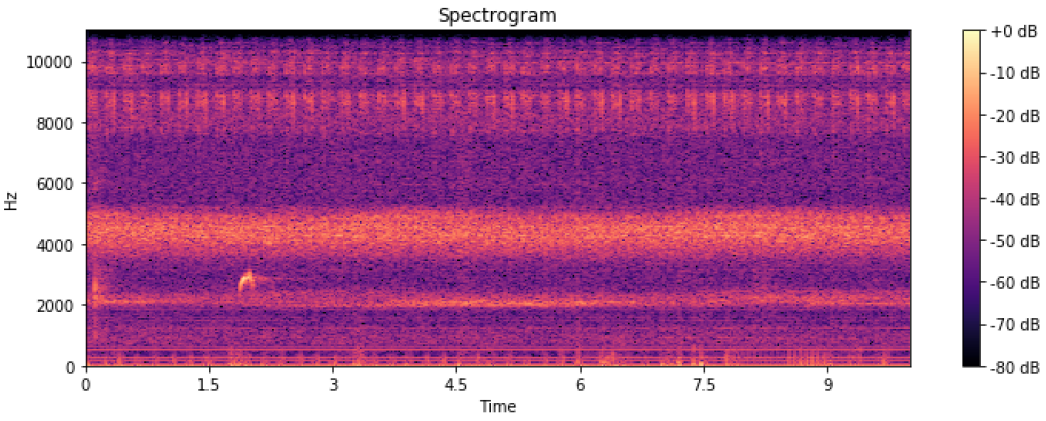

Figure 4.1: Spectrogram of a flight call

Figure 4.1 contains a flight call between 1.5 seconds to 3 seconds. Figure 4.2 contains the same flight call in Figure 4.1 after applying noise reduction techniques. From the human eye observation, the flight call is much better identified in the clip after the noise reduction. So that the computers will too.

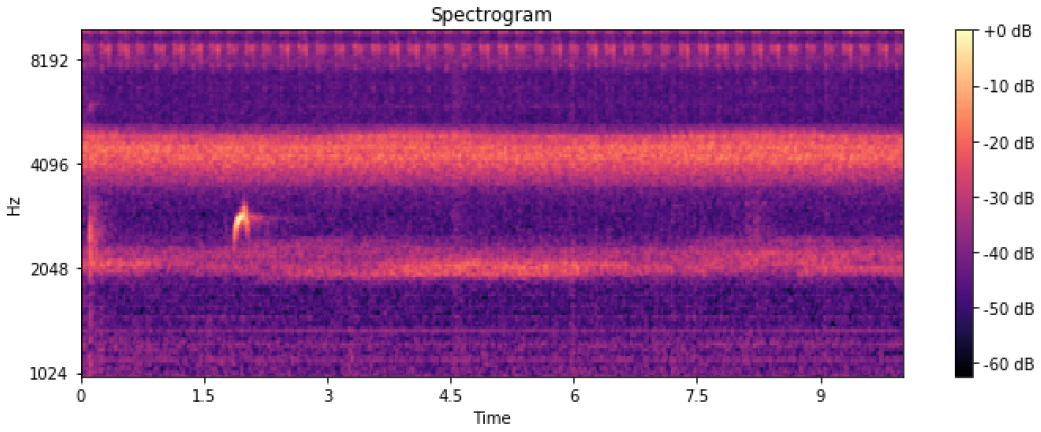

Figure 4.2: Spectrogram of a flight call after noise reduction

4.3.2 Evaluation of RoI Extraction Techniques

The purpose of using the RoI Extraction techniques was to improve the feature quality by focusing only on the sound event. Instantaneous Energy Threshold was one of the techniques. Even though we could set a threshold based on frequency range of flight calls, it is not enough for the task since we can't limit the sound events by a threshold. Some sound events could happen in between threshold levels. Therefore by setting a fixed threshold we would miss some of the features of those sounds, so that we will miss classify them. Second technique tried out was Short Time Energy Threshold which sets the threshold dynamically, by observing the signal for a short time. This technique was good enough for the particular scenario where we had short clips containing either flight calls or non-flight calls.

4.3.3 Evaluation of Feature Selection and Reduction

Feature selection and reduction is one of the most important aspects in the process. 3 popular techniques were tried out.

Recursive Feature Elimination(RFE)

Before running recursive feature elimination, it is really important to remove the highly correlated features because it provides the same information. Out of 40 features in the *Feature set 1*, 8 features showed a correlation coefficient of more

than 80%.

8 Features : MFCC_Mean_1, MFCC_Mean_2, MFCC_Mean_3, MFCC_Mean_4, spectral_flatness_Mean, spectral_rolloff_Mean, spectral_centroid_Mean, spectral_flatness_Std

All the above features which had a correlation coefficient above 80% were removed. In order to execute the RFE algorithm we need to choose an estimator, number of features to remove at each step, cross validation method and the scoring method. For the estimator Random Forest Classifier and SVM Linear Kernel was tested out. Due to the lack of computing power only the SVM Linear Kernel was tested, other kernels took long periods to process. For the step value of 1 was used, therefore in each iteration 1 feature is eliminated. For the cross validation method Stratified K fold was used with K=10. For the scoring method "Accuracy" was used. The Figure 4.3 shows percentage of correct classifications against the number of features of the two estimators used, in two lines for the *Feature set 1*. With RF Classifier it reaches the maximum accuracy with only 5 features. But for SVM with Linear Kernel it reaches the max with 10 features. Therefore as a result of RFE it can be seen that the number of features between 4 - 11 are required to achieve the maximum accuracy of RF and SVM with Linear kernel classifiers.

Running Recursive Feature Elimination algorithm for *Feature set 1* with Random Forest classifier showed the optimal number of features are 5. Recursive feature importance of the top 5 features are indicated in the bar chart in Figure 4.4.

Random Forest Algorithm (RF)

First the RF model was created using 100 estimators. Then trained the model with a training dataset. Then visualized the feature importance variable of the algorithm for the *Feature set 1*. The results are shown below in the bar chart in Figure 4.5. The Recursive Feature Elimination with RF's top most features are also in the top list here as well but not exactly in the same order.

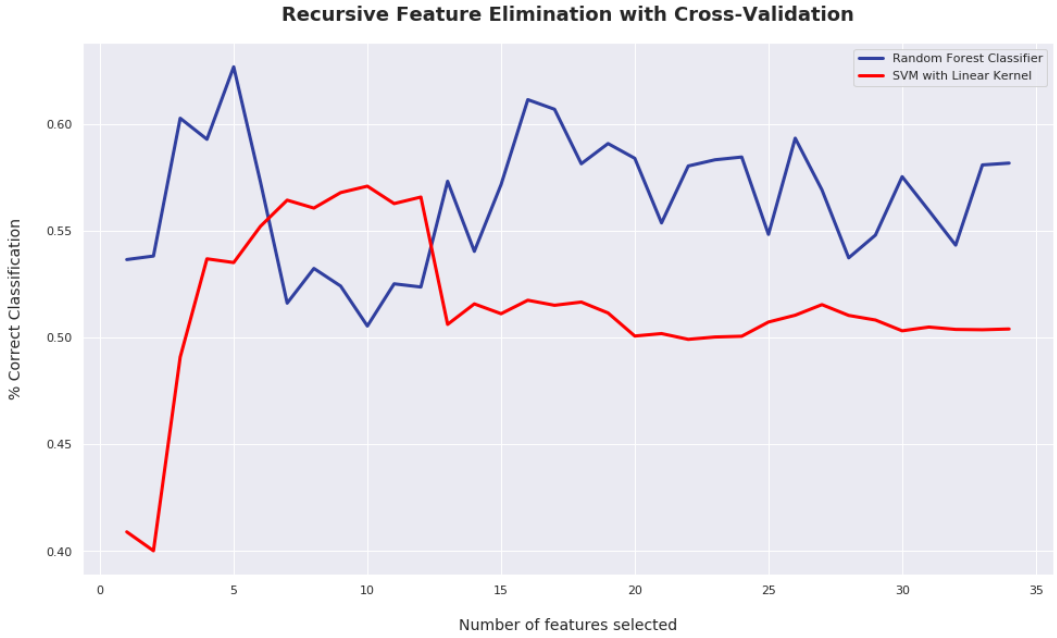

Figure 4.3: Recursive Feature Elimination with Cross-Validation with RF and SVM classifiers for Feature set 1

Principal Component Analysis(PCA)

Figure 4.6 shows how 2 component PCA against each other for the *Feature set 1* was conducted. Ground truth indicating the color of each data point. There was no clear margin that could separate the two categories apart. Therefore 3 component PCA for *Feature set 1* also visualised to observe a more detailed view which is shown in Figure 4.7. But both results were showing a complex distribution of data.

4.3.4 Evaluation of Supervised Classification Models

Supervised classification models were evaluated using all the Feature Sets.

Support Vector Machines(SVM)

SVM is one of the most popular algorithms in the category of Supervised classification models. SVM could be run with 4 different kernels. Polynomial kernel has a parameter called degree, the Figure 4.8 shows how its performance varied according to the value of degree parameter for the *Feature set 1*. 3 degrees were

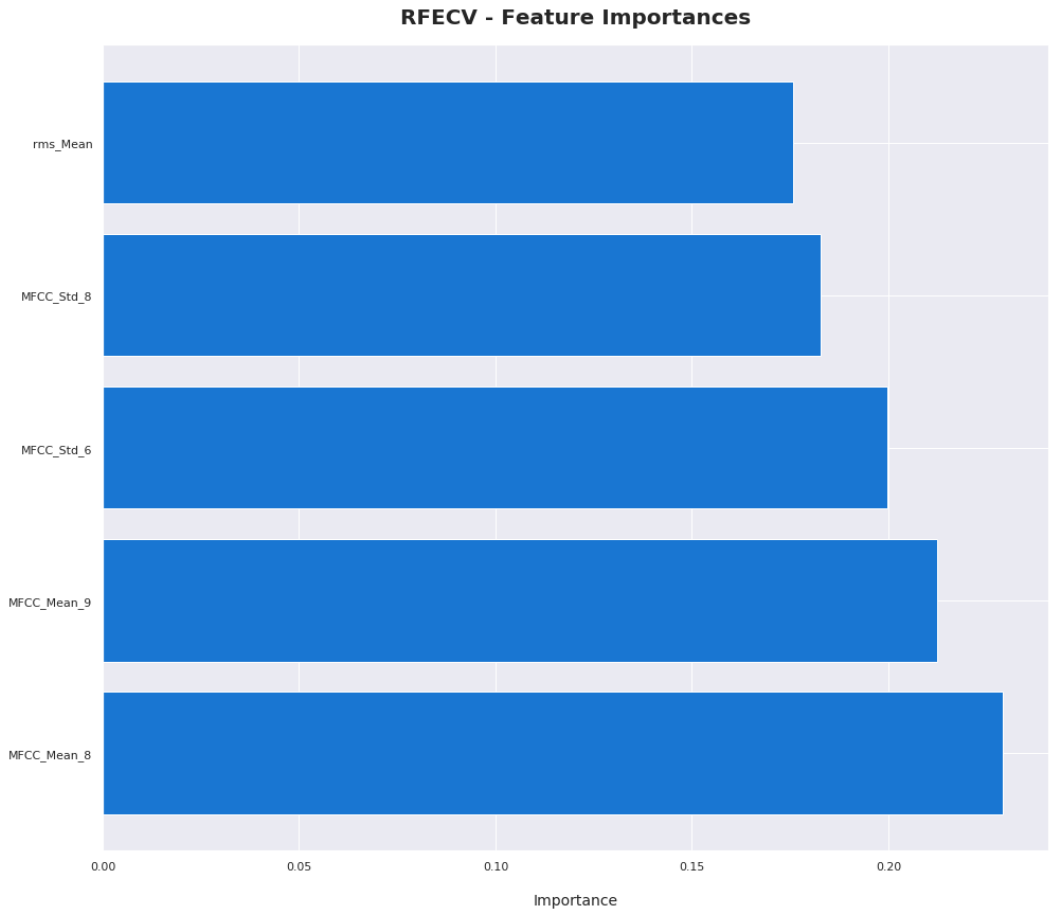

Figure 4.4: Feature Importance Score

the ideal number compared to all the other values. It will maximise the accuracy of the classification. Soon after selecting the parameter value of Polynomial kernel degree as 3. All other 3 kernels were evaluated with the *Feature set 1*.

Table 4.1: Classification Accuracy of Feature set 1 from 4 different SVM kernels

Metric	Linear	Sigmoid	Polynomial	RBF
Accuracy	57.1%	49.6%	61.3%	61.7%

The maximum accuracy that could be achieved is shown in Table 4.1. Sigmoid kernel performed the worst even though sigmoid kernel is more capable for binary classification tasks. The RBF kernel and polynomial kernel both have performed really well. Overall the RBF kernel out performed all the kernels. For the *Feature Set 1* used in this research RBF kernel will be a better option out of other kernels.

From the above results, the best performing Support Vector Machine classifier

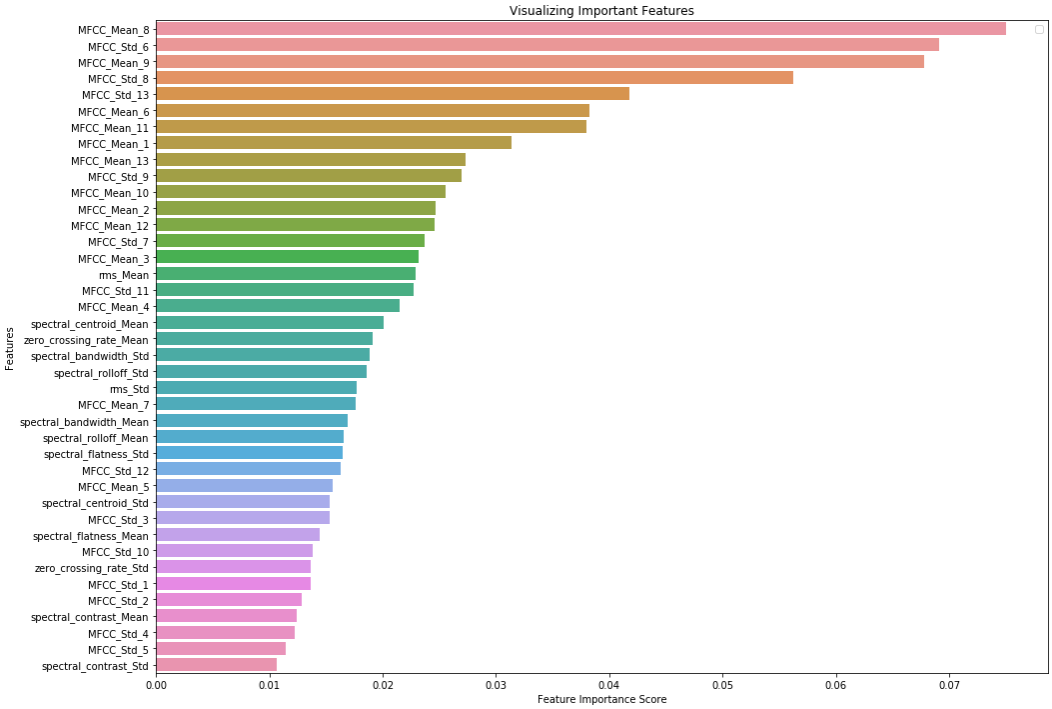

Figure 4.5: Visualizing the Feature Importance Variable in RF for the Feature set 1

with rbf kernel was used to evaluate the Feature set 2 and Feature set 3. Table 4.2 shows the results. Overall accuracy of the classifier has been improved with additional features.

Table 4.2: Classification Accuracy of All the Feature sets from SVM RBF kernel

Metric	Feature set 1	Feature set 2	Feature set 3
Accuracy	61.7%	65.16%	69.3%

Random Forest(RF)

RF was evaluated with Feature Set 1 for different numbers of trees. The Figure 4.9 shows how classification accuracy varied against the number of trees in RF. The maximum accuracy rate of 69.9% was observed when the number of trees was 180. Removing all the least important features reduced the training time. But it didn't affect the overall accuracy of the classification model.

After evaluating the *Feature set 1*, 180 numbers of trees were giving the best

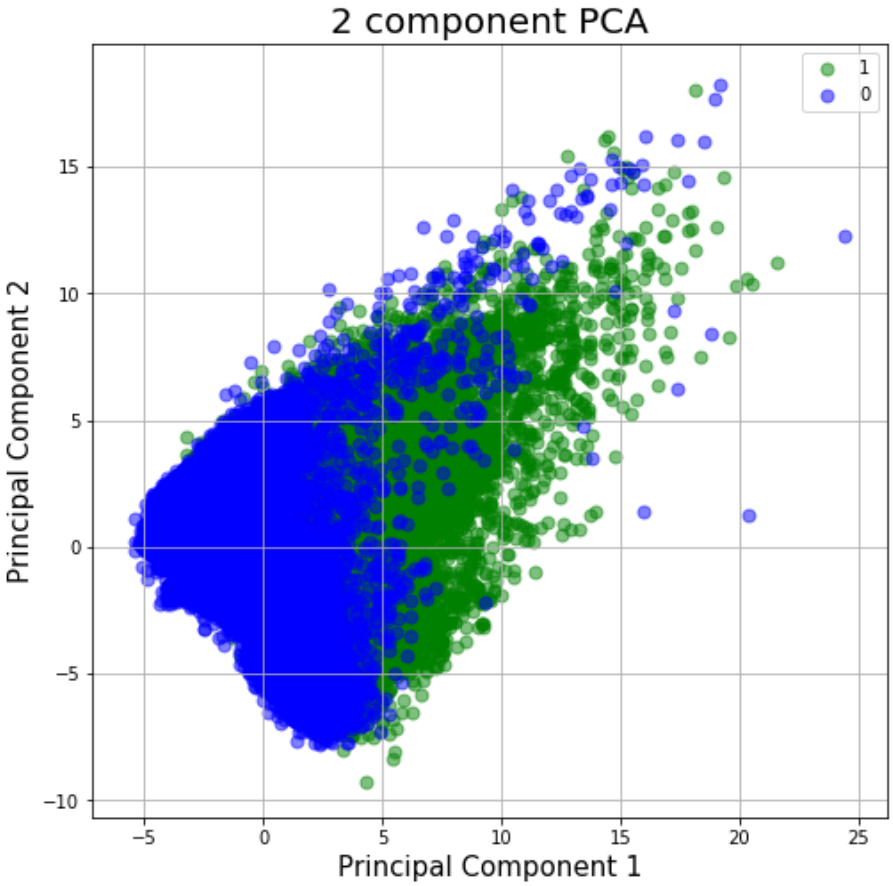

Figure 4.6: 2 Component PCA for Feature set 1

results. So that same classifier was used to evaluate the *Feature set 2* and *Feature set 3*. Table 4.3 shows the results.

Table 4.3: Classification Accuracy of All the Feature sets from RF Classifier with 180 trees

Metric	Feature set 1	Feature set 2	Feature set 3
Accuracy	69.9%	72.4%	76.6%

k-Nearest Neighbor(kNN)

kNN was evaluated with *Feature set 1*. When running the kNN Algorithm the value of k should be determined. Accuracy rate of the classifier differs when the value of k and the used number of features change. The Figure 4.10 shows how accuracy rate fluctuated according to the values chosen. Initially the optimal

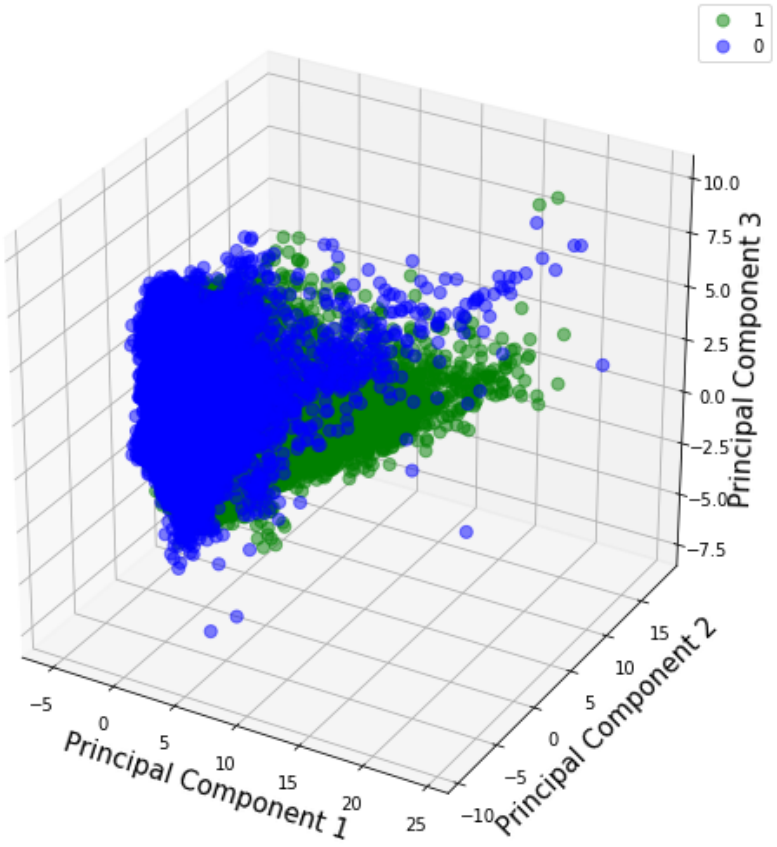

Figure 4.7: 3 Component PCA for Feature set 1

number of features suggested in the Feature Selection process was used to evaluate accuracy against different values of k. It showed a decent accuracy level where the maximum accuracy was observed at k=11.

The top 5 features with a high feature importance score were, MFCC_Mea_8, MFCC_Mean_9, MFCC_Std_6, MFCC_Std_8, rms_Mean. Figure 4.11 shows how MFCC_Mean_8 plots against other features will visualize indicating the classification results of kNN in two different colours. Then respectively tried to increase the number of features to 19 and 25. That showed relatively less accuracy values compared to the optimal number of features(5). Then tried out the algorithm with all the 40 features and 35 features. It showed the best accuracy rate of 59.99% at the value k=11 and with 40 features. The Figure 4.10 shows the results of all instances.

From the above results, the best accuracy for the Feature set 1 was generated

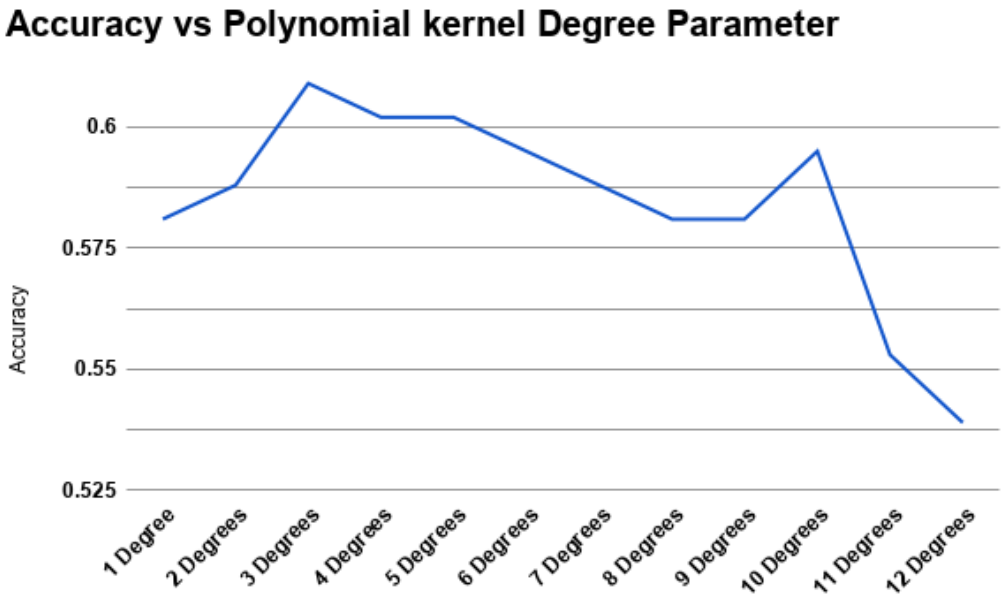

Figure 4.8: Accuracy against SVM with Polynomial kernel Degree Parameter for Feature set 1

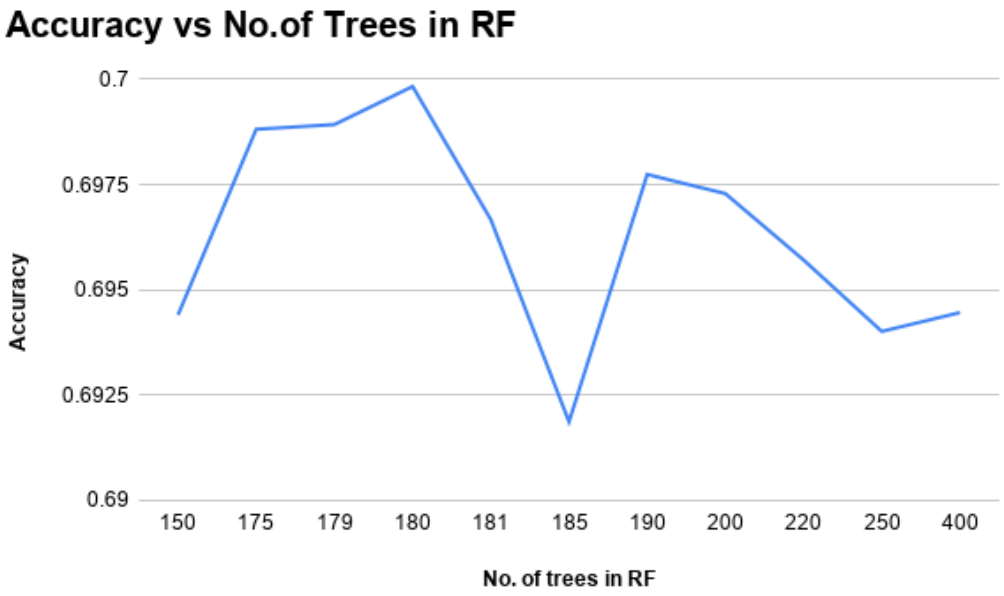

Figure 4.9: Accuracy against No. of trees in RF algorithm for Feature set 1

when the value of k=11 and when all the features are available. Assuming that condition is valid for the rest of the data sets the data was evaluated. The Table 4.4 shows the results.

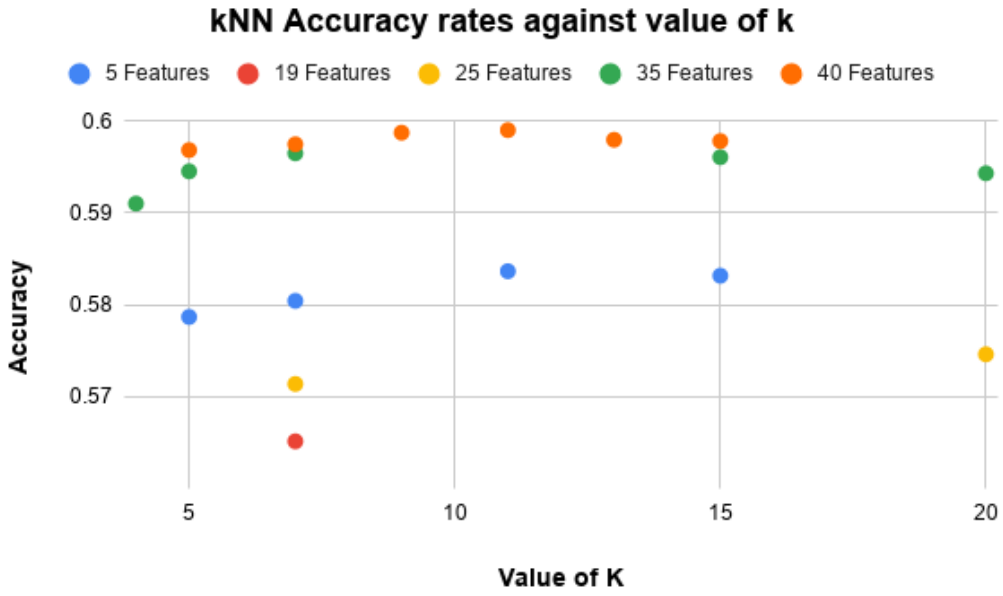

Figure 4.10: Accuracy against Value of k in kNN algorithm for Feature set 1

Table 4.4: Classification Accuracy of All the Feature sets from kNN Classifier with k = 11

Metric	Feature set 1	Feature set 2	Feature set 3
Accuracy	59.9%	61.6%	64.6%

4.3.5 Evaluation of Deep Learning Classification Models

Deep Neural Network was created and tested by changing multiple parameters to get the best accuracy. All the Feature Sets were evaluated.

A deep neural network was trained by changing the number of epochs trained and evaluated with the accuracy rates for the *Feature set 1*, results are shown in Figure 4.12. Observation was when increasing the number of epochs to 150 the Train Accuracy goes up to 83.68% but the Test Accuracy goes down to 72.73%. Where the network is being over-fitted to the training data. By reducing the number of epochs until the Test Accuracy becomes the maximum. Best Test Accuracy of 75.31% was reached when the number of epochs was 12. Therefore the best accuracy that could be achieved by this classification model for this data set was 75.31%.

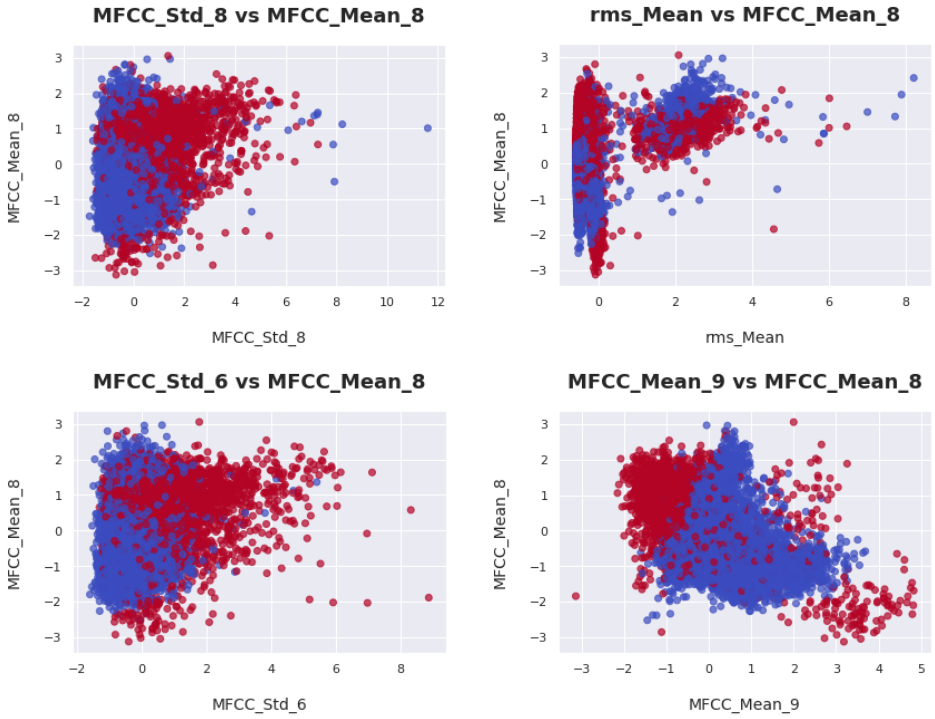

Figure 4.11: kNN results of MFCC_Mean_8 against other high importance features

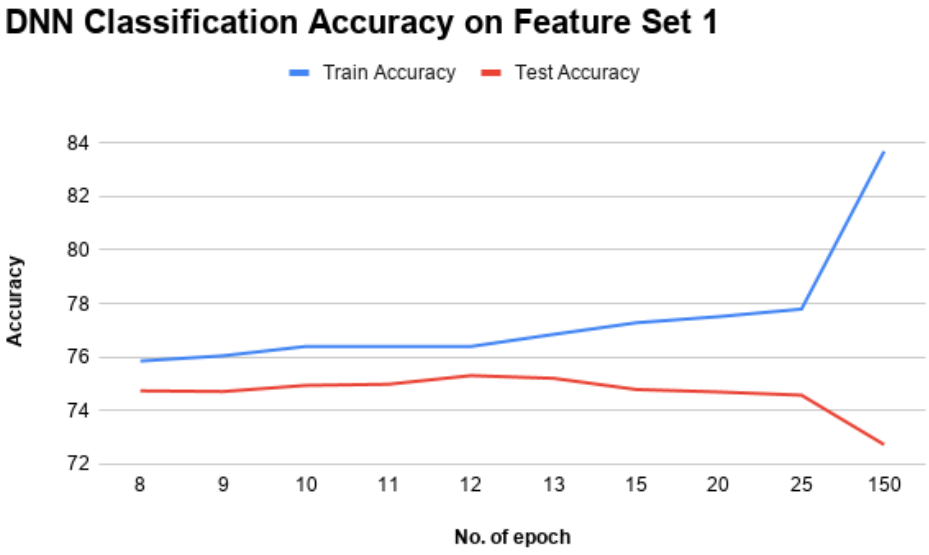

Figure 4.12: Accuracy against No. of epoch for DNN algorithm for Feature set 1

The fine tuned deep neural network used to classify *Feature set 1* was used to evaluate the classification accuracy of DNN from *Feature set 2* and *Feature set 3*. Table 4.5 shows the results.

Table 4.5: Classification Accuracy of All the Feature sets from DNN Classifier

Metric	Feature set 1	Feature set 2	Feature set 3
Accuracy	76.67%	80.67%	87.67%

4.4 Discussion

The dataset was divided into two parts as a Train set and Test set. Test segments were classified using the created classifiers, which are optimised to train segments. Three models were created for the three feature sets. Classification results of the 3 feature sets for all the four different classifiers are shown in Fig.4.13. DNN has performed best with all 3 Feature sets. The best classification accuracy has been reached from Feature set 3. This work aimed to find out a less computationally intensive detection technique to classify the flight calls. To compare the work done in [2] and current work, consider Table 4.6. According to the results, it is clear that the approach taken in work done in [2] is far better provided the high computational power. Even though the current work is behind classification accuracy, that result is reached by a smaller network that requires less computational power. In more specific terms, the current work was executed by an Intel Core i5 3.0GHz central processing unit with 16GB DDR3 Random Access Memory.

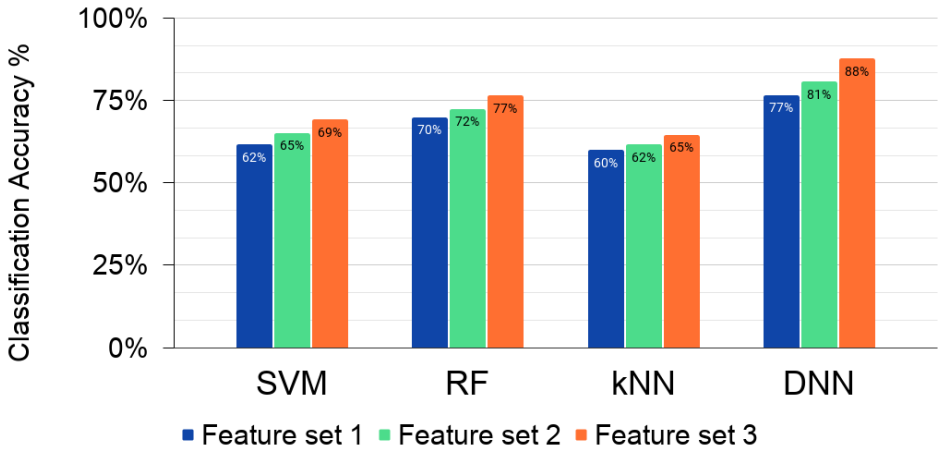

Figure 4.13: Accuracy vs feature sets along with the classifier used

Table 4.6: Comparison of Related and Current work of flight call classification

Metric	Related Work	Current Work
Dataset	Bird-vox-fullnight	Bird-vox-fullnight
Computational power	GPU	CPU
Classification Technique	CNN (667k hidden units)	DNN (10k hidden units)
Test-set classification accuracy	90.48%	87.67%

Chapter 5

CONCLUSION

This research presented a technique that can be used to detect multiple occurrences of similar acoustic events in a real-world dataset in a continuous recording with varying background noise conditions using signal processing and classification techniques with low computational power. As a use case, detection of flight calls of migratory birds has been focused on the research. Identifying flight calls of migratory birds, the count of species and evaluation of flight call counts from several locations will be a useful insight to biologists to judge the behavioural changes of birds. Furthermore, it is really important to reduce the number of collisions with wind turbines. Therefore detecting species early can reduce the number of collisions by stopping the wind turbine. To study the use case of flight calls CLO-43SD, CLO-WTSP, CLO-SWTH and BirdVox-full-night datasets were identified. The birdVox-full-night dataset is used in this research because it matches the requirement to study this use case. Since the chosen dataset contains real-world audio recordings, low signal to noise ratio was observed in most of the clips. Therefore it was important to consider noise reduction techniques.

ROI extraction is really important to capture the sound event in a continuous recording. Several ROI extraction techniques have been used. The feature extraction process has been explored deeper. Three different Feature sets have been discussed. Spectrogram-based Maximally Stable Extremal Regions(SMSER) feature extraction technique was developed in this work. It shows that these features work better with Spectral Features, Temporal Features and Spectrogram based Frequency Statistics Features. MSER image processing technique which has been used previously in other domains has been used to create the SMSER Features.

Four different classifiers, including SVM, RF, kNN, and DNN, have been used to evaluate 3 Feature sets' performance. Classification results from this new feature extraction technique seem to be better with DNN classifiers. Comparing

all three datasets, Feature set 3 shows better results when classified with DNN.

As previous research shows, to increase the accuracy of a classification work, there are three options to try out. Increasing the number of data is the first option. The second would be to increase the dimensionality of the data(number of features). And the last option is to increase the computational power. One objective was to achieve better accuracy with low computational power: The increased number of features have improved the accuracy in this scenario. Achieving better accuracy with low computational power was one of the main objectives of this research. When considering the work results, an increased number of features helped improve the accuracy when high computational power is a constraint.

For example, in case if computational power is a constraint, the accuracy of the classification process can be increased by using the novel feature extraction method proposed in this research. This work will help biologists and computer scientists in developing countries implement new ways to mitigate the collisions with the wind turbines and get senses of flight call counts of migratory birds on different continents.

References

[1] New York Cornell Lab of Ornithology, Ithaca. The nine most important things to know about bird song, 2014.

[2] Vincent Lostanlen, Justin Salamon, Andrew Farnsworth, Steve Kelling, and Juan Pablo Bello. Birdvox-full-night: A dataset and benchmark for avian flight call detection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 266–270. IEEE, 2018.

[3] Sara Keen, Jesse C Ross, Emily T Griffiths, Michael Lanzone, and Andrew Farnsworth. A comparison of similarity-based approaches in the classification of flight calls of four species of north american wood-warblers (parulidae). *Ecological informatics*, 21:25–33, 2014.

[4] Benjamin Van Doren, Andrew Farnsworth, Thomas G Dietterich, Kevin Webb, Steve Kelling, Jed Irvine, Jeffrey Geevarghese, and Daniel Sheldon. Reconstructing velocities of migrating birds from weather radar–a case study in computational sustainability. 2014.

[5] Daniel Fink, Theodoros Damoulas, Nicholas E Bruns, Frank A La Sorte, Wesley M Hochachka, Carla P Gomes, and Steve Kelling. Crowdsourcing meets ecology: hemisphere-wide spatiotemporal species distribution models. *AI magazine*, 35(2):19–30, 2014.

[6] Justin Salamon, Juan Pablo Bello, Andrew Farnsworth, Matt Robbins, Sara Keen, Holger Klinck, and Steve Kelling. Towards the automatic classification of avian flight calls for bioacoustic monitoring. *PloS one*, 11(11):e0166866, 2016.

[7] Ronald P Larkin, William R Evans, and Robert H Diehl. Nocturnal flight calls of dickcissels and doppler radar echoes over south texas in spring. *Journal of Field Ornithology*, 73(1):2–9, 2002.

[8] Jason Wimmer, Michael Towsey, Paul Roe, and Ian Williamson. Sampling environmental acoustic recordings to determine bird species richness. *Ecological Applications*, 23(6):1419–1428, 2013.

[9] Andrew Farnsworth and Irby J Lovette. Evolution of nocturnal flight calls in migrating wood-warblers: apparent lack of morphological constraints. *Journal of Avian Biology*, 36(4):337–347, 2005.

[10] John T Emlen and Michael J DeJong. Counting birds: The problem of variable hearing abilities (contando aves: El problema de la variabilidad en la capacidad auditiva). *Journal of Field Ornithology*, pages 26–31, 1992.

[11] Yves Bas, Vincent Devictor, Jean-Pierre Moussus, and Frédéric Jiguet. Accounting for weather and time-of-day parameters when analysing count data from monitoring programs. *Biodiversity and Conservation*, 17(14):3403–3416, 2008.

[12] Wesley M. Hochachka, Daniel Fink, Rebecca A. Hutchinson, Daniel Sheldon, Weng-Keen Wong, and Steve Kelling. Data-intensive science applied to broad-scale citizen science. *Trends in Ecology Evolution*, 27(2):130 – 137, 2012.

[13] Miguel A. Acevedo, Carlos J. Corrada-Bravo, Héctor Corrada-Bravo, Luis J. Villanueva-Rivera, and T. Mitchell Aide. Automated classification of bird and amphibian calls using machine learning: A comparison of methods. *Ecological Informatics*, 4(4):206 – 214, 2009.

[14] T. Damoulas, S. Henry, A. Farnsworth, M. Lanzone, and C. Gomes. Bayesian classification of flight calls with a novel dynamic time warping kernel. In *2010 Ninth International Conference on Machine Learning and Applications*, pages 424–429, Dec 2010.

[15] Selin Bastas, Mohammad Wadood Majid, Golrokh Mirzaei, Jeremy Ross, Mohsin M Jamali, Peter V Gorsevski, Joseph Frizado, and Verner P Bingman. A novel feature extraction algorithm for classification of bird flight

calls. In *2012 IEEE International Symposium on Circuits and Systems*, pages 1676–1679. IEEE, 2012.

[16] Dan Stowell and Mark D Plumbley. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. *PeerJ*, 2:e488, 2014.

[17] Juha T Tanttu, Jari Turunen, Arja Selin, and MIRKO OJANEN. Automatic feature extraction and classification of crossbill (loxia spp.) flight calls. *Bioacoustics*, 15(3):251–269, 2006.

[18] T Mitchell Aide, Carlos Corrada-Bravo, Marconi Campos-Cerqueira, Carlos Milan, Giovany Vega, and Rafael Alvarez. Real-time bioacoustics monitoring and automated species identification. *PeerJ*, 1:e103, 2013.

[19] Olivier Dufour, Thierry Artieres, Hervé Glotin, and Pascale Giraudet. Clusterized mel filter cepstral coefficients and support vector machines for bird song identification. In *Soundscape Semiotics-Localization and Categorization*. IntechOpen, 2014.

[20] Rolf Bardeli, Daniel Wolff, Frank Kurth, Martina Koch, K-H Tauchert, and K-H Frommolt. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. *Pattern Recognition Letters*, 31(12):1524–1534, 2010.

[21] Todor D Ganchev, Olaf Jahn, Marinez Isaac Marques, Josiel Maimone de Figueiredo, and Karl-L Schuchmann. Automated acoustic detection of vanellus chilensis lampronotus. *Expert systems with applications*, 42(15-16):6098–6111, 2015.

[22] Andrew Digby, Michael Towsey, Ben D Bell, and Paul D Teal. A practical comparison of manual and autonomous methods for acoustic monitoring. *Methods in Ecology and Evolution*, 4(7):675–683, 2013.

[23] Andrew Farnsworth. Flight calls and their value for future ornithological studies and conservation research. *The Auk*, 122(3):733–746, 2005.

[24] P. Somervuo, A. Harma, and S. Fagerlund. Parametric representations of bird sounds for automatic species recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):2252–2263, 2006.

[25] Z. Chen and R. C. Maher. Semi-automatic classification of bird vocalizations using spectral peak tracks. *The Journal of the Acoustical Society of America*, 120(5):2974–2984, 2006.

[26] Jose F Ruiz-Muñoz, Mauricio Orozco Alzate, and Germán Castellanos-Domínguez. Multiple instance learning-based birdsong classification using unsupervised recording segmentation. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[27] Seppo Fagerlund and Unto K Laine. New parametric representations of bird sounds for automatic classification. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8247–8251. IEEE, 2014.

[28] Iosif Mporas, Todor Ganchev, Otilia Kocsis, Nikos Fakotakis, Olaf Jahn, Klaus Riede, and Karl L Schuchmann. Automated acoustic classification of bird species from real-field recordings. In *2012 IEEE 24th International Conference on Tools with Artificial Intelligence*, volume 1, pages 778–781. IEEE, 2012.

[29] Peter Jančovič and Münevver Köküer. Automatic detection and recognition of tonal bird sounds in noisy environments. *EURASIP Journal on Advances in Signal Processing*, 2011(1):982936, 2011.

[30] Judith C Brown and Patrick JO Miller. Automatic classification of killer whale vocalizations using dynamic time warping. *The Journal of the Acoustical Society of America*, 122(2):1201–1207, 2007.

[31] S. S. Londhe and S. S Kanade. Bird species identification using support vector machine. *International Journal of Computer Science and Mobile Computing*, 4(9):125–135, 2015.

[32] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[33] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[34] C. W. Maina. Bioacoustic approaches to biodiversity monitoring and conservation in kenya. In *IEEE IST-Africa Conference*, pages 1–8, 2015.

[35] P. Rai, V. Golchha, A. Srivastava, G. Vyas, and S. Mishra. An automatic classification of bird species using audio feature extraction and support vector machines. In *IEEE International Conference on Inventive Computation Technologies (ICICT)*, volume 1, pages 1–5, 2016.

[36] X. Dong, J. Xie, M. Towsey, J. Zhang, and P. Roe. Generalised features for bird vocalisation retrieval in acoustic recordings. In *IEEE International Workshop onMultimedia Signal Processing (MMSP)*, pages 1–6, 2015.

[37] M. A. Raghuram, N. R Chavan, R. Belur, and S. G. Koolagudi. Bird classification based on their sound patterns. *International Journal of Speech Technology*, 19(4):791–804, 2016.

[38] S. Fagerlund. Bird species recognition using support vector machines. *Eurasip Journal on Advances in Signal Processing*, 2007.

[39] Thiago LF Evangelista, Thales M Priolli, Carlos N Silla Jr, Bruno A Angelico, and Celso AA Kaestner. Automatic segmentation of audio signals for bird species identification. In *Multimedia (ISM), 2014 IEEE International Symposium on*, pages 223–228. IEEE, 2014.

[40] J. J. Noda, C. M. Travieso, D. Sánchez-Rodríguez, M. K. Dutta, and A. Singh. Using bioacoustic signals and support vector machine for auto-

matic classification of insects. In *IEEE International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 656–659, 2016.

[41] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K. H. Tauchert, and K. H. Frommolt. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. *Pattern Recognition Letters*, 31(12):1524–1534, 2010.

[42] J. C. Brown and P. J. O. Miller. Automatic classification of killer whale vocalizations using dynamic time warping. *The Journal of the Acoustical Society of America*, 122(2):1201–1207, 2007.

[43] T. S. Brandes. Feature vector selection and use with hidden markov models to identify frequency-modulated bioacoustic signals amidst noise. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6):1173–1180, 2008.

[44] K. Kaewtip, A. Alwan, C. O'Reilly, and C. E. Taylor. A robust automatic birdsong phrase classification: A template-based approach. *The Journal of the Acoustical Society of America*, 140(5):3691–3701, 2016.

[45] Miguel A Acevedo, Carlos J Corrada-Bravo, Héctor Corrada-Bravo, Luis J Villanueva-Rivera, and T Mitchell Aide. Automated classification of bird and amphibian calls using machine learning: A comparison of methods. *Ecological Informatics*, 4(4):206–214, 2009.

[46] S Gunasekaran and K Revathy. Automatic recognition and retrieval of wild animal vocalizations. *International Journal of Computer Theory and Engineering*, 3(1):136, 2011.